SurroundSDF: Implicit 3D Scene Understanding Based on Signed Distance Field

Lizhe Liu[1*] Bohua Wang[2*§] Hongwei Xie[1*] Daqi Liu[1] Li Liu[1] Zhiqiang Tian[2]
Kuiyuan Yang[1] Bing Wang[1†]
[1]Xiaomi EV [2]School of Software Engineering, Xi'an Jiaotong University

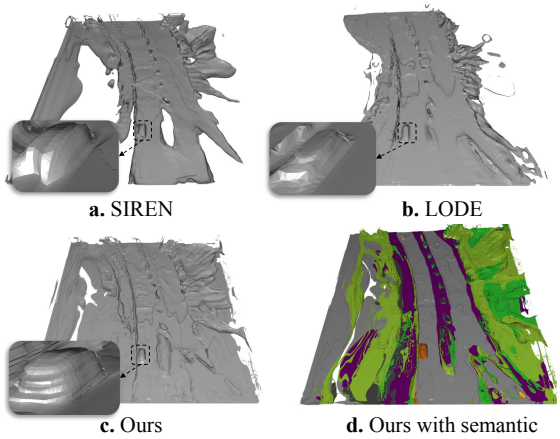
Abstract

Vision-centric 3D environment understanding is both vital and challenging for autonomous driving systems. Recently, object-free methods have attracted considerable attention. Such methods perceive the world by predicting the semantics of discrete voxel grids but fail to construct continuous and accurate obstacle surfaces. To this end, in this paper, we propose ***SurroundSDF*** *to implicitly predict the signed distance field (SDF) and semantic field for the continuous perception from surround images. Specifically, we introduce a query-based approach and utilize SDF constrained by the Eikonal formulation to accurately describe the surfaces of obstacles. Furthermore, considering the absence of precise SDF ground truth, we propose a novel weakly supervised paradigm for SDF, referred to as the* ***Sandwich Eikonal*** *formulation, which emphasizes applying correct and dense constraints on both sides of the surface, thereby enhancing the perceptual accuracy of the surface. Experiments suggest that our method achieves SOTA for both occupancy prediction and 3D scene reconstruction tasks on the nuScenes dataset.*

1. Introduction

With the recent advancement of 3D object detection algorithms [10, 11, 16, 18, 21, 24], vision-centric autonomous driving system become more practicable. Nevertheless, the persisting challenges related to the long-tail problem and the coarse depiction of the 3D scene underscore its insufficiency. Consequently, a deeper comprehension of 3D geometry and semantics is needed for safety and reliability. This paper delves into a novel vision-centric paradigm of dense and continuous 3D scene understanding.

Current approaches of dense 3D prediction can be classified into two categories, *3D reconstruction* and *3D perception*. Specifically, *3D reconstruction* algorithms [6, 33] generate dense point clouds enriched with semantic information by projecting depth maps and 2D semantic maps into 3D space. *3D perception* methods [3, 12, 17, 25, 35, 41, 46] predict the occupied status and the semantic of 3D voxel grids. However, either method proves to be redundant or can only predict coarse-grained discrete grids.



a. SIREN **b.** LODE **c.** Ours **d.** Ours with semantic

Figure 1. The scene perspective results from surround images input. **a, b.** The result of SIREN and LODE supervision with camera-only input (the original methods are point cloud-based). **c.** Our result. **d.** Our result with semantics.

Instead, we develop a vision-centric framework to describe 3D scenes leveraging neural implicit Signed Distance Function (SDF) representation, which we refer to as **SurroundSDF**. Generally, we aim to: (1) Continuously describe the 3D scene by reconstructing smooth surfaces; (2) Explore the difficulty associated with the utilization of SDF representation and pose an appropriate training strategy; (3) Exploit the strong representation to kill two birds with one stone, addressing 3D semantic segmentation and continuous 3D geometry reconstruction within one framework.

In pursuit of the first and the third objectives, we employ SDF to represent the canonical 3D scene in terms of distance from the surface. Furthermore, we construct an implicit field that encompasses both semantics and geometry by exploiting this strong representation. However, it's still nontrivial to accomplish the SDF modeling, where complete surface and normal vectors, which are crucial supervisions, cannot be accurately computed. To this end, SIREN [34] proposed a weakly supervised implicit per-

*Equal Contribution.
§Work done during an internship at Xiaomi EV.
†Corresponding Author.

ception method, which accomplishes 3D mesh completion from point clouds. While LODE [14] proposed to utilize occupancy ground truth (GT) to supervise SDF. However, these methods rely on the input of point clouds, and only alleviate this difficulty by mimicking the surface distance but are either sparse or inaccurate.

To overcome the aforementioned challenges, we propose the **Sandwich Eikonal** formulation, a novel weak supervision paradigm for SDF modeling. Figure 1 gives some reconstruction results and shows the benefit of our method. This method emphasizes applying correct and dense constraints on both sides of the surface to enhance the geometric accuracy and continuity of the surface. Moreover, we revisit current pipelines for training the perception branch and design a novel loss that enhances the integration of geometry and semantics thereby reducing inconsistencies. Our contributions can be summarized as follows:

- We propose a vision-centric implicit semantic SDF perception method, achieving accurate and continuous 3D perception. To the best of our knowledge, we are the first to utilize SDF for surround-view 3D perception.
- We introduce the Sandwich Eikonal formulation, a novel weak supervision paradigm for SDF.
- We demonstrate how to employ this representation for the reconstruction of continuous 3D geometry with precise semantic information.
- We achieve state-of-the-art results on 3D dense semantics perception tasks. Comprehensive experiments on the NuScenes dataset [2] provide extensive validation of our approach.

2. Related Work

Occupancy Prediction Recently, researches [3, 12, 17, 25, 32, 32, 35, 41, 46] on occupancy prediction have demonstrated advantages in 3D scene understanding. Compared to the traditional object detection paradigm [5, 10, 11, 15, 16, 18, 21, 22, 24, 39, 42], occupancy perception has the following advantages. First, it can express dense 3D geometry. Second, it can accurately provide spatial locations for objects beyond predefined categories. Third, it can describe the shapes of irregular obstacles.

Based on these advantages, occupancy prediction tasks have attracted significant research. These methods predict the occupancy status in the region of interest around the ego vehicle from point clouds or images. Specifically, the space is first divided into voxel grids at a specific resolution. Then the occupancy status and semantics of each grid are estimated. SurroundOcc [41] proposed a surround-view 3D occupancy perception method that utilizes spatial 2D-3D attention to lift image features into 3D space. In addition, to realize the dense occupancy prediction, SurroundOcc designed a pipeline to convert the point cloud to dense occupancy ground truth. VoxFormer [17] employed an MAE-like [9] approach to achieve camera-based semantic occupancy prediction. FB-OCC [20, 20] proposed a novel forward-backward projection method to compensate for the insufficient BEV feature density of the forward projection method and the large number of mismatches in 2D and 3D space caused by the backward projection. Despite the impressive results, they are limited by the specific resolution of the occupancy annotations, which limit the continuous perception of the scene.

Implicit Scene Perception Scene reconstruction refers to the task of predicting the 3D geometry structures from some incomplete representations, e.g., images, point clouds, and voxel grids. Implicit neural scene reconstruction methods [1, 4, 14, 26, 30, 34, 45] have demonstrated advantages in accuracy and substantial potential. Generally, they train a neural network to predict a continuous field. Thus, it is possible to query occupancy information for any point within the 3D space. DeepSDF [30] utilizes the SDF to achieve the implicit reconstruction of 3D objects at the instance level. NeRFs [26] and its numerous variants [4, 7, 43] achieve implicit 3D scene reconstruction and novel view synthesis from multi-view images. NeuS [38] and VolSDF [44] introduce the signed distance function into neural rendering, enabling more precise object surface reconstruction. SIREN [34] utilizes the Eikonal equation to achieve semi-supervised SDF reconstruction for point clouds. However, these methods lack generalization to novel scenes. To achieve online implicit scene perception in driving scenes, LODE [14] proposes the Locally Conditioned Eikonal formulation and introduces a dense occupancy ground truth for supervision, which significantly improves the effect of SDF reconstruction in the driving scenes. However, it is limited by the sparsity of input LiDAR points or inaccurate ground truth. To solve these problems, we propose a camera-only implicit scene understanding method, which aims at continuous and accurate surface perception.

3. Formulation

This section analyzes the SDF supervision paradigm based on the Eikonal formulation with its variations and introduces our **Sandwich Eikonal** supervision approach.

3.1. Eikonal-based SDF Constraints

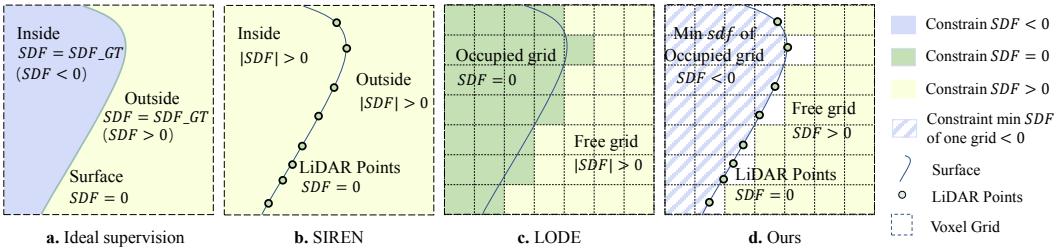
Given a coordinate $\mathbf{x}$ in the 3D scene, our goal is to construct a function ϕ such that $\phi(\mathbf{x})$ provides the SDF value at point $\mathbf{x}$. The Eikonal-based optimization objectives [14] is

$$\mathcal{O}_E = \int_{\Omega_0} O_0 d\mathbf{x} + \int_{\Omega_1} O_1 d\mathbf{x} + \int_{\Omega_2} O_2 d\mathbf{x}, \tag{1}$$

and

$$\begin{cases} O_0 = \left| |\nabla_{\mathbf{x}} \phi(\mathbf{x})| - 1 \right|, & \mathbf{x} \in \Omega_0 \\ O_1 = \left| \nabla_{\mathbf{x}} \phi(\mathbf{x}) - \mathbf{n}(\mathbf{x}) \right|, & \mathbf{x} \in \Omega_1 \\ O_2 = \left| \phi(\mathbf{x}) - SDF(\mathbf{x}) \right|, & \mathbf{x} \in \Omega_2 \end{cases}, \tag{2}$$

where Ω_0 is the set of the whole 3D space of interest, Ω_1 is the set of points on the surface, Ω_2 is the set of points



a. Ideal supervision **b.** SIREN **c.** LODE **d.** Ours

Figure 2. **a.** Ideal supervision for SDF where the SDF GT is provided. **b.** SIREN supervision which uses LiDAR points to supervise the surface. **c.** LODE supervision based on occupancy GT. **d.** Our supervision paradigm which combines the LiDAR GT and occupancy GT, it is closer to the ideal supervision.

with SDF GT annotations, $\nabla_{\mathbf{x}}\phi(\mathbf{x})$ is the gradient on point $\mathbf{x}$, $\mathbf{n}(\mathbf{x})$ is the normal, and $SDF(\mathbf{x})$ is the real SDF value at point $\mathbf{x}$. One major obstacle to this optimization objective lies in the design of O_2, which provides direct supervision for SDF as shown in Figure 2.a. This objective requires the annotation of SDF values for each anchor point in the scene, which is monumental and challenging.

In the absence of precise SDF annotations, SIREN [34] and LODE [14] have implemented weakly supervised paradigm through the following variations of O_2,

$$\begin{cases} O'_{2-0} = |\phi(\mathbf{x})|, & \mathbf{x} \in \Omega_1 \\ O'_{2-1} = \psi(|\phi(\mathbf{x})|), & \mathbf{x} \in \Omega_3, \end{cases} \tag{3}$$

where $\psi(\cdot)$ is a monotonically decreasing function and $\Omega_3 \subseteq \Omega_0/\Omega_1$ is the whole space except the surface.

The objective O'_{2-0} aims to constrain the SDF values on the surface to 0. Considering the absence of precise supervision, O'_{2-1} adopts a fuzzy constraint that pushes the absolute SDF value away from zero in Ω_3. While the constraint in Ω_3 is loose, it is adequate to capture the geometry information of the surface.

In practical implementation, SIREN [34] samples on-surface LiDAR points in Ω_1 and considers the space excluding the LiDAR points as Ω_3, as is shown in Figure 2.b. When the LiDAR points are dense enough, this strategy can effectively achieve good constraint effects [34]. However, LODE [14] has demonstrated that the sparsity of LiDAR points in the driving scenes results in discontinuities of the estimated mesh. To achieve denser supervision, they use the dense occupancy GT as supervision. This approach treats both the surface and interior of the obstacles as the Ω_1. Specifically, they sample the Ω_1 points as the center of "occupied" grids, and the Ω_3 points as the center of "free" grids, as is shown in 2.c. However, this paradigm remains imprecise, as considering the center of the occupied grid to be the surface of objects would introduce errors.

3.2. Sandwich Eikonal Formulation

To address the above concerns, we require a supervision paradigm that has the following characteristics: 1) sampling points as the supervision on the surface must strictly adhere to the surface, 2) for regions beyond the surface, refrain from supervision in uncertain areas to ensure precision, and 3) dense supervision signals. Considering the above factors, we introduce our variants of Eikonal formulation as follows:

$$\begin{cases} O_{2-0} = \phi(\mathbf{x}), & \mathbf{x} \in \Omega_1 \\ O_{2-1} = \psi(-min(\phi(\mathbf{x}))), & \mathbf{x} \in \Omega^i_{occ} \\ O_{2-2} = \psi(\phi(\mathbf{x})), & \mathbf{x} \in \Omega^j_{free}, \end{cases} \tag{4}$$

where Ω^i_{occ} is the i-th grid of occupied grids and Ω^j_{free} is the j-th grid of free grids.

To this end, we incorporate the LiDAR points and the occupancy GT into the SDF supervision. As is shown in Figure 2.d, for precise sampling points on the surface, we follow SIREN and constrain the SDF value of LiDAR points to be 0. To apply accurate and dense supervision to regions beyond the surface, we introduce the occupancy GT into the supervision, as is shown in optimization objectives O_{2-1} and O_{2-2}. According to the generation process of occupancy GT [35, 36], if a grid is labeled as "occupied", it implies that at least part of the grid area falls inside the object. Therefore, as the objective O_{2-1} shows, we constrain the minimum SDF value of each occupied grid to be less than zero and push it to the negative range. On the other hand, if a grid is labeled as "free", the entire grid is outside the object and we push the SDF value of free grids to the positive range, as is shown in objective O_{2-2}. We employ a multi-frame point cloud fusion strategy to densify the sampling for surface supervision, as described in Section 4.2. Compared to previous supervision paradigms, our formulation emphasizes applying appropriate and dense supervision on both sides of the surface to enhance the geometric constraint based on the Eikonal formulation. Therefore, we name our supervision strategy the "**Sandwich Eikonal**" formulation.

4. Method

Given multi-camera images $\mathbf{I} = \{I_1, I_2, ..., I_n\}$, we aim to predict the 3D SDF $\phi(\mathbf{x}) : \mathbb{R}^3 \Rightarrow \mathbb{R}$ and semantic field $S(\mathbf{x}) : \mathbb{R}^3 \Rightarrow \mathbb{R}^s$, where s indicates the number of classes. The following will introduce our architecture and detailed constraints based on our Sandwich Eikonal formulation.

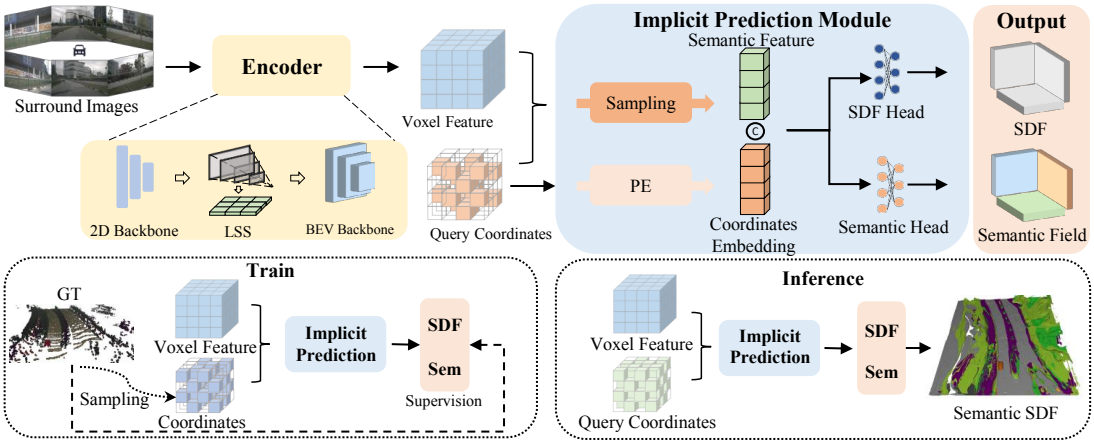


Figure 3. The architecture of our SurroundSDF. Given the surround images as input, an encoder composed of a 2D backbone, LSS module, and BEV backbone, is employed to extract voxel features. We adopt a query-based approach to sample features from the voxel features. Specifically, first, a set of query coordinates in the region of interest is selected. Subsequently, using trilinear interpolation, semantic features are queried from the voxel features. Finally, after concatenation with the positional embeddings from the query coordinates, the features pass through the SDF head and semantic head respectively, yielding SDF and semantic fields. For training, the query coordinates are sampled according to the GT, and the SDF and semantic field are supervised by the losses introduced in Section 4.2. In the inference phase, based on appropriate sampling and post-processing, continuous and accurate scene perception results are obtained (see Section 4.3).

4.1. Architecture

The overall architecture is shown in Figure 3. First, a CNN-based backbone is used to obtain the image features. Then, the LSS [31] is used to project the image features to BEV space, and a BEV encoder is adopted to obtain the voxel features V of dimension $h \times w \times d \times c$ (c indicates the number of channels). Further, for each point $\mathbf{x}$ in Ω_0, the corresponding features can be obtained by querying V with the trilinear sampling method. Therefore, $\phi(\mathbf{x})$ and $S(\mathbf{x})$ can be expressed as:

$$\begin{aligned} \phi(\mathbf{x}) &= H_{sdf}(C(Q(V,\mathbf{x}),PE(\mathbf{x})),\theta),\mathbf{x}\in\Omega_0, \\ S(\mathbf{x}) &= H_{sem}(C(Q(V,\mathbf{x}),PE(\mathbf{x})),\theta),\mathbf{x}\in\Omega_0, \end{aligned} \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^3$, C and Q represent the concatenation and query process, respectively. H_{sdf} and H_{sem} indicate the SDF head and semantic head, implemented with three-layer MLPs with sine activation functions. θ represents the learnable parameters. PE represents the positional encoding, which aims to capture high-frequency information. Specifically, for a coordinate $\mathbf{x} \in \mathbb{R}^3$, PE encodes each of its dimensions into a 2n-dimensional vector, it can be expressed as:

$$PE(\mathbf{x}) = (\zeta(x),\zeta(y),\zeta(z)) \in \mathbb{R}^{6n}, \tag{6}$$

where x, y, z denote the coordinates, $\zeta(\cdot)$ represents an encoding scheme using trigonometric function mapping:

$$\zeta(x) = (sin(2^0\pi x), cos(2^0\pi x), ..., cos(2^{n-1}\pi x)). \tag{7}$$

In training, we design a joint supervision method that utilizes both coarse-grained voxel and fine-grained point cloud supervision based on our Sandwich Eikonal formulation. In inference, we query the voxel features to estimate the SDF values and semantic logits of each query point.

4.2. Supervision

This section will elaborate on deriving the SDF loss from our Sandwich Eikonal formulation. Then, we will introduce the semantic loss and our joint supervision strategy.

SDF Supervision Following objectives O_0, O_1 in Equation 2 and O_{2-0} to O_{2-2} in Equation 4, we derive the losses for SDF supervision, based on our Sandwich Eikonal formulation. Considering that continuous space supervision is unfeasible, we sample the discrete points in the corresponding space (Ω_0, Ω_1, Ω_{occ} and Ω_{free}) for supervision.

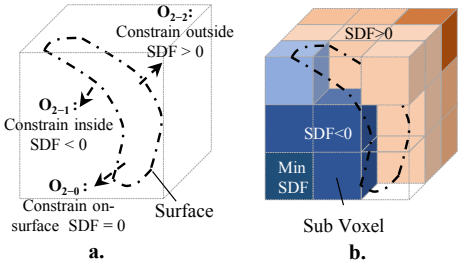


Figure 4. **a.** SDF constraints with Sandwich Eikonal formulation in continuous form. **b.** SDF sampling in discrete form.

The correctness and density of the supervision sampling are of paramount importance. Since our approach is based on the surround cameras, we only sample from the camera-visible space in Ω_0 (space of interest) [36]. For Ω_1 (surface points), we sample exclusively from the LiDAR points. To enhance the sampling density, temporal point cloud fusion is adopted. Specifically, $\Omega_1 = \{p_{t-k}, p_{t-(k-1)}, ..., p_{t+(k-1)}, p_{t+k}\}$, where p_{t+j} is the LiDAR points of the $t+j$ frame and t is the current frame. Based on occupancy GT [36], Ω_{occ} and Ω_{free} are the sets of

voxel grids labeled as "occupied" and "free", respectively.

Additionally, in our objective O_{2-1}, the computation of the minimum SDF within the occupied grid is required. As is shown in Figure 4, we approximate this minimum value through discrete sampling. The minimum value among the SDF values at the centers of each sub-grid is regarded as the minimum SDF value for the grid. Based on Equation 2, 4 and the above sampling strategy, our SDF loss $\mathcal{L}_{sdf}$ is expressed as follows:

$$\begin{aligned}\mathcal{L}_{sdf} = \gamma_1 \frac{1}{n_{total}} \sum_{i=1}^{n_{total}} (|\nabla_{\mathbf{x}} \phi(\mathbf{x}^i)| - 1) \\ + \gamma_2 \frac{1}{n_{1 \cup occ}} \sum_{i=1}^{n_{1 \cup occ}} (\nabla_{\mathbf{x}} \phi(\mathbf{x}^i) - n(\mathbf{x}^i)) \\ + \gamma_3 \frac{1}{n_1} \sum_{i=1}^{n_1} |\phi(\mathbf{x}^i)| \\ + \gamma_4 \frac{1}{n_{occ}} \sum_{i=1}^{n_{occ}} e^{\alpha \times min(\phi(\mathbf{x}^i), \mathbf{x}^i \in \Omega^i_{occ})} \\ + \gamma_5 \frac{1}{n_{free}} \sum_{i=1}^{n_{free}} e^{-\alpha \times random(\phi(\mathbf{x}^i), \mathbf{x}^i \in \Omega^i_{free})},\end{aligned} \tag{8}$$

where α is a hyperparameter that determines the extent of the deviation for the interior and exterior space of the object and we fix α to 100 in experiments. $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$ are loss weights, $n(\mathbf{x})$ represents the normal vector of the current point on surface. n_1, n_{occ}, and n_{free} represent the number of points or voxel grids sampled from Ω_1, Ω_{occ} and Ω_{free}, respectively. n_{total} is the total number of samples: $n_{total} = n_1 + n_{occ} + n_{free}$.

Semantic Supervision A widely used cross-entropy loss is adopted to supervise the semantic field as follows:

$$\mathcal{L}_{sem} = \frac{1}{n_{occ}} \sum_{i=1}^{n_{occ}} CE(S(\mathbf{x}|\mathbf{x} \in \Omega^i_{occ}), \hat{y}^i_{sem}), \tag{9}$$

where CE represents cross-entropy loss and $\hat{y}^i_{sem}$ is the semantic ground truth of i-th voxel grid.

Joint Supervision The above losses optimize the SDF and semantic field separately, which leads to a non-negligible issue of ambiguity between geometric and semantic optimization. To illustrate this issue, we train a model and evaluate the occupancy IoU representing geometric accuracy and the semantic mIoU representing overall semantic accuracy following the occupancy prediction metrics [35, 36]. For each voxel grid, if its minimum SDF is less than a certain threshold, the voxel is considered 'occupied'. As is shown in Figure 5, the optimal thresholds producing the highest peak of these curves are significantly different. This indicates that greater geometric accuracy does not necessarily lead to improved overall semantic accuracy.

To mitigate this issue, we propose a joint supervision strategy. First, for the minimum SDF value $min(\phi(\mathbf{x}))$ of each voxel grid, a SoftMax function is used to convert it to the free probability: $p_{free} = softmax(\beta t, \beta min(\phi(\mathbf{x})))$, where t is fixed to 0.005, β is fixed to 100. Second, we map the free probability p_{free} to logits using an inv-sigmoid function: $l_{free} = log(p_{free}/(1 - p_{free}))$. Finally, l_{free} is combined with semantic logits l_{sem}, and the joint loss $\mathcal{L}_{joint}$ is expressed as:

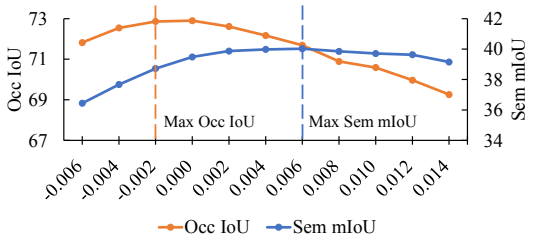


Figure 5. Variation of occupancy IoU and semantic mIoU with SDF Threshold. Note that the peak values of these two indicators correspond to different SDF thresholds.

$$\mathcal{L}_{joint} = Dice(C(l_{sem}, l_{free}), C(\hat{y}_{sem}, \hat{y})), \tag{10}$$

where C is the concatenation operation and $\hat{y}_{sem}$ is the semantic GT of voxel grid. $Dice$ represents the Dice loss [27]. Finally, the total loss is calculated as:

$$\mathcal{L} = \gamma_6 \mathcal{L}_{sdf} + \gamma_7 \mathcal{L}_{sem} + \gamma_8 \mathcal{L}_{joint} \tag{11}$$

where $\gamma_6, \gamma_7, \gamma_8$ are loss weights.

4.3. Inference

Based on the output SDF and semantic field, diverse outputs for different tasks can be obtained through different sampling strategies and post-processing methods.

Semantic Mesh Generation To generate the semantic mesh, we uniformly query the SDF value in the space of interest and use the marching cubes algorithm [28] to obtain the mesh. Subsequently, at the vertex positions of each triangular face, we query the corresponding semantic information, resulting in a mesh enriched with semantics.

Occupancy Prediction The objective of the occupancy prediction is to anticipate the semantics of the voxel grid in space. We achieve this through the joint prediction based on the predicted SDF and semantic field as follows:

$$\mathcal{S}_i = argmax(C(l^i_{sem}, l^i_f)), \tag{12}$$

where $\mathcal{S}_i$ is the i-th voxel grid, l^i_{sem} is the logits of the i-th voxel grid,and l^i_f the logits value from SDF.

Semantic Scene Reconstruction Our SurroundSDF demonstrates a notable advantage by achieving semantic scene reconstruction concurrently with scene perception. As is shown in Figure 6, we sequentially query the semantic and SDF values for the region of interest frame by frame. The SDF values and semantic logits of static objects are projected onto the sampling points in the world coordinate

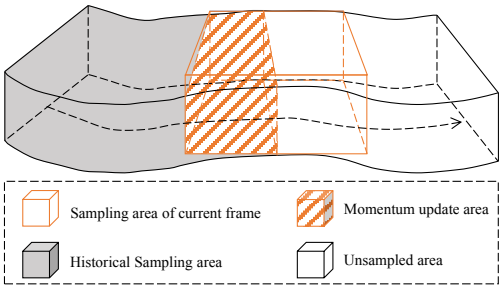


Figure 6. The process of semantic scene reconstruction. SDF values and semantic logits are sampled long the vehicle trajectory into the world coordinate system. For regions where the historical sampling overlaps with the current frame, the historical SDF and semantic logits are updated based on a momentum update strategy.

system. For the overlapping of the historical sampling area and the current frame sampling area, we employ a momentum update strategy to refine the historical SDF values and semantic logits. Figure 7 shows the final scene reconstruction result with the marching cubes algorithm.

5. Experiments

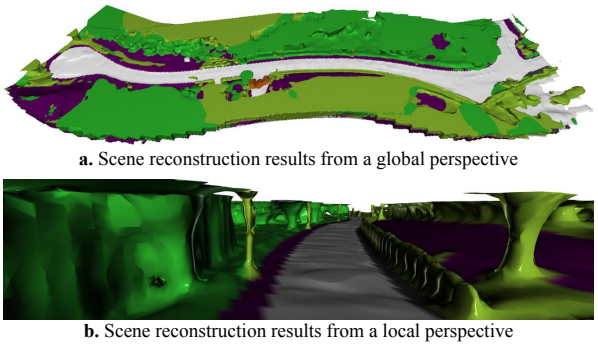
5.1. Experimental Setup

Version	Backbone	Image size	Channels	Frames
Small	ResNet50 [8]	256 × 704	32	9
Medium	ResNet101 [8]	640 × 1600	128	6
Large	ConvNext-B [23]	640 × 1600	256	3

Table 1. Experimental settings of different versions.

Implementation details We follow BEVStereo [15] to construct our 3D encoder, where the depth net predicts 88 discrete depth values, the voxel feature resolution is 200 × 200 × 16, and the temporal fusion strategy is applied. We pre-train our architecture on semantic segmentation, depth estimation, and 3D object detection tasks on the NuScenes training set. To compare with methods under different settings, we train three versions: *Small*, *Medium*, and *Large* with different backbone, input image size, voxel feature channels, and the number of temporal frames, as is shown in Tabel 1. For the medium and large versions, we reduce the number of temporal frames to save GPU memory. We use the Adam optimizer [13], set the batch size to 32, fix the learning rate to 1e-4, and train for around 32 epochs for each version. The loss weights $\gamma_1 \sim \gamma_8$ are set to 1.0, 1.0, 30.0, 0.05, 0.05, 1.0, 1.0, and 1.0 respectively.

Dataset We evaluate our method on Occ3D-nuScenes dataset [35, 36], which is an occupancy dataset with dense voxel grid label based on the nuScenes dataset [2]. It processes dynamic and static objects separately, uses multi-frame aggregation to achieve point cloud densification, and then reconstructs the dense point cloud to obtain a 3D occupancy label. It has voxel-level semantic labels of 17 classes,



a. Scene reconstruction results from a global perspective

b. Scene reconstruction results from a local perspective

Figure 7. Visualization of semantic scene reconstruction.

which contain 16 common classes and an additional general object class. Each frame covers a range of [-40m, -40m, -1m, 40m, 40m, 5.4m] with a voxel size of 0.4m, resulting in a voxel grid resolution with 200 × 200 × 16.

Evaluation Metrics To evaluate our performance on dense 3D scene perception, we follow the 3D semantic occupancy prediction task [36] and use the widely adopted mean intersection over union (mIoU).

$$mIoU = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i + FN_i}, \tag{13}$$

where C represents the number of categories (set to 17 in the experiment), and TP, FP, and FN represent true positive, false positive, and false negative respectively.

Metrics	Formula
Abs Rel	$\frac{1}{n}\sum\|d - d^*\|/d^*$
Sq Rel	$\frac{1}{n}\sum\|d - d^*\|^2/d^*$
RMSE	$\sqrt{\frac{1}{n}\sum\|d - d^*\|^2}$
$\delta < 1.25$	$\frac{1}{n}\sum(max(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25)$

Table 2. Depth metrics for 3D scene reconstruction

To evaluate our scene reconstruction performance, we render the 2D depth image for the surround cameras based on the reconstructed scene and validate using 2D depth evaluation. As is shown in Table 2, we use the metrics commonly used for depth estimation, which aim to measure the difference between the predicted depth obtained by mesh rendering and GT depth obtained by LiDAR points.

5.2. 3D Semantic Occupancy Prediction

Following the widely adopted experimental setup [19, 36], we conduct the semantic occupancy prediction experiment on the Occ3D-nuScenes dataset. For a fair comparison, in the inference phase, we divide the space into voxel grids with the same size as the other methods (0.4m) and predict the occupancy semantics following Section 4.3. As is shown in Tabel 3, the mIoU score of our *Small* version using ResNet50 as the backbone and without any test-time-augmentation (TTA) achieves 42.37, surpassing all the

methods with the same image backbone. Additionally, our *Medium* version significantly outperforms BEVFormer [18] and VoxFormer [17] using the ResNet101 backbone. Moreover, our *Large* version surpasses UniOCC [29] with the same ConvNext-B backbone by a 1.5 mIoU score, and even outperforms FB-OCC [19, 20], which utilizes a larger image backbone and TTA strategy. These results demonstrated the effectiveness of SurroundSdf in the perception task.

5.3. 3D Scene Reconstruction

To evaluate the performance of scene reconstruction, we compare our method with state-of-the-art (SOTA) surround-view-based depth estimation methods [33, 40] and occupancy prediction methods [19, 41]. To acquire the surround depth, we utilize the semantic mesh generation steps described in Section 4.3 to generate the mesh and then render the depth maps for each camera based on the camera parameters. To fairly compare with the occupancy-based methods, we follow a similar process to render the depth maps with predicted occupied voxel grids, and all methods use RestNet50 as the image backbone. As is shown in Table 4, our method exhibits significant advantages over occupancy-based approaches across all metrics. Moreover, on the primary metric of absolute relative error, we achieve SOTA results, surpassing SurroundDepth [40] and R3D3 [33], which are the SOTA methods for surround-view depth estimation.

Methods	**Backbone**	**Image Size**	**mIoU**
SurroundOcc [41]	ResNet50	256 × 704	36.32
BEVStereo [15]	ResNet50	384 × 704	39.1
UniOCC [29]	ResNet50	256 × 704	39.7
MiLO [37]	ResNet50	256 × 704	40.49
FB-OCC [19]	ResNet50	256 × 704	40.69
FB-OCC* [19]	ResNet50	256 × 704	42.06
Ours	ResNet50	256 × 704	**42.37**
BEVFormer [18]	ResNet101	900 × 1600	40.6
VoxFormer [17]	ResNet101	900 × 1600	40.6
Ours	ResNet101	640 × 1600	**46.0**
SurroundOcc [41]	InterImage-B	640 × 1600	40.7
BEVFormer [18]	InterImage-XL	640 × 1600	43.3
BEVDet [10]	Swin-B	640 × 1600	43.1
UniOCC [29]	ConvNext-B	640 × 1600	51.5
FB-OCC* [19]	InternImage-H	960 × 1769	52.79
Ours	ConvNext-B	640 × 1600	**53.01**

Table 3. Comparsion on the Occ3D-nuScenes val set. The superscript * denotes using test time augmentation.

5.4. Ablation Study

In this section, we study the proposed strategies and the impact of GT quality.

Methods	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25$ ↑
SurroundOcc	0.274	2.072	5.327	0.482
FB-OCC	0.342	2.047	5.970	0.290
SurroundDepth	0.224	2.102	4.573	**0.751**
R3D3	0.259	2.328	5.577	0.583
Ours	**0.174**	**1.097**	**4.402**	0.747

Table 4. 3D scene reconstruction performance. AbsRel is the main metric. "↓" means less is better.

Ablation Study on Proposed Strategies We conduct ablation experiments based on the settings of the *Small* version in Tabel 1 and the results are shown in Tabel 5. These studies aim to analyze the effectiveness of different 3D representations and supervisions. Specifically, we first construct a baseline model (model A) that is supervised by the occupancy GT with the cross-entropy loss and output the occupancy results. Then, to investigate the 3D perception capability of implicit SDF and semantic field, we build the subsequent models based on the architecture in Section 4.1 and output the SDF and semantic field. To analyze the proposed weak supervision paradigm with previous methods, we reproduce the supervision methods of SIREN (model B) and LODE (model C). In model D, we employ our Sandwich Eikonal for SDF supervision. Lastly, model E incorporates the joint supervision strategy in section 4.2.

Comparing the results of Model B, and C with A, we observe a significant advantage in the AbsRel metric for both models B and C. This highlights the robust reconstruction capability of SDF-based 3D representation. However, we also observe a significant decline in the semantic mIoU metric for model B, attributed to the sparse supervision with LiDAR points. LODE mitigates this issue by introducing occupancy supervision, but its mIoU score is still lower than the baseline. Using the proposed Sandwich Eikonal supervision paradigm, the results of model D not only surpass the baseline in semantic mIoU but also outperform models B and C in AbsRel. This indicates the effectiveness of our supervision approach, which successfully integrates the supervision from LiDAR points and occupancy GT. Furthermore, by incorporating of the proposed joint supervision strategy, model E exhibits a 1.85 improvement in semantic mIoU, accompanied by a slight regression of 0.003 in the AbsRel metric. For some long-tail categories, such as bicycle, motorcycle, and traffic cone, our joint supervision strategy yields a noticeable improvement.

Ablation Study on Ground Truth Quality To further validate the roles of OCC GT and LiDAR GT, we conducted ablation experiments on the resolution of OCC GT and the density of LiDAR points respectively. As is shown in Table 6, the impact of OCC GT resolution variations in our method's OCC GT is mitigated with LiDAR points supervision. When the voxel grid resolution increased from 0.4m to 1.6m, the mIoU only decreased by 5.2%, and Abs-Rel increased by a mere 4.0%. In comparison, the SOTA OCC

Model	3D Representation	SDF Supervision	Joint Sup	others	barrier	bicycle	bus	car	vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation	mIoU	Abs-Rel
A	OCC	—	—	8.2	47.0	7.6	43.6	51.4	25.6	18.4	21.7	22.4	32.2	38.7	83.6	43.0	55.3	57.7	50.6	44.6	38.33	0.342
B	SDF	SIREN	—	8.0	41.4	18.6	25.9	41.0	19.5	19.0	15.6	23.0	19.8	27.8	72.3	38.6	48.7	46.5	36.6	33.6	31.51	0.188
C	SDF	LODE	—	8.7	46.0	18.6	38.5	47.2	25.3	22.2	20.4	26.3	23.7	34.5	81.2	44.8	54.0	53.8	40.3	37.8	36.66	0.192
D	SDF	Sandwich	—	12.3	47.9	25.4	43.0	51.7	27.4	24.9	24.9	27.4	34.0	39.4	83.6	42.9	54.4	58.1	48.3	43.2	40.52	**0.171**
E	SDF	Sandwich	✓	13.9	49.7	27.8	44.6	53.0	30.0	29.0	28.3	31.1	35.8	41.2	83.6	44.6	55.3	58.9	49.6	43.8	**42.37**	0.174

Table 5. Results of the ablation experiments. "3D Representation" indicates whether the scene is described by occupancy voxels or SDF. "SDF Supervision" indicates the supervised paradigm. "Joint Sup" indicates whether the joint supervision strategy in Section 4.2 is used.

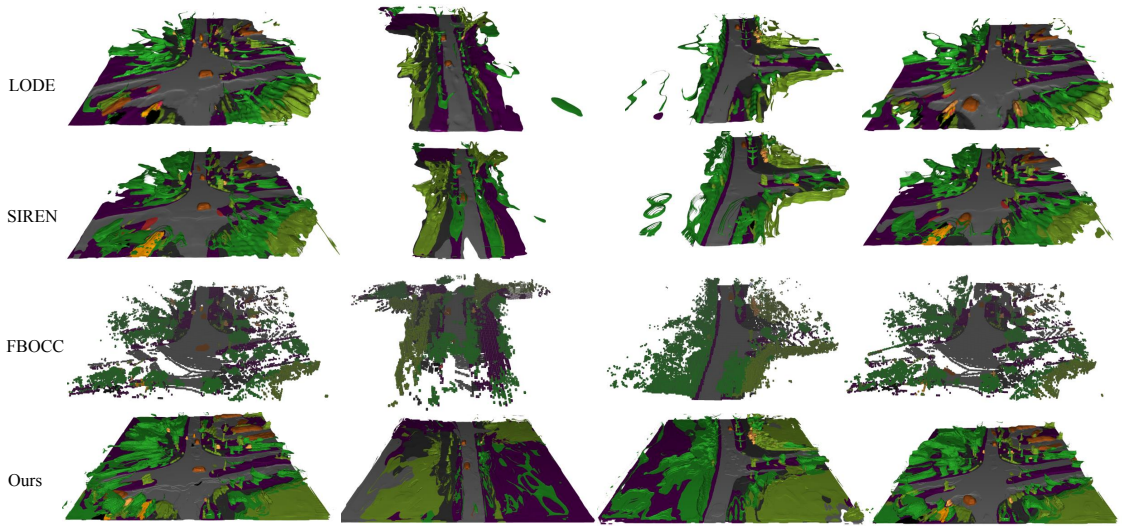


Figure 8. Comparison of the 3D scene perspective results with SOTA methods.

OCC GT size	0.4m		0.8m		1.6m	
Target	mIoU	Abs-Rel	mIoU	Abs-Rel	mIoU	Abs-Rel
FBOCC	40.69	0.342	36.01	0.370	26.96	0.445
SurroundSDF	42.37	0.174	40.74	0.176	40.16	0.181

Table 6. Results of the experiments on the resolution of OCC GT.

Sampling rate	1	0.5	0.2	0.1
mIoU	42.37	42.37	42.16	41.07
Abs-Rel	0.174	0.170	0.186	0.186

Table 7. Ablation study on the density of LiDAR points GT.

prediction method FBOCC experienced a substantial mIoU decrease of 33.7%, coupled with a 30.1% increase in Abs-Rel. As shown in Table 7, utilizing only 10% of the point cloud resulted in a marginal decrease of 3.1% in mIoU and a modest increase of 6.9% in Abs-Rel. These experiments further substantiate the robustness and efficacy of the proposed Sandwich Eikonal supervision paradigm.

5.5. Comparison of Visual Results

We visualize our 3D scene perception results and compare them with the reproduced SIREN [34] and LODE [14] (models B and C in Section 5.4), and the SOTA occupancy prediction method FBOCC [19, 20], as illustrated in Figure 8. Our SurroundSDF demonstrates significant advantages in the continuity and accuracy of the geometric structure.

6. Conclusion

In this work, we propose SurroundSDF, a novel vision-centric 3D scene understanding framework. We introduce SDF to address the continuity and accuracy of perception from surround cameras. Moreover, in the absence of SDF GT, we propose Sandwich Eikonal formulation, a novel SDF supervision paradigm, which emphasizes imposing appropriate constraints on both sides of the surface to enhance geometric accuracy. Furthermore, to alleviate the inconsistency between geometric optimization objectives and semantic optimization objectives, we introduce a joint supervision strategy. Based on the generated SDF and semantic field, different 3D representations can be obtained, including the scene mesh, occupancy voxel grids, and semantic reconstruction mesh of the whole scene. Comprehensive experiments validate the performance of our method.

References

[1] Ben Agro, Quinlan Sykora, Sergio Casas, and Raquel Urtasun. Implicit occupancy flow fields for perception and prediction in self-driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1379–1388, 2023. 2

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6

[3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2

[4] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations*, 2019. 1

[7] 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. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2, 7

[11] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. 2022. 1, 2

[12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 1, 2

[13] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 6

[14] Pengfei Li, Ruowen Zhao, Yongliang Shi, Hao Zhao, Jirui Yuan, Guyue Zhou, and Ya-Qin Zhang. Lode: Locally conditioned eikonal implicit scene completion from sparse lidar. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 8269–8276. IEEE, 2023. 2, 3, 8

[15] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 2, 6, 7

[16] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1, 2

[17] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1, 2, 7

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 7

[19] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 6, 7, 8

[20] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 2, 7, 8

[21] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 1, 2

[22] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[24] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 2

[25] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. 1, 2

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

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016. 5

[28] Timothy S Newman and Hong Yi. A survey of the marching cubes algorithm. *Computers & Graphics*, 30(5):854–879, 2006. 5

[29] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shaoqing Xu, Zhiyi Lai, Shanghang Zhang, Kuiyuan Yang, and Xiaomi Car. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. 7

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

[31] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 4

[32] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, 2020. 2

[33] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3216–3226, 2023. 1, 7

[34] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 1, 2, 3, 8

[35] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 1, 2, 3, 5, 6

[36] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 3, 4, 5, 6

[37] Thang Vu, Jung-Hee Kim, Myeongjin Kim, Seokwoo Jung, and Seong-Gyun Jeong. Milo: Multi-task learning with localization ambiguity suppression for occupancy prediction cvpr 2023 occupancy challenge report. 7

[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. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 2

[39] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, , and Justin M. Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *The Conference on Robot Learning (CoRL)*, 2021. 2

[40] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on Robot Learning*, pages 539–549. PMLR, 2023. 7

[41] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 2, 7

[42] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M ^{2} bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 2

[43] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. In *The Eleventh International Conference on Learning Representations*, 2022. 2

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

[45] Weihao Yuan, Xiaodong Gu, Heng Li, Zilong Dong, and Siyu Zhu. Monocular scene reconstruction with 3d sdf transformers. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[46] Yunpeng Zhang, Phigent Robotics, and Zheng Zhu. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9433–9443, 2023. 1, 2