2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

LODE: Locally Conditioned Eikonal Implicit Scene Completion from Sparse LiDAR

Pengfei Li[1,2], Ruowen Zhao[1,3], Yongliang Shi[1], Hao Zhao[1], Jirui Yuan[1], Guyue Zhou[1✉] and Ya-Qin Zhang[1]

***Abstract*— Scene completion refers to obtaining dense scene representation from an incomplete perception of complex 3D scenes. This helps robots detect multi-scale obstacles and analyse object occlusions in scenarios such as autonomous driving. Recent advances show that implicit representation learning can be leveraged for continuous scene completion and achieved through physical constraints like Eikonal equations. However, former Eikonal completion methods only demonstrate results on watertight meshes at a scale of tens of meshes. None of them are successfully done for non-watertight LiDAR point clouds of open large scenes at a scale of thousands of scenes. In this paper, we propose a novel Eikonal formulation that conditions the implicit representation on localized shape priors which function as dense boundary value constraints, and demonstrate it works on SemanticKITTI and SemanticPOSS. It can also be extended to semantic Eikonal scene completion with only small modifications to the network architecture. With extensive quantitative and qualitative results, we demonstrate the benefits and drawbacks of existing Eikonal methods, which naturally leads to the new locally conditioned formulation. Notably, we improve IoU from 31.7% to 51.2% on SemanticKITTI and from 40.5% to 48.7% on SemanticPOSS. We extensively ablate our methods and demonstrate that the proposed formulation is robust to a wide spectrum of implementation hyper-parameters. Codes and models are publicly available at `https://github.com/AIR-DISCOVER/LODE`.**

I. Introduction

Representing 3D data with neural implicit functions is actively explored recently due to its strong modeling capability and memory efficiency [1]–[7]. Meanwhile, it can be easily meshed and rendered to facilitate human viewing. While most methods are fully supervised [1]–[3], SIREN [6] proposes an Eikonal implicit scene completion method with weak supervision needed. It learns a signed distance function (SDF), which measures the nearest distance to the scene surface, through the process of solving an Eikonal differential equation with only points on the surface and without knowing SDF values in free space. This scheme is promising for large-scale sparse LiDAR data because (1) for non-water-tight scenes, it is hard to define signed distance in free space and (2) it requires less supervision than non-Eikonal formulations and thus is simpler, especially when considering the completion of thousands of scenes.

However, even after an exhaustive parameter search, SIREN fails to fit sparse LiDAR data (Fig. 1-a), which limits its application in many important scenarios such as autonomous driving. This is understandable as SIREN is a pure generative model and LiDAR point clouds are extremely sparse. Specifically, the reasons are three-fold: (1) The sparsity of on-surface points amplifies the negative impact of wrongly sampled off-surface anchors. (2) The normal orientations of sparse points cannot be estimated accurately from their neighbors, which serve as a necessary boundary value constraint for SIREN fitting. (3) Without trustworthy boundary values, enforcing a hard Eikonal constraint leads to even inaccurate SDF values. As shown in Fig. 1-b, the SIREN fitting result is fragmented.

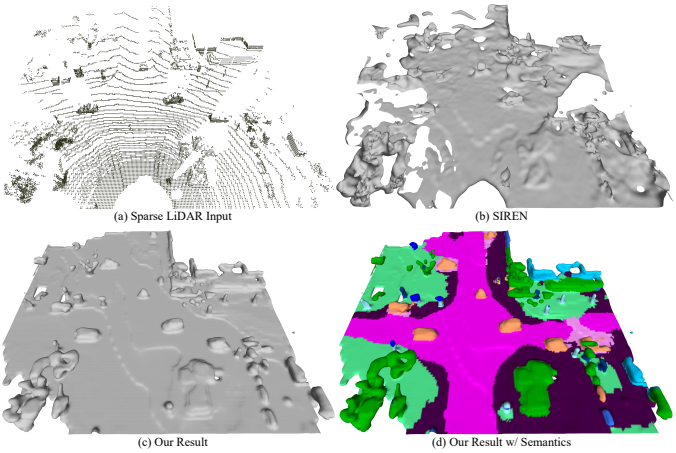


Fig. 1. (a) The input is a sparsity-variant point cloud of road scenes captured by LiDAR. (b) The implicit fitting result of SIREN [6]. Note that this is well tuned by an exhaustive parameter search. (c) The output of our method is a neural signed distance function of arbitrary resolution, i.e., implicit scene completion. (d) Our result with extended semantic parsing.

To overcome these limitations, we develop a novel Eikonal implicit formulation by introducing an intermediate embedding domain, where localized shape priors are contained. Instead of directly fitting a function to map 3D Cartesian coordinates to signed distances, we first map the Euclidean space to a corresponding high-dimensional shape embedding space, and then the signed distance space. These shape embeddings function as dense boundary values that entangle both zeroth-order (on-surface points) and first-order (normal directions) constraints, in a data-driven manner. Naturally, the issue of enforcing a hard Eikonal constraint is also alleviated. This proposed formulation is named **L**ocally **Cond**itioned **E**ikonal Formulation and abbreviated as **LODE**. The result of LODE is significantly better than SIREN (Fig. 1-c). And the supplementary video demonstrates LODE performs well on in-the-wild sequential LIDAR inputs.

[1] Institute for AI Industry Research (AIR), Tsinghua University, China {Shiyongliang,zhaohao,yuanjirui,zhouguyue,zhangyaqin}@air.tsinghua.edu.cn.
[2]Department of Computer Science and Technology, Tsinghua University, China, li-pf22@mails.tsinghua.edu.cn.
[3] University of Chinese Academy of Sciences, China zhaorewen20@mails.ucas.ac.cn.

Specifically, to implement LODE, we propose a novel hybrid architecture combining a discriminative model and a generative model. The discriminative part of our method exploits the strong representation learning power of sparse convolution, generating latent shape embeddings from sparse point cloud input. The generative model takes as input the ground truth point cloud coordinates along with pointwise latent shape embeddings retrieved by trilinear sampling and predicts SDF values of these points. During inference, the ground truth points are replaced with the points of interest.

Furthermore, to demonstrate the flexibility of LODE, we extend our method to implicit semantic completion in two ways: (1) by adding a dense discriminative head to predict semantic labels which can be mapped to the implicit function using K-Nearest-Neighbors; (2) by adding a parallel implicit generative head to directly model the implicit semantic field. We evaluate them on SemanticKITTI and achieve results (Fig. 1-d) comparable to state-of-the-art methods.

To summarize, our contributions are as follows:

- We develop a locally conditioned Eikonal implicit scene completion formulation that incorporates learned shape priors as dense boundary value constraints.
- We apply the formulation in road scene understanding, leading to the first Eikonal implicit road scene completion method without knowing SDF values in free space.
- We achieve state-of-the-art completion results on SemanticKITTI and SemanticPOSS, outperforming the best Eikonal completion results by +19.5% and +8.2% IoU. Code, data, and models will be released.

II. Related Works

Neural Implicit Representation. The general principle of neural implicit representation is to train a neural network to approximate a continuous function that is hard to parameterize otherwise. [1] proposes to learn deep signed distance functions conditioned on shape embeddings. [2] approximates occupancy functions with conditional batch-norm networks. [8] introduces data-driven shape embeddings into occupancy networks for indoor scene completion. [3] uses hyperplanes as compact implicit representations to reconstruct shapes sharply and compactly. [6] shows that using gradient supervision allows Eikonal SDF learning with only on-surface SDF value supervision and sine activations are critical to its success. [9] combines Gaussian ellipsoids and implicit residuals to represent shapes accurately. Some recent works exploit 3D implicit representations for instance-level understanding from point cloud [10] or RGB [11] inputs. Despite these advances, there is no work yet on implicit scene completion on LiDAR point clouds where the data is extremely sparse with heterogeneous distribution. Our method bridges this gap with the proposed locally conditioned Eikonal formulation.

LiDAR-based scene understanding. While there are many advances in camera-based cognitive scene understanding [12]–[16], LiDAR point cloud provides reliably accurate 3D structural information and thus has been mainly leveraged in geometric scene understanding. Numerous LiDAR-based SLAM methods are proposed to improve the quality and real-time performance [17]–[21]. [22]–[24] further incorporate semantic understanding into LiDAR SLAM systems. These methods explicitly complete a scene with multiple frames of LiDAR data, while our goal is to achieve implicit completion with a single frame. Sparse LiDAR points are also efficiently leveraged in depth completion [25], [26], object detection [27], [28], segmentation [29]–[34]. We believe the performance of these methods will be boosted with LODE completing the original sparse representation. Moreover, by providing fine geometric details, LODE may aid in the detection of anomalous obstacles [35] and some other tasks.

III. Formulation

A. Eikonal Implicit Completion Formulation

Eikonal completion methods aim at fitting the signed distance function (SDF) of a scene. The signed distance is the nearest distance from a point of interest to the scene surface, with the sign denoting whether the point is located outside (positive) or inside (negative) of the surface. The iso-surface where the signed distance equals zero implicitly delineates the scene. Formally, the goal of Eikonal implicit completion is to find a function $\Phi(\mathbf{x})$, which satisfies a set of M constraints $\mathcal{C}_m$, to approximate the underlying SDF. Each constraint relates the function $\Phi(\mathbf{x})$ or its gradient to certain input quantities $\mathbf{a}(\mathbf{x})$ on the corresponding domain Ω_m:

$$\begin{aligned} \mathcal{C}_m(\mathbf{a}(\mathbf{x}),&\Phi(\mathbf{x}), \nabla_{\mathbf{x}}\Phi(\mathbf{x})) = 0, \\ &\forall \mathbf{x} \in \Omega_m, m = 0, ..., M-1. \end{aligned} \tag{1}$$

Specifically, these constraints are required:

$$\mathcal{C}_0 := |\nabla_{\mathbf{x}}\Phi(\mathbf{x})| - 1, \mathbf{x} \in \Omega_0. \tag{2}$$

$$\mathcal{C}_1 := \nabla_{\mathbf{x}}\Phi(\mathbf{x}) - \mathbf{n}(\mathbf{x}), \mathbf{x} \in \Omega_1. \tag{3}$$

$$\mathcal{C}_2 := \Phi(\mathbf{x}) - \mathrm{SDF}(\mathbf{x}), \mathbf{x} \in \Omega_2. \tag{4}$$

Here, $\mathcal{C}_0$ guarantees $\Phi(\mathbf{x})$ satisfies the Eikonal equation in the whole physical space of interest (Ω_0), which is a intrinsic property of SDF. $\mathcal{C}_1$ forces that the gradients of $\Phi(\mathbf{x})$ equal the normal vectors for input on-surface points (Ω_1). $\mathcal{C}_2$ constrains the values of $\Phi(\mathbf{x})$ equal the ground truth SDF for labeled anchor points (Ω_2). In this way, the problem can be regarded as an Eikonal boundary value problem, where the differential equation $\mathcal{C}_0$ is to be solved under the first-order constraint $\mathcal{C}_1$ and the zeroth-order constraint $\mathcal{C}_2$.

However, the ground truth SDF values in free space are difficult to obtain. A recent method named SIREN [6] proposes an intriguing variant where the domain of $\mathcal{C}_2$ is limited to on-surface points in Ω_1. As the ground truth SDF values of points in Ω_1 are zero, $\mathcal{C}_2$ is reduced to:

$$\mathcal{C}_2 := \Phi(\mathbf{x}), \mathbf{x} \in \Omega_1. \tag{5}$$

To remedy the lack of constraints on off-surface points, SIREN introduces another constraint:

$$\mathcal{C}_3 := \psi(\Phi(\mathbf{x})), \mathbf{x} \in \Omega_3. \tag{6}$$

Here, ψ pushes $\Phi(\mathbf{x})$ values away from 0, for randomly and uniformly sampled off-surface points ($\Omega_3 \subseteq \Omega_0 \setminus \Omega_1$).

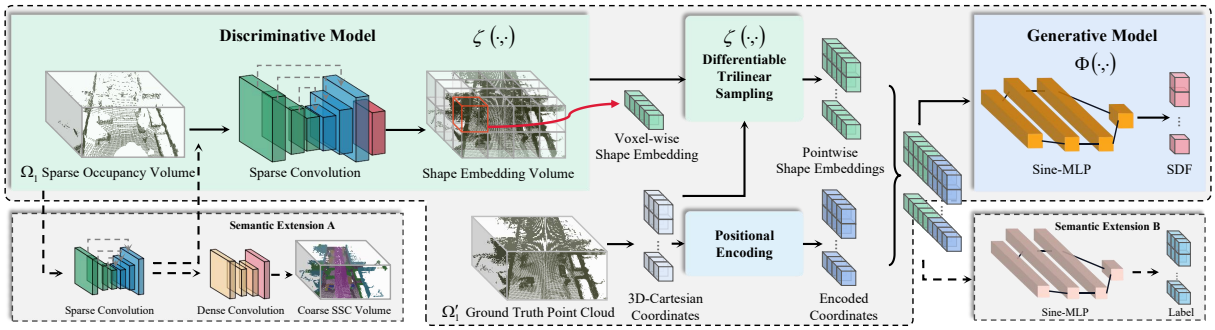


Fig. 2. **Overview of our architecture.** The discriminative model extracts shape priors and the generative model predicts SDF values. They are bridged by differentiable triliner sampling. Positional Encoding is used to represent more details. Two semantic extension options are outlined in small dashed boxes.

Nevertheless, these constraints fail to cope with the scenario where on-surface points in Ω_1 are sampled from sparse LiDAR point cloud data. Reasons are three-fold: (1) The sparsity of on-surface points in Ω_1 amplifies the negative impact of $\mathcal{C}_3$ on the wrongly sampled off-surface anchors in Ω_3 (i.e., located on or near the surface). (2) The normal orientations of sparse points in Ω_1 cannot be estimated accurately from their neighbors, leading to an incorrect constraint $\mathcal{C}_1$. (3) Without trustworthy boundary value constraints $\mathcal{C}_3$ and $\mathcal{C}_1$, enforcing the hard Eikonal constraint $\mathcal{C}_0$ leads to even inaccurate SDF values in free space.

B. Locally Conditioned Eikonal Formulation (LODE)

To overcome the aforementioned limitations, we propose a locally conditioned Eikonal formulation $\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}$ to approximate SDF. Here, we use $\zeta(\cdot,\cdot)$ to first map the Euclidean space to a high-dimensional shape embedding space. It functions as a dense boundary value constraint for the differential equation. Then $\Phi(\cdot,\cdot)$ maps the shape embedding space to the signed distance space. As a result, the constraints to be satisfied are formally re-written as:

$$\mathcal{C}_0' := |\nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}| - 1, \mathbf{x} \in \Omega_0. \tag{7}$$

$$\mathcal{C}_4 := \rho(\zeta(\mathbf{x},\Omega_1)), \mathbf{x} \in \Omega_0. \tag{8}$$

We use $\rho(\zeta(\mathbf{x},\Omega_1))$ to represent the underlying dense constraint contained in the shape embedding space, which implicitly entangles correct $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ constraints of the modified formulations:

$$\mathcal{C}_1' := \nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)} - \mathbf{n}(\mathbf{x}), \mathbf{x} \in \Omega_1'. \tag{9}$$

$$\mathcal{C}_2' := \Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}, \mathbf{x} \in \Omega_1'. \tag{10}$$

$$\mathcal{C}_3' := \psi(\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}), \mathbf{x} \in \Omega_3'. \tag{11}$$

Here, Ω_1' contains the dense ground truth on-surface points and $\Omega_3' \subseteq \Omega_0 \setminus \Omega_1'$. Hence the aforementioned problem of trustworthy boundary values is resolved. Naturally, the issue of enforcing a hard Eikonal constraint is also alleviated.

We implement the proposed LODE formulation in a data-driven manner. The acquisition of functions $\zeta(\cdot,\cdot)$ and $\Phi(\cdot,\cdot)$ can be cast in a loss function that penalizes deviations from the constraints $\mathcal{C}_0'$, $\mathcal{C}_1'$, $\mathcal{C}_2'$, and $\mathcal{C}_3'$ on their domain:

$$\begin{aligned}\mathcal{L}_{\text{LODE}} = \lambda_1 &\int_{\Omega_0} \big\| |\nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}| - 1 \big\| \, d\mathbf{x} \\ + \lambda_2 &\int_{\Omega_1'} (1 - \langle \nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}, \mathbf{n}(\mathbf{x}) \rangle) d\mathbf{x} \\ + \lambda_3 &\int_{\Omega_1'} \big\| \Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)} \big\| \, d\mathbf{x} \\ + \lambda_4 &\int_{\Omega_3'} \psi(\Phi(\mathbf{x},\mathbf{e})|_{\mathbf{e}=\zeta(\mathbf{x},\Omega_1)}) d\mathbf{x},\end{aligned} \tag{12}$$

where λ_1 - λ_4 are constant weight parameters and $\langle \cdot,\cdot \rangle$ calculates cosine similarity between two vectors.

IV. Method

To realize LODE, we propose a hybrid neural network architecture combining a discriminative model with a generative model, as shown in Fig. 2. The discriminative part exploits the strong representation learning power of sparse convolution, generating latent shape embeddings from sparse input Ω_1. It together with the differentiable trilinear sampling module works as function $\zeta(\cdot,\cdot)$. The generative model consists of an MLP, functioning as $\Phi(\cdot,\cdot)$. It takes as input the encoded coordinates of ground truth points Ω_1' along with pointwise latent shape embeddings and predicts SDF values of these points. Using gradient descent, we can get the optimized $\zeta(\cdot,\cdot)$ and $\Phi(\cdot,\cdot)$ in the parameterized form.

A. Discriminative Model

Intuitively, road scenes have the characteristic of repetition. Thus convolutional neural network can be employed as the discriminative model to exploit the translation invariance.

Taking LiDAR points Ω_1 as input, we first conduct voxelization to obtain 3D occupancy volume V_{occ} with size $1 \times D_{\text{occ}} \times W_{\text{occ}} \times H_{\text{occ}}$. Then the discriminative model maps it into a shape embedding volume V_{se} with size $d_{\text{se}} \times D_{\text{se}} \times W_{\text{se}} \times H_{\text{se}}$, where d_{se} is the dimension of the shape embedding outputs. To tackle the sparsity of V_{occ}, we employ the sparse operations of the Minkowski Engine [36] to build the model, which is a multiscale encoder-decoder network. It extracts shape priors via a shape completion process: the encoder consisting of convolutional blocks aggregates localized features, and the decoder involving generative deconvolutional blocks generates dense results.

Yet the constant generation of new voxels will destroy the sparsity just as the *submanifold dilation problem* [37]. To avoid this, we use a pruning block to prune off redundant voxels. It contains a convolutional layer to determine the binary classification result of whether a voxel should be pruned, which is supervised with binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{com}} = -\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}[y_{i,j}\log(p_{i,j}) \\ + (1-y_{i,j})\log(1-p_{i,j})],\end{aligned} \tag{13}$$

where m is the count of supervised blocks, n_i denotes the count of voxels in the i-th block, $y_{i,j}$ and $p_{i,j}$ are the true and predicted existence probabilities for voxel i respectively.

B. Differentiable Trilinear Sampling Module

After generating V_{se}, pointwise shape embedding $\mathbf{e}_i \in \mathbb{R}^{d_{\text{se}}}$ for query point $\mathbf{x}_i \in \Omega_0$ is needed. We use trilinear interpolation to sample $\mathbf{e}_i$ for $\mathbf{x}_i$ from its 8 nearest voxel centers to maintain the continuity of the latent shape field at the voxel borders. Formally, with the length of voxel edge normalized, the trilinear sampling for $\mathbf{e}_i$ can be written as:

$$\begin{aligned}e_i^c = \sum_{m}^{D_{\text{se}}}\sum_{n}^{W_{\text{se}}}\sum_{k}^{H_{\text{se}}} e_{mnk}^c \times \max(0, 1-|x_i - x_m|) \\ \times \max(0, 1-|y_i - y_n|) \times \max(0, 1-|z_i - z_k|),\end{aligned} \tag{14}$$

where e_i^c and e_{mnk}^c are shape embeddings on channel c for $\mathbf{x}_i = (x_i, y_i, z_i)$ and voxel center $\mathbf{x}_{mnk} = (x_m, y_n, z_k)$. Then the gradient with respect to $\mathbf{e}_{mnk}$ for backpropagation is:

$$\begin{aligned}\frac{\partial e_i^c}{\partial e_{mnk}^c} = \sum_{m}^{D_{\text{se}}}\sum_{n}^{W_{\text{se}}}\sum_{k}^{H_{\text{se}}} \max(0, 1-|x_i - x_m|) \\ \times \max(0, 1-|y_i - y_n|) \times \max(0, 1-|z_i - z_k|).\end{aligned} \tag{15}$$

This differentiable trilinear sampling mechanism allows loss gradients to flow back to V_{se} and further back to the discriminative model, making it possible to train discriminative model and the following generative model cooperatively.

C. Positional Encoding Module

Positional encoding has proved an effective technique in neural rendering [4] [38] for its capacity to capture high-frequency information. Thus we leverage it to represent more geometric details of the signed distance field. Specifically, the 3D Cartesian coordinate $\mathbf{x}_i$ is encoded into high-dimensional feature $\mathbf{y}_i = (\gamma_{\text{enc}}(x_i), \gamma_{\text{enc}}(y_i), \gamma_{\text{enc}}(z_i)) \in \mathbb{R}^{d_{\text{enc}}}$, where

$$\begin{aligned}\gamma_{\text{enc}}(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \cdots, \\ \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)).\end{aligned} \tag{16}$$

L is the number of frequency octaves and thus $d_{\text{enc}} = 6L$.

D. Generative Model

We use an MLP as the generative model for implicit SDF representation and use sine as a periodic activation function to better model details [6]. Thus Φ can be formalized as:

$$\begin{aligned}\Phi(\mathbf{x}) = &\mathbf{W}_n(\phi_{n-1} \circ \phi_{n-2} \circ \cdots \circ \phi_0)(\mathbf{x}) + \mathbf{b}_n, \\ &\mathbf{x}_j \mapsto \phi_j(\mathbf{x}_j) = \sin(\mathbf{W}_j\mathbf{x}_j + \mathbf{b}_j),\end{aligned} \tag{17}$$

where $\phi_j : \mathbb{R}^{M_j} \mapsto \mathbb{R}^{N_j}$ is the j^{th} layer of the model. Given $\mathbf{x}_j \in \mathbb{R}^{M_j}$, the layer applies the affine transform with weights $\mathbf{W}_j \in \mathbb{R}^{N_j \times M_j}$ and biases $\mathbf{b}_j \in \mathbb{R}^{N_j}$ on it, and then pass the resulting vector to the sine nonlinearity.

In our implementation, the concatenated vector $[\mathbf{y}_i, \mathbf{e}_i]$ is taken as input. And model weights are shared for all scenes. We use the proposed loss function (12) to optimize model weights and shape embeddings. Note that during training, $\mathbf{x}_i$ is sampled from dense ground truth Ω_1' instead of sparse input Ω_1. Thus, in a data-driven manner, our generative model can effectively map the shape embedding space to the signed distance space with abundant geometric information.

E. Semantic Extension

To demonstrate the flexibility of LODE, we extend our method to implicit semantic completion in two ways.

Semantic Extension A. We first semantically segment V_{occ} with a sparse CNN. Then a dense CNN is used to predict coarse semantic completion results. Mapping it to our implicit representation using K-Nearest-Neighbor, we get the refined implicit semantic results.

Semantic Extension B. We add a parallel implicit generative head to directly model the implicit semantic label field. Its structure is similar to aforementioned generative model, except that it outputs the probabilities of label classification.

The results are supervised with a cross-entropy loss:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N_{\text{seg}}}\sum_{i=1}^{N_{\text{seg}}}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}), \tag{18}$$

where $y_{i,c}$ and $p_{i,c}$ are the true and predicted probabilities. N_{seg} points and C categories are considered.

F. Training and Inference

During training, we randomly sample N_{on} on-surface points from Ω_1' and N_{off} off-surface points from Ω_3', optimizing the whole neural network with loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LODE}} + \lambda_5\mathcal{L}_{\text{com}} + \lambda_6\mathcal{L}_{\text{seg}}. \tag{19}$$

Here, λ_5 and λ_6 are constant weight parameters. Note that $\lambda_6 = 0$ when the semantic extension is not included.

During inference, we uniformly sample N_{inf}^3 points from Ω_0 at a specified resolution. And we use a threshold v_{th} close to zero to select the points with estimated SDF values smaller than v_{th} as explicit surface points for evaluation.

V. Experiments

Dataset. We evaluate the proposed LODE on SemanticKITTI [39] and SemanticPOSS [40]. There are 22 sequences (8550 scans) and 6 sequences (2988 scans) of road scene LiDAR data in the two datasets respectively. Each scan covers a range of 51.2m ahead of the LiDAR, 25.6m to each side, and 6.4m in height. We follow the official split for training and validation. **Metric.** We use the interactions over union (IoU) metric for evaluation. **Implementation Details.** For the discriminative model, we set $D_{\text{occ}} = 256$, $W_{\text{occ}} = 256$, $H_{\text{occ}} = 32$, $m = 5$. For the generative model,

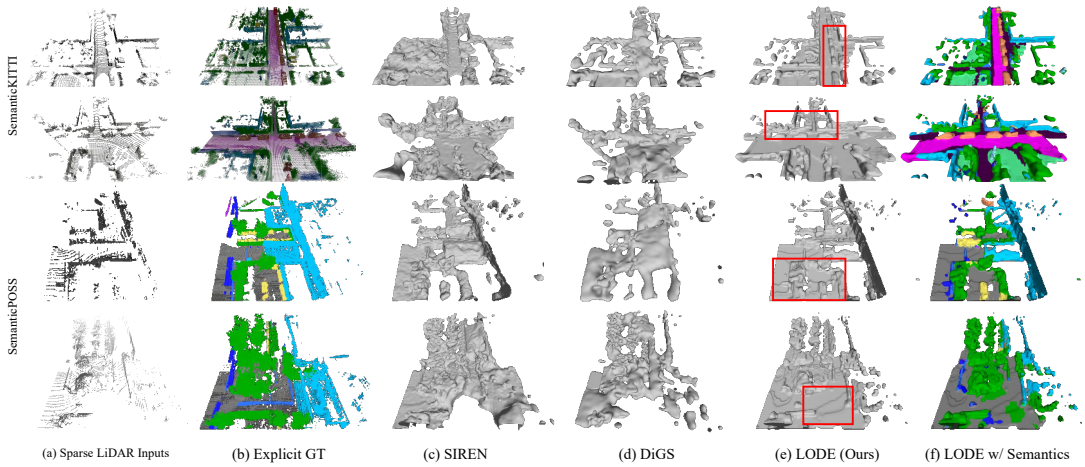


Fig. 3. Qualitative results of Eikonal implicit road scene completion on the SemanticKITTI and SemanticPOSS validation set.

we use $N_{\text{on}} = N_{\text{off}} = 16000$ and $N_{\text{inf}} = 256$. For training, we set $\lambda_1 = 3000$, $\lambda_2 = 100$, $\lambda_3 = 100$, $\lambda_4 = 50$, $\lambda_5 = 100$ and use the Adam optimizer with an initial learning rate of 10^{-4}. When the semantic extension is included, $\lambda_6 = 50$.

A. Scene Completion Effectiveness of LODE

We compare our LODE with other recent strong Eikonal completion methods on the validation sets of SemanticKITTI and SemanticPOSS. The results of these Eikonal methods are obtained by applying them to every LiDAR sweep and taking the average. In Table.I, the first row shows that directly comparing the input sparse point cloud with completion ground truth yields 10.3% and 13.0% IoU. The existing Eikonal methods improve the IoU to 31.7% and 40.5%. Thanks to the new locally conditioned Eikonal formulation, our approach further improves IoU to 51.2% and 48.7%. These results demonstrate the effectiveness of LODE.

TABLE I
SCENE COMPLETION RESULTS MEASURED IN IOU (%).

Method	Reference	SemanticKITTI	SemanticPOSS
Input	-	10.3	13.0
SIREN [6]	NeurIPS 2020	26.3	36.0
Fourier Features [41]	NeurIPS 2020	28.6	30.9
BACON [42]	CVPR 2022	30.5	40.5
DiGS [43]	CVPR 2022	31.7	37.4
LODE (Ours)	-	**51.2** (+19.5)	**48.7** (+8.2)

This large improvement is better demonstrated with qualitative results in Fig. 3. Though the existing Eikonal methods are successful for clean synthetic point cloud data uniformly sampled on watertight meshes as shown in [6], fitting large-scale outdoor scenes captured by LiDAR (Fig.3-a) is much more difficult. On the one hand, many regions are not sampled thus missing in the point cloud. On the other hand, caused by the mechanism of LiDAR, data sparsity increases with distance and it is extremely sparse at the far end. These result in the lack of effective boundary values for solving the Eikonal differential equation. For this reason, as pure generative models, these methods fail to fit road scenes and produce lots of artifacts (Fig.3-c,d). Our LODE, on the contrary, takes data-driven shape priors generated by a strong sparse convolutional network as the dense boundary values and successfully completes the scenes. As shown in Fig.3-e and highlighted in red boxes, both occluded and incomplete regions are better reconstructed than other Eikonal methods.

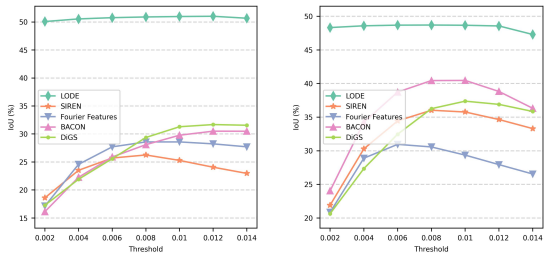


Fig. 4. IoU comparisons under different thresholds.

In order to show that the significant margins reported in Table. I are robust to Marching Cubes thresholds, we provide an exhaustive evaluation in Fig.4. It is clear that our method outperforms other methods under all inspected thresholds.

B. Implementation Robustness of LODE

To better understand LODE and demonstrate its robustness to hyper-parameters in implementation, we provide a series of ablation studies on SemanticKITTI as follows.

Discriminative model design. We investigate two factors: (1) Where to add pruning blocks; (2) Conv layer number in the output block that generates shape embeddings. As shown in Table.II, LODE is robust to these design choices.

Does generative model capacity matter? Deeper and wider models usually achieve better results for recognition. To explore whether generative model capacity matters for LODE, we ablate the width, depth, and activations of the MLP. As shown in Table.III, different configurations produce similar results. It demonstrates the capacity of generative model is not a performance bottleneck. Interestingly, using ReLU instead of Sine activation only brings a performance drop of 1.75%. It suggests that in challenging scenarios like ours, using Sine activation is not as critical as in SIREN [6].

Which dimension of shape volume matters? To study which factor is the deciding one for the representation power of the shape volume, we evaluate different shape embedding dimensions and scale sizes. By scale size, we mean the down-sampling ratio $\frac{D_{\text{occ}}}{D_{\text{se}}}$. The results are summarized in

Table.IV, showing that using shape embeddings of dimension 128 is already capable of representing our scenes well. But increasing scale size leads to a sharp drop of IoU, which reflects the importance of the locality of shape priors.

TABLE II

DISCRIMINATIVE MODEL.

Pruning Blocks	Output Block	IoU (%)
Last 1	2 convs	49.5
Last 2	2 convs	49.1
Last 3	2 convs	50.6
Last 4	2 convs	51.0
All	2 convs	**51.1**
All	4 convs	50.9

TABLE IV

SHAPE EMBEDDING.

Shape Dimension	Scale Size	IoU (%)
128	4	50.9
512	4	**51.2**
256	2	50.3
256	4	51.0
256	8	49.2
256	16	44.8

TABLE III

GENERATIVE MODEL.

Width	Depth	Activations	IoU (%)
128	4	Sine	**51.0**
256	4	Sine	51.0
512	4	Sine	50.9
256	3	Sine	49.6
256	5	Sine	50.9
256	4	ReLU	49.3

TABLE V

POSITIONAL ENCODING.

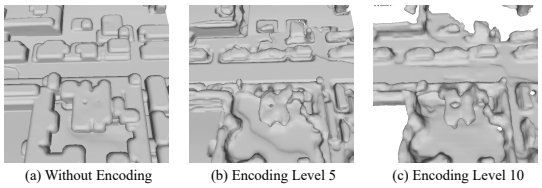
Positional Encoding	Include xyz	Encoding Level L	IoU (%)
×	-	-	40.4
✓	✓	5	40.3
✓	✓	10	51.0
✓	✓	15	50.9
✓	×	10	**51.1**

TABLE VI

SAMPLING STRATEGY.

Sample Strategy	IoU (%)
Trilinear	**51.0**
Nearest	48.1

Is trilinear sampling necessary? We justify the necessity of trilinear sampling in our method using Table. VI. A trivial nearest neighbor sampling leads to a performance drop of 2.9%. This is a clear margin that shows the benefit of smoothly interpolating shape embeddings.



(a) Without Encoding (b) Encoding Level 5 (c) Encoding Level 10

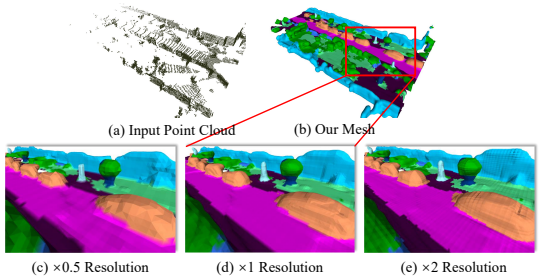
Fig. 5. Qualitative results with different positional encoding strategies.

How to encode positional information? The goal of positional encoding is to represent fine geometric details of the scene. We investigate positional encoding levels and whether to concatenate original coordinates. As shown in Table.V, when positional encoding is not used or the encoding level is low, the completion IoU decreases dramatically. Through the qualitative results in Fig. 5-a, it is clear that leaving out positional encoding leads to the loss of details.

With these ablations and analyses, we demonstrate the role of each module and the impact of hyper-parameters in detail. Meanwhile, despite the performance fluctuation under different hyper-parameters, our quantitative results are always better than the existing Eikonal methods shown in the second to the fifth row of Table.I, which demonstrates the effectiveness and robustness of LODE.

C. Flexibility of LODE

Table.VII shows semantic scene completion results on the SemanticKITTI validation set, which is evaluated on 19 categories. With little impact on original completion



(a) Input Point Cloud (b) Our Mesh

(c) ×0.5 Resolution (d) ×1 Resolution (e) ×2 Resolution

Fig. 6. Scene completion results at multiple resolutions.

results, the semantic extension A and B achieve 20.2% and 18.0% mIoU, respectively. Although they under-perform the state-of-the-art method JS3C-Net [44], our models allow implicit completion and models the signed distance field. Qualitative results shown in Fig. 3-f demonstrate faithful semantic implicit completion. Last but not least, we map explicit semantic completion results from JS3C-Net to our implicit completion using K-Nearest-Neighbors, achieving 23.4% mIoU. These results show the flexibility of LODE as it can be easily extended to provide semantic information.

TABLE VII

SEMANTIC SCENE COMPLETION RESULTS ON SEMANTICKITTI.

Approach	ext.A	ext.B	JS3C	LODE w/ JS3C
mIoU (%)	20.2	18.0	22.7	**23.4**

D. Other Potential Benefits of LODE

As illustrated in Fig.6, with LODE, we can get mesh reconstructions at any resolution, which is due to its **continuous representation** of the signed distance field. Meanwhile, as shown in Fig.1 and Fig.3, LODE enables the indiscernible LiDAR data to be converted into **human-friendly visualizations**, which demonstrates its capacity to enhance human understanding of robot-perceived information. Moreover, the proposed LODE has **compatibility with implicit planning algorithms** [45]–[49], and thus can be leveraged to facilitate downstream robotic manipulation tasks.

VI. CONCLUSION

In this study, we propose a novel locally conditioned Eikonal formulation named LODE for implicit scene completion. Learned shape embeddings are treated as dense boundary values that constrain signed distance function learning. We implement the formulation as a hybrid neural network combining discriminant and generative models. The network is trained to implicitly fit road scenes captured by sparse LiDAR point clouds, without accessing exact SDF values in free space. Large-scale evaluations on SemanticKITTI and SemanticPOSS show that our method outperforms existing Eikonal methods by a large margin. We also extend the proposed method for semantic implicit completion in two ways, achieving strong qualitative and quantitative results.

ACKNOWLEDGEMENTS

This work was sponsored by Tsinghua-Toyota Joint Research Fund (20223930097) and Baidu Inc. through Apollo-AIR Joint Research Center.

REFERENCES

[1] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[2] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[3] Z. Chen, A. Tagliasacchi, and H. Zhang, "Bsp-net: Generating compact meshes via binary space partitioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 45–54.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[5] B. Lee, C. Zhang, Z. Huang, and D. D. Lee, "Online continuous mapping using gaussian process implicit surfaces," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6884–6890.

[6] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[7] Y. Duan, H. Zhu, H. Wang, L. Yi, R. Nevatia, and L. J. Guibas, "Curriculum deepsdf," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 2020, pp. 51–67.

[8] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 523–540.

[9] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser, "Local deep implicit functions for 3d shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4857–4866.

[10] Y. Nie, J. Hou, X. Han, and M. Nießner, "Rfd-net: Point scene understanding by semantic instance reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4608–4618.

[11] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, and S. Liu, "Holistic 3d scene understanding from a single image with implicit representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8833–8842.

[12] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler, "Learning to act properly: Predicting and explaining affordances from images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 975–983.

[13] X. Chen, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Cerberus transformer: Joint semantic, affordance and attribute parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 649–19 658.

[14] C. Ye, Y. Yang, R. Mao, C. Fermüller, and Y. Aloimonos, "What can i do around here? deep functional scene understanding for cognitive robots," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4604–4611.

[15] P. Li, B. Tian, Y. Shi, X. Chen, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Toist: Task oriented instance segmentation transformer with noun-pronoun distillation," *arXiv preprint arXiv:2210.10775*, 2022.

[16] Y. Li, Y. Tu, X. Chen, H. Zhao, and G. Zhou, "Distance-aware occlusion detection with focused attention," *IEEE Transactions on Image Processing*, vol. 31, pp. 5661–5676, 2022.

[17] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1271–1278.

[18] D. Droeschel and S. Behnke, "Efficient continuous-time slam for 3d lidar-based online mapping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5000–5007.

[19] I. Vizzo, X. Chen, N. Chebrolu, J. Behley, and C. Stachniss, "Poisson surface reconstruction for lidar odometry and mapping," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5624–5630.

[20] Y. Pan, P. Xiao, Y. He, Z. Shao, and Z. Li, "Mulls: Versatile lidar slam via multi-metric linear least square," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 633–11 640.

[21] M. Ramezani, G. Tinchev, E. Iuganov, and M. Fallon, "Online lidar-slam for legged robots with robust registration and deep-learned loop closure," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 4158–4164.

[22] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[23] X. Chen, A. Milioto, E. Palazzolo, P. Giguere, J. Behley, and C. Stachniss, "Suma++: Efficient lidar-based semantic slam," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4530–4537.

[24] L. Li, X. Kong, X. Zhao, W. Li, F. Wen, H. Zhang, and Y. Liu, "Saloam: Semantic-aided lidar slam with loop closure," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7627–7634.

[25] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3288–3295.

[26] K. Choi, S. Jeong, Y. Kim, and K. Sohn, "Stereo-augmented depth completion from a single rgb-lidar image," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 641–13 647.

[27] D. Maturana and S. Scherer, "3d convolutional neural networks for landing zone detection from lidar," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 3471–3478.

[28] J. Dou, J. Xue, and J. Fang, "Seg-voxelnet for 3d vehicle detection from rgb and lidar data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4362–4368.

[29] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss, "Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6529–6536, 2021.

[30] S. Li, Y. Liu, and J. Gall, "Rethinking 3-d lidar point cloud segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[31] L. Yi, B. Gong, and T. Funkhouser, "Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 363–15 373.

[32] A. Milioto, J. Behley, C. McCool, and C. Stachniss, "Lidar panoptic segmentation for autonomous driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8505–8512.

[33] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall, "Multi-scale interaction for real-time lidar data segmentation on an embedded platform," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 738–745, 2021.

[34] J. Behley, A. Milioto, and C. Stachniss, "A benchmark for lidar-based panoptic segmentation based on kitti," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 596–13 603.

[35] A. Asvadi, C. Premebida, P. Peixoto, and U. Nunes, "3d lidar-based static and moving obstacle detection in driving environments: An approach based on voxels and multi-region ground planes," *Robotics and Autonomous Systems*, vol. 83, pp. 299–311, 2016.

[36] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[37] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.

[38] M. Niemeyer and A. Geiger, "Giraffe: Representing scenes as compositional generative neural feature fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 453–11 464.

[39] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[40] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, and H. Zhao, "Semanticposs: A point cloud dataset with large quantity of dynamic instances," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 687–693.

[41] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[42] D. B. Lindell, D. Van Veen, J. J. Park, and G. Wetzstein, "Bacon: Band-limited coordinate networks for multiscale scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 252–16 262.

[43] Y. Ben-Shabat, C. H. Koneputugodage, and S. Gould, "Digs: Divergence guided shape implicit neural representation for unoriented point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 323–19 332.

[44] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," *arXiv preprint arXiv:2012.03762*, 2020.

[45] D. Driess, J.-S. Ha, M. Toussaint, and R. Tedrake, "Learning models as functionals of signed-distance fields for manipulation planning," in *Conference on Robot Learning*. PMLR, 2022, pp. 245–255.

[46] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, "3d neural scene representations for visuomotor control," in *Conference on Robot Learning*. PMLR, 2022, pp. 112–123.

[47] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, "Vision-only robot navigation in a neural radiance world," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4606–4613, 2022.

[48] J.-S. Ha, D. Driess, and M. Toussaint, "Deep visual constraints: Neural implicit models for manipulation planning from visual input," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 857–10 864, 2022.

[49] Y. Pan, Y. Kompis, L. Bartolomei, R. Mascaro, C. Stachniss, and M. Chli, "Voxfield: Non-projective signed distance fields for online planning and 3d reconstruction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5331–5338.