Instance Completion and Motion Estimation with Deep Shape Priors for Autonomous Driving

Shashank Dammalapati Panyawat Rattana

Abstract— **Accurate pose estimation of surrounding vehicles is crucial for robust autonomous driving. Existing methods often suffer from outliers and inaccuracies, particularly in challenging environments. Inspired by recent object-oriented SLAM approaches that use shape priors, this work presents a novel method for robustly estimating the poses and shapes of surrounding vehicles in challenging autonomous driving scenarios. This work proposes a novel approach that jointly optimizes a single deep shape code and multiple transformation parameters for each vehicle across multiple frames. While relying on a simple Intersection-Over-Union (IoU)-based tracking algorithm to maintain associations. Our multi-frame pose and shape optimization approach leverages the temporal consistency of vehicle shapes and object tracking information and tries to demonstrate the effectiveness of using deep shape priors to improve the reconstruction, detection, and tracking quality of the cars in the scene.**

I. Introduction

Precise understanding of the surrounding environment, particularly the presence and position of other vehicles, is paramount for the success of autonomous driving systems. In this context, 3D bounding boxes (3D BBoxes) derived from LiDAR data have emerged as a powerful tool for object detection, tracking, and motion prediction. By leveraging the inherent 3D measurement capabilities of LiDAR, 3D BBoxes directly capture the size, shape, and pose of objects in the scene, providing crucial information for safe and efficient navigation.

However, relying solely on LiDAR data for 3D bounding box generation [8] [1] [9] presents inherent limitations. The sparsity of LiDAR point clouds, particularly at longer distances, can lead to incomplete or inaccurate bounding boxes, especially for small or intricately shaped objects. Occlusions caused by other objects or the environment further compound this issue, masking crucial information and leading to misinterpretations of object shapes within the bounding box [6]. Additionally, sensor noise and environmental factors can introduce artifacts and inaccuracies into the point cloud, resulting in jittery or distorted bounding boxes, as shown in Fig. 1. These limitations impact the overall performance of object detection and tracking algorithms, potentially leading to misinterpretations and unsafe situations on the road.

Therefore, while 3D BBoxes represent a valuable tool for exploiting LiDAR's strengths in autonomous driving, overcoming their limitations is crucial for robust and reliable object perception. This necessitates the exploration of alternative or complementary methods that can enhance the accuracy and robustness of 3D object understanding. Integrating additional data sources, such as camera images [13], or incorporating prior knowledge [5] [12] [2] about object shapes through deep learning techniques are promising avenues for further research. Ultimately, addressing these challenges will lead the way for more accurate and robust object perception, contributing to the safety and efficiency of autonomous driving systems.

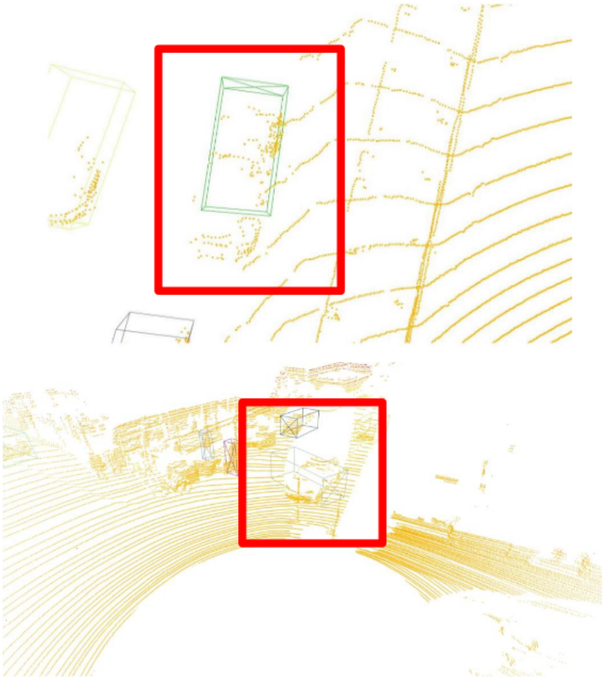


Fig. 1: This image contains two LiDAR frames with inaccurate detections from the network.

For object tracking, achieving robust perception remains a challenge. While object-aware methodologies offer promise, traditional approaches still struggle with incomplete object representations and limited adaptability. Our work introduces a novel approach that integrates learned shape priors [5], represented by a latent code $\mathbf{z}$ and a generative model $G(\mathbf{z})$, into the tracking framework. This enables detailed and complete object reconstructions, even with occlusions, surpassing the limitations of pre-defined databases or scratch-built models. Moreover, our shape code-based representation facilitates back propagation-based optimization, which promises to lead to improvement in tracking accuracy compared to baseline methods.

The main contribution of this paper is a novel object tracking and reconstruction approach that combines learned

This MSRP is supervised by Xingguang Zhong and Yue Pan

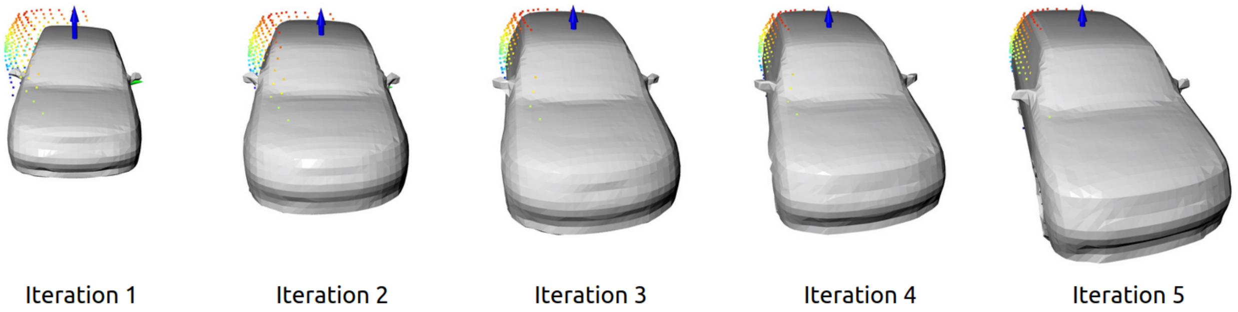


Fig. 2: Illustration emphasizing optimization of pose over time

shape priors with Gauss-Newton optimization for robust dynamic perception in autonomous driving. In our method, we maintain a DeepSDF [5] based vehicle representation, along with an IOU-based data association method that facilitates the implementation of a multi-frame sliding window Gauss-Newton optimization for refining object poses across multiple frames, along with a single associated shape code, which can be converted into a high-fidelity mesh of the tracked vehicle.

In sum, we make three key claims: (i) Our approach enables detailed and complete object reconstructions, surpassing the limitations of existing methods.; (ii) The integration of learned shape priors enhances object tracking accuracy compared to baseline methods.; (iii) We demonstrate improvements in both qualitative and quantitative metrics, validating the effectiveness of our approach. These claims are backed up by the paper and our experimental evaluation.

II. Related Work

While accurate sensor localization, like that offered by Kiss-ICP [10], is crucial for understanding the global position in autonomous driving, precise object localization within the sensor frame remains a challenge. Shape priors have emerged as a powerful tool to address this, with frameworks like DeepSDF [5] introducing learned 3D geometry representations enabling high-quality shape completion. This has opened doors to tackling optimization problems in object pose regression and map reconstruction.

Recent works have built upon this foundation, integrating deep shape priors into optimization frameworks for object tracking and mapping. DSP-SLAM [11], for example, leverages them to refine object reconstructions within a map, achieving real-time performance and high accuracy across diverse datasets. Similarly, FRODO [7] employs deep shape embeddings for incremental object pose refinement, leading to more robust scene understanding.

The benefits of this approach extend beyond autonomous driving. In agricultural robotics, Panoptic Mapping with Fruit Completion [4] utilizes deep shape priors to estimate the complete 3D shapes and poses of fruits in cluttered environments. This demonstrates the wider potential of shape priors in optimizing object pose and map reconstruction tasks.

Overall, these advancements showcase the immense potential of deep shape priors in addressing optimization challenges within object tracking, mapping, and pose estimation. Inspired by these developments, we propose a novel method that simultaneously optimize poses of the particular car in consideration over multiple frames, which are part of a sliding window. The exact method and its implementation are discussed in the next section.

III. Our Approach

Our system is a sequential object state estimation pipeline. Along with the bounding box pose of the object, we also estimate the complete shape of the object represented as a shape code z. We use the DeepSDF [5] network to implicitly represent the shape of the car. This gives us 2 main advantages: 1. Ability to model previously unseen cars, 2. Ability to reconstruct complete and detailed car models. It takes as input the shape code $z \in \mathbb{R}^{64}$ and a query point location $x \in \mathbb{R}^3$ and outputs a signed distance function value (SDF) G, i.e. $s = G(x, z)$ to the car surface represented by the network. We designed our system to work with Lidar data. In this project, We aim to estimate the full dense shape z and 7-DoF pose T_{so}, represented as a homogeneous transformation matrix $T_{SO} = [sR_{SO}, t_{SO}; 0, 1] \in$ Sim(3), for every detected car over all the lidar frames it exists in. In this paper, the Sensor frame is denoted by S, and the Object Frame (i.e. Canonical Frame) is denoted by O. We formulate this as a joint optimization problem, which iteratively refines the shape code and object pose T_{SO} from an initial estimate. We use E_{surf} as the primary loss function and implement a familiar Gauss-Newton solver with analytical Jacobians. This will give us the optimised T_{SO}, i.e. Transformation from Object frame to the Sensor Frame. 3D Bounding boxes generated by any detection networkReferences can be used as initial guess for the T_{SO} in our method, as shown in 4 A. Each bounding box is parameterized by 7 parameters $(x, y, z, l, b, h, \theta)$, where θ is the yaw of the bounding box with respect to the sensor frame. Using our method, we optimize this initial estimates and improve the bounding boxes accuracy.

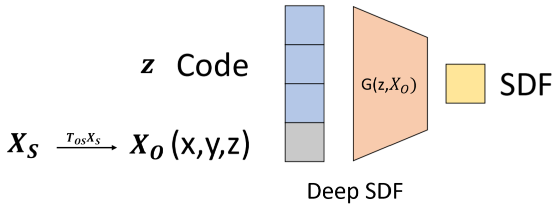


Fig. 3: This image illustrates our optimization frame work for a single Lidar frame

A. Loss Function

Our pipeline leverages the DeepSDF [5] network's capability to represent the shape of arbitrary, unseen cars. We exploit this to create an optimization pipeline, as illustrated in Figure 3. We measure the loss function using two main components:

1. E_{surf}, based on how well the car points in the LiDAR scan align with the reconstructed car surface after the transformation T_{OS}, formulated as follows:

$$E_{\text{surf}} = \frac{1}{|P|}\sum \mathbf{G^2}(p_0, Z) \tag{1}$$

where $|P|$ represents the number of points on the car surface, p_O represents points of the car surface in the Object/Canonical Frame as defined in DeepSDF [5].

2. $E_{\text{z-regularization}}$, formulated as:

$$E_{\text{z-regularization}} = \|Z\|^2 \tag{2}$$

Thus, the overall loss function (E) is given by:

$$E = \lambda_s E_{\text{surf}} + \lambda_c \|Z\|^2 \tag{3}$$

where λ_s and λ_c are tunable hyperparameters for these individual terms.

B. Optimization

Given a single lidar frame with a known detection. We use our method to improve the detection by minimizing the overall loss function. We employ a Gauss-Newton optimization approach with analytical Jacobians, leveraging the quadratic nature of all terms.

$$E = \lambda_s \frac{1}{|P|}\sum \mathbf{G^2}(p_0, Z) + \lambda_c \|Z\|^2 \tag{4}$$

We already have the relation between p_0 and p_0 as a function of the T_{OS}. So, we write the Energy term as follows,

$$E = \lambda_s \frac{1}{|P|}\sum \mathbf{G^2}(T_{OS}\cdot p_S, Z) + \lambda_c \|Z\|^2 \tag{5}$$

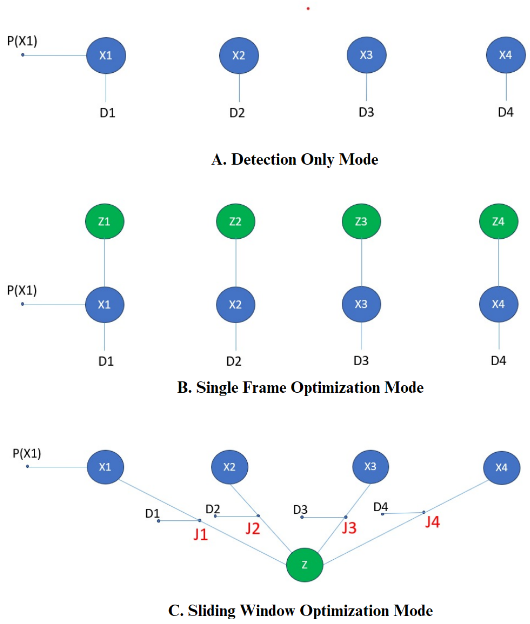


Fig. 4: This image shows graphical representation of how each unknown variable is linked to each other, in different implementations of our optimization process.Xi, Zi are the pose, and shape code of the car in ith frame

By, minimizing the loss function 5, we estimate the unknown parameters Z, T_{OS}. These new estimates of these parameters are optimized in the sense that they represent the underlying data more accurately, i.e. the surface points now align well with the reconstructed car and the reconstructed car is close to the shape of the real car. The first version of our optimization is implemented on a per-frame basis, as shown in 4 B. Where we limited the optimization to information from a single frame. We later, implemented instance association of cars over multiple frames using an IOU based method. This lets us utilize information from multiple frames to further theoretically improve our optimization process. We exploit the fact that the shape of the car remains same over time as it moves in the environment. Using this, we implemented two new optimization pipelines:

1. Optimization consecutive frames using the shape code that is an outcome of the previous frame's optimization, as shown in 4 B, except the shape code Z is initialized from the Z of the previous timestamp 2. A novel method to jointly optimize multiple frames at the same time and outputs a single shape code Z and multiple T_{OS} corresponding to each frame, as shown in 4 C

Using the above three mentioned optimization pipelines that utilize deep shape priors, we expect to improve the quality of detection and tracking of the vehicles in the scene.

IV. Experimental Evaluation

The main focus of this work is to show that using deep shape priors can improve our bounding box

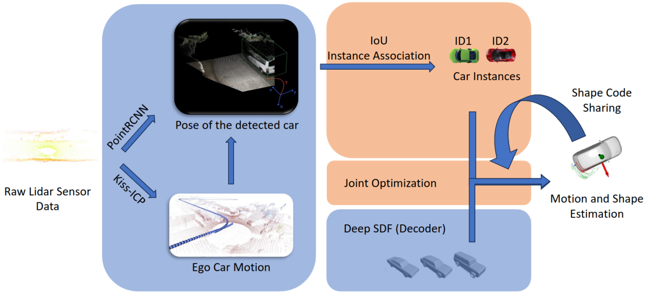


Fig. 5: The pipeline for multi-object global tracking

We present our experiments to show the capabilities of our method. The results of our experiments also support our key claims, which are: (i) Our approach enables detailed and complete object reconstructions, surpassing the limitations of existing methods.; (ii) The integration of learned shape priors enhances object tracking accuracy compared to baseline methods.; (iii) We demonstrate improvements in both qualitative and quantitative metrics, validating the effectiveness of our approach. These claims are evaluated by the paper and our experimental results.

A. Experimental Setup

Dataset: In our work, we leverage Argoverse2 [14], a dataset recorded across 6 U.S. cities. This dataset is equipped with two LiDAR sensors, seven ring cameras, and two front-facing stereo cameras. Notably, two 32-beam LiDARs rotate at 10 Hz in the same direction but are oriented 180° apart. Consequently, LiDAR returns from the two stacked 32-beam sensors are consolidated into a single sweep, with the reference frame set to the ego vehicle reference frame rather than the individual LiDAR reference frame. The dataset furnishes 10 Hz 3D cuboid annotations across 30 classes. Our primary focus lies in the regular vehicle class, encompassing cars, vans, pickup trucks, SUVs, and more. Annotations for the same object instance include consistent track identifiers over time, presented in the ego vehicle's reference frame. Each sequence within the dataset spans approximately 15 seconds, comprising an average of 150 LiDAR sweeps. Due to the inherent limitations of LiDAR observations in capturing the complete point cloud of objects, we selectively choose sequences where a car is observed from various angles over time, such as when a car turns at an intersection or when it passes in front of the ego car. We also used Kitti [3] sequences to test our optimization process, but due to the unavailability of the ground truth 3D bounding boxes for those sequences, we only resorted to qualitative assessment of our method's performance on this dataset.

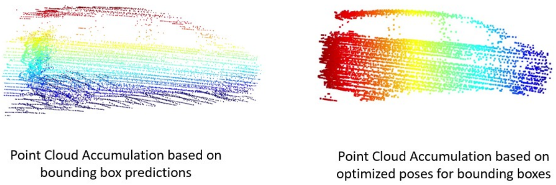
B. Implementation

We implemented our three versions of the optimization process on the Argoverse [14] dataset. We extracted 10 sequences with 20-100 frames, each focused on a particular car. So, for each sequence, we implemented

1. Single frame Optimization, as shown in 4 B 2. Single frame Optimization with Code sharing, as shown in 5 3. Sliding Window based optimization, as shown in 4 C

C. Qualitative Results

1) Accumulation Quality: In this project, our aim was to demonstrate the effectiveness of shape priors in enhancing detection quality. Essentially, we seek to track either multiple cars or a single car over time. One method to validate this is by assessing the accumulation quality of various bounding box generation techniques. Figure 7 depicts the accumulated point cloud of a car surface. On the left, the accumulation performed using a detection network [9] is evidently less accurate compared to the accumulation on the right, which is achieved using optimized bounding box transformations.


Point Cloud Accumulation based on bounding box predictions

Point Cloud Accumulation based on optimized poses for bounding boxes

Fig. 7: This shows the qualitative results of our optimization process with respect to Accumulation

2) Visualization of Optimization: In the context of cars, shape completion and bounding box regression can also be evaluated visually in extreme cases. To verify that our method is functioning properly, we deliberately provided grossly incorrect initialization values for bounding box poses and visually observed the optimization process. For this purpose, we visualized multiple iterations during the optimization. Figure 6 illustrates the optimization of the shape, where we can clearly observe that the reconstructed car's shape gradually matches more closely with the optimization. Similarly, in Figure 2, we can observe the correction of large yaw errors in the bounding box over the optimization process..

D. Quantitative Analysis

1) Detection Vs Optimization: In the quantitative analysis, we analyze how the integration of learned shape priors enhances object tracking accuracy compared to baseline methods. This evaluation involves tracking objects over time and comparing the tracking results with ground truth annotations. This includes measures such as Intersection over Union (IOU), Yaw error rates, and other relevant performance indicators.

The performance metrics reported in the table are based on experiments conducted on a dataset comprising 10 diverse sequences, each capturing instances of cars from various angles. These sequences aim to represent a range of real-world scenarios, providing a comprehensive evaluation of the methods under different conditions. We implemented our single frame based optimization methods (with and without

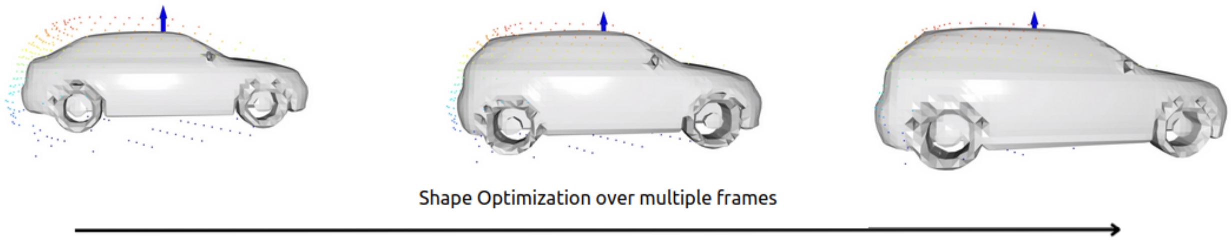


Fig. 6: Illustration emphasizing optimization of shape over time

IOU (Detection)	IOU (Single Frame Opt)	IOU (SFO with Code Sharing)	Yaw (Detection)	Yaw (Single Frame Opt)	Yaw (SFO with Code Sharing)
0.691 (0.082)	0.7126 (0.082)	0.710 (0.082)	2.124 (1.673)	1.656 (1.357)	1.739 (1.487)

TABLE I: Comparison of two different optimization methods in IOU and Yaw error performance along with a baseline detection network. All values in this table are averaged over 10 different sequences. Values inside the ”()” are standard deviation values.

TABLE II: Performance Metrics of Sliding Window Optimization Vs Baseline Detection Network with different Sliding Window Sizes All values in this table are averaged over 10 different sequences. Values inside the ”()” are standard deviation values..

	IOU		Error		Window
	Detector	Opt Pose	Detector	Opt Pose	
	0.705(0.054)	0.685(0.047)	1.769(1.002)	1.206(0.683)	5
	0.709(0.054)	0.690(0.054)	1.835(1.583)	1.402(0.982)	15
	0.706(0.055)	0.703(0.058)	1.860(1.645)	1.525(1.021)	20

code sharing), the results in Table I are metrics averaged from these 10 sequences. The IOU metric assesses the accuracy of object detection by measuring the overlap between the predicted bounding boxes and ground truth annotations. There's only a slight improvement in IOU from detection to optimization techniques, indicating moderate performance enhancements. Yaw error quantifies the accuracy of estimating the yaw angle of the detected cars. It measures the deviation between the predicted and ground truth yaw angles. There's a notable reduction in yaw error from detection to optimization methods, suggesting improved accuracy in estimating the orientation of the cars. Both single frame optimization(SFO) and SFO with code sharing demonstrate similar trends in performance improvement compared to the baseline detection method. However, the addition of code sharing in SFO doesn't lead to a significant enhancement in performance over single frame optimization alone, indicating a limited impact on the overall results.

2) Ablation Studies on Sliding Window Optimization: The table (II) includes a different optimization process called sliding window optimization, and this table contains ablation studies on this method. We analyze the Intersection over Union (IOU) and Yaw Error values, as well as the effect of window size on the results.

Contrary to our expectations, the sliding window method performed worse compared to the detector. However, the simple optimization method discussed earlier, which is just a single frame optimization method, had better results. Despite this, the yaw estimation is still good with the sliding window compared to the detector.

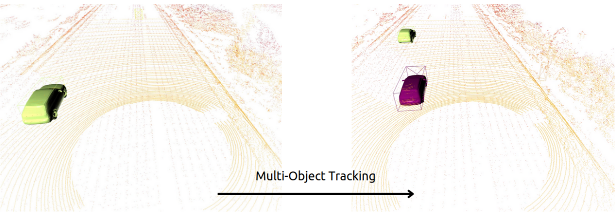


Fig. 8: Visualization of Multi-Object Tracking with Shape Completion

The table reveals that as the window size increases, there is a tendency for the IOU to slightly decrease, while the yaw error tends to increase. This suggests that larger window sizes might lead to less accurate object localization and orientation estimation.

Overall, these findings indicate that while the sliding window optimization method may not be as effective as expected in terms of IOU, it still shows promise in terms of yaw estimation, albeit not surpassing the performance of the baseline detector. Further investigation into the optimization parameters and strategies may be warranted to improve the overall performance of the sliding window approach.

V. Conclusion

The main focus of this work is to show that using deep shape priors can improve our bounding box accuracy. We present our experiments to show the capabilities of our method. The results of our experiments also support our key claims, which are:

(i) Our approach enables detailed and complete object reconstructions by using DeepSDF. A shape code can generate complete car geometry, as shown in 8

(ii) The integration of learned shape priors enhances object tracking accuracy compared to baseline methods. We observed this especially in the yaw estimation.

(iii) We demonstrate improvements in both qualitative and quantitative metrics, validating the effectiveness of our approach.

These claims are evaluated by the paper and our experimental results. Specifically, our quantitative analysis on diverse sequences showcases the benefits of single-frame optimization methods, both with and without code sharing. However, we also introduce an alternative approach, sliding window optimization, which, contrary to expectations, yielded inferior results compared to the detector. Although sliding window optimization did not perform as well in terms of IOU, it showed promise in yaw estimation. Further investigation into optimization parameters and strategies may lead to improvements in its overall performance.

In conclusion, our study underscores the importance of leveraging deep shape priors for enhanced object tracking accuracy, offering insights into the effectiveness of different optimization methods and contributing to advancements in computer vision research.

REFERENCES

[1] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia. VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking, 2023.

[2] F. Engelmann, J. Stuckler, and B. Leibe. SAMP: Shape and Motion Priors for 4D Vehicle Reconstruction. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar 2017.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss. Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots, 2023.

[5] J.J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation, 2019.

[6] R. Qian, X. Lai, and X. Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130:108796, October 2022.

[7] M. Rünz, K. Li, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, and R. Newcombe. FroDO: From Detections to 3D Objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14717, 2020.

[8] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection, 2021.

[9] S. Shi, X. Wang, and H. Li. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud, 2019.

[10] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023.

[11] J. Wang, M. Rünz, and L. Agapito. DSP-SLAM: Object Oriented SLAM with Deep Shape Priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371. IEEE, 2021.

[12] R. Wang, N. Yang, J. Stueckler, and D. Cremers. Directshape: Direct photometric alignment of shape priors for visual vehicle pose and shape estimation, 2020.

[13] Y. Wang, Q. Mao, H. Zhu, Y. Zhang, J. Ji, and Y. Zhang. Multi-modal 3D object detection in autonomous driving: A survey. *arXiv preprint arXiv:2106.12735*, 2021.

[14] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J.K. Pontes, D. Ramanan, P. Carr, and J. Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting, 2023.