Towards Robust LiDAR-based Perception in Autonomous Driving

Jiachen Sun jiachens@umich.edu
Yulong Cao yulongc@umich.edu
Qi Alfred Chen [†] alfchen@uci.edu
Z. Morley Mao zmao@umich.edu

University of Michigan [†] University of California, Irvine

Abstract

Perception plays a pivotal role in autonomous vehicles (AVs), which utilizes onboard sensors like cameras and LiDARs (Light Detection and Ranging) to assess surroundings. Recent studies have demonstrated that LiDAR-based perception is vulnerable to spoofing attacks, in which adversaries spoof a fake vehicle in front of a victim AV by strategically injecting laser signals to the victim's LiDAR sensor. In this work, we first explore a general vulnerability of current LiDAR-based perception architectures that the ignored occlusion and distancing patterns in point clouds make AVs vulnerable to spoofing attacks. We construct the first black-box spoofing attack based on our identified vulnerability, which universally achieves around 80% mean success rates on all target models. We further take a first step towards designing a general architecture for robust LiDAR-based perception, and propose sequential view fusion (SVF) which reduces the mean attack success rate to around 2.3%.

1. Introduction

In autonomous driving perception, 3D object detection is indispensable to ensure safe and correct driving decisions, which takes point clouds generated by LiDAR sensors as input and yields 3D bounding boxes of target objects. Point cloud data contains location information of each reflected point along with its intensity (reflectance). Due to a heavy reliance on LiDAR, a few prior studies have explored the security of LiDAR and its usage in autonomous driving [13, 15, 5]. Among them, Cao *et al.* are the first to discover that the deep learning model for LiDAR-based perception used in a real-world autonomous driving system can be fooled to detect a fake vehicle by strategically injecting a small number of spoofed LiDAR points [5]. However, the attack proposed was evaluated on only one specific model (*i.e.*, Baidu Apollo 2.5) assuming white-box access, which may be unrealistic. Moreover, we find that existing LiDAR spoofing attacks [15, 5] cannot directly generalize to all three state-of-the-art models.

In this work, we perform the first study to systematically explore, discover, and defend against a *general* vulnerability existing among three state-of-the-art LiDAR-based 3D object detection model designs: bird's-eye view (BEV)-based (Baidu Apollo 5.0 [1]), voxel-based (PointPillars [11]), and point-wise (PointRCNN [14]). To explore the vulnerability, we validate two potential false positive situations based on our empirical observations of deep learning models and unique physical features of LiDAR, and discover that *all* the three state-of-the-art 3D object detection model designs above generally ignore the *occlusion and distancing patterns* in point clouds, which are two physical invariants for LiDAR. This allows an adversary to spoof almost two magnitudes fewer points into the victim's LiDAR but still can deceive the perception model into detecting a fake front-near vehicle (§3.1). We construct the first black-box spoofing attack based on our identified vulnerability and demonstrate its effectiveness (§3.2). We further take a first step to design a general architecture for robust LiDAR-based perception, and show that it can effectively defend existing spoofing attacks (§4).

Overall, this work makes the following contributions:

- We perform the first study to explore the general vulnerability of current LiDAR-based perception architectures. We discover that current LiDAR-based perception models ignore several physical features of LiDAR which result in the success of spoofing attacks. We construct the first black-box spoofing attack based on this vulnerability. Large-scale evaluations show that attackers can achieve around 80% mean success rates on all target models.
- We design a general architecture for robust LiDAR-based perception in AVs by embedding the front view (FV) representation of LiDAR point clouds. We find that existing view fusion-based models are still vulnerable to LiDAR spoofing attacks. To address their limitations, we propose sequential view fusion (SVF). SVF leverages a semantic segmentation module to better utilize FV features. Evaluations show that SVF can further reduce the mean attack success rate to 2.3%.

2. Threat Model

Sensor attack capability. We adopt the formulation in Adv-LiDAR [5] to describe the sensor attack capability ($\mathcal{A}$):

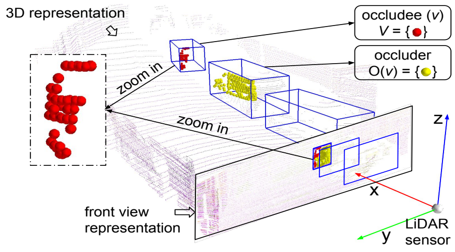


Figure 1: Illustration of an occluded vehicle in LiDAR point clouds from two different representations (*i.e.*, 3D and front view).

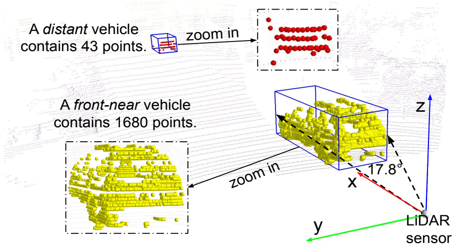

A *distant* vehicle contains 43 points.
zoom in
A *front-near* vehicle contains 1680 points.
zoom in
17.8°
z
x
y
LiDAR sensor

Figure 2: Illustration of a distant vehicle and a front-near vehicle in LiDAR point clouds, where the front-vehicle occupies 17.8° horizontal angles.

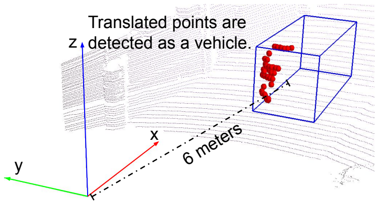


Figure 3: V (red points) in Figure 1 are still detected as a valid vehicle when directly translated to a front-near location.

1) attackers are able to inject at most 200 points into the victim LiDAR; 2) attackers are able to modify the *distance*, *altitude*, and *azimuth* of a spoofed point to the victim LiDAR by changing the delay intervals of the attack devices. Especially, the *azimuth* of a spoofed point can be modified within a horizontal viewing angle of 10°.

Black-box spoofing attack. We consider LiDAR spoofing attacks as our threat model [15, 13]. We adopt the attack goal of Adv-LiDAR [5]: to spoof a *front-near* vehicle located 5-8 meters in front of the victim AV. We assume that attackers can control the spoofed points within the observed sensor attack capability ($\mathcal{A}$). Note that attackers are not required to have access to the machine learning model nor the perception system.

Defense against general spoofing attacks. We also consider defending such LiDAR spoofing attacks and assume a stronger attack model that adversaries have white-box access to the machine learning models in AVs.

3. Black-box Spoofing Attack

We find that existing spoofing attacks suffer from *effectiveness*, *generality*, and *white-box access* limitations, as introduced in §1. Motivated by the above limitations, in this section, we leverage an in-depth understanding of the intrinsic physical nature of LiDAR to identify a general design-level vulnerability for current LiDAR-based perception, and further construct the first black-box spoofing attack on state-of-the-art models.

3.1. Vulnerability Identification

Despite a lack of generality, Adv-LiDAR was able to spoof a fake front-near vehicle by injecting much fewer amount of points than it is required for a valid vehicle representation. For example, Cao *et al.* have demonstrated that an attack trace with merely *60* points and *8*° of horizontal angles is sufficient to deceive Apollo 2.5 [5]. However, a valid front-near vehicle (§2) contains around *2000* points and occupies about *15*° of horizontal angles in KITTI point clouds [9]. It remains unclear why such spoofing attacks can succeed despite a huge gap in the amount of points between that of a fake and a valid vehicle. To answer this question, we identify two situations where a valid vehicle contains a small number of points: 1) **an occluded vehicle** and 2) **a distant vehicle** as shown in Figure 1-2, which are similar to human visual perception where occluded and distant objects contain much fewer pixels in our retinas. Though LiDAR sensors share similarities with human visual perception, all three state-of-the-art classes of LiDAR-based perception models operate object detection tasks in the 3D Euclidean space (§1) different from 2D vision recognition pipelines. We find that such small difference actually leaves a potential attack surface for adversaries to launch spoofing attacks (§2). More specifically, we discover and validate two false positive (**FP**) conditions that apply to all models, which could contribute to the success of Adv-LiDAR [5]:

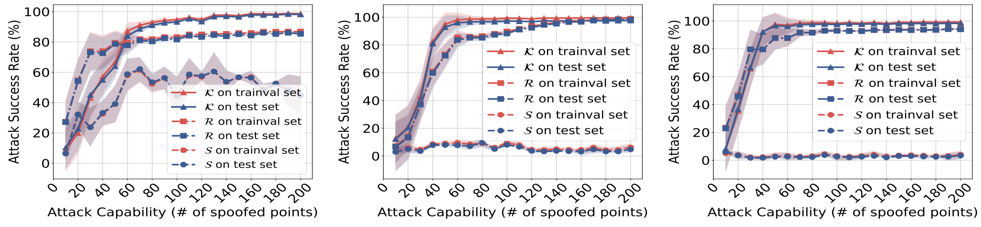
FP1: *If an occluded vehicle is detected in the pristine point cloud by the model, its* ***point set*** *will still be detected as a vehicle when directly moved to a front-near location.*

FP2: *If a distant vehicle is detected in the pristine point cloud by the model, its* ***point set*** *will still be detected as a vehicle when directly moved to a front-near location.*

As mentioned earlier, the sensor attack capability $\mathcal{A}$ is far from spoofing a fully exposed front-near vehicle's point set. However, **FP1** and **FP2** provide two strategies for adversaries to launch spoofing attacks with fewer points and horizontal angles. As a result, attackers can directly spoof a vehicle imitating various occlusion (**FP1**) and distancing (**FP2**) patterns that satisfy the sensor attack capability $\mathcal{A}$ to fool the state-of-the-art models. For example, the V (red points) in Figure 1 only contains 38 points and occupies 4.92° horizontally when translated to 6 meters in front of the AV. We confirm that it can deceive all three target models successfully, as visualized in Figure 3.

3.2. Attack Construction

Constructing black-box attacks on deep learning models is non-trivial. Our methodology attempts to closely represent realistic physical attacks using traces from real-world datasets (*e.g.* KITTI [9]). In order to test different sensor attack capability, we extract occluded vehicles' point sets with varying numbers of points (5-200 points) from the KITTI validation set. We then construct a small dataset $\mathcal{K}$ containing 100 point sets with different attack capabilities. Besides collecting existing real-world traces, the identified vulnerability also supports adversaries in generating customized attack traces, which are more efficient for pipelining the attack process. We utilize a 3D car mesh and implement a renderer [6] simulating the function of a LiDAR sensor that probes the car mesh by casting lasers, as shown



(a) ASR of Apollo 5.0.

(b) ASR of PointPillars.

(c) ASR of PointRCNN.

Figure 4: Attack Success Rate (ASR) of proposed black-box spoofing attacks on three target state-of-the-art models.

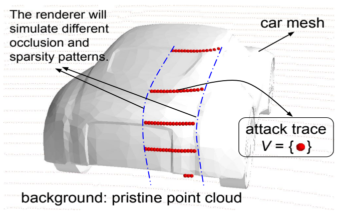

The renderer will simulate different occlusion and sparsity patterns.
car mesh
attack trace
V = {●}
background: pristine point cloud

Figure 5: The process of generating attack traces for $\mathcal{R}$ from the implemented renderer.

in Figure 5. We also follow the same procedure to build a small dataset $\mathcal{R}$ containing 100 rendered point sets.

We further leverage a global translation matrix $H(\theta, \tau)$ (Equation 1) [5] to move every attack trace (V_i) to a front-near location (*i.e.* 5-8 meters in front of the victim AV) in the pristine point cloud, where θ and τ correspond to the *azimuth* and *distance* of the translation (§2), respectively:

$$V'_{i\,\mathbf{w_i}} = V_{i\mathbf{w_i}}$$

$$\begin{bmatrix} V'_{i\,\mathbf{w_x}} \\ V'_{i\,\mathbf{w_y}} \\ V'_{i\,\mathbf{w_z}} \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & \tau\cos(\theta+\alpha) \\ \sin\theta & \cos\theta & 0 & \tau\sin(\theta+\alpha) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{i\mathbf{w_x}} \\ V_{i\mathbf{w_y}} \\ V_{i\mathbf{w_z}} \\ 1 \end{bmatrix} \tag{1}$$

$(V_{i\mathbf{w_x}}, V_{i\mathbf{w_y}}, V_{i\mathbf{w_z}}, V_{i\mathbf{w_i}})$ denotes the $xyz\text{-}i$ feature vectors (introduced in §1) of all points in V_i, and $\alpha = arctan(V_{i\mathbf{w_y}}/V_{i\mathbf{w_x}})$. Such a translation matrix has been demonstrated to satisfy both the physical constraints of LiDAR [5] and attack capability ($\mathcal{A}$).

3.3. Attack Evaluation and Analysis

Experimental setup. The evaluations are performed on the KITTI trainval and test sets [9]. We test all the generated attack traces from $\mathcal{K}, \mathcal{R}$ on all point cloud samples on three target models by simulated spoofing. We also utilize attack traces ($\mathcal{S}$) generated by the blind sensor-level spoofing attack with no control of the points as a baseline.

Evaluation metrics. We leverage the default thresholds used by three target models to measure the Attack Success Rate (ASR). We label an attack successful as long as the model detects a vehicle at the target location whose confidence score exceeds the default threshold:

$$ASR = \frac{\text{\# of successful attacks}}{\text{\# of total point cloud samples}} \tag{2}$$

Figure 4 shows the ASR of the simulated spoofing attack with different attack capabilities (*i.e.* number of points). As expected, the ASR increases with more spoofed points. The ASRs are able to universally achieve higher than 80% in all target models with more than 60 points spoofed. Notably, the attack traces from $\mathcal{R}$ achieve comparable ASR with $\mathcal{K}$ on all target models, which demonstrate that adversaries can efficiently leverage a customized renderer to generate attack traces (Figure 5). Such rendered traces can be directly programmed into hardwares for physical spoofing attacks.

4. Sequential View Fusion

In this section, we take the first step towards exploring the feasibility of embedding the ignored physical features into end-to-end learning that provides better robustness for LiDAR-based perception.

4.1. Why should FV Representations help?

We observe that LiDAR natively measures range data. Thus, projecting the LiDAR point cloud into the perspective of the LiDAR sensor will naturally preserve the physical features of LiDAR. Such projecting is also known as the FV of LiDAR point clouds [12]. Given a 3D point $\vec{p} = (x, y, z)$, we can compute its coordinates in FV $\vec{p}_{FV} = (r, c)$ by:

$$\begin{aligned} c &= \lfloor \arctan(y, x)/\Delta\theta \rfloor \\ r &= \lfloor \arctan(z, \sqrt{x^2 + y^2})/\Delta\phi \rfloor \end{aligned} \tag{3}$$

where $\Delta\theta$ and $\Delta\phi$ are the horizontal and vertical fire angle intervals. As shown in Figure 1, since the *occluder* $\mathsf{O}(v)$ and *occludee* V neighbor with each other in the FV, deep learning models have opportunities to identify the occlusion. The abnormal sparsity of a fake "distant" vehicle will be also exposed, as valid vehicles' points are clustered, while the spoofed points scatter in the FV. Thus, the FV representation of point clouds embeds both ignored features.

4.2. SVF Architecture

We find existing view fusion schemes [19, 8] that utilize symmetric designs cannot provide better robustness because the 3D (or BEV) representation dominates the model making the FV representation not critical in the end-to-end architectures [10].

Based on the above understandings, we propose a new view fusion schema called sequential view fusion (SVF). SVF comprises of three modules (Figure 6), which are: 1) semantic segmentation: a semantic segmentation network that utilizes the FV representation to computes the point-wise confidence scores (*i.e.*, the probability that one point belongs to a vehicle). 2) view fusion: the 3D representation is augmented with semantic segmentation scores. 3) 3D object detection: a LiDAR-based object detection network that takes the augmented point clouds to predict bounding

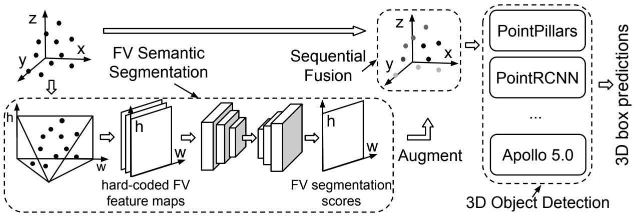
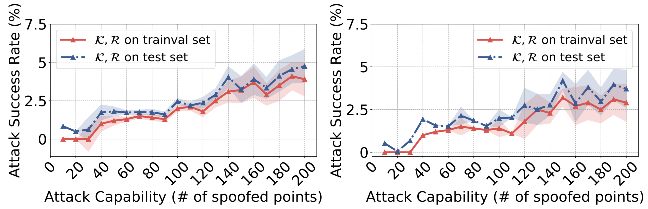


Figure 6: Sequential view fusion (SVF) architecture.



(a) ASR of SVF-PointPillars.

(b) ASR of SVF-PointRCNN.

Figure 7: Attack Success Rate (ASR) of proposed black-box spoofing attack on SVF models.

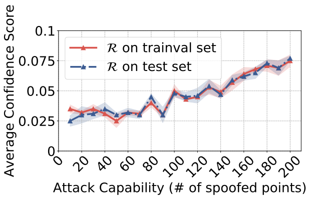


Figure 8: Average confidence score of Adv-LiDAR on the semantic segmentation network.

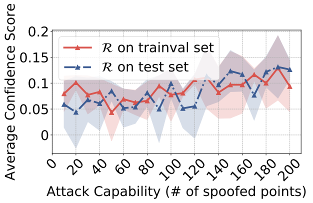


Figure 9: Average confidence score of the adaptive attack on the semantic segmentation network.

boxes. Instead of leaving the models to learn the importance of different representations by themselves, we attach a semantic segmentation network to the raw FV data. By doing so, we enforce the end-to-end learning to appreciate the FV features, so that the trained model will be resilient to LiDAR spoofing attacks.

Semantic segmentation. The semantic segmentation networks accept the FV represented point clouds and associate each point in FV with a probability score that it belongs to a vehicle. These scores provide aggregated information on the FV representation. Compared to 3D object detection or instance segmentation, which is intractable over FV, semantic segmentation is an easier task as it does not need to estimate object-level output. Moreover, there are extensive studies on semantic segmentation over FV represented point clouds [17, 18, 3, 16, 4], and the segmentation networks achieve much more satisfactory results than the 3D object detection task over FV. In our implementation, we adopt the high-level design in LU-Net [4]. It is worth noting that the end-to-end SVF architecture is agnostic to the semantic segmentation module.

View fusion. The fusion module re-architects existing symmetric designs which integrate the 3D representation with the confidence scores generated by the semantic segmentation module. Specifically, we use Equation 3 for mapping between $\vec{p} = (x, y, z)$ and $\vec{p}_{FV}(r, c)$, and augment each $\vec{p}$ with the point-wise confidence score from its corresponding $\vec{p}_{FV}$.

3D object detection. SVF is also agnostic to the 3D object detection module. This module takes the augmented point clouds and output the 3D box predictions. In this paper, we utilize PointPillars and PointRCNN models.

4.3. SVF Evaluations

Experimental setup. We train SVF-PointPillars and SVF-PointRCNN on the KITTI training set, and evaluate them against Adv-LiDAR [5] on Apollo 5.0.

Evaluation metrics. We evaluate the average precision (AP) of SVF-PointPillars and SVF-PointRCNN on the KITTI validation set, and leverage ASR to test their robustness against LiDAR spoofing attacks.

Figure 7 shows the ASR of our proposed spoofing attacks. As shown, the attacks are no longer effective in SVF models. The ASR reduces from more than 95% (original models) to less than 4.5% on both models with the maximum attack capability. The mean ASR also drops from 80% to around 2.3%. We also perform ablation study on SVF, and demonstrate that the FV features are more important in SVF models. Both SVF models also achieve comparable AP compared to the original models.

4.3.1 Defense against White-box and Adaptive Attacks

We test whether the state-of-the-art white-box attack, Adv-LiDAR can succeed in both the semantic segmentation and 3D object detection modules. We first directly apply the attack traces that successfully fool Apollo 5.0 to the segmentation network and record the mean confidence score of all the points belonging to the attack trace. Figure 8 shows that the mean confidence scores are consistently below 0.08 which is too low to be classified as a valid vehicle whose mean confidence score is around 0.73 in our trained model.

Model-level defenses are usually vulnerable to simple adaptive attacks [2, 7]. We assume that the adversaries are aware of the SVF architecture. The attack goal is to both fool the semantic segmentation and 3D object detection modules. We also leverage the formulation in [5] to utilize the global transformation matrix $H(\theta, \tau)$ to control the spoofed points. Figure 9 shows that none of the attack traces' average confidence score reaches 0.2 in the segmentation module, which is still far from the mean average confidence score of valid vehicle 0.73. Therefore, the adaptive attacks also cannot break the robustness of SVF.

5. Conclusion

In this paper, we perform the first study to explore the general vulnerability of LiDAR-based perception architectures. We construct the first black-box spoofing attack based on the identified vulnerability, which universally achieves an 80% mean success rate on target models. We further present SVF, the first general architecture for robust LiDAR-based perception that reduces the mean spoofing attack success rate to 2.3%.

References

[1] Baidu Apollo. `http://apollo.auto`, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9297–9307, 2019.

[4] Pierre Biasutti, Vincent Lepetit, Jean-Francois Aujol, Mathieu Brédif, and Aurélie Bugeau. Lu-net: An efficient network for 3d lidar point cloud semantic segmentation based on end-to-end-learned 3d features and u-net. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[5] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2267–2281. ACM, 2019.

[6] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

[7] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[10] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. *arXiv preprint arXiv:1906.04691*, 2019.

[11] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[12] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016.

[13] Jonathan Petit, Bas Stottelaar, Michael Feiri, and Frank Kargl. Remote attacks on automated vehicles sensors: Experiments on camera and lidar. *Black Hat Europe*, 11:2015, 2015.

[14] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

[15] Hocheol Shin, Dohyun Kim, Yujin Kwon, and Yongdae Kim. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications. In *International Conference on Cryptographic Hardware and Embedded Systems*, pages 445–467. Springer, 2017.

[16] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. *arXiv preprint arXiv:1807.06288*, 2018.

[17] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893. IEEE, 2018.

[18] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4376–4382. IEEE, 2019.

[19] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. *arXiv preprint arXiv:1910.06528*, 2019.