Towards Robust LiDAR-based Perception in Autonomous Driving: General Black-box Adversarial Sensor Attack and Countermeasures

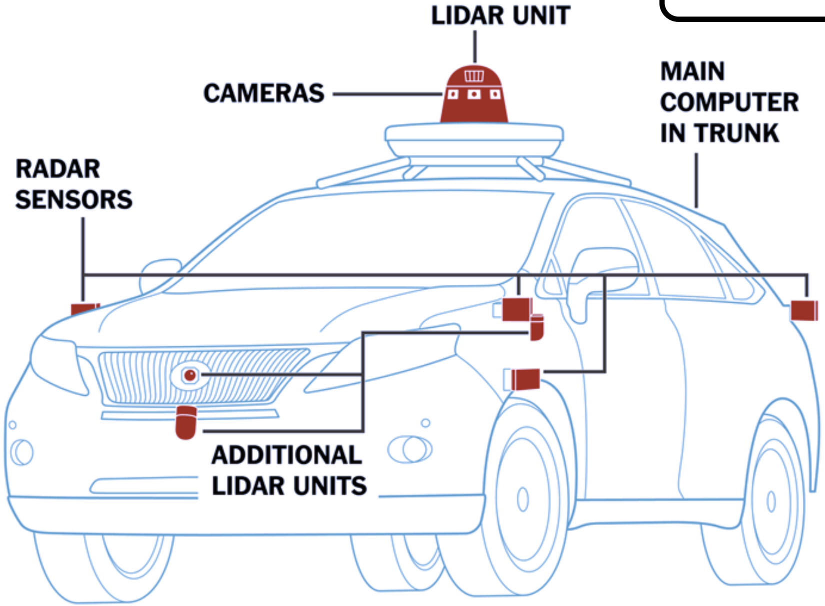
Jiachen Sun[1], Yulong Cao[1],
Qi Alfred Chen[2], and Z. Morley Mao[1]

1 UNIVERSITY OF MICHIGAN

2

29TH USENIX SECURITY SYMPOSIUM

Autonomous Vehicle (AV) Perception



Sensors → **Perception**: Position, Speed, **Object Detection**

↓

Prediction: Object Future Path

↓

Planning: AV Future Path

↓

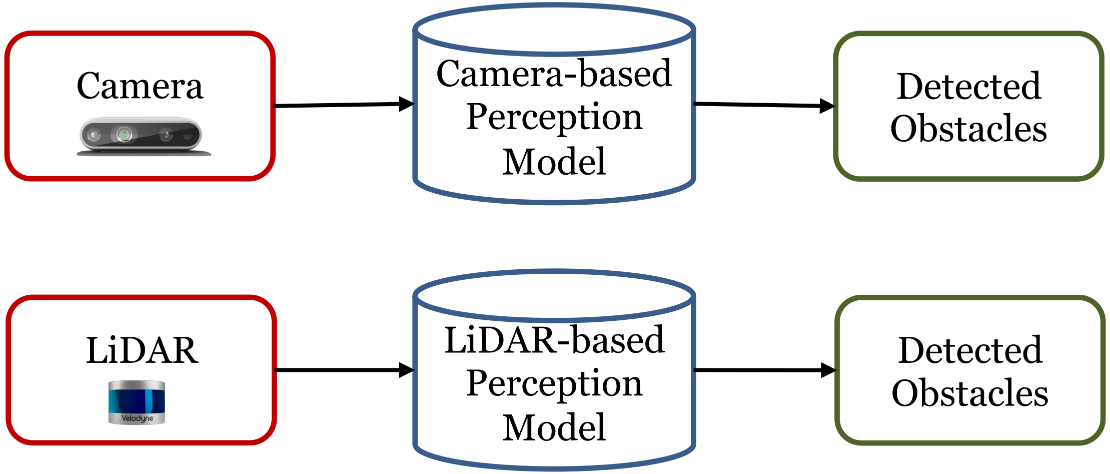
Control: Breaking, Steering, ...

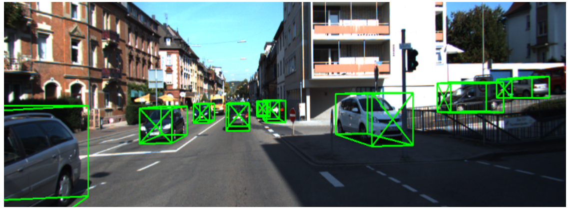
LiDAR: Light Detection And Ranging

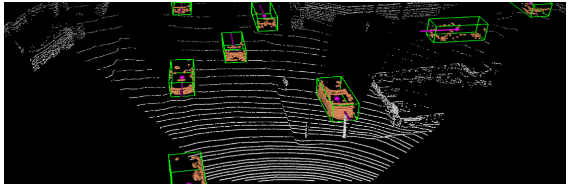
Picture ref: https://softwareengineeringdaily.com/2017/07/28/self-driving-deep-learning-with-lex-fridman/

Autonomous Vehicle (AV) Perception

- Machine learning, especially **deep learning**, is heavily adopted in state-of-the-art AV perception pipelines.

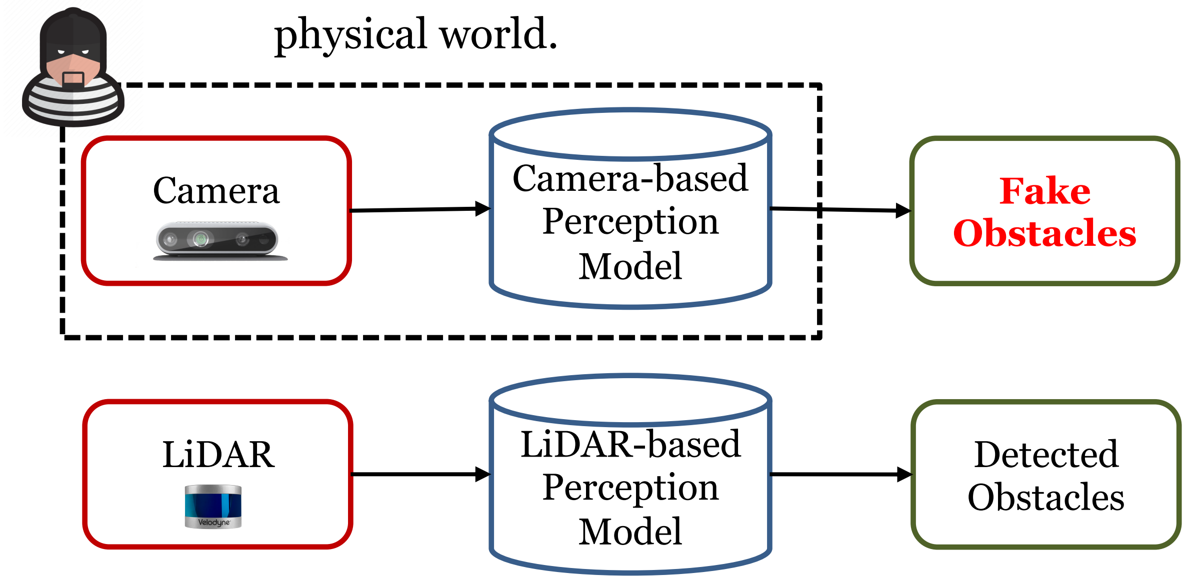


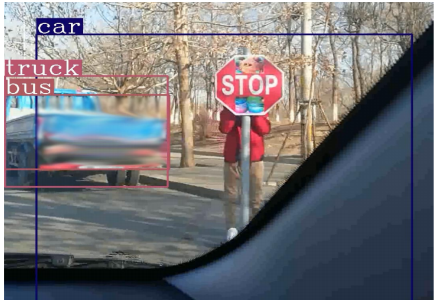




Related Work: Security of AV Perception

- Security of camera-based perception is well studied
  - Found to be vulnerable to adversarial machine learning (AML) attacks in the physical world.

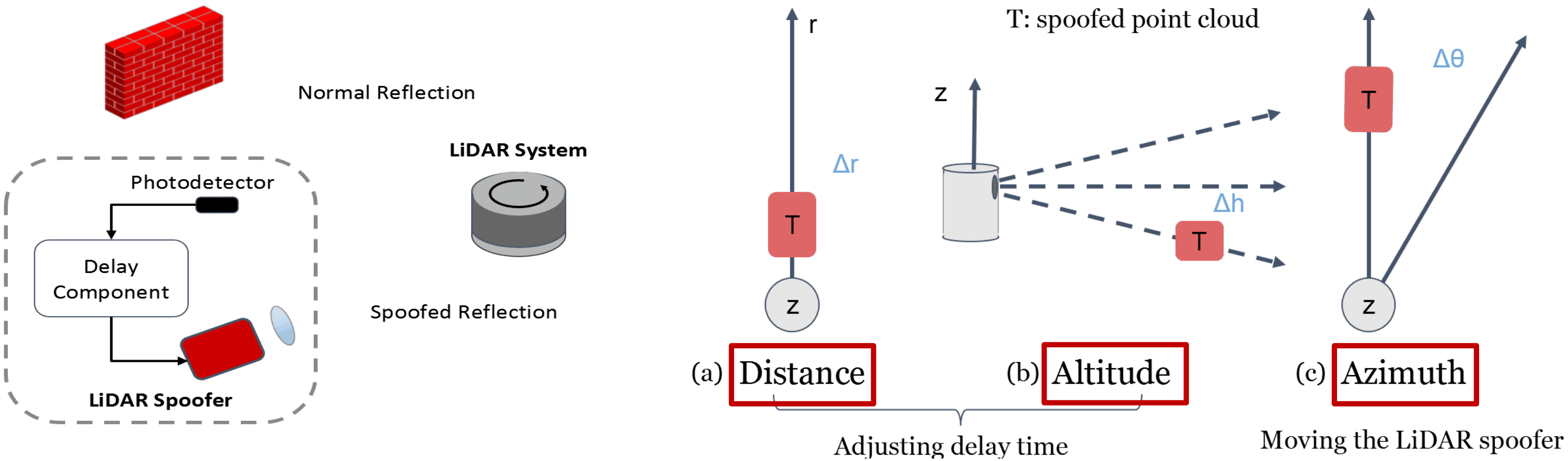




1. Eykholt, Kevin, et al. "Physical adversarial examples for object detectors." arXiv preprint arXiv:1807.07769 (2018).
2. Zhao, Yue, et al. "Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

Related Work: Security of LiDAR-based AV Perception

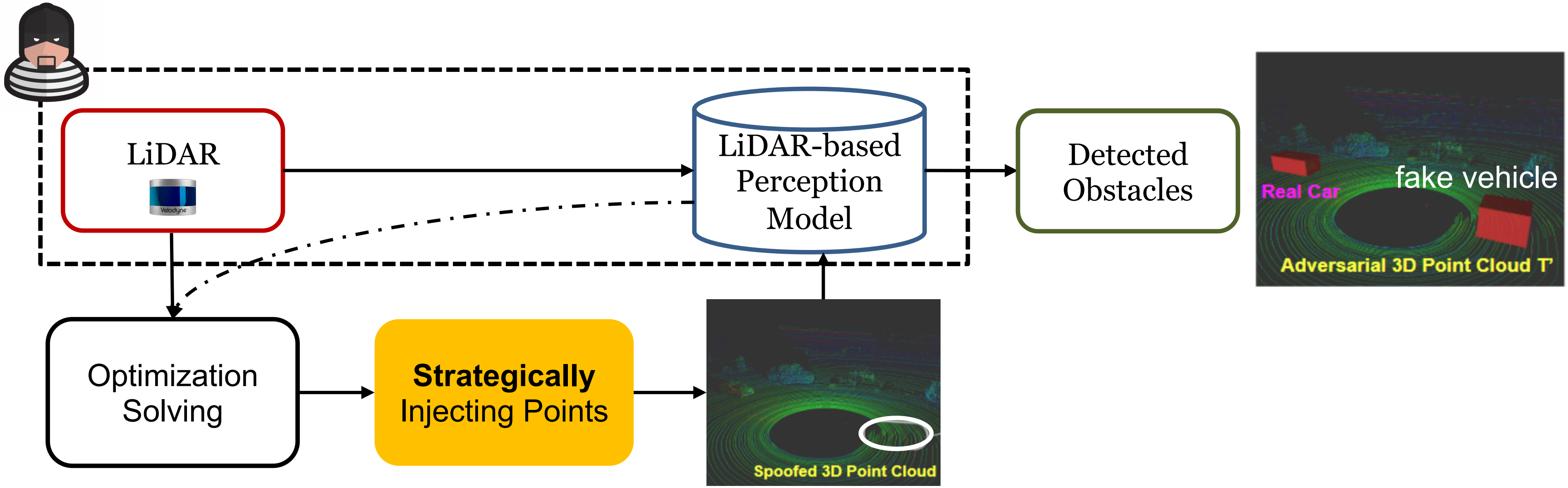
- Adv-LiDAR [1] demonstrated LiDAR-based perception is vulnerable to sensor attack with the help of **AML**.
  - **Formulation of the sensor attack capability.**
  - **Strategically** injecting points.



[1] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

Related Work: Security of LiDAR-based AV Perception

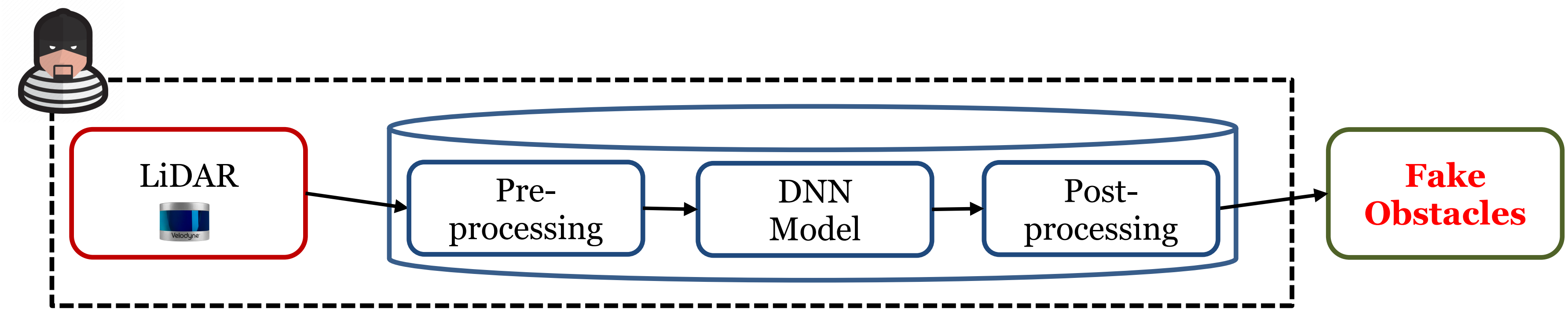
- Adv-LiDAR [1] demonstrated LiDAR-based perception is vulnerable to sensor attack with the help of **adversarial machine learning**.



[1] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

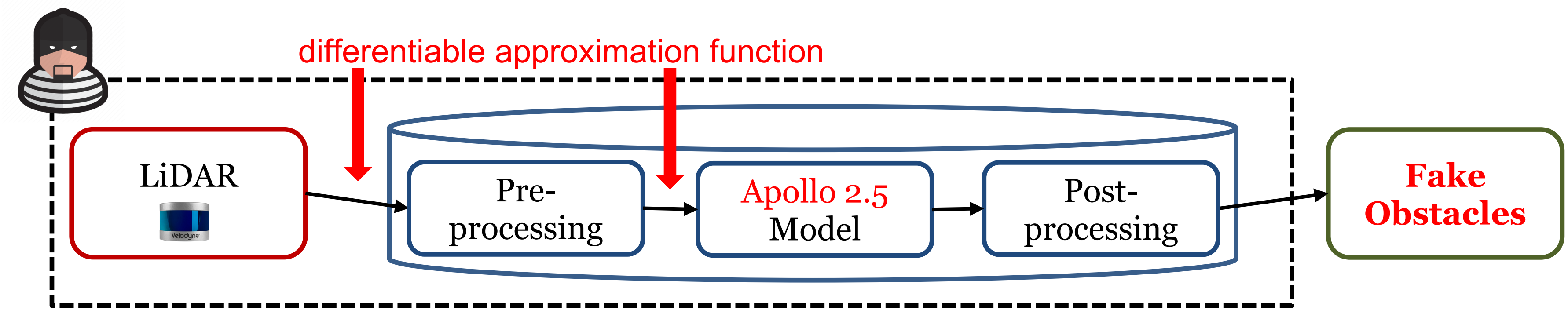
Motivation: Limitations of Existing Work

- **White-box attack limitation**
  - Adv-LiDAR assumes that attackers have **full** knowledge of LiDAR-based perception model along with its pre- and post-processing modules.



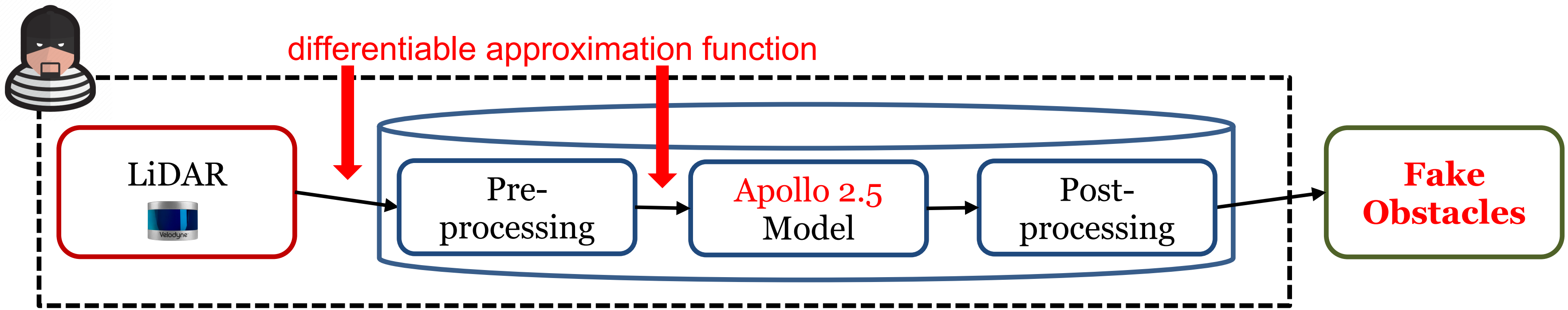
Motivation: Limitations of Existing Work

- White-box attack limitation
- **Attack generality limitation**
  - Adv-LiDAR **only** targets Apollo 2.5 model. The designed differentiable approximation function cannot generalize to other models.
  - Optimized adversarial examples generated by Adv-LiDAR **cannot** attack other models.



Motivation: Limitations of Existing Work

- White-box attack limitation
- Attack generality limitation
- **No practical defense solution**
  - There is no countermeasure proposed, making AVs still open to LiDAR spoofing attacks.



Contributions

- Explore a ***general*** vulnerability of current LiDAR-based perception architectures.
  - Construct the ***first black-box*** attacks and achieve ~80% mean attack success rates on all target models .

Contributions

- Explore a ***general*** vulnerability of current LiDAR-based perception architectures and construct the ***first black-box*** spoofing attack.

- Perform the ***first*** defense study, proposing CARLO as an anomaly detection module that can be stacked on LiDAR-based perception models.
  - Reduce the mean attack success rate to ~5.5% without sacrificing the detection accuracy.

Contributions

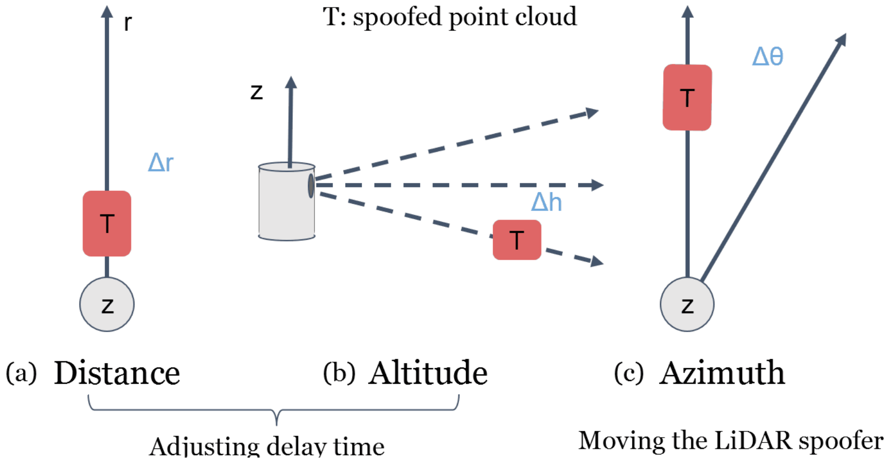
- Explore a ***general*** vulnerability of current LiDAR-based perception architectures and construct the ***first black-box*** spoofing attack.

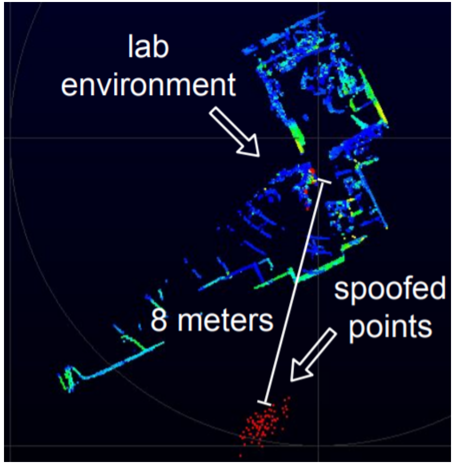
- Perform the ***first*** defense study, proposing CARLO as an anomaly detection module that can be stacked on LiDAR-based perception models.

- Design the ***first*** end-to-end ***general*** architecture for robust LiDAR-based perception.
  - Reduce the mean attack success rate to ~2.3% with similar detection accuracy to the original model.

Threat Model

- **Physical sensor attack capability**[1]
  - *Number of points*. Attackers can spoof at most 200 points into the LiDAR point clouds.
  - *Location of points*. Attackers can modify the *distance, altitude, and azimuth* of a spoofed point. *Azimuth* is within 10°.





[1] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

Threat Model

- Physical sensor attack capability[1]
  - *Number of points:* 200 points.
  - *Location of points:* distance, altitude, and azimuth (10°).
- **Attack model**
  - Goal: spoofing fake vehicles right in front of the victim AV [1] .
  - Attackers can control the spoofed points within the described sensor attack capability.
  - Attackers are **not** required to have access to the perception systems.

[1] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

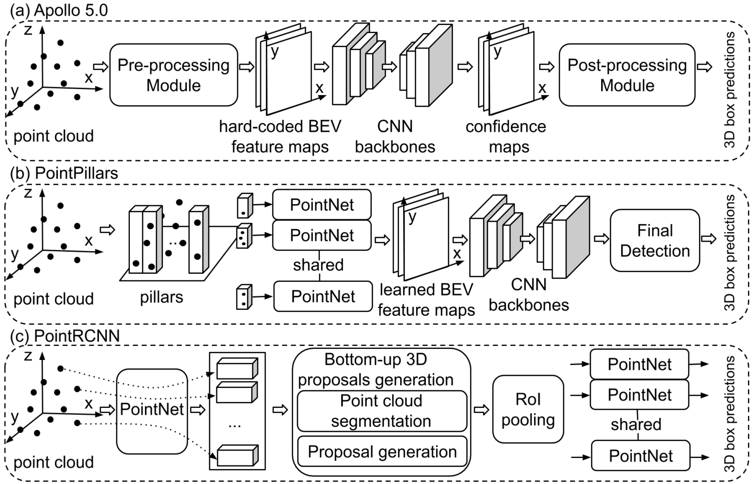
Threat Model

- Physical sensor attack capability[1]
  - *Number of points:* 200 points.
  - *Location of points:* distance, altitude, and azimuth (10°).
- Attack model
  - Goal: spoofing fake vehicles right in front of the victim AV [1] .
  - Within the described sensor attack capability.
  - Black-box access assumption.
- **Defense model**
  - We consider defending LiDAR spoofing attacks under **both** white- and black-box settings.
  - We focus on software-level countermeasures due to cost concerns.

[1] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.

State-of-the-art LiDAR-based Perception Models

- Bird's-eye view (BEV)-based Model
  - **Baidu Apollo 5.0**[1] (latest version)
  - Baidu Apollo 2.5 (model attacked in [2])
- Voxel-based Model
  - **PointPillars**[3] (CVPR'19, used by AutoWare [4])
  - VoxelNet[5] (CVPR'18)
- Point-wise Model
  - **PointRCNN**[6] (CVPR'19)
  - Fast PointRCNN[7] (ICCV'19)

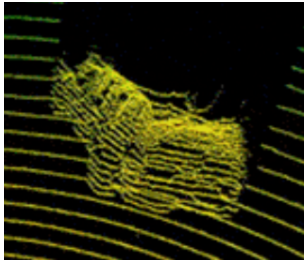


[1] Baidu Apollo. https://apollo.auto, 2020.
[2] Cao, Yulong, et al. "Adversarial sensor attack on lidar-based perception in autonomous driving." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.
[3] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[4] AutoWare.ai. https://gitlab.com/autowarefoundation/autoware.ai, 2020.
[5] Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[6] Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. "Pointrcnn: 3d object proposal generation and detection from point cloud." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
[7] Chen, Yilun, et al. "Fast point r-cnn." Proceedings of the IEEE International Conference on Computer Vision. 2019.

A General Vulnerability &
Black-box Adversarial Sensor Attack

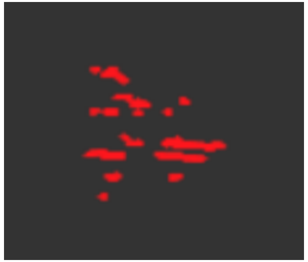
Behind the Scenes of Adv-LiDAR

- A valid front-near vehicle (located 5-8 meters right in front of the AV) should contain ~**2000** reflected points and occupy **15°** in azimuth[1] .



A valid front-near vehicle

- However, Adv-LiDAR was able to spoof a fake front-near vehicle by injecting much fewer amount of points (**80** points).

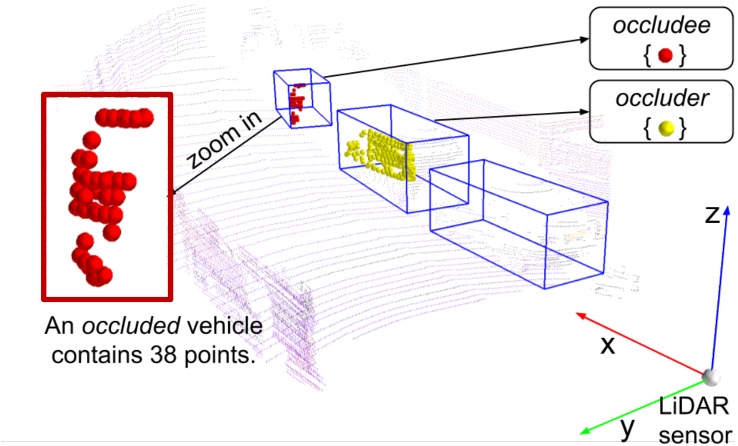


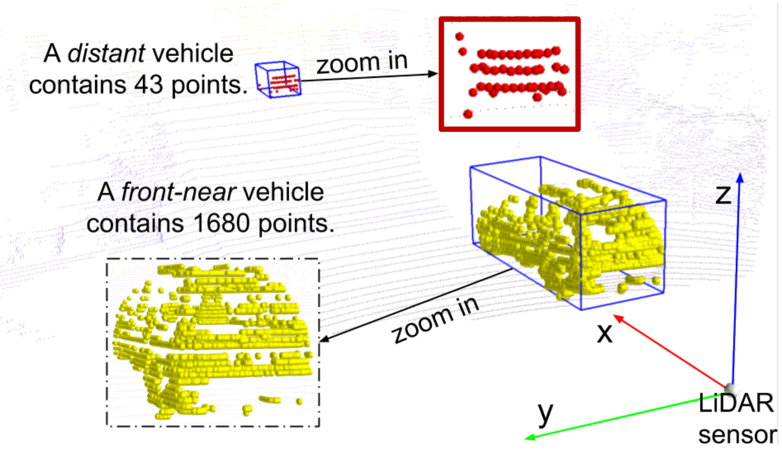
An attack trace generated by Adv-LiDAR

[1] Statistical study on KITTI dataset (64-beam LiDAR) KITTI Vision Benchmark: 3D Object Detection. http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d, 2020.

Behind the Scenes of Adv-LiDAR

- Two situations that a **valid** vehicle contains much fewer points in a LiDAR point cloud:
  - An **occluded** vehicle
  - A **distant** vehicle





False Positives

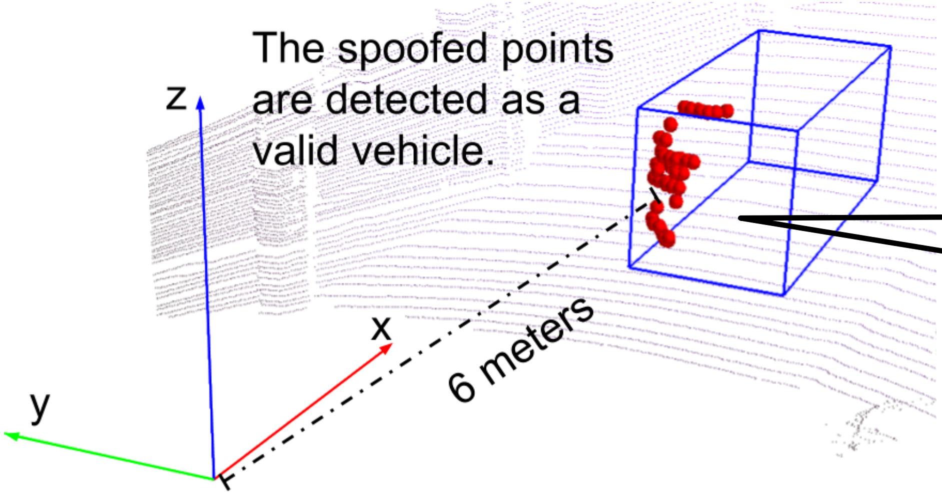
- Based on these observations, we find and validate two **false positive** (FP) conditions for the models:

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

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

Vulnerability Identification

Attackers can directly exploit such two **FP** conditions to fool the LiDAR-based perception models and spoof a fake vehicle with much fewer points.

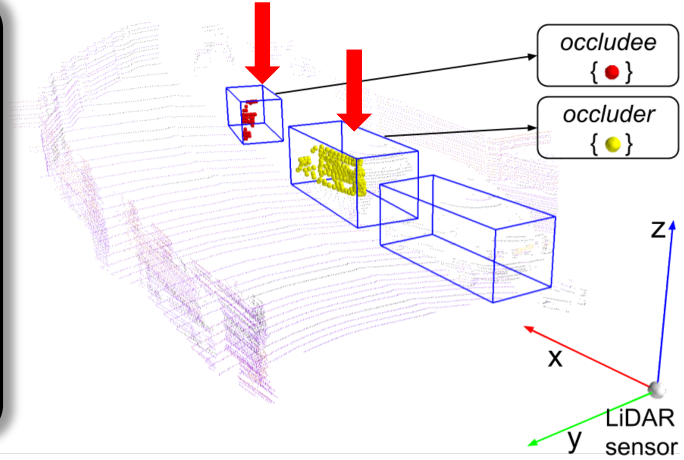


38 points
4.92° in azimuth

Vulnerability Identification

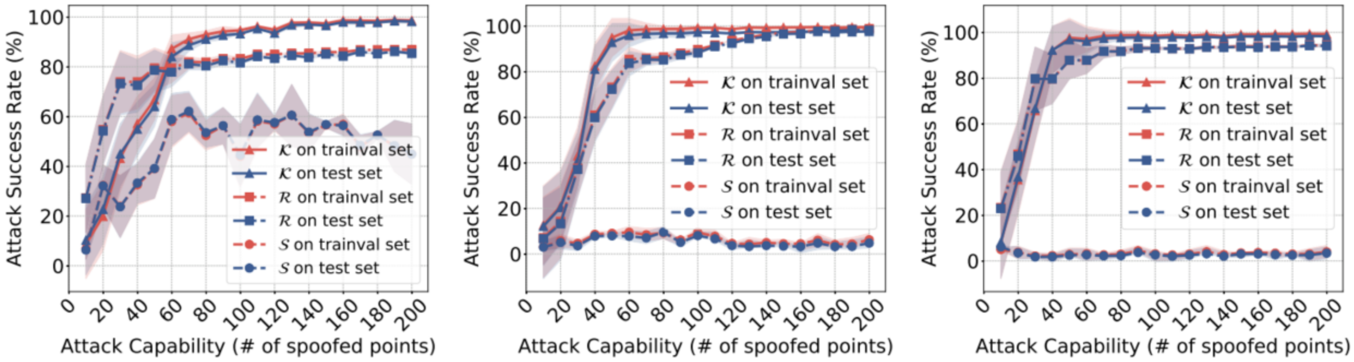
Attackers can directly exploit such two **FP** conditions to fool the LiDAR-based perception models and spoof a fake vehicle with much fewer points.

- FP1 ⟹ State-of-the-art models perform detection in the 3D space where the *occluder* and *occludee* stands **apart** with each other. However, DNN models prefer **local** features.
- FP2 ⟹ Object detection models are designed to be **insensitive** to the locations of objects.



Attack Evaluation

- Evaluation setup
  - Environments: KITTI[1] point clouds.
  - Combination of digital spoofing and physical spoofing.
- Black-box attacks universally achieve ~80% mean attack success rate (ASR) on all target models.



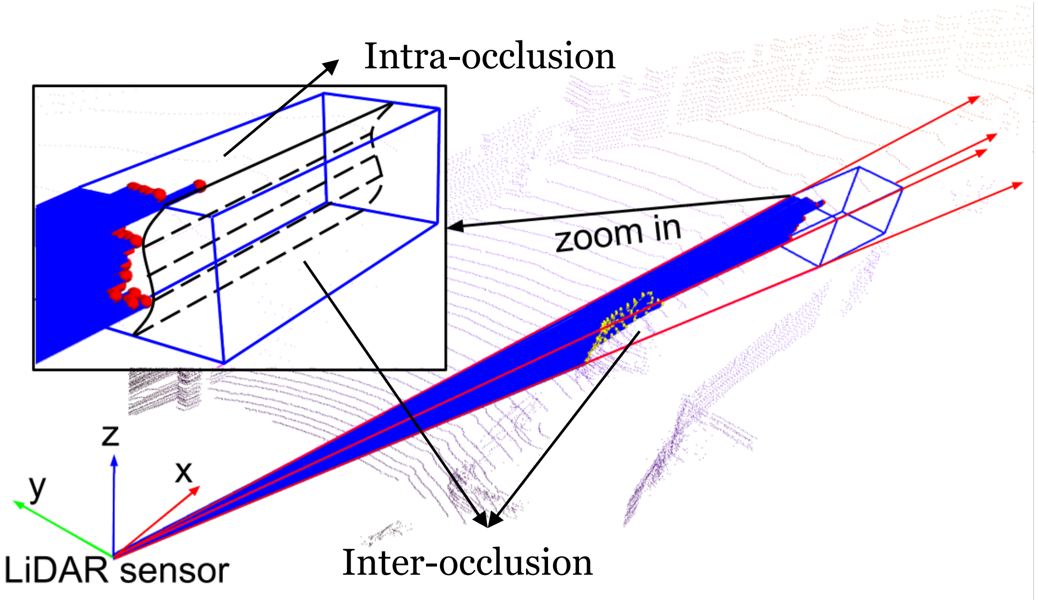
(a) *ASR* of Apollo 5.0. (b) *ASR* of PointPillars. (c) *ASR* of PointRCNN.

[1] KITTI Vision Benchmark: 3D Object Detection http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d, 2020.
Please refer to our paper for more detailed robustness analysis.

CARLO: oCclusion-Aware hieRarchy anomaLy detectiOn

Free Space Detection

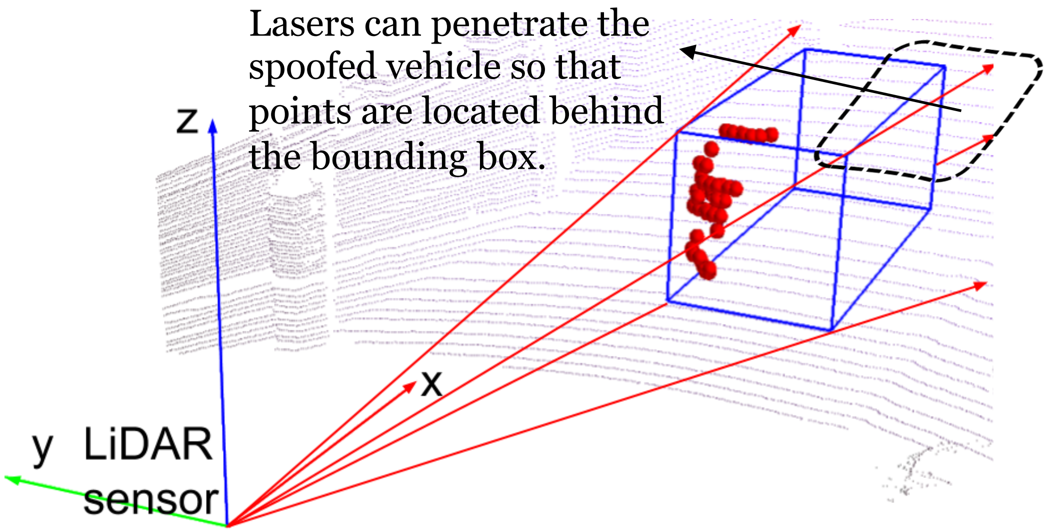
- **Free space**: the frustum (the straight-line path) from the LiDAR sensor and any point in the point cloud.



Due to intra-occlusion and inter-occlusion, there is **limited** free space inside a **valid** vehicle's bounding box.

Free Space Detection

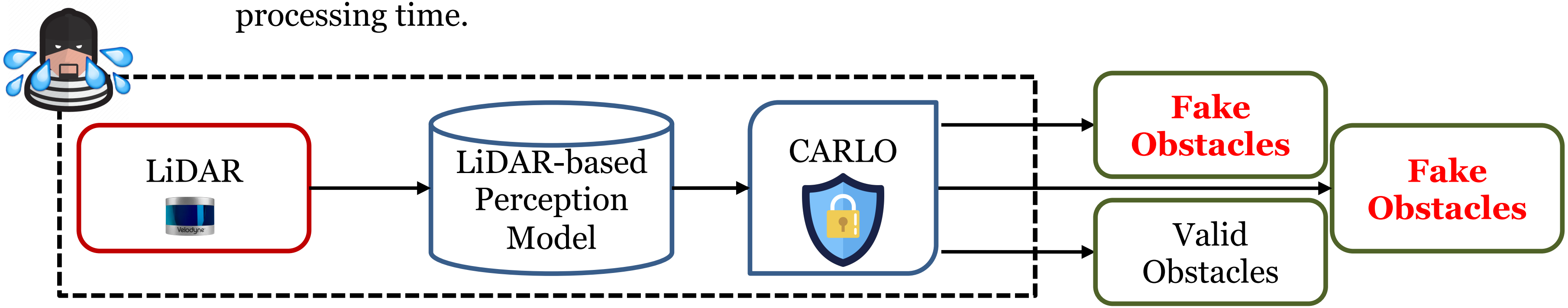
- **Free space**: the frustum (the straight-line path) from the LiDAR sensor and any point in the point cloud.



Due to the limited sensor attack capability, there is a **large** portion of free space inside a **fake** vehicle's bounding box.

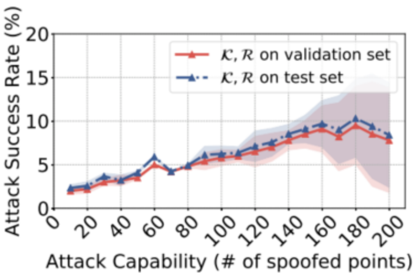
CARLO

- CARLO serves as a **post-processing** module leveraging free space as a **physical invariant** to detect spoofed vehicles.
- CARLO can be efficiently stacked onto existing LiDAR-based perception architectures.
  - **No** need for model re-training.
  - Consists of another **GPU-friendly** submodule to achieve around **8.5ms** per-vehicle processing time.

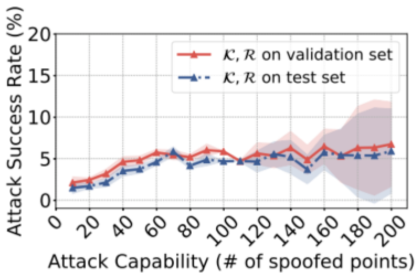


CARLO Evaluation

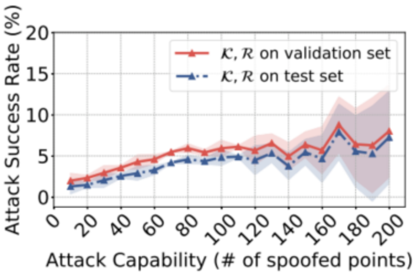
- CARLO overall reduces the mean attack success rate from ~80% to 5.5%.
- The accuracy of CARLO achieves at least 99.5%.
  - The 0.5% detection errors comes from some faraway vehicles.
  - No immediate impacts on AV's current driving decisions.
- CARLO can also defend the white-box attack, Adv-LiDAR, and its adaptive attack.



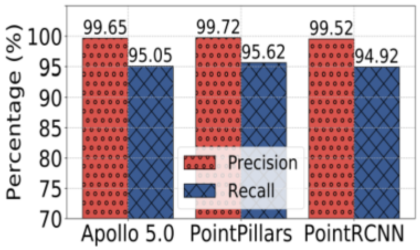
(a) CARLO-guarded Apollo 5.0.



(b) CARLO-guarded PointPillars.



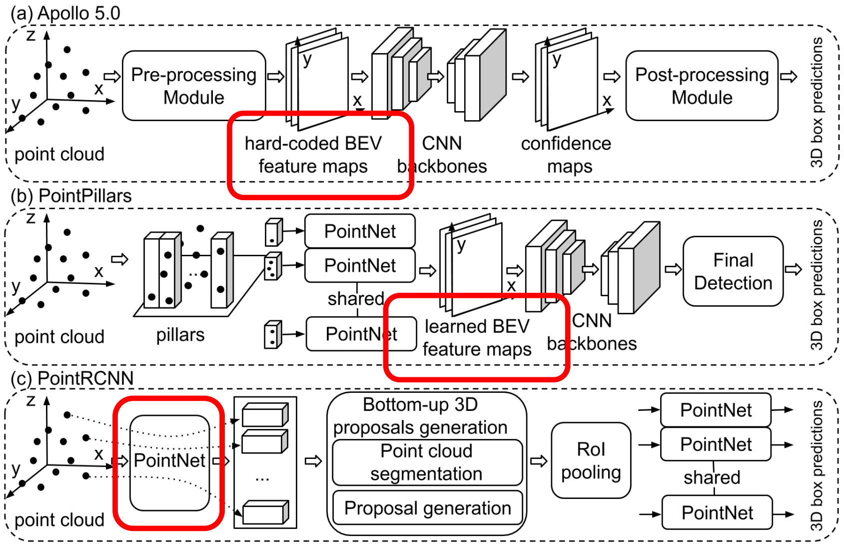
(c) CARLO-guarded PointRCNN.

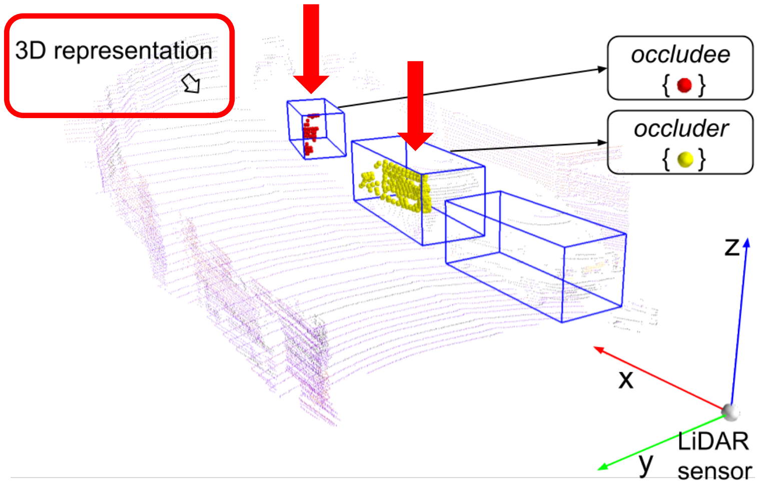


(d) Precision and recall of CARLO.

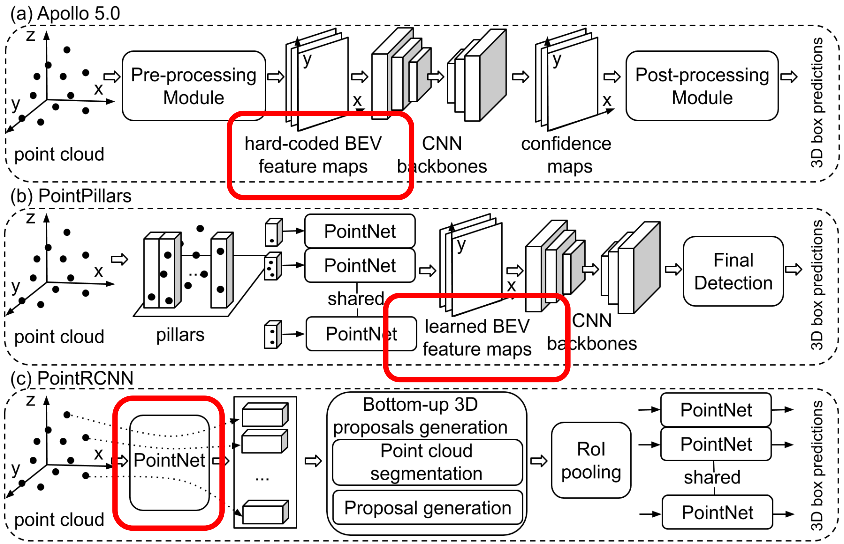
SVF: Sequential View Fusion
A Robust LiDAR-based Perception Architecture

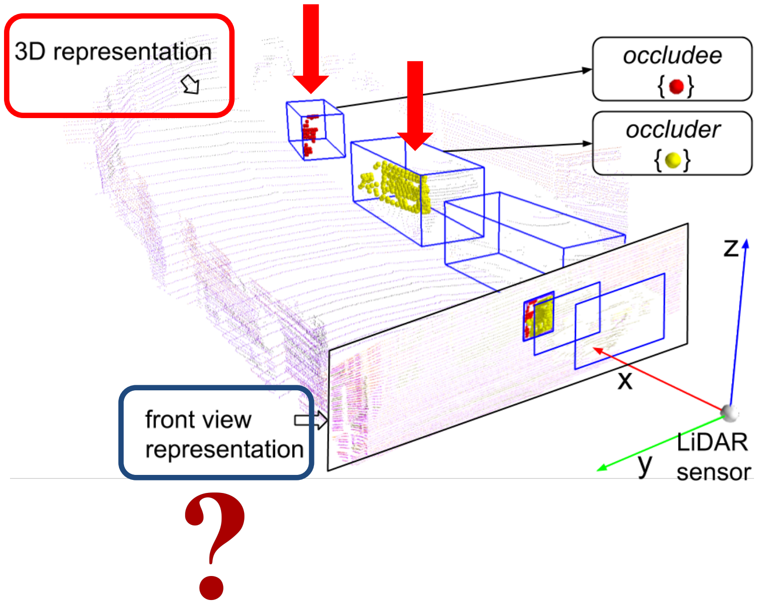
Existing Architectures Revisit





Existing Architectures Revisit

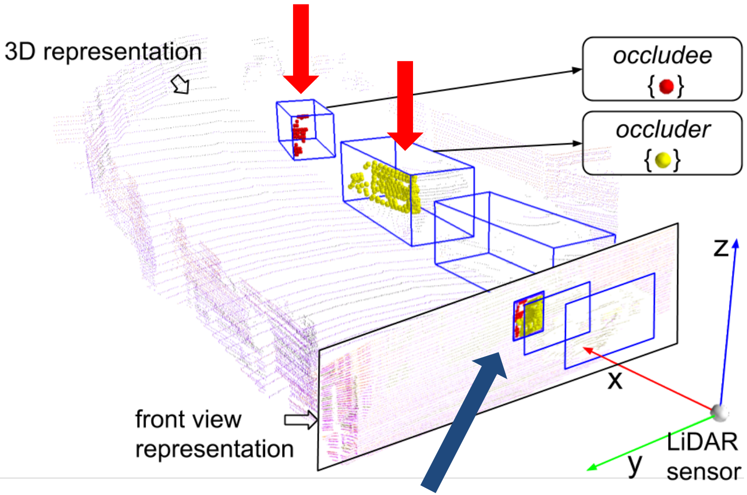




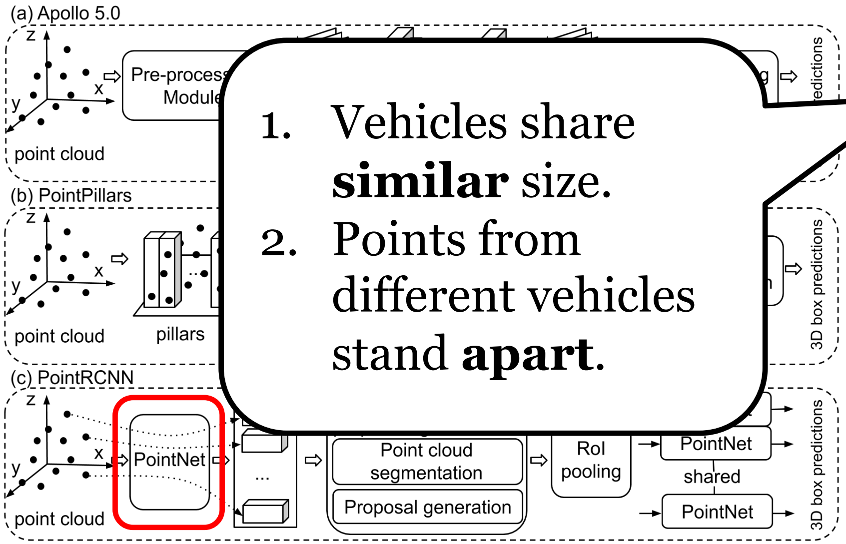
Front View (FV) Should Help!

- The *occluder* and *occludee* **neighbor** with each other in the FV, making it possible for DNN models to learn the **local** correlations. $\Longrightarrow$ FP1

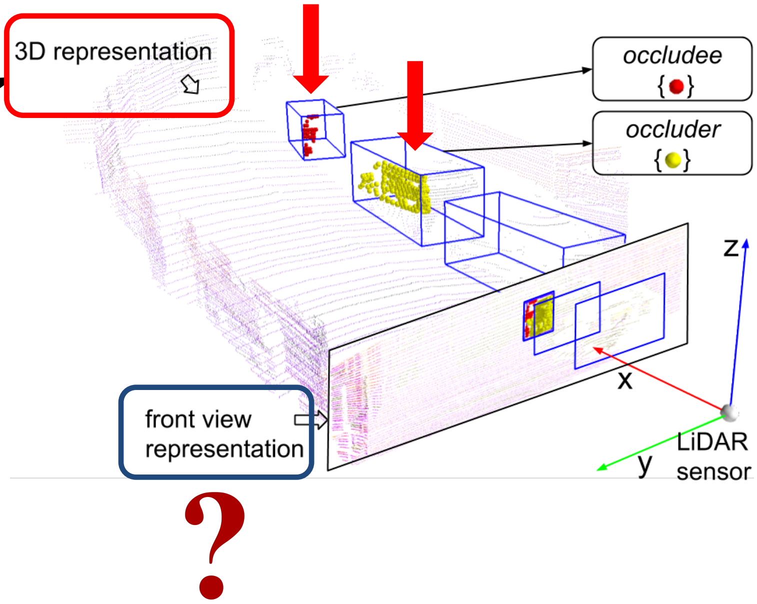
- A valid vehicle's points are **clustered** in the FV. However, due to the limited sensor attack capability, attack traces will **scatter** in the FV. $\Longrightarrow$ FP2



Front View (FV) Should Help!

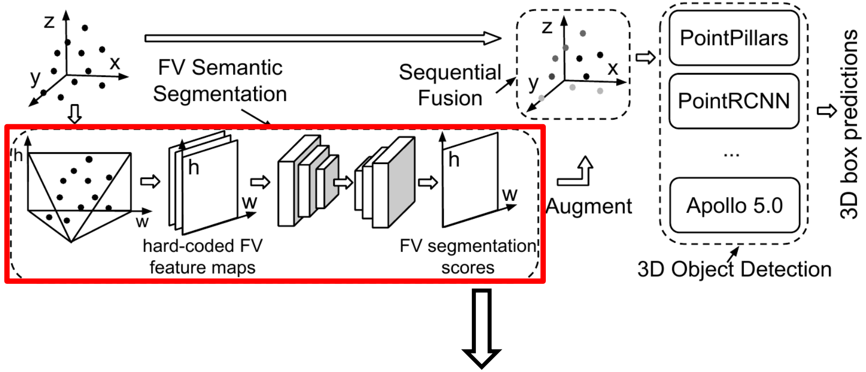


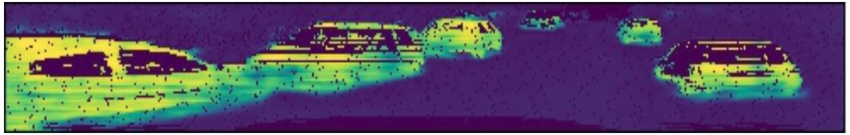
1. Vehicles share **similar** size.
2. Points from different vehicles stand **apart**.



Sequential View Fusion (SVF)

- Attach a semantic segmentation network to the FV representation.
  - Output the probability score of each point that it belongs to a vehicle.
  - An **easier** task as it does not need to estimate object-level output.
  - Achieve much more **satisfactory** results than the 3D object detection task over FV[1,2].

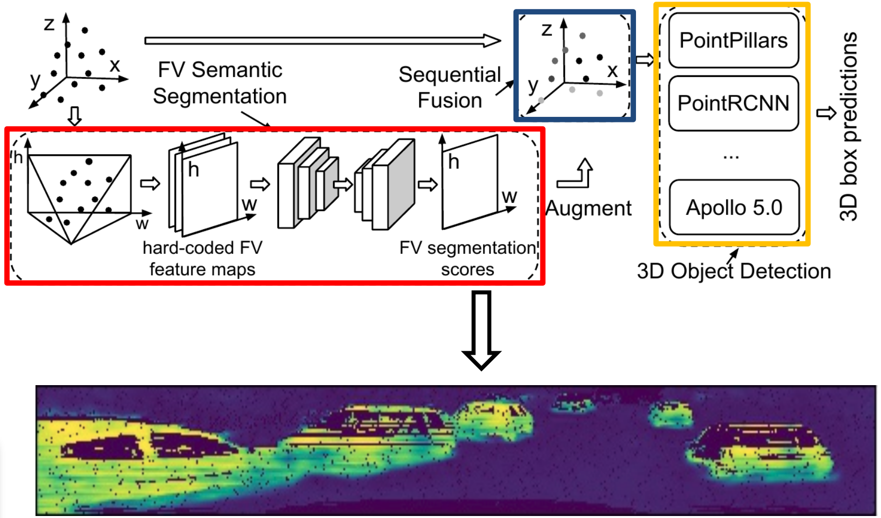




[1] Biasutti, Pierre, et al. "LU-Net: An Efficient Network for 3D LiDAR Point Cloud Semantic Segmentation Based on End-to-End-Learned 3D Features and U-Net." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.
[2] B. Wu, et al. Squeezeseg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud. In International Conference on Robotics and Automation,

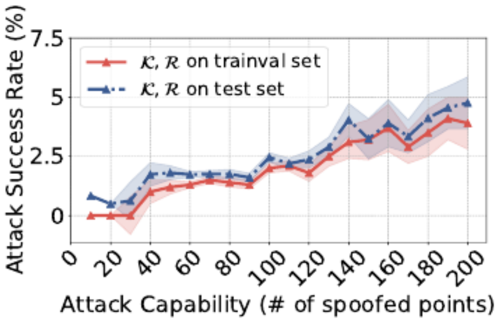
Sequential View Fusion (SVF)

- Attach a semantic segmentation network to the FV representation.
- The original point cloud is **augmented** with the scores from the FV.
- The final 3D object detection module takes the augmented point cloud as input.
  - **Reserve the advantages of detection on 3D representations with useful information from FV.**

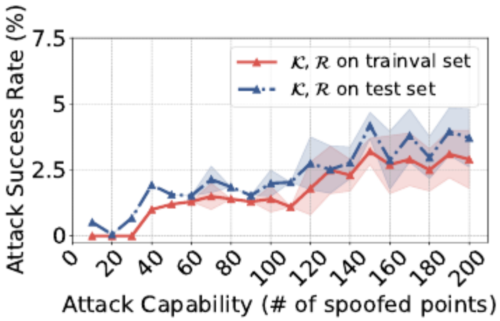


SVF Evaluation

- SVF models are shown to be robust against LiDAR spoofing attacks, where the mean success rates are merely ~2.3%.
  - Similar detection accuracy with the original models.
- SVF models are also resilient to the state-of-the-art white-box attack, Adv-LiDAR, and its adaptive attack.



(a) *ASR* of SVF-PointPillars.



(b) *ASR* of SVF-PointRCNN.

Please refer to our paper for white-box robustness evaluation of SVF.

Limitations

- Attack Practicality
  - Large-scale evaluations are based on digital LiDAR spoofing.
  - Physical LiDAR spoofing is performed in in-lab environments.
  - No real road test due to cost concerns.
- Vulnerability Completeness
  - The identified vulnerability only partially explains the success of LiDAR spoofing attacks.
- Defenses Guarantees
  - Both defense solutions cannot provide strong guarantees.
  - Defenses may fail when the sensor attack capability improves dramatically (e.g., injecting 1500 points).

Conclusion

- Explore a ***general*** vulnerability of current LiDAR-based perception architectures and construct the ***first black-box*** spoofing attack.
- Perform the ***first*** defense study, proposing CARLO as an anomaly detection module that can be stacked on LiDAR-based perception models.
- Design the ***first*** end-to-end ***general*** architecture for robust LiDAR-based perception.

Thank you !

Q & A

Contact us!
jiachens@umich.edu