Safety21

INNOVATING FOR SAFETY

US DOT National
University Transportation Center for Safety

Carnegie Mellon University

Cost-Effective Radar-Based Depth Perception and Scene Interpretation

John M. Dolan (PI) (https://orcid.org/0000-0003-2062-100X)

Final Report – August 31, 2026

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under [grant number 69A3552344811 / 69A3552348316] from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

Technical Report Documentation Page

<table>
<tr><td>1. Report No.

Enter the report number assigned by the sponsoring agency.</td><td>2. Government Accession No.</td><td>3. Recipient’s Catalog No.</td></tr>
<tr><td colspan="2" rowspan="2">4. Title and Subtitle

Cost-Effective Radar-Based Depth Perception and Scene Interpretation</td><td>5. Report Date

August 31, 2026</td></tr>
<tr><td>6. Performing Organization Code

Enter any/all unique numbers assigned to the performing organization, if applicable.</td></tr>
<tr><td colspan="2">7. Author(s)

John M. Dolan, Ph.D. https://orcid.org/0000-0003-2062-100X</td><td>8. Performing Organization Report No.

Enter any/all unique alphanumeric report numbers assigned by the performing organization, if applicable.</td></tr>
<tr><td colspan="2" rowspan="2">9. Performing Organization Name and Address

The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213</td><td>10. Work Unit No.</td></tr>
<tr><td>11. Contract or Grant No.

Federal Grant # 69A3552344811 / 69A3552348316</td></tr>
<tr><td colspan="2" rowspan="2">12. Sponsoring Agency Name and Address

US Department of Transportation, 1200 New Jersey Avenue, SE Washington, DC 20590</td><td>13. Type of Report and Period Covered

Final Report (July 1, 2025 – July 31, 2026)</td></tr>
<tr><td>14. Sponsoring Agency Code

USDOT</td></tr>
<tr><td colspan="3">15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.</td></tr>
</table>

16. Abstract

One of the current challenges for autonomous vehicles (AV) is providing provable safety in urban driving and highway driving. LIDAR provides accurate depth and shape information, but remains quite expensive. In this research we explore the viability of fusing 4D radar images with camera images as an alternative to using LIDAR. Our research aims to establish the data fusion of radar and camera images in autonomous driving as a replacement for data fusion using traditional LIDAR, radar, and camera images. Our current approach demonstrates that a trained radar detector can achieve approximately 3.5 meters of Unidirectional Chamfer Distance (UCD) against ground truth LiDAR data on a dataset recently released by Delft University of Technology (DUT) in the Netherlands, which is 28% better than the state of the art (SOTA).

Our approach integrates the images obtained from deep neural network (DNN) 4D radar detectors with conventional camera RGB images. It introduces a novel pixel positional encoding algorithm inspired by Bartlett's spectrum estimation technique. This algorithm transforms radar depth maps and RGB images into a unified pixel image subspace facilitating the learning of their potential correspondence. Our method effectively leverages high-resolution camera images to train radar depth map generative models, addressing the limitations of conventional radar detectors in complex vehicular environments while sharpening the radar output. We develop spectrum estimation algorithms tailored for radar depth maps and RGB images, a comprehensive training framework for data-driven generative models, and a camera-radar deployment scheme for AV operation.

17. Key Words	**18. Distribution Statement**
Autonomous vehicles, 4D radar, sensor fusion, depth estimation	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161. Enter any other agency mandated distribution statements. Remove NTIS statement if it does not apply.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	14	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

1. Executive Summary

One of the current challenges for autonomous vehicles (AV) is providing provable safety in urban driving and highway driving. LIDAR provides accurate depth and shape information, but remains quite expensive. In this research we explore the viability of fusing 4D radar images with camera images as an alternative to using LIDAR. Our research aims to establish the data fusion of radar and camera images in autonomous driving as a replacement for data fusion using traditional LIDAR, radar, and camera images. Our current approach demonstrates that a trained radar detector can achieve approximately 3.5 meters of Unidirectional Chamfer Distance (UCD) against ground truth LiDAR data on a dataset recently released by Delft University of Technology (DUT) in the Netherlands, which is 28% better than the state of the art (SOTA).

In this project we improved radar sensing in four respects: 1) improved performance of the sensor-LiDAR fusion method pioneered by the University of Delft; 2) application of the results to adverse weather, especially fog; 3) development of a simulation environment for additional experiments; and 4) object classification on real data. We have also made progress in neural representation-based methods that yield higher fidelity and compactness in sensing.

2. Improving radar-camera sensor performance

We created a cost-effective perception framework for autonomous vehicles [1] that integrates 4D radar depth maps with RGB camera images through a novel spectrum-learning-based fusion. Conventional 4D radars, though robust and affordable, produce sparse depth maps compared to lidars. To address this, we propose a pixel positional encoding algorithm inspired by Bartlett's spatial spectrum estimation, transforming both radar and camera modalities into a shared spatial spectrum subspace. This transformation aligns their frequency components, enabling spectrum-based supervision of radar data using high-resolution camera imagery.

The proposed pipeline consists of radar and camera preprocessing, spectrum estimation, and a ResNet-based generative model trained to reconstruct dense radar depth maps from the encoded data. Using the RaDelft dataset, the method achieves substantial performance gains, improving Unidirectional Chamfer Distance (UCD) by 52.6 % and Mean Absolute Error (MAE) by 24.2 % over the state-of-the-art deep (SOTA) radar detector. Moreover, we report a 3.9x increase in Pearson correlation and 76.7x improvement in mutual information between the two modalities, demonstrating stronger radar-camera alignment (see Fig. 1 for qualitative examples). Ablation studies confirm the importance of the number of spectral segments, semantic segmentation quality, and spectral representation for optimal learning. The proposed system aims to eliminate the need for costly lidars during real-time operation by training radar detectors with camera groundtruth offline, yielding sharper and denser radar depth maps. Spectrum learning offers a principled, signal-processing-informed pathway toward affordable, high-quality perception in autonomous vehicles, paving the way for future work on variance reduction and real-time implementation.

Figure 1. Qualitative results from the training module for four example frames. From left to right for each example frame: Scene in RGB, ground truth, original radar depth map, output from trained ResNet101. In each of the 4 frames, observe that our approach leads to sharper depth maps.

We found something counterintuitive for the Deep Learning (DL) and Autonomous Vehicle communities: bigger does not mean better when it comes to fast, predictive DL models used in predicting point clouds from raw 4D Radar signals. Evaluating on the RaDelft dataset, we do find substantive gains when transitioning from ResNet18 panoptic segmentation backbone to ResNet50 [2][3]. However, there is a significant drop-off in performance when scaling to deeper networks such as ResNet101 or ResNet152. Still, with the optimal result being found with ResNet50, there remain significant questions that need to be addressed: (1) How well will deeper models scale with larger datasets? (2) How can we further improve the overall detection quality of 4D radar-derived 3D point clouds? (3) Is it possible to achieve denser, more accurate spatial representations using 4D Radar, comparable to those of the rich scene representations the AV community relies on through LiDAR? We have begun to investigate some of these questions by leveraging recent advancements in transformer-based technologies, which have enabled numerous applications requiring spatiotemporal understanding through context. By leveraging the attention mechanisms present in the transformer architecture, we believe it is possible to not only improve the overall prediction quality of point clouds using 4D Radar, but also reconstruct them to achieve denser representations of a scene [4][5][6][7]. In hopes of testing and evaluating such a model, we require larger sums of data beyond what is provided within the RaDelft dataset. As a result, we aim to utilize additional data from new sources like NVIDIA's Physical AI dataset containing 1700+ hours of time-synchronized 4D Radar, camera, and LiDAR frames across 20+ countries [8]. We hope to produce such a model capable of offering a reliable alternative to LiDAR at a fraction of the cost while retaining Radar's inherent robustness to adverse weather conditions, something LiDAR and camera sensor suites lack.

3. Robust radar-camera fusion in adverse weather

We propose a camera-guided radar labeling pipeline that transfers semantic priors from camera segmentation to radar, followed by class-aware clustering to refine the transferred labels. This cross-modal supervision significantly improves the detection probability compared to a radar-only setup. Third, we study robustness to adversarial weather by simulating fog at varying intensities in the image domain and measuring how fog degrades image semantic segmentation and final radar labeling performance. This analysis clarifies how camera-based labeling benefits radar under degraded visual conditions.

Fig. 2 illustrates the camera-radar fusion labeling pipeline. Camera image and radar depth maps are segmented separately, and fused into a joint semantic map. This map is used to assign per-point labels after projecting the radar point cloud into the camera frame. The resulting labeled radar points are further spatially refined.

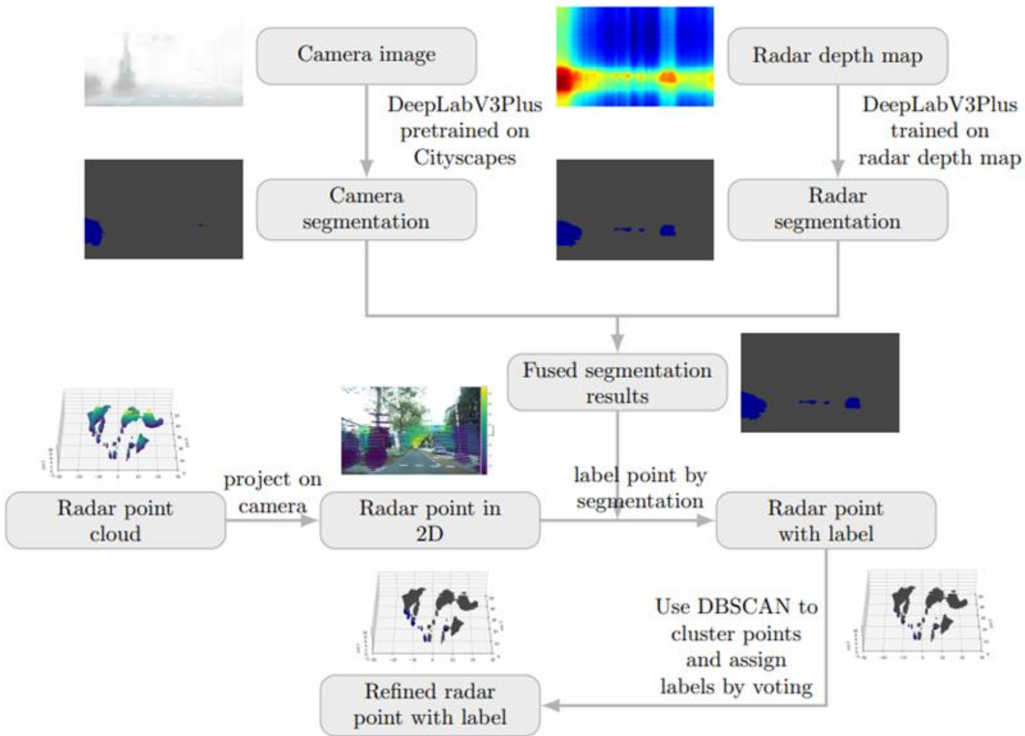

Figure 2. Radar–camera fusion and radar point cloud labeling pipeline

4. Radar simulator

We developed and validated a 4D radar simulation pipeline in the CARLA autonomous driving simulator to reproduce the statistical and structural properties of real-world radar datasets such as RaDelft [2]. The proposed system integrates synchronized LiDAR, radar, and RGB camera streams and transforms them into a unified radar representation aligned with real-world 4D radar data. Sensor measurements are converted into spherical coordinates consisting of range, azimuth, elevation, and Doppler velocity, while additional preprocessing steps generate structured radar cubes for downstream analysis and model evaluation. The pipeline also includes coordinate frame transformations, Doppler approximation, and voxelized radar representations designed to emulate real-world radar sensing characteristics while remaining compatible with CARLA's simulation framework.

We then extended the work beyond sensor emulation to evaluate the simulation-to-real domain gap between CARLA-generated radar data and the RaDelft dataset. We evaluate similarity based on semantic distributions, voxel occupancy analysis, statistical divergence metrics, and geometric alignment. Using instance segmentation, we extracted semantic object distributions for classes such as vehicles, pedestrians, riders, trucks, and buses from both simulated and real-world data. Experimental results showed that dominant object classes exhibit similar distributions across datasets, indicating that the simulator preserves core traffic structure, while discrepancies remained for rare object classes such as pedestrians and riders. We computed Jensen-Shannon Divergence (JSD), bidirectional Kullback-Leibler (KL) divergence, and L1 distance between simulated and real-world radar distributions. The CARLA-to-RaDelft comparison achieved a JSD of 0.0411 and bidirectional KL divergences of 0.1755 and 0.1716, which are lower than the natural scene-to-scene variation observed within RaDelft itself (JSD 0.0695, KL 0.2915). We additionally evaluated geometric consistency using Chamfer distance, where CARLA Town02 achieved a distance of 2.54, comparable to the range observed across real-world RaDelft scenes. Finally, we replicated Delft's ResNet-based radar detector [2] in the CARLA simulation environment to evaluate downstream generalization. Detection performance remained consistent across domains, with probability of detection (PD) values of 0.6213 on RaDelft and 0.6104 on 20k simulated CARLA frames.

5. Radar object classification

A key challenge in roadside radar perception is the reliable classification of detected vehicles, particularly when radar point cloud signatures overlap across object classes at extended ranges. Our contribution to this work focused on designing and analyzing a principled classification framework to address this problem. We developed a binary vehicle classifier based on the Neyman-Pearson (NP) criterion, which provides a framework for hypothesis testing. The discriminating feature is the 2D positional standard deviation of each radar-tracked object's point cloud cluster, a measure of spatial spread that differs characteristically between passenger cars and motorcycles. The baseline NP classifier, however, yielded a total classification error rate of 52.1%, driven by the substantial overlap between the two class distributions.

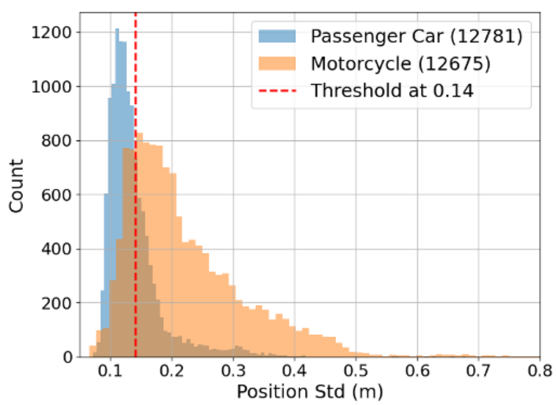

Fig 3. Class distribution with respect to position standard deviation.

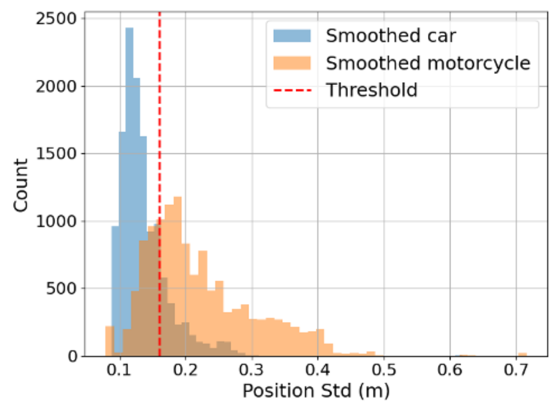

Fig 4. Class distribution after temporal smoothing with respect to position standard deviation.

To address this, we proposed a temporal averaging technique that applies a uniform moving average filter over a window of 100 frames to the per-track feature time series. This effectively suppresses short-term measurement noise, which arises from multipath interference, RCS fluctuations, and aspect-angle variation, while preserving the underlying class signature. The smoothing reduced the variance of both class distributions, yielding tighter and more separable distributions. A subsequent linear search over the threshold space identified an optimal decision boundary at $\eta^* = 0.16$ m, reducing the total classification error rate to 42.7%, which is a 9.4% improvement over the baseline. Figures 3 and 4 show the class distributions before and after temporal smoothing with optimal decision thresholds.

Project Publications

1. Alsakabi, M., Erickson, A., Dolan, J.M., & Tonguz, O. (2025). Toward a Low-Cost Perception System in Autonomous Vehicles: A Spectrum Learning Approach. Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), Australia. https://arxiv.org/abs/2502.01940
2. Alsakabi, M., Dolan, J.M., & Tonguz, O. (2026). FM-SIREN & FM-FINER: Implicit Neural Representation Using Nyquist-based Orthogonality. To appear in IEEE Workshop for Machine Learning in Signal Processing (MLSP). https://arxiv.org/abs/2606.06671

References

[1] M. Alsakabi, A. Erickson, J. M. Dolan, and O. K. Tonguz, “Toward a low-cost perception system in autonomous vehicles: A spectrum learning approach,” arXiv preprint arXiv:2502.01940, 2025.

[2] I. Roldan, A. Palffy, J. F. P. Kooij, D. M. Gavrila, F. Fioranelli, and A. Yarovoy, “A Deep Automotive Radar Detector Using the RaDelft Dataset,” IEEE Transactions on Radar Systems, vol. 2, pp. 1062–1075, 2024, doi: 10.1109/TRS.2024.3485578.

[3] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-Maadeed, “Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches,” Computer Science Review, vol. 53, p. 100645, Aug. 2024, doi: 10.1016/j.cosrev.2024.100645.

[4] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 77–85. doi: 10.1109/CVPR.2017.16.

[5] J. Shi, L. Xu, P. Li, X. Chen, and S. Shen, “Temporal Point Cloud Completion With Pose Disturbance,” IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 4165–4172, Apr. 2022, doi: 10.1109/LRA.2022.3146585.

[6] M. Seo, H. Lim, K. Lee, L. Carlone, and J. Park, “BUFFER-X: Towards Zero-Shot Point Cloud Registration in Diverse Scenes,” Mar. 11, 2025, arXiv: arXiv:2503.07940. doi: 10.48550/arXiv.2503.07940.

[7] J. Yin, J. Shen, C. Guan, D. Zhou, and R. Yang, “LiDAR-Based Online 3D Video Object Detection With Graph-Based Message Passing and Spatiotemporal Transformer Attention,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020, pp. 11492–11501. doi: 10.1109/CVPR42600.2020.01151.

[8] NVIDIA Corporation. Physical AI Autonomous Vehicles Dataset. Hugging Face, 2026, huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles. Accessed 26 May 2026.

[9] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. Van Gool, “Semantic understanding of foggy scenes with purely synthetic data,” in IEEE International Conference on Intelligent Transportation Systems (ITSC), 2019.

[10] B. Sun, I. Roldan, and F. Fioranelli, “Automatic labelling & semantic segmentation with 4d radar tensors,” in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.