NATIONAL CENTER FOR UNDERSTANDING FUTURE
TRAVEL BEHAVIOR AND DEMAND

Final Project Report

Deep Learning with LiDAR Point Cloud Data for Automatic Roadway Health Monitoring

BY

Rafael Trinidad
Email: retrinidad@cpp.edu

Ardavan Sherafat
Email: sherafat@cpp.edu

Hao Ji, PhD
Email: hji@cpp.edu

Yongping Zhang, PhD
Email: yongpingz@cpp.edu

Wen Cheng, PhD
Email: wcheng@cpp.edu

Omar Mora, PhD
Email: oemora@cpp.edu

California State Polytechnic University, Pomona
3801 W Temple Ave, Pomona, CA 91768

September 2025

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. N/A	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Deep Learning with LiDAR Point Cloud Data for Automatic Roadway Health Monitoring		**5. Report Date** September 25, 2025 **6. Performing Organization Code** N/A
7. Author(s) Rafael Trinidad, https://orcid.org/0009-0003-6234-1310 Ardavan Sherafat Hao Ji, PhD, https://orcid.org/0000-0001-8303-1491 Yongping Zhang, PhD, https://orcid.org/0000-0002-5935-3834 Wen Cheng, PhD, https://orcid.org/0000-0001-7225-6169 Omar Mora, PhD, https://orcid.org/0000-0002-5884-9205		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address California State Polytechnic University, Pomona 3801 W Temple Ave, Pomona, CA 91768		**10. Work Unit No. (TRAIS)** N/A **11. Contract or Grant No.** 69A3552344815 and 69A3552348320
12. Sponsoring Agency Name and Address U.S. Department of Transportation, University Transportation Centers Program, 1200 New Jersey Ave, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report, 2024-2025 **14. Sponsoring Agency Code** USDOT OST-R
15. Supplementary Notes Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.		

16. Abstract

Maintaining roadway safety requires accurate monitoring of both pavement conditions and traffic sign infrastructure. Traditional methods, such as manual inspection or LiDAR-based systems, are costly, time-consuming, and difficult to scale. This work introduces a cost-effective, image-based framework that leverages dense 3D point cloud data, geometric modeling, and deep learning for automatic roadway health monitoring.

The first study addresses pothole detection using monocular high-definition video processed through photogrammetry to generate dense 3D point clouds. After outlier removal and road surface isolation, the M-estimator Sample Consensus (MSAC) algorithm fits local planar models to identify irregularities. Potholes are visualized with elevation heatmaps, enabling severity and location assessment. Experiments show that higher video resolution and frame rates significantly improve reconstruction quality and detection accuracy, making this a practical, low-cost solution for surface monitoring.

The second study focuses on traffic sign localization using image-based depth data. A Reprojection Loss Network (RLN) minimizes reprojection errors to refine 3D bounding boxes, while a geometric-aware post-processing step aligns them with the planar features of traffic signs. Evaluations on a custom dataset derived from KITTI confirm accurate 3D localization without LiDAR, supporting robust comprehension and legibility analysis in autonomous driving.

Together, these contributions demonstrate that monocular imaging combined with advanced computational methods offers scalable and affordable alternatives to LiDAR. By unifying pothole detection and traffic sign localization, this study highlights the potential of image-based point cloud analysis to improve roadway safety, infrastructure maintenance, and intelligent transportation systems.

17. Key Words Potholes; 3D Point Cloud; Plane Fitting; 3D Detection; 3D Localization; Autonomous Driving; Geometric-aware; Projection Loss		**18. Distribution Statement** No restrictions.	
19. Security Classif.(of this report) Unclassified	**20. Security Classif.(of this page)** Unclassified	**21. No. of Pages** 41	**22. Price** N/A

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 No. 69A3552344815 and 69A3552348320 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.

ACKNOWLEDGMENTS

This research was partially supported by the National Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the U.S. Department of Transportation (USDOT) under grant numbers 69A3552344815 and 69A3552348320. The authors would like to thank the TBD National Center, USDOT for their support of university-based research in transportation, particularly for the funding provided for this project.

TABLE OF CONTENTS

EXECUTIVE SUMMARY……. 1

STUDY 1: ROAD POTHOLE EXTRACTION FROM MONOCULAR VIDEO IN THE WILD

INTRODUCTION ……...……...…...….2

LITERATURE REVIEW …...……4

METHODS ……...……..6

DATA …...……..8

ANALYSIS ……...……....10

RESULTS ……...……..14

CONCLUSIONS AND POLICY IMPLICATIONS ……....15

APPENDIX A ……17

APPENDIX B ……20

APPENDIX C …….……...22

REFERENCES ...…….……..23

STUDY 2: GEOMETRIC-AWARE 3D OBJECT DETECTION FOR TRAFFIC SIGNS

INTRODUCTION 25

LITERATURE REVIEW 26

METHODS ……...……....27

RESULTS AND DISCUSSION 31

CONCLUSIONS AND POLICY IMPLICATIONS 34

REFERENCES 35

LIST OF FIGURES

STUDY 1: ROAD POTHOLE EXTRACTION FROM MONOCULAR VIDEO IN THE WILD

Figure 1 Stages of proposed pothole extraction system……..……7
Figure 2 Select frames from the Sudbury.com video.……...……..…....9
Figure 3 3D point cloud road model, after cleaning and segmentation.……...…..……9
Figure 4 Detected pothole points concatenated to form a new point cloud model.…….……..10
Figure 5 Elevation heatmap of potholes……....11
Figure 6 An additional 3D point cloud road model....……...12
Figure 7 Select frames from the Vatti Github video……..12
Figure 8 Elevation heatmap of potholes from the 3D point cloud road model of the second part of the experiment….……...13
Figure B1 Using Total Curvature Estimation tool on provided example models…….…….20
Figure B2 Using Total Curvature Estimation tool on a partition of our road data ……...21
Figure B3 Our road data partition before Total Curvature Estimation tool is applied …….21
Figure C Sensor Configuration …….22

STUDY 2: GEOMETRIC-AWARE 3D OBJECT DETECTION FOR TRAFFIC SIGNS

Figure 1 Initial bounding boxes, corresponding masks and their contours 28
Figure 2 Training loss of RLN model over 30 epochs …….. 31
Figure 3 The original reprojection in red and the model output in green ……..……... 32
Figure 4 Red bounding boxes show the initial approximations, while blue boxes represent the results after geometric-aware refinement in 3D point cloud ……... 33

EXECUTIVE SUMMARY

Roadway safety and infrastructure maintenance represent ongoing challenges for cities worldwide. In the U.S. alone, pothole-related vehicle damage cost drivers an estimated $26.5 billion in 2021. Meanwhile, traffic sign visibility and accuracy remain critical for both human drivers and autonomous vehicles. Traditional methods for roadway monitoring often rely on costly LiDAR systems or labor-intensive manual inspections, limiting scalability. This report explores affordable, image-based alternatives that leverage deep learning and 3D point cloud data to monitor both pavement conditions and traffic sign infrastructure.

The first study presents a scalable pothole detection system that requires only a standard high-definition video camera. Road footage is converted into dense 3D models, cleaned to remove noise, and analyzed using robust geometric algorithms to identify depressions consistent with potholes. The system generates heatmaps that highlight severity and location, providing actionable insights for maintenance planning. Real-world testing demonstrates promising detection accuracy across diverse conditions, with higher-resolution video improving results. Its affordability and lightweight design enable deployment across fleets or roadside units, supporting continuous monitoring that reduces emergency repair costs and improves driver safety.

The second study focuses on accurate 3D traffic sign detection. Using image-derived depth data, the system employs a Reprojection Loss Network (RLN) to minimize spatial errors and a geometric-aware refinement process to align bounding boxes with the planar surfaces of signs. Evaluations on a custom dataset derived from KITTI show sub-1% error rates, with significant improvements in orientation and dimensional accuracy. This approach reduces reliance on LiDAR while ensuring precise localization is necessary for autonomous navigation.

Together, these contributions highlight the potential of image-based roadway monitoring systems to provide three key benefits:

- **Cost savings** through elimination of expensive LiDAR hardware.
- **Proactive safety** via real-time detection of hazards and sign degradation.
- **Scalability** through integration with municipal fleets, roadside units, or autonomous vehicles.

By combining pothole detection with traffic sign localization, this report demonstrates how deep learning and image-based point cloud analysis can form a unified, affordable, and data-driven framework for roadway health monitoring. Future work will explore higher-quality video inputs, automated segmentation, and hybrid approaches to further enhance accuracy and extend applicability to other roadway features.

STUDY 1: ROAD POTHOLE EXTRACTION FROM MONOCULAR VIDEO IN THE WILD

INTRODUCTION

Roads can accumulate damage over time with heavy use, leading to various defects such as potholes. A recent AAA study reveals that pothole damage has cost United States drivers $26.5 billion in 2021 alone (2022). Potholes can cause significant vehicle damage for drivers, including tire punctures, bent wheels, and even suspension damage. If left unmitigated, these potholes can become deeper or wider with continued use, posing an increased risk to public safety. This makes it crucial to detect and resolve potholes promptly to prevent further damage and ensure road safety.

While many methods have been developed to detect road potholes, there is more to be done to improve cost-effectiveness, accuracy, and speed. Traditional approaches often rely on manual inspections, which are prone to being labor-intensive and time-consuming. More recent technologies, such as vibration sensors and LiDAR vehicle scanning, offer automated solutions, but can involve high costs or complex implementations. LiDAR-radar systems, such as those used by Waymo Jaguar I-Pace autonomous SUVs, can reach costs of up to $9,300 compared with the $400 camera system of a Tesla Model 3 (Hull, D., & Trudell, C., 2025). As cities continue to expand and traffic volumes increase, there is a growing need for efficient, scalable, and affordable pothole detection systems that can help authorities maintain road networks more proactively and minimize risks for all road users.

Inexpensive camera-based solutions have proved to be a viable alternative to expensive hardware solutions, in particular, a semantic segmentation method being featured in the paper by Rateke and von Wangenheim (2021). It introduced a low-cost, passive-vision approach that classifies road surfaces and identifies damage types (including potholes) from monocular images, supported by their RTK dataset of annotated frames.

In this paper, a system is proposed that is cost-effective and swift, requiring only a high-definition video camera for operation. This makes it significantly more affordable than many other pothole detection approaches. While using a stereo camera setup can generate enhanced results, as demonstrated by works like Zhao et al. (2024) and Du et al. (2020), our paper's proposed system uses a single camera to achieve high-quality pothole detection outcomes.

The proposed system of this paper utilizes a variant of the Random Sample Consensus (RANSAC) algorithm, specifically the m-estimator sample consensus (MSAC) (Torr, P. H. S., & Zisserman, A., 2000). This algorithm is well known for its robustness in fitting geometric models, such as planes or surfaces, to data that contain a high proportion of noise outliers, which makes it particularly effective for real-world applications involving noisy data, such as identifying road surfaces and pothole detection from 3D point clouds generated using road images. By applying MSAC to the video data captured by the high-definition camera, the system can extract information about the road surface and identify deviations that indicate potholes.

Compared to traditional RANSAC, MSAC offers improved accuracy by minimizing a modified cost function that penalizes points according to their residual errors rather than using a strict inlier/outlier threshold. This leads to more reliable detection results, even in challenging lighting or weather conditions. Furthermore, leveraging video data allows for continuous, frame-by-frame analysis, enhancing the spatial and temporal resolution of pothole detection. As a result, the proposed approach not only achieves high detection accuracy but also enables real-

time implementation, making it practical for integration into vehicle-based or roadside monitoring systems.

Experimentally, it is shown that using monocular videos can effectively extract and detect road potholes for road condition assessment. The process involves several key steps designed to transform simple video footage into actionable structural information. First, video data, sourced from two external sources, is continuously captured using a single high-definition camera mounted on a moving vehicle. The recorded frames are then processed to generate a dense 3D point cloud model of the road surface through photogrammetry methods (RealityScan in the case of this paper).

This initial point cloud often contains noise, outliers, and non-road elements such as vehicles, shadows, or roadside objects. To address this, a cleaning and preprocessing stage is performed to remove irrelevant points and improve overall data quality. Next, ground area extraction is performed to precisely isolate the actual road surface from its surroundings. This step is crucial for minimizing false positives and ensuring accurate analysis.

Once the road surface has been isolated, pothole detection is carried out by fitting a planar model to the extracted ground points using the MSAC algorithm. Deviations from this fitted plane, which appear as local depressions or irregularities, are then identified as potential potholes. By quantifying these deviations in terms of depth and area, the system can assess the severity of each defect, providing valuable data for prioritizing maintenance efforts.

The results demonstrate the effectiveness of the proposed system in various real-world conditions, including different lighting, weather, and surface textures. The combination of a single high-definition camera and the robust MSAC algorithm offers a cost-effective, scalable, and efficient solution for pothole detection, eliminating the need for expensive LiDAR systems or complex sensor arrays, such as the one shown in Figure C (RSXD, n.d.). Moreover, the simplicity of the setup allows for easy deployment on a large scale, enabling frequent and automated road condition assessments.

By facilitating timely identification and repair of potholes, the system has the potential to greatly enhance road safety and reduce vehicle maintenance costs for drivers. In addition, the data collected can support city planning and infrastructure management by providing continuous, up-to-date information on road health.

LITERATURE REVIEW

Wu et al. (2019) proposed a hybrid pothole detection framework that integrates image-based deep learning with three-dimensional point cloud analysis to overcome the limitations of traditional single-modality approaches. Their method first applies the DeepLabV3+ semantic segmentation network to pavement images to identify candidate pothole regions, which are then mapped onto a mobile LiDAR point cloud for geometric validation. By separating interior and exterior points around each candidate, the system fits a road plane to exterior points while using interior points to calculate pothole depth. This two-step process not only distinguishes potholes from filled patches but also provides quantitative measurements of severity, achieving a mean depth accuracy of 1.5-2.7 cm in real-world testing on a 26.4 km expressway in Shanghai. Compared to image-only methods, which lack precise depth information, and LiDAR-only methods, which struggle with edge localization, this fusion approach demonstrates how combining semantic segmentation with point cloud geometry can deliver both accurate detection and practical severity assessment, making it particularly relevant for road maintenance applications.

Du et al. (2020) proposed a pothole detection method using 3D point cloud segmentation. Instead of relying on accelerometers, 2D image-based approaches, which are prone to noise, or costly LiDAR scanners, the authors use binocular stereo vision to reconstruct a 3D point cloud of the road surface. In the point cloud, potholes appeared as depressions. A plane is fit to the road surface using least-squares plane fitting. By then subtracting this road plane, candidate pothole regions are left remaining. K-means clustering is then employed to group the pothole points and remove outliers. A region growing segmentation algorithm is used to expand from seed points, ensuring that the entire pothole boundary is extracted accurately. Without relying on specialized equipment, Du et al. are able to combine plane fitting, k-means clustering, and region growing segmentation for robust pothole detection. While stereo vision provides accuracy advantages, it still increases the complexity of camera calibration and computational overhead, highlighting the trade-off between accuracy and efficiency in vision-based detection.

Traditional monocular depth estimation and stereo matching methods in perspective view struggle with accurate fine-grained road elevation from perspective images, creating a need for Bird's-Eye-View (BEV) reconstruction (Zhao et al., 2024). Zhao et al.'s RoadBEV, a vision-based approach for road surface reconstruction uses BEV perception to improve elevation estimation for autonomous vehicles. Traditional monocular depth estimation and stereo matching approaches in perspective view struggle to capture fine-grained road geometry due to sparse and biased depth curves. RoadBEV addresses these limitations by projecting voxel features from the camera view into a BEV representation, where elevation estimation is treated as a classification problem over predefined height bins. The authors propose two models: RoadBEV-mono, using monocular input, and RoadBEV-stereo, using stereo image pairs to exploit multi-view correspondence. Evaluations on the Road Surface Reconstruction Dataset, collected by the authors, demonstrate significant improvements over conventional monocular depth and stereo matching methods. The study highlights that BEV-based Road Surface Reconstruction provides dense, top-down road features, enabling more accurate and robust elevation reconstruction, showing potential for enhancing planning and control in autonomous driving.

Dhiman et al. (2018) proposed a stereo vision-based pothole detection system that uses multi-frame accumulation to improve road surface reconstruction accuracy. Their method starts with disparity estimation from stereo images and plane fitting using RANSAC to isolate candidate pothole regions, which are then refined through the construction of a digital elevation

model by aligning multi-frame 3D point clouds to a road-centered coordinate system. By identifying valleys in the digital elevation model through connectedness and region-growing analysis, their approach achieves more reliable detection compared to single-frame stereo methods. The system demonstrated over 30% improvement in accuracy while remaining cost-effective by relying on only on stereo cameras, making it a great alternative to expensive LiDAR solutions. Building on this paper's reliance on RANSAC, our work employs an improved RANSAC variant to reduce overall error (Torr, P. H. S., & Zisserman, A., 2000).

More recently, camera-based semantic segmentation methods have demonstrated the feasibility of detecting road surface damage using inexpensive hardware. Motivated by the lack of research done to explore the road conditions of developing countries, Rateke and von Wangenheim (2021) introduced a low-cost, passive-vision approach that classifies road surfaces and identifies damage types from monocular images, supported by their RTK dataset of annotated frames of road imperfections of Brazilian roadways. Their work differentiates between asphalt, paved, and unpaved roads, and detects various types of surface damage beyond potholes, including cracks and bumps. Using U-NET with Resnet34 and Resnet50 encoders, they employed a two-stage training process, pretraining on unweighted classes and fine-tuning with class weights, to address severe class imbalance. Their approach reliably detects most key road features and damages, except for rare classes such as speed bumps and cracks that remain a challenge. The results of this paper prove that monocular camera-based approaches can more than sufficiently extract useful road surface.

Taken together, these studies illustrate the evolution of pothole detection from multimodal fusion and stereo-based reconstruction to BEV perception and deep learning segmentation. Yet a clear trade-off emerges between robustness, computational cost, and the ability to capture a fine-grained geometry. While stereo and BEV methods improve elevation accuracy, they require more complex sensing and processing. Segmentation-based methods enable broader surface classification but struggle with small-scale defects. This motivates our approach: a lightweight, monocular vision pipeline that leverages MSAC-based geometric reconstruction to detect and quantify potholes with lower computational overhead, while maintaining the accuracy necessary for practical road monitoring.

METHODS

As Stage I of the proposed pothole detection system illustrates in Figure 1, 3D point cloud road data is obtained with the assistance of RealityScan 2.0 (previously known as RealityCapture) photogrammetry software (Epic Games, n.d.). At the start of this stage, a high-definition camera is mounted securely on a vehicle, positioned to capture a wide field of view that includes both the road surface and surrounding non-road areas. This setup ensures comprehensive coverage of the environment directly in front of the vehicle during movement. The vehicle then travels along road segments, continuously recording high-resolution video footage of the road conditions. Once the video data is acquired, RealityScan software can be used to process the captured video frames to reconstruct a detailed 3D point cloud representation of the scene. This reconstruction integrates multiple viewpoints extracted from the video to model the spatial geometry of both road and surrounding objects. The resulting point cloud provides a dense and precise digital model that serves as the foundation for further analysis and pothole detection. LiDAR could be used to collect 3D point clouds but is not recommended given that it is an expensive sensor to acquire.

Stage II describes the crucial process of data cleaning and preprocessing of the road model before it can be effectively utilized for pothole detection tasks. The open-source 3D point cloud editor software CloudCompare is employed to visualize and refine the raw 3D point cloud generated in the previous stage (CloudCompare, n.d.). During this phase, a Statistical Outlier Filter is applied to identify and remove noisy or isolated points that may result from errors in the photogrammetry reconstruction or environmental noise, thereby improving the overall accuracy and smoothness of the model. In addition, the point clouds are reoriented so that the road plane aligns perpendicularly to the z-axis of the coordinate system. This standardization simplifies subsequent analysis and facilitates easier plane fitting and height calculations. The preprocessing stage ensures that the point cloud is cleaner, more consistent, and geometrically well-aligned, which is vital for the precise identification of road surface anomalies in later stages.

After data cleaning, stage III focuses on isolating the ground areas from the comprehensive 3D point clouds. This involves segmenting and removing parts of the point cloud model that corresponds to non-road features, such as roadside vegetation, vehicles, curbs, or other infrastructure elements, since these are not relevant to the objective of pothole detection. The refined point cloud now consists solely of the road surface, allowing for more targeted analysis. To manage and analyze this large dataset more efficiently, the point cloud is then divided into smaller, manageable rectangular sections or “tiles” along the longitudinal axis of the road. Each partition represents a specific segment of the road, enabling localized plane fitting and detailed surface analysis within each section. This tiling strategy supports scalable processing and facilitates the detection of small-scale surface deformations that may otherwise be overlooked in a large continuous model.

For the final stage, Stage IV, the M-estimator Sample Consensus (MSAC) algorithm is utilized to robustly fit a plane to each road partition derived in the previous step (MathWorks, n.d.). This algorithm is an improved variant of the well-known RANSAC method, designed to handle noisy data and outliers effectively. The implementation of the MSAC algorithm and subsequent data analysis, including inlier and outlier extraction as well as elevation heatmap generation, were performed using MATLAB. The elevation heatmap effect is produced by coloring points based on their z-coordinate. Other approaches for generating heatmaps were also considered, including total curvature estimation proposed by Chen (2023a, 2023b). As shown in Figure B1 of Appendix B, the tool performed well on Chen’s example data models. However,

when applied to our road data, the results were mixed. It effectively highlighted deeper potholes but struggled to detect shallower ones, as depicted in Figure B2 of Appendix B. Figure B3 depicts the same road data partition from Figure B2, but before the total curvature estimation tool is applied. By fitting a plane to each tile, the algorithm models the expected flat surface of the road, while simultaneously identifying points that deviate significantly from this model (MathWorks, n.d.). The algorithm outputs a set of inlier points, which closely adhere to the fitted plane, and a set of outlier points, which represent deviations such as potholes or other surface anomalies. These outlier points are then extracted to form a new point cloud that specifically highlights the pothole regions. From this extracted data, an elevation heatmap can be generated in MATLAB, visually emphasizing variations in road surface height and providing quantitative information about the depth and severity of each pothole. This final analysis not only facilitates precise localization of potholes but also supports maintenance prioritization and detailed road condition assessments.

See Appendix A for the MATLAB code that executes Stage IV of the pothole extraction system for both data sources used in this paper.

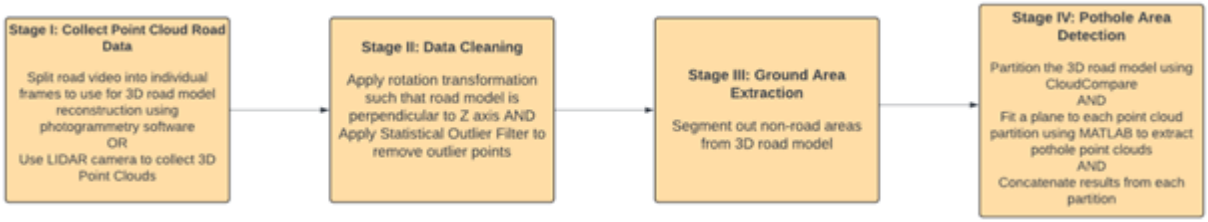

Figure 1: Stages of proposed pothole extraction system

DATA

In the first part of the experiment conducted for this paper, the system commences with acquiring a 3D point cloud model of a road surface using publicly available video data. Specifically, a 15-second video with a 1920 x 1080-pixel resolution, recorded by a high-definition camera mounted on a moving vehicle driving over a pothole-filled road, was sourced from YouTube (Sudbury.com, 2019). This video footage captures various sections of the road that exhibit prominent surface anomalies. Selected keyframes from the video are shown in Figures 2a, 2b, and 2c, each highlighting visible and distinct potholes that will later be observed in the generated 3D model. Because the video is recorded at 30 frames per second (fps), a total of approximately 450 individual frames is extracted and processed. These frames are then fed into RealityScan to reconstruct a detailed 3D model of the scene.

The second part of the experiment mirrors the procedure used in the first part but introduces a different data source to evaluate the system's adaptability and robustness. For this phase, a new point cloud is generated using a 27-second video recorded at 1280 x 720 resolution and 25 fps, sourced from the "potholesinaruralroad.mp4" video shared in the README file from a GitHub repository (Vatti, 2024).

Figure 2: Select frames from the Sudbury.com video (2019) and their matching pothole detection results; from top to bottom, the left images will be referred to as Figure 2a, 2b, and 2c and the right images will be referred to as Figure 2d, 2e, and 2f

Figure 3: 3D point cloud road model, after cleaning and segmentation; Note the depressions in the center area of the model, which represent potholes in the road

ANALYSIS

The point cloud generated through the RealityScan 3D reconstruction process initially includes both road and non-road elements due to the camera's field of view, which could encompass curbs, sidewalks, or roadside vegetation. To address this, the model undergoes a series of preprocessing steps, including outlier removal, rotational alignment with the z-axis, and segmentation to exclude non-road regions. The resulting refined model, as shown in Figure 3, retains the road surface while excluding irrelevant geometry. Notably, potholes are distinguishable in this cleaned model, appearing as localized depressions within the relatively flat road plane. These visible cavities provide early confirmation that the reconstruction has preserved essential surface deformations critical for pothole detection.

To follow, the experiment proceeds to the plane fitting stage, where the road surface is analyzed on a per-tile basis. Each tile is individually processed using the M-estimator Sample Consensus (MSAC) algorithm, which attempts to fit a best-fit plane to each partition while minimizing the impact of outliers. For this experiment, the maximum allowed distance from an inlier point to the fitted plane (Mathworks, n.d.) is set to 0.071 units. Figure 4 presents the results of this detection process, showing how points that significantly deviate from the road plane—likely representing potholes—are isolated and visualized.

A visual comparison of the original video frames (Figure 2) with the detection results (Figure 4) reveals strong alignment between the observed potholes and the extracted features. The real-world potholes visible in Figure 2a are detected near coordinates (-10, 3) and extracted in Figure 2d, those from Figure 2b are visible near coordinates (0, 2) and extracted in Figure 2e, and those from Figure 2c are visible around coordinates (15, 2) and extracted in Figure 2f. These spatial correlations confirm that the plane fitting approach effectively localizes potholes, especially the more pronounced or wider ones.

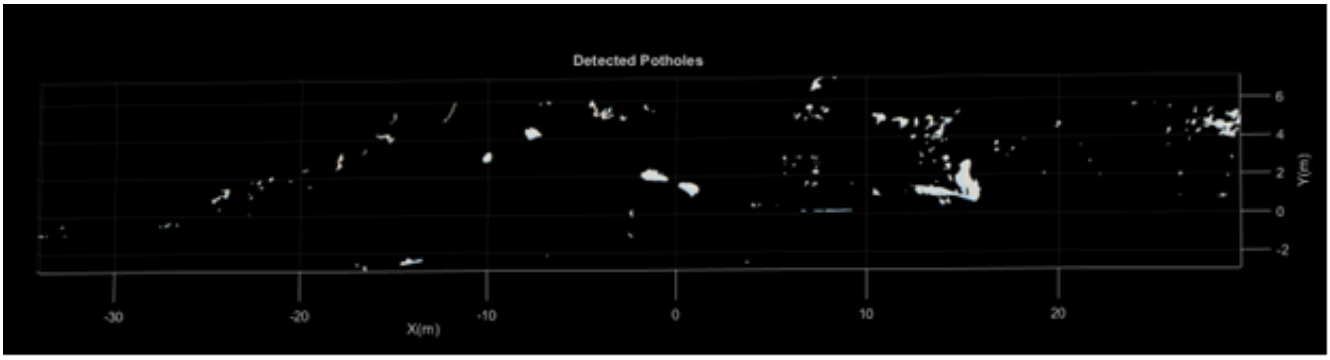

Figure 4: Detected pothole points concatenated to form a new point cloud model

With the detected pothole points now isolated into a new point cloud model, an elevation heatmap is generated to visualize the depth and severity of the potholes. This map, shown in Figure 5, enables intuitive identification of areas requiring urgent road maintenance. The elevation data highlights not only the location but also the approximate depth of each detected anomaly, making this an informative tool for road condition assessment.

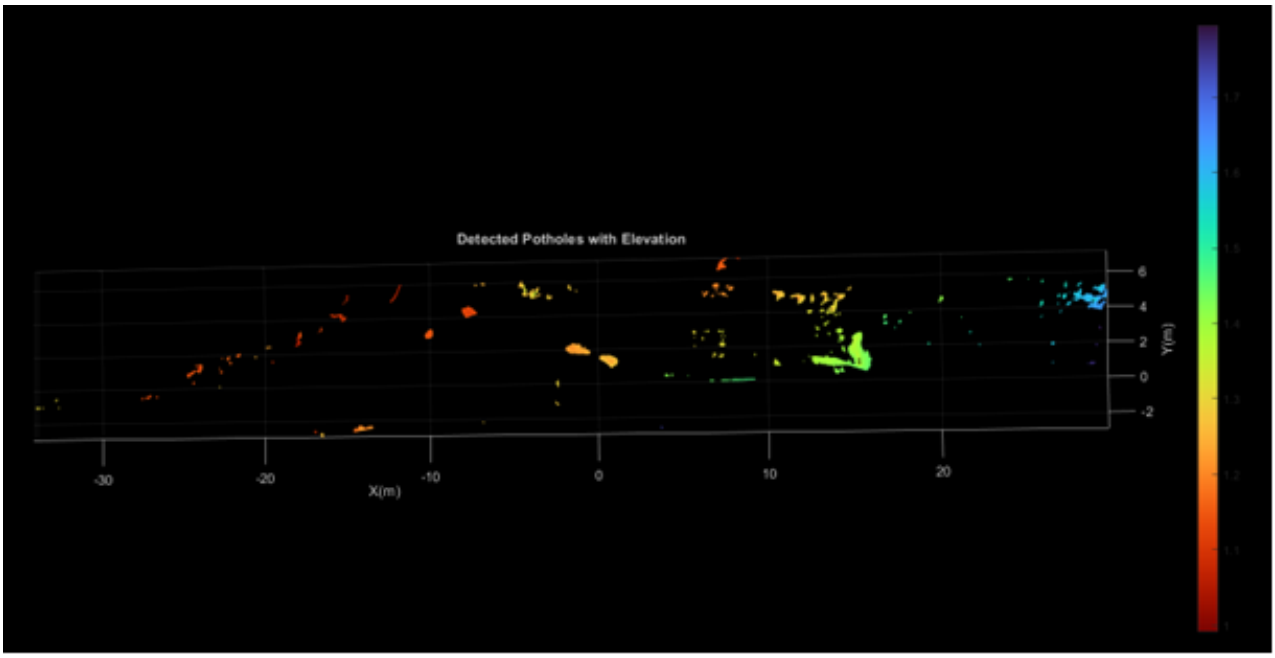

Figure 5: Elevation heatmap of potholes, with red indicating the deepest depth and blue indicating the shallowest depth

The second video from the next part of the experiment captures a different road segment, and a notable distinction is the presence of deep water-filled potholes, which complicates the reconstruction process. These puddle-filled depressions are more challenging for RealityScan to model accurately, often resulting in sparse or incomplete sections in the generated point cloud. Figures 7a and 7b present selected frames from the video, displaying several prominent potholes. The point cloud reconstructed from this dataset is shown in Figure 6, and due to surfaces like water, some regions of the model contain holes or inconsistencies. To account for the denser noise and shallower detail, the MSAC plane fitting algorithm is used again but with a stricter inlier distance threshold of 0.01.

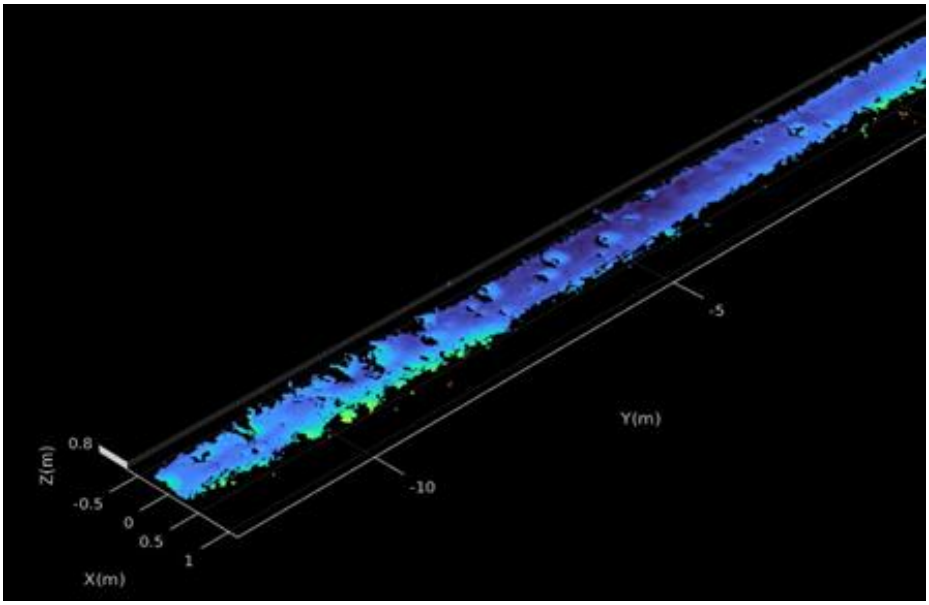

Figure 6: An additional 3D point cloud road model, after cleaning and segmentation

Figure 7: Select frames from the Vatti Github video (2024) that was used to generate the point cloud model in Figure 6; from top to bottom, the left images will be referred to as Figure 7a and 7b and the right image will be referred to as 7c and 7d

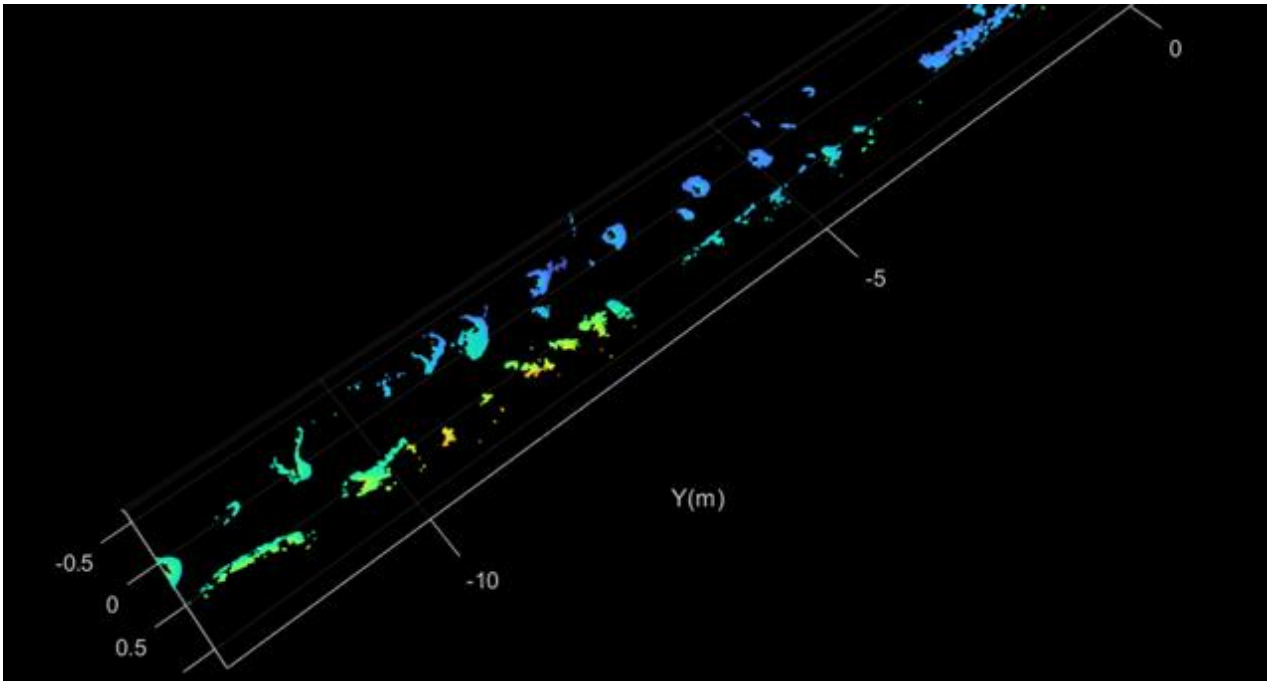

Figure 8: Elevation heatmap of potholes from the 3D point cloud road model of the second part of the experiment

The results of pothole detection for this lower-resolution model are displayed in Figure 8. The lighter-colored regions in this figure correspond to potholes with more significant vertical displacement.

RESULTS

The spatial correlations from the first part of the experiment confirm that the plane fitting approach effectively localizes potholes, especially the more pronounced or wider ones. However, the detections also include some extraneous artifacts, especially near the upper edge of the detections of Figure 4. Inspection of the full model in Figure 3 suggests that this region contains sparse data and reconstruction artifacts, likely due to the limitations of the original video's resolution and frame rate. These artifacts introduce gaps in the point cloud, which the system occasionally misclassifies as depressions. This indicates that increasing the source video's resolution and frame rate would likely enhance the density and accuracy of the reconstructed model, reduce false positives, and improve detection quality.

The second, lower-resolution model produced mixed results. Despite the reduced input quality of the source data, the algorithm successfully detects several large potholes, especially those on the left side of the road. In contrast, the right side exhibited significantly fewer detections. A likely explanation is that the camera in this video was positioned closer to the left side of the vehicle, offering a clearer view of that portion of the road while the right side remained partially occluded. This asymmetric perspective reduced the reconstruction quality across the entire model.

To assess detection quality, the potholes from the original frames (Figure 7) were mapped to their corresponding regions in the 3D coordinate space of Figures 6 and 8. The potholes in Figure 7a align near coordinates (0, -11), while those in Figure 7b appear around coordinates (0, -6). This mapping confirms that the major features visible in the video are captured by the detection system. Nevertheless, several false positives remain, attributable to the lower-resolution input video. The sparser point cloud increases uncertainty in plane fitting and heightens the risk of misclassifying holes or gaps as potholes.

Overall, these findings demonstrate that while the system can detect potholes across different video qualities, detection accuracy and robustness improve considerably with higher-resolution and higher frame-rate data. This underscores the importance of source video quality in maximizing the effectiveness of the proposed detection framework.

CONCLUSIONS AND POLICY IMPLICATIONS

With extensive use over time, potholes frequently accumulate on roads, causing a range of issues, including driver inconvenience, vehicle collisions, and significant vehicle degradation. Addressing these potholes promptly is essential to prevent serious accidents and costly repairs. Detecting and resolving these potholes efficiently can mitigate potential hazards and expenses.

This paper introduces a comprehensive, four-stage pothole extraction system designed to be both swift and cost-effective compared to existing methods. The proposed system's stages include:

1. 3D Point Cloud Road Model Acquisition: Utilizing RealityScan, a detailed 3D point cloud model of the road surface is created from captured video frames. This model provides high-resolution data that is crucial for accurate detection of surface anomalies.
2. Data Model Cleaning and Preprocessing: The raw point cloud data undergoes a thorough cleaning process to remove noise and irrelevant data. Preprocessing steps ensure that the data set is optimized for subsequent analysis, enhancing the accuracy of pothole detection.
3. Ground Area Extraction: In this stage, the road surface is isolated from the surrounding environment. This involves distinguishing the ground area from non-road areas, which is critical for focusing the analysis on the actual road surface where potholes may form.
4. Pothole Area Detection: The final stage involves fitting a plane to the road model to identify deviations that indicate the presence of potholes. By extracting these pothole points, an elevation heatmap is generated, providing a visual representation of each detected point's elevation relative to the road surface.

Experimental evaluation using two road models highlights both the system's strengths and limitations. The results show that pothole detection improves substantially when higher-resolution and higher-frame-rate video data are used, while lower-quality inputs lead to sparser reconstructions and more false positives. These findings suggest that the proposed system is a promising foundation for pothole detection, but its full potential depends on the availability of higher-quality input data.

The implementation of this four-stage system offers significant advantages, including reduced costs and improved speed of pothole detection compared to traditional methods. By addressing potholes before they lead to severe damage or accidents, this system contributes to enhanced road safety and longevity. Furthermore, the ability to generate precise elevation heatmaps allows for a detailed analysis of road conditions, facilitating proactive maintenance strategies.

We evaluate this system using two different road models created from video frame data. From the comparison of the results of our two-part experiment, we reveal that pothole detection is improved when utilizing higher resolution and higher frame rate video data. Finally, we also suggest further enhancements to this four-stage pothole detection system to achieve improved results quicker.

Future work may explore integrating this pothole detection approach with other sensing modalities and real-time cloud-based or edge-based processing to further improve accuracy and operational efficiency. In addition, acquiring higher-quality 3D models for more extensive experimentation would be beneficial; for example, using a high-definition camera capable of capturing at least 60 frames per second could help generate more detailed and accurate road models. Implementing a general "sliding window" approach, rather than the static partitioning

used in the current system, may also enhance pothole detection by allowing for more flexible and adaptive analysis of the point cloud data. Furthermore, investigating a deep learning-based method for ground area extraction could automate the manual segmentation currently performed in stage III, potentially increasing both speed and consistency. Finally, combining this 3D detection system with a 2D pothole detection system could create a hybrid framework that reduces the likelihood of missed detections and improves overall robustness (Vatti, 2024).

Beyond its technical contributions, this system offers practical benefits for transportation agencies and policymakers. By leveraging inexpensive video data to generate accurate 3D road models, the framework supports proactive maintenance strategies that reduce repair costs and extend roadway lifespans. Its scalability makes it suitable for deployment in both developed and developing regions, where budget constraints often limit access to high-cost sensing technologies. Integration into smart transportation or city management systems could further support data-driven decision-making, improving road safety, optimizing maintenance schedules, and ultimately reducing accident risk and vehicle damage associated with neglected potholes.

APPENDIX A

MATLAB Pothole Extraction code for first road model

```
filepath = 'road_model for Video 1\road_model_segmented_and_cleaned.las';
lasReader = lasFileReader(filepath);
ptCloud = readPointCloud(lasReader);
figure
pcshow(ptCloud.Location)
colormap(flipud(turbo))
xlabel("X(m)")
ylabel("Y(m)")
zlabel("Z(m)")
title("Original Point Cloud")

maxDistance = 0.071;

ptCloudOut = [];
filepath = 'road_model for Video 1\tiled_road_model';
files = dir(fullfile(filepath,'*.las'));
for i = 1:length(files)
   filename = files(i).name;
   fullfilename = fullfile(filepath,filename);
   lasReader = lasFileReader(fullfilename);
   ptCloud = readPointCloud(lasReader);
   [model1,inlierIndices,outlierIndices] = pcfitplane(ptCloud,maxDistance);
   plane1 = select(ptCloud,inlierIndices);
   remainPtCloud = select(ptCloud,outlierIndices);
   [model2,inlierIndices,outlierIndices] = pcfitplane(remainPtCloud,maxDistance);
   plane2 = select(remainPtCloud,inlierIndices);
   remainPtCloud = select(remainPtCloud,outlierIndices);
   ptCloudOut = [ptCloudOut plane2];
end

potholes = pccat(ptCloudOut);
figure
pcshow(potholes)
xlabel("X(m)")
ylabel("Y(m)")
zlabel("Z(m)")
title("Detected Potholes")

figure
pcshow(potholes.Location)
colormap(flipud(turbo))
xlabel("X(m)")
ylabel("Y(m)")
```

```
zlabel("Z(m)")
title("Detected Potholes with Elevation")
```

MATLAB Pothole Extraction code for second road model

```
filepath = 'road_model for Video 2\road_model_2_cleaned_and_segmented.las';
lasReader = lasFileReader(filepath);
ptCloud = readPointCloud(lasReader);
figure
pcshow(ptCloud.Location)
colormap(flipud(turbo))
xlabel("X(m)")
ylabel("Y(m)")
zlabel("Z(m)")
title("Original Point Cloud")

maxDistance = 0.01;

ptCloudOut = [];
filepath = 'road_model for Video 2\tiled_road_model_2';
files = dir(fullfile(filepath,'*.las'));
for i = 1:length(files)
   filename = files(i).name;
   fullfilename = fullfile(filepath,filename);
   lasReader = lasFileReader(fullfilename);
   ptCloud = readPointCloud(lasReader);
   [model1,inlierIndices,outlierIndices] = pcfitplane(ptCloud,maxDistance);
   plane1 = select(ptCloud,inlierIndices);
   remainPtCloud = select(ptCloud,outlierIndices);
   [model2,inlierIndices,outlierIndices] = pcfitplane(remainPtCloud,maxDistance);
   plane2 = select(remainPtCloud,inlierIndices);
   remainPtCloud = select(remainPtCloud,outlierIndices);
   ptCloudOut = [ptCloudOut plane2];
end

potholes = pccat(ptCloudOut);
figure
pcshow(potholes)
xlabel("X(m)")
ylabel("Y(m)")
zlabel("Z(m)")
title("Detected Potholes")

figure
pcshow(potholes.Location)
colormap(flipud(turbo))
```

```
xlabel("X(m)")
ylabel("Y(m)")
zlabel("Z(m)")
title("Detected Potholes with Elevation")
```

APPENDIX B

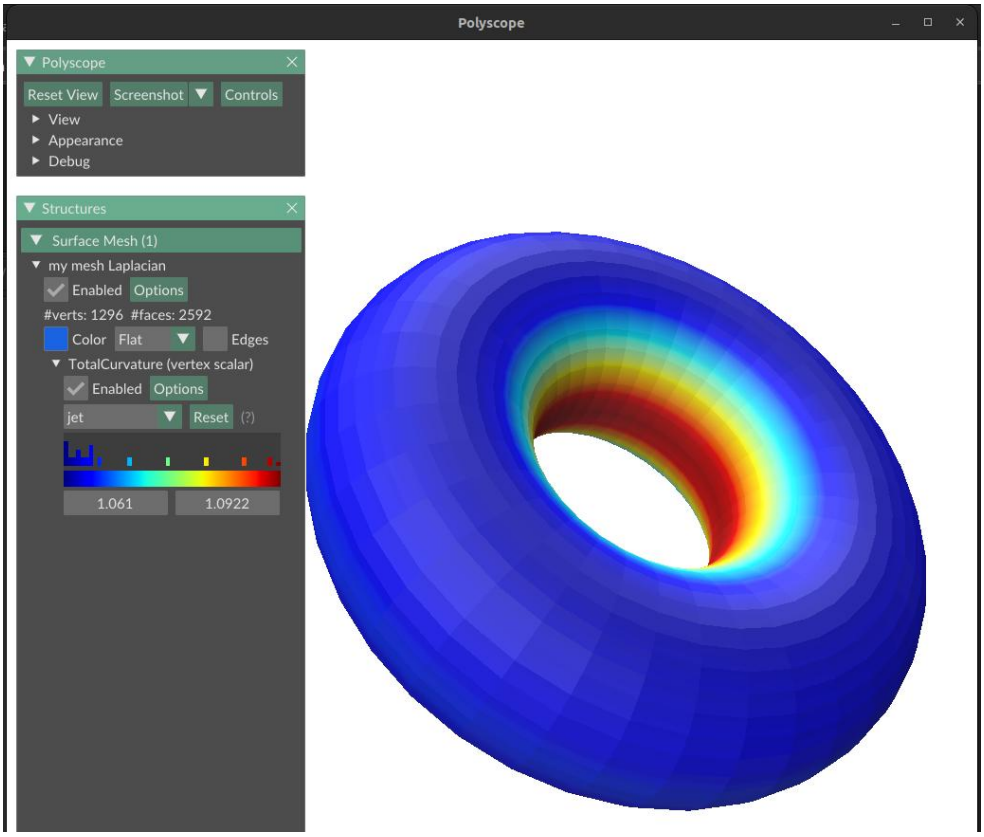

Figure B1: Using Total Curvature Estimation tool on provided example models, all produced by Crane He Chen work

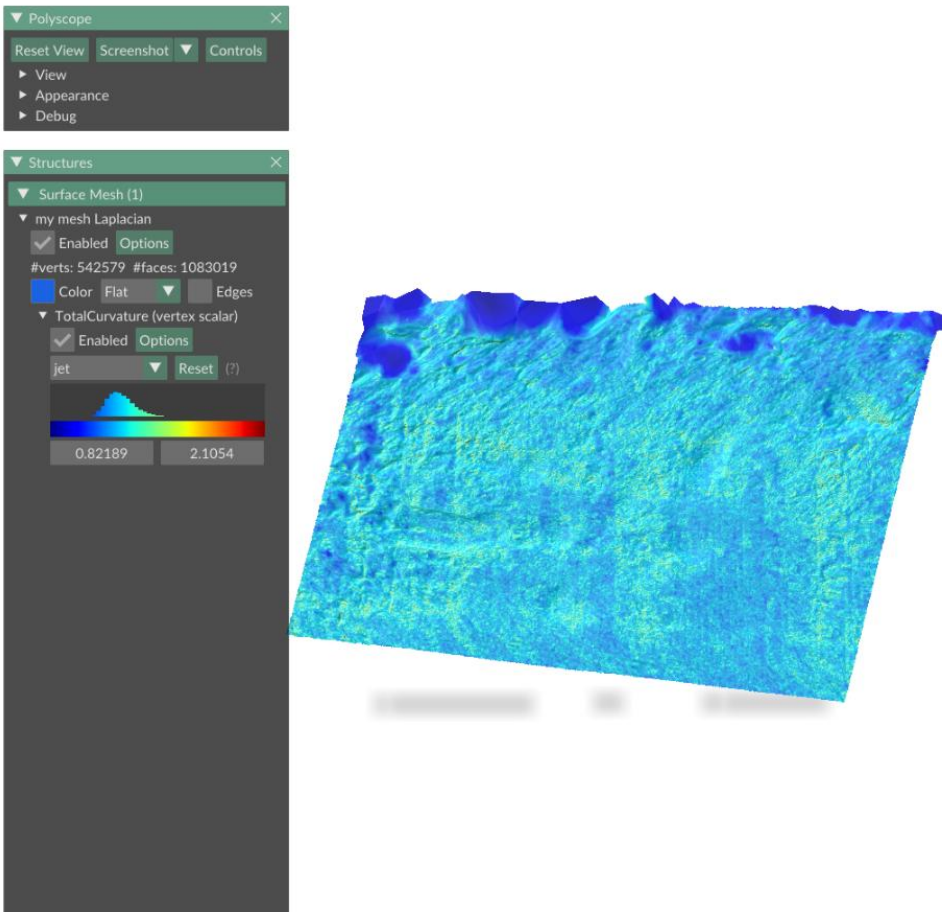

Figure B2: Using Total Curvature Estimation tool on a partition of our road data

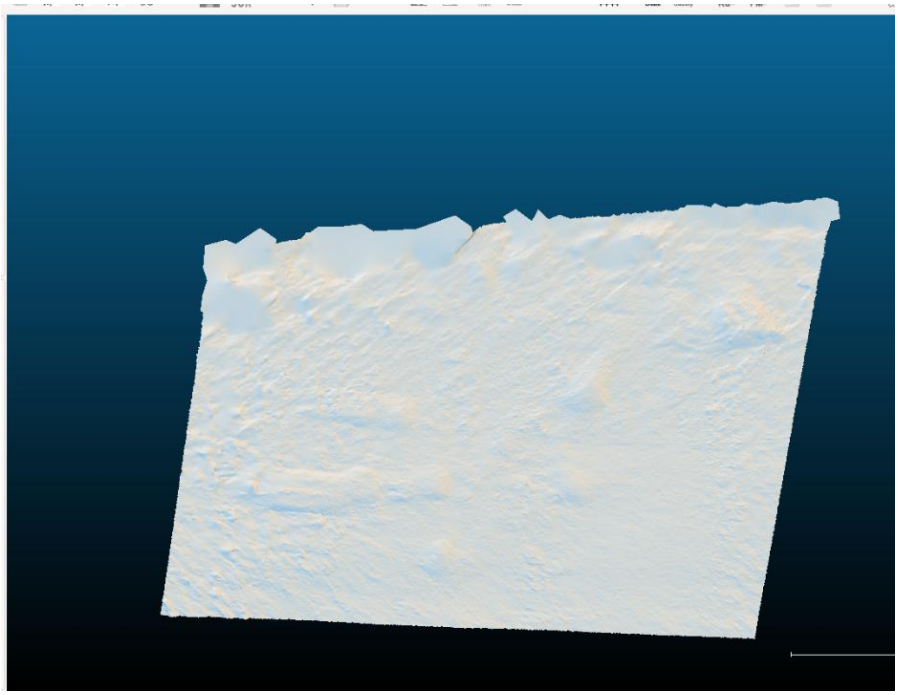

Figure B3: Our road data partition before Total Curvature Estimation tool is applied

APPENDIX C

Figure C: Sensor Configuration from (RSXD, n.d.), composed of a LI-AR023ZWDR camera, Hesai XT32 LiDAR sensor, UBlox F9P GNSS-RTK, XSENS MTi670 IMU, and an ADXL345 Accelerometer

REFERENCES

AAA Newsroom. (2022, March). *AAA: Potholes pack a punch as drivers pay $26.5 billion in related vehicle repairs*. AAA Newsroom. https://newsroom.aaa.com/2022/03/aaa-potholes-pack-a-punch-as-drivers-pay-26-5-billion-in-related-vehicle-repairs/

Chen, C. H. (2023a). Estimating discrete total curvature with per triangle normal variation. In *ACM SIGGRAPH 2023 Talks* (pp. 1-2). https://doi.org/10.48550/arXiv.2305.12653

Chen, C. H., (2023b) *total_curvature_estimation* [Github repository]. https://github.com/HeCraneChen/total-curvature-estimation

CloudCompare. (n.d.). *CloudCompare*. https://www.cloudcompare.org/

Dhiman, A., Chien, H., & Klette, R. (2018). A Multi-frame Stereo Vision-Based Road Profiling Technique for Distress Analysis. *2018 15th International Symposium on Pervasive Systems, Algorithms and Networks (I-SPAN)*, 7-14. https://doi.org/10.1109/I-SPAN.2018.00012

Du, Y., Zhou, Z., Wu, Q., Huang, H., Xu, M., Cao, J., & Hu, G. (2020). A pothole detection method based on 3D point cloud segmentation. In *Proceedings of SPIE: Vol. 11519. Twelfth International Conference on Digital Image Processing (ICDIP)* (p. 1151909). SPIE. https://doi.org/10.1117/12.2573124

Epic Games. (n.d.). *RealityScan*. https://www.realityscan.com/en-US

Hull, D., & Trudell, C. (2025, July 31). *A fatal Tesla crash shows the limits of full self driving*. Bloomberg. https://www.bloomberg.com/features/2025-tesla-full-self-driving-crash/.

MathWorks. (n.d.). *pcfitplane*. https://www.mathworks.com/help/vision/ref/pcfitplane.html#busqbp7-1-referenceVector.

Rateke, T., & von Wangenheim, A. (2021). Road surface detection and differentiation considering surface damages. *Autonomous Robots, 45*(2), 299–312. https://doi.org/10.48550/arXiv.2006.13377

RSXD. (n.d.). *Hardware platform*. https://thu-rsxd.com/sensors/

Sudbury.com. (2019, March 21). *Take a pothole tour of Greater Sudbury* [Video]. https://www.youtube.com/watch?v=SyIQirLZB7A&t=31s.

Torr, P. H. S., & Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding, 78*(1), 138–156. https://doi.org/10.1006/cviu.1999.0832

Vatti, M. (2024). *pothole_detection_yolov8* [GitHub repository]. https://github.com/mounishvatti/pothole_detection_yolov8?tab=readme-ov-file.

Wu, H., Yao, L., Xu, Z., Li, Y., Ao, X., Chen, Q., Li, Z., & Meng, B. (2019). Road pothole extraction and safety evaluation by integration of point cloud and images derived from mobile mapping sensors. *Advanced Engineering Informatics, 42,* 100936. https://doi.org/10.1016/j.aei.2019.100936.

Zhao, T., Yang, L., Xie, Y., Ding, M., Tomizuka, M., & Wei, Y. (2024). RoadBEV: Road surface reconstruction in bird's eye view. *arXiv Preprint*, arXiv:2404.06605. https://arxiv.org/abs/2404.06605

STUDY 2: GEOMETRIC-AWARE 3D OBJECT DETECTION FOR TRAFFIC SIGNS

INTRODUCTION

Accurate detection and localization of traffic signs are crucial for the safe operation of autonomous vehicles (Yu et al., 2024; Fang et al., 2022). While traditional 3D object detection methods often rely on LiDAR technology (Fang et al., 2022; Zhang et al., 2019; Huang et al., 2017), they can be expensive and complex. Alternative image-based depth estimation techniques offer a cost-effective solution, but achieving high precision remains challenging.

This paper uses image-derived depth data to explore a method that detects and refines 3D bounding boxes for traffic signs. Central to this method are two key components: the Reprojection Loss Network (RLN) and a geometric-aware refinement process. The RLN focuses on reducing reprojection errors, thereby improving the spatial precision of detected objects. This process is critical for ensuring the detected 3D positions align accurately with real-world objects, a necessary condition for reliable autonomous navigation.

The geometric-aware refinement further enhances the bounding boxes by adjusting them to match the planar surfaces of traffic signs. This step is crucial for achieving accurate orientation and dimension representation, which is vital for autonomous systems' correct interpretation of traffic signs.

The method's effectiveness is demonstrated using a custom dataset annotated for traffic signs, showing significant improvements in detection accuracy. This research highlights the potential of image-based depth estimation methods for cost-effective and accurate 3D object detection in autonomous driving applications.

LITERATURE REVIEW

Accurate depth estimation is crucial for 3D object detection in autonomous driving, particularly when seeking cost-effective alternatives to LiDAR. Stereo-based approaches, monocular depth estimation, and ensemble models have all contributed to advancing image-based 3D perception.

You et al. (2019) advanced the pseudo-LiDAR framework by addressing depth estimation errors for distant objects in stereo-based detection. They proposed a stereo depth network that optimizes depth directly and a graph-based correction method using sparse LiDAR to de-bias predictions. Their method outperforms prior stereo-based approaches and improves far-distance detection by up to 40%, approaching LiDAR-level accuracy at lower cost. These advancements motivate the use of stereo-derived depth for precise traffic sign localization in our work, particularly for objects with small spatial footprints.

Li et al. (2023) introduced a cross-cue fusion module to enhance depth estimation in dynamic scenes, integrating monocular and multi-view cues within a unified volumetric representation. By applying cross-cue attention, their approach improves robustness in both static and dynamic regions, reducing errors in motion-sensitive areas. This work highlights the importance of incorporating multiple depth cues and inspires our refinement strategy to account for variability in traffic scenes.

Yin et al. (2023) addressed the challenge of recovering metric 3D structure from a single image, overcoming limitations of affine-invariant monocular methods that lack metric scale. They proposed a canonical camera space transformation and a random proposal normalization loss to improve local depth accuracy. Trained on over eight million images, their model achieves competitive zero-shot performance across multiple benchmarks and mitigates scale drift in monocular SLAM. While their approach excels at general monocular metric depth recovery, it does not specifically target small, planar objects like traffic signs, a gap our method addresses.

Cantrell et al. (2020) explored monocular RGB-based depth estimation as an alternative to traditional range sensors, proposing a modular ensemble of U-Nets that integrates pretrained features such as semantic segmentation. Their W-Net Connected variant achieved lower mean squared error than individual U-Nets, demonstrating that combining segmentation features with RGB input can enhance depth estimation accuracy. Despite higher computational costs, their work demonstrates the potential of feature-enriched networks for precise, cost-effective depth prediction, supporting the use of segmentation-guided refinement in our pipeline.

Yang et al. (2024) presented Depth Anything, a foundational model for monocular depth estimation designed to generalize across diverse visual conditions. Leveraging large-scale unlabeled and labeled image data in a teacher-student framework, and integrating semantic priors via feature alignment, their model achieves superior zero-shot depth estimation and state-of-the-art performance on benchmark datasets. Their approach underscores the value of scalable, task-agnostic pretraining for robust depth perception, reinforcing the benefits of incorporating pretrained features in our traffic sign detection pipeline.

Collectively, these studies demonstrate the potential of image-based depth estimation for 3D scene understanding. However, few methods specifically address accurate, cost-effective 3D localization of planar traffic signs. Our work builds on these foundations by integrating a Reprojection Loss Network (RLN) and geometric-aware refinement to precisely detect and align traffic sign bounding boxes in 3D space, providing a practical alternative to expensive LiDAR-based systems.

METHODS

Our methodology comprises several distinct stages: first, we acquire a depth map from a pair of stereo images to reconstruct the 3D scene. This is followed by the detection and segmentation of the object of interest within the left image. Using the segmented area, we calculate the object's depth and project its bounding box onto the reconstructed 3D scene. The refinement process is divided into two stages. The first stage, part of the training phase, involves optimizing the bounding boxes based on reprojection errors. The second stage, a post-processing step, employs a geometric-aware approach to further refine the bounding boxes of traffic signs, correcting their rotation and adjusting their dimensions for enhanced precision.

A. 3D Scene Reconstruction

In the 3D reconstruction phase, we adopt the methodology outlined in Pseudo-Lidar++ (You et al., 2019) and utilize their Stereo Depth Network (SDN), which is specifically designed to optimize accurate depth estimation rather than traditional disparity estimation. The optimization is expressed as follows:

$$\sum_{(u,v)\in A} \ell(Z(u,v) - Z^*(u,v)) \quad (1)$$

where $Z(u,v)$ is the estimated depth, $Z^*(u,v)$ is the ground truth depth, ℓ is the smooth L1 loss, and A includes pixels for which ground truths are available.

This method emphasizes reducing depth estimation errors for distant objects by constructing a cost volume on a depth grid, thereby enabling 3D convolutions and the loss function to operate at the appropriate scale for depth accuracy. It is important to note that while this phase can be substituted with other depth estimation methods capable of delivering accurate scene depth (Li et al., 2023; Yin et al., 2023; Cantrell et al., 2020; Yang et al., 2024), the precision of subsequent steps significantly relies on the robustness of this phase.

B. Object Detection

In this phase, we fine-tune an existing pre-trained object detection model to identify traffic signs within the image. The bounding boxes from the fine-tuned model serve as prompt encoders for the Segment Anything Model (SAM) (Kirillov, 2023), which produces a segmentation mask. We further refine this mask by applying outlier removal techniques and identifying the rectangular contour of each detected object, thereby achieving more precise bounding boxes. Additionally, we use the segmented mask to calculate depth. This step is crucial, as the bounding boxes generated by the object detection model may not be entirely accurate and are axis-aligned. Moreover, for signs that are non-rectangular, considering the entire area within the bounding box for depth estimation could lead to inaccuracies due to surrounding areas that do not correspond to the actual sign. Therefore, for each sign, we calculate the depth using only the pixels that correspond to the mask. This process is illustrated in **(Figure 1)**.

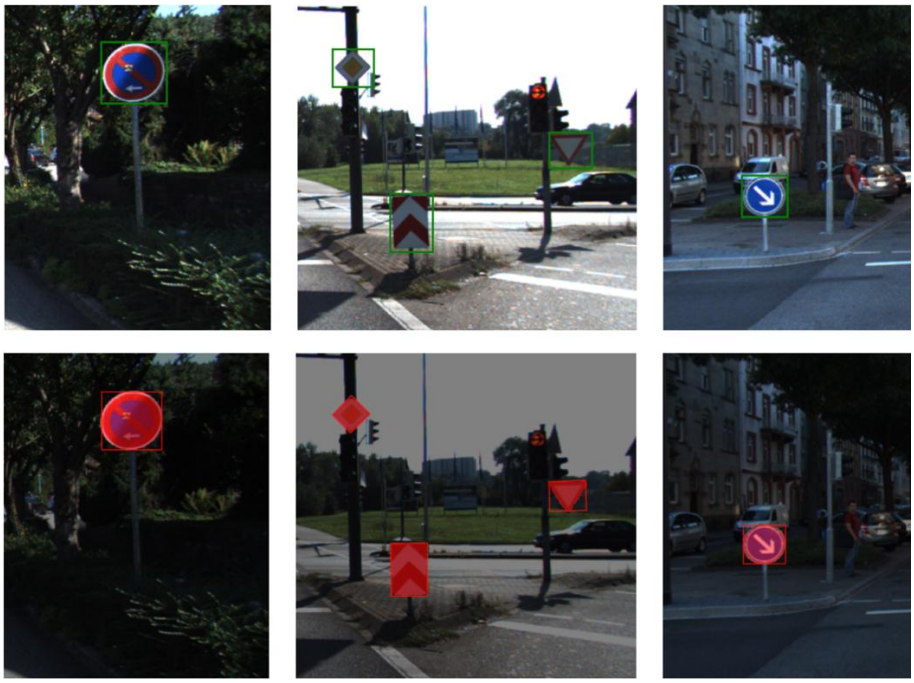

Figure 1 Initial bounding boxes (top), corresponding masks and their contours (bottom). The images are from the Kitti dataset (Geiger et al., 2013)

C. Projection

During the projection phase, the refined bounding boxes are translated into the 3D scene using the stereo projection matrix P. The focal lengths along the x and y axes (f_x and f_y), as well as the optical center coordinates (c_x and c_y), are extracted from P. For each pixel within the bounding box, we examine the associated mask to determine valid points. Only pixels where the mask is true are considered, and their depth z from the depth map is used. The valid z values are collected and processed to exclude the lower and upper quartiles, focusing on the central range of depth values to mitigate outliers. We then compute the average z value from these middle quartiles and convert the bounding box corners into 3D coordinates using the following transformations:

$$x = (u - c_x).\frac{z}{f_x} \tag{2}$$

$$y = (v - c_y).\frac{z}{f_y} \tag{3}$$

These coordinates form the initial 3D bounding boxes. Subsequent refinements of these initial projections will be addressed in two stages. The first stage corrects for reprojection errors, and the second stage utilizes a geometric-aware approach as post-processing. This phased approach ensures detailed and precise alignment of the bounding boxes with the objects in the scene, which will be elaborated on in the following sections.

D. Reprojection Loss

In this phase, the objective is to enhance the accuracy of 3D bounding boxes by minimizing reprojection errors, crucial for applications where precise object detection and positioning are essential for operational safety and efficiency.

The RLN (Reprojection Loss Net) model, a specialized convolutional neural network, was developed to refine the spatial positioning of bounding boxes by integrating image and depth information. This network, constructed with multiple convolutional, pooling, and fully connected layers, is designed to meticulously extract spatial features from the image and depth inputs, enabling the model to adjust the 3D bounding boxes to better match their 2D projections.

During training, the Generalized Intersection over Union (GIoU) loss (Rezatofighi, 2019) is employed to measure the refinement effectiveness:

$$GIoU\ Loss = 1 - \frac{|A \cap B|}{|A \cup B|} + \frac{|C - (A \cup B)|}{|C|} \tag{4}$$

where A and B denote the areas of the predicted and ground truth bounding boxes, respectively; $A \cap B$ is their intersection; $A \cup B$ is their union; and C is the smallest enclosing box that contains both A and B. This loss function evaluates how well the predicted bounding box covers the ground truth while minimizing unnecessary coverage beyond the actual object, which is crucial for accurate object recognition and localization.

The model is optimized using the Adam optimizer, chosen for its ability to efficiently manage sparse gradients and adapt parameters based on the data, ensuring the bounding boxes are refined to meet stringent accuracy requirements. This refinement process significantly enhances the detection and positioning accuracy of objects, which is critical for systems relying on reliable data for decision-making.

E. Geometric Aware Refinement

In the post-processing phase, we leverage the planar nature of physical traffic signs to enhance the precision of their 3D bounding boxes. Given that real-world traffic signs are essentially planes, our focus is on optimizing the alignment and dimensions of the bounding boxes around these planes. To achieve this, we also shift the bounding box so that the calculated plane is centered, followed by constraining the box to fit closely around the traffic sign.

The process involves the following key steps:

1. **PCA Computation**: For each bounding box generated during the projection phase, we identify a subset of points from the 3D scene corresponding to the traffic sign. These points form a point cloud within the bounding box. We then apply Principal Component Analysis (PCA) to this point cloud to extract the principal components, which represent the variance and directionality of the data.
2. **Plane Formation**: Using the eigenvectors corresponding to the two largest eigenvalues, we define the plane that best fits the traffic sign. These eigenvectors form the basis vectors of the plane.
3. **Angle Calculation**: We calculate the angle between the newly formed plane and the XY plane by computing the arccosine of the dot product of the normal to the PCA plane and the Z-axis unit vector, normalized by the product of their magnitudes:

$$\theta = \cos^{-1}\left(\frac{n_{PCA}.k}{\|n_{PCA}\|}\right) \tag{5}$$

where n_{PCA} is the normal vector of the PCA plane, and k is the unit vector along the Z-axis.

4. **Rotation**: The bounding boxes, initially axis-aligned, are rotated by this angle to align them with the identified PCA plane.
5. **Bounding Box Constraining**: We shift the bounding box so that the plane calculated is centered within it. Then, we adjust the dimensions of the bounding box to fit tightly around the PCA plane, ensuring it encloses only the area of the traffic sign without including excess background or empty space.

This geometric-aware refinement ensures that the bounding boxes are not only accurately positioned but also properly oriented and scaled relative to the traffic signs' actual dimensions in the 3D space. This method significantly enhances the fidelity of the bounding boxes, aligning them more closely with the physical properties of the traffic signs.

RESULTS AND DISCUSSION

To evaluate our approach, we constructed a custom dataset from the Kitti 3D Object Detection dataset (Geiger et al., 2013) by specifically selecting images that contain at least one traffic sign. The Kitti dataset, renowned in the computer vision community for autonomous driving research, provides a wealth of data, including LiDAR point clouds for detailed 3D scene representation, calibration parameters for accurate sensor data alignment, and stereo images from left and right viewpoints to facilitate depth perception and 3D reconstruction. Initially, the dataset's training labels primarily cover categories like cars, pedestrians, and cyclists, and notably omit traffic signs.

To address this gap and tailor the dataset to our needs, we manually annotated the traffic signs in the 2D left images and 3d point cloud. This modification ensures that our dataset is specifically equipped to test our model's performance in detecting and refining traffic sign bounding boxes.

A. Reprojection Loss

To evaluate our method, we employ reprojection error as our primary metric. Reprojection error is calculated by projecting the refined 3D bounding boxes back onto the 2D image plane using the camera's calibration parameters and then comparing these projected bounding boxes against the manually annotated ground truth in the 2D images. The discrepancy is quantified as the pixel distance between the edges of the projected 2D bounding boxes and those of the ground truth, providing a direct measure of the accuracy with which the 3D models align with their 2D counterparts.

We trained the Reprojection Loss Network (RLN) model on 210 annotated images from our custom training dataset. **(Figure 2)** shows the model loss over 30 epochs on the training dataset.

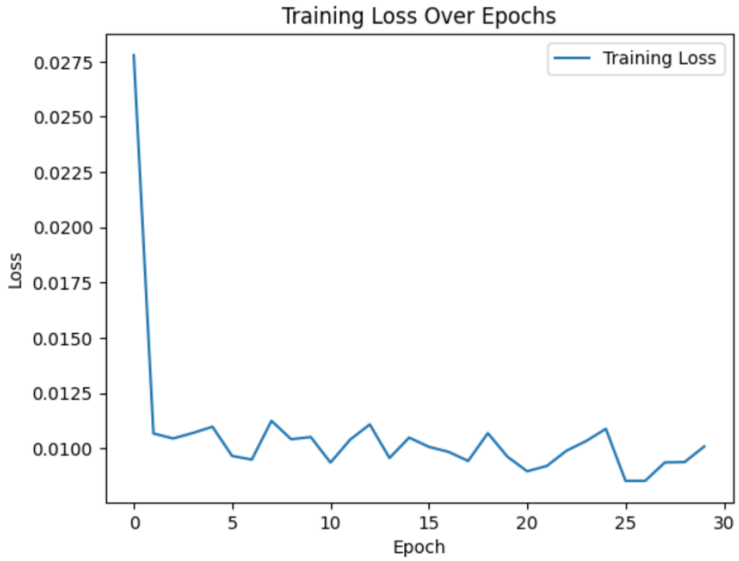

Figure 2 Training loss of RLN model over 30 epochs

The initial reprojection error in these training images was 18.6476%, which the model successfully reduced to 0.9658%. Testing on 49 testing images, the model further demonstrated

its generalization capability by decreasing the initial error from 16.3048% to 0.5694%. Figure 3 shows examples of the reprojection improvements, highlighting the effectiveness of our model.

Figure 3 The original reprojection in red and the model output in green

B. Geometric-Aware Refinement

The proposed method can be applied to 3D scenes reconstructed from images or to real LiDAR data. In our experiments, we initially used bounding boxes obtained solely from the 3D scene reconstruction to approximate the bounding boxes in the LiDAR sensor space. We then employed our geometric-aware refinement process to enhance these approximations.

As demonstrated by **(Figure 4)**, our method successfully identified and refined the bounding boxes for the LiDAR point cloud data, accurately aligning them with the actual dimensions and orientations of the traffic signs. This refinement process not only improved the positioning and orientation of the bounding boxes but also ensured they closely matched the physical properties of the objects detected in the LiDAR data, highlighting the method's effectiveness in different 3D data sources.

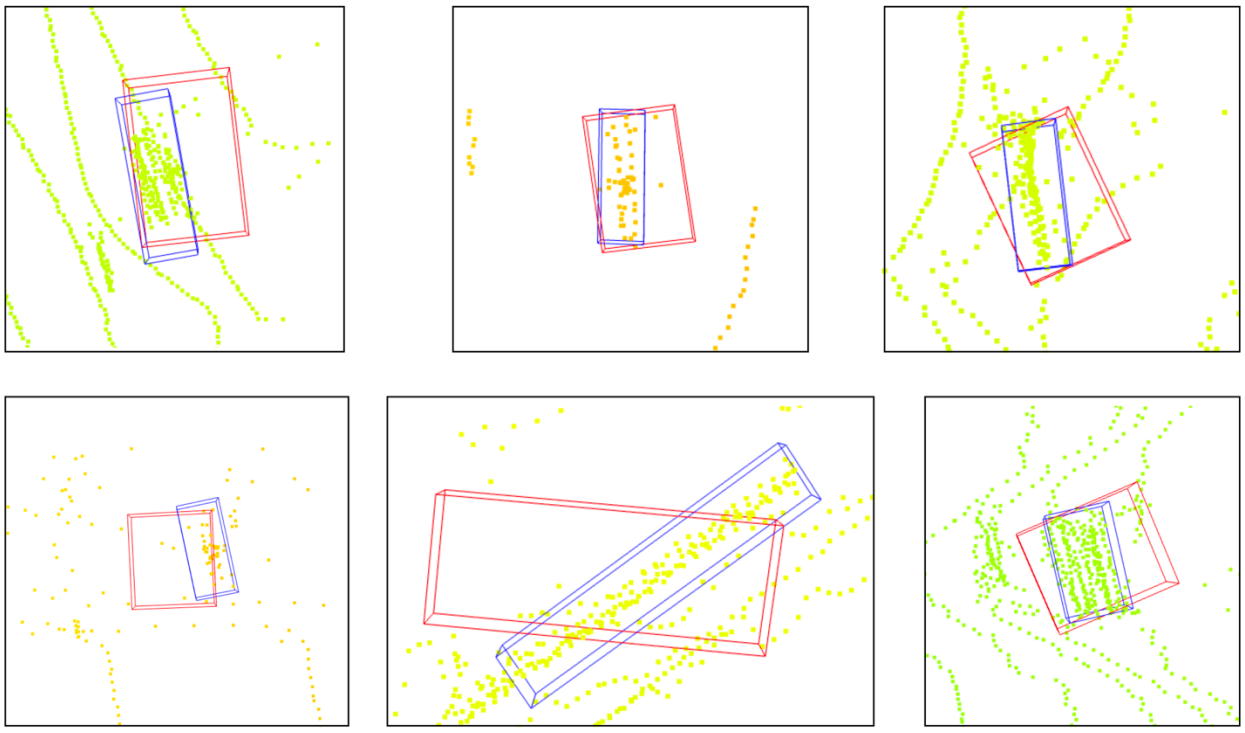

Figure 4 Red bounding boxes show the initial approximations, while blue boxes represent the results after geometric-aware refinement in 3D point cloud (Bird's eye view). The refinement process significantly improves the alignment and accuracy of the 3D bounding boxes for Kitti's LiDAR data.

CONCLUSIONS AND POLICY IMPLICATIONS

This study presented a method for 3D object detection of traffic signs using image-based depth data, avoiding the reliance on expensive LiDAR sensors. The approach combines a Reprojection Loss Network (RLN) and a geometric-aware refinement technique to enhance the precision of 3D bounding boxes. The RLN focuses on minimizing reprojection errors, while the geometric-aware refinement aligns the bounding boxes with the planar features of traffic signs, ensuring accurate orientation and dimensions.

Our experiments showed significant improvements in reprojection accuracy and alignment with actual objects, validated using a custom dataset based on the Kitti 3D Object Detection dataset specifically annotated for traffic signs.

This method offers a cost-effective and precise solution for 3D traffic sign detection and localization, which is crucial for autonomous driving. Future work could explore further improvements, such as the integration of this approach with other sensory data and its application to other objects important for autonomous navigation.

The findings of this study carry important policy implications for the deployment of autonomous vehicles. By reducing dependence on expensive LiDAR systems, the proposed image-based depth estimation approach lowers the overall cost of perception technology, making autonomous navigation more accessible to a broader range of communities. Policymakers could leverage such cost-effective methods to encourage wider adoption of autonomous systems, particularly in regions with limited infrastructure funding. Integrating reliable traffic sign detection into regulatory standards could further enhance roadway safety while supporting all users in transportation through more affordable autonomous vehicle designs and broader societal benefits.

REFERENCES

Cantrell, K. J., Miller, C. D., & Morato, C. (2020). Practical depth estimation with image segmentation and serial U-Nets. In *Proceedings of the 6th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)* (pp. 406–414). SCITEPRESS. https://doi.org/10.5220/0009781804060414

Fang, L., You, Z., Shen, G., Chen, Y., & Li, J. (2022). A joint deep learning network of point clouds and multiple views for roadside object classification from LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing, 193,* 115–136. https://doi.org/10.1016/j.isprsjprs.2022.08.022

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research, 32*(11), 1231–1237. https://doi.org/10.1177/0278364913491297

Huang, P., Cheng, M., Chen, Y., Luo, H., Wang, C., & Li, J. (2017). Traffic sign occlusion detection using mobile laser scanning point clouds. *IEEE Transactions on Intelligent Transportation Systems, 18*(9), 2364–2376. https://doi.org/10.1109/TITS.2016.2639582

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., & Dollár, P. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4015–4026). IEEE. https://doi:10.1109/ICCV51070.2023.00371

Li, R., Gong, D., Yin, W., Chen, H., Zhu, Y., Wang, K., Chen, X., Sun, J., & Zhang, Y. (2023). Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 21539–21548). IEEE. https://doi.org/10.48550/arXiv.2304.08993

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 658–666). IEEE. https://doi.org/10.48550/arXiv.1902.09630

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10371–10381). IEEE. https://doi.org/10.48550/arXiv.2401.10891

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., & Shen, C. (2023). Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9043–9053). IEEE. https://doi.org/10.48550/arXiv.2307.10984

You, Y., Wang, Y., Chao, W. L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., & Weinberger,

K. Q. (2019). Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving. *arXiv*. https://doi.org/10.48550/arXiv.1906.06310

Yu, C., Cai, Y., Zhang, J., Kong, H., Sui, W., & Yang, C. (2024). VRSO: Visual-centric reconstruction for static object annotation. *arXiv*. https://doi.org/10.48550/arXiv.2403.15026

Zhang, S., Wang, C., Lin, L., Wen, C., Yang, C., Zhang, Z., & Li, J. (2019). Automated visual recognizability evaluation of traffic signs based on 3D LiDAR point clouds. *Remote Sensing, 11*(12), 1453. https://doi.org/10.3390/rs11121453