An In-Depth Analysis of a Multi-Sensor System for Smart City Road Maintenance: Detailed Design, Implementation, and Validation of a LiDAR and AI-Driven Approach

Giovanni Nardini*, Roberto Nucera*, Alessandro Ulleri*, Stefano Cordiner†,
Eugenio Martinelli†, Arianna Mencattini† and Iulian Gabriel Coltea*
*Key To Business s.r.l., Rome, Italy
e-mail: g.nardini@key2.it, r.nucera@key2.it, a.ulleri@key2.it, i.coltea@key2.it
†Department of Industrial Engineering, University of Rome Tor Vergata, Rome, Italy
e-mail: stefano.cordiner@uniroma2.eu, eugenio.martinelli@uniroma2.eu, mencattini@eln.uniroma2.it

***Abstract*—This paper details a complete, vehicle-mounted system for automated road surface inspection, developed to enhance the efficiency and safety of large-scale urban infrastructure management. As an extended version of our previous work presented at the SMART 2025 conference, this study provides an in-depth analysis of the system architecture, a refined Artificial Intelligence (AI) pipeline, and a detailed performance evaluation under challenging real-world scenarios. The system operates on a multi-sensor fusion principle, integrating High-Resolution Light Detection and Range (LiDAR) point clouds for precise 3D geometry, camera imagery for visual texture analysis, and high-accuracy Global Navigation Satellite Systems (GNSS) and inertial data for robust georeferencing. Its AI capabilities are driven by custom models: a fine-tuned Convolutional Neural Networks (CNNs) model detects and classifies road defects like potholes and cracks in images, while a Visual Transformer (ViT) semantic segmentation model provides comprehensive semantic scene understanding to avoid false positives. Through a precise LiDAR-camera calibration, these 2D detections are then projected into the 3D domain of the point clouds. This critical step isolates each defect, allowing for the creation of a three-dimensional model and the precise quantification of its physical properties, such as surface area, depth, and volume. A significant contribution of this work is the extensive validation conducted across dozens of kilometers in a complex urban road environment in Rome, Italy. We present key quantitative results that achieve high detection accuracy and centimeter-level measurement precision. Furthermore, we discuss the iterative tuning process that overcame operational challenges like motion blur, misclassification of manholes, shadows, and road markings. The findings confirm that the system is a robust and scalable solution, with a pipeline optimized for edge computing to enable real-time analysis, delivering actionable data through a map-based web portal to facilitate proactive urban road management.**

Keywords-smart cities; road maintenance; LiDAR; AI; edge computing; sensor fusion.

I. Introduction

As urban areas grow, the need to monitor road conditions efficiently becomes crucial for keeping infrastructure intact and promoting road safety. The conventional methods of inspecting roads are laborious, time-consuming, and frequently fall short of providing the accuracy required for proactive repairs. However, recent progress in sensor technology, artificial intelligence, and data integration presents fresh opportunities for transforming road condition monitoring.

Our approach overcomes this challenge by basing AI inference solely on standard Red-Green-Blue (RGB) images and then projecting the 2D detection information into the 3D domain provided by the LiDAR. This is achieved through a meticulous camera-LiDAR calibration process, as shown in Figure 1.

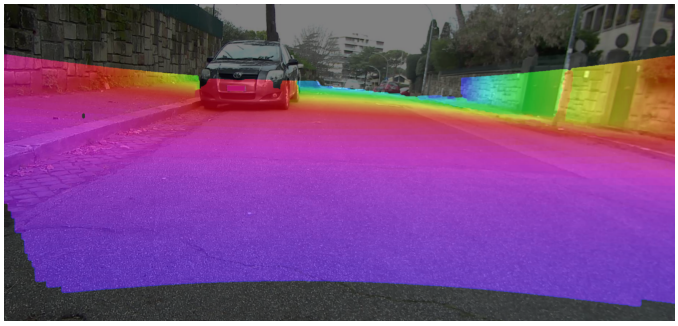

Figure 1. Camera-LiDAR Registration.

This strategy allows us to leverage large, pre-existing public datasets for training, significantly reducing development time and cost. As an extension of our preliminary work [1], this paper details the complete system that utilizes LiDAR technology along with RGB imaging and GNSS/INS data within a Robot Operating System (ROS) based framework to identify and map road surface issues efficiently. We provide a deep dive into the system's architecture, the full AI pipeline, and a comprehensive report on its real-world performance, including a transparent discussion of the operational challenges we overcame.

From an economic standpoint, the system's adaptability to city vehicles, including public transport, could potentially transform routine operations into continuous, cost-effective road monitoring. Combining this distributed sensing with on-the-ground human supervision, such as cleaning personnel, creates a hybrid model that optimizes resource use and enhances data accuracy, leading to efficient urban road maintenance.

The paper is structured as follows. Section II discusses the state of the art. Section III outlines the specific advancements over our previous conference paper. Section IV describes the system's hardware and software architecture. Section V details the AI pipeline. Section VI presents the experimental validation results, and Section VII concludes the paper.

II. Related Work

Over the past few years, many approaches have been explored for automated road inspection. Some methods rely on inertial data [2], while others utilize pure machine learning and computer vision techniques [3][4]. More sophisticated approaches exploit deep learning models [5] or combine vision and depth sensing together with spatial AI [6][7].

The technologies that have been tested for depth estimation are based on stereoscopy, Red-Green-Blue-Depth (RGB-D) cameras, and LiDAR. However, each has its own disadvantages: stereoscopy generally does not work well with feature-poor surfaces. RGB-D cameras based on Time of Flight (ToF) technology, while achieving good accuracy, drop their performance in outdoor environments and are limited to a range of a few meters. Conversely, LiDAR provides the most long-range and accurate measurements but at the expense of lower point density and the need for an additional imaging system to obtain the scene picture. Furthermore, approaches using RGB-D images as input for AI detection models, while achieving good performance due to depth information, are strongly affected by the context, sensor position, and framing of the training data. Therefore, they require the acquisition of huge amounts of images from every possible angle and distance in order to replicate all possible setups.

III. Extension of Previous Work

This journal article represents a substantial extension of the preliminary research presented in our conference paper [1]. While the original work introduced the concept of the multi-sensor fusion architecture, this paper incorporates significant technical advancements and a more rigorous validation methodology.

Firstly, the AI pipeline has been refined. In [1], the focus was primarily on detection feasibility. In this work, we present a consolidated dual-model approach (YOLOv8-seg and SegFormer) with a dedicated section on the dataset curation process, including the integration of negative examples for manholes to reduce false positives.

Secondly, the experimental validation has been vastly expanded. The previous work relied on limited datasets. Here, we present results from an extensive on-site campaign covering approximately 70 km of urban roads in Rome. This includes a new quantitative analysis of telemetry accuracy (area, volume, depth) and geolocation precision compared to ground truth.

Thirdly, we include a detailed discussion on the iterative tuning process required to handle real-world environmental challenges, such as shadows and motion blur, which were not addressed in the initial study. This comparison underscores the transition from a proof-of-concept to a field-validated prototype.

IV. System Architecture

The system was engineered as a modular, vehicle-mounted unit designed for robust data acquisition in dynamic urban environments. Its architecture integrates carefully selected hardware components with a sophisticated software pipeline built on ROS to ensure interoperability and scalability.

A. Hardware Configuration

The hardware setup was chosen to balance high-performance data acquisition with resilience to on-road conditions and suitability for edge computing. An NVIDIA Jetson AGX Orin 64GB module serves as the central computing unit, providing the necessary power for real-time AI inference. For 3D perception, an HESAI Technology AT128 Hybrid Solid-State LiDAR was selected for its optimal Field of View (FOV) and mechanical resilience. Visual information is captured by a 4K global shutter camera, a critical choice made to eliminate motion blur. A Microstrain 3DM-GQ7 module provides precise geolocalization and orientation by fusing data from a dual-antenna GNSS receiver and a 9-axis IMU. The entire system is supported by a 4G/LTE router, a network switch, and a dedicated power management system. The overall setup is shown in Figure 2.

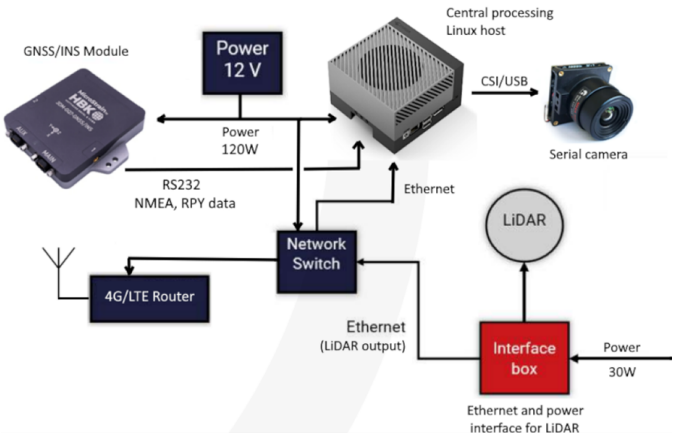

Figure 2. Hardware setup.

B. Software Architecture

The software was built entirely on ROS 2, a flexible framework for communication and data processing. The architecture functions as a pipeline that transforms raw sensor data into actionable insights. The process begins with dedicated driver nodes that interface with each sensor and publish data onto specific ROS topics. A core AI Inference Node, developed for this project, subscribes to these topics and uses an internal synchronizer to create a coherent snapshot of the environment from different sensor inputs. This synchronized data is then fed into the AI models. The resulting output is packaged into Safetensors files for later use. A separate Post-Processing and Reporting Node operates independently on these files. This agent performs the final data fusion and analysis, projecting 2D detections onto the 3D LiDAR point cloud, calculating the geometric properties of each defect, and assigning precise geographic coordinates. Finally, it formats the data into a JSON payload for transmission to a map-based web portal. This decoupled architecture makes the process resilient to network connectivity issues.

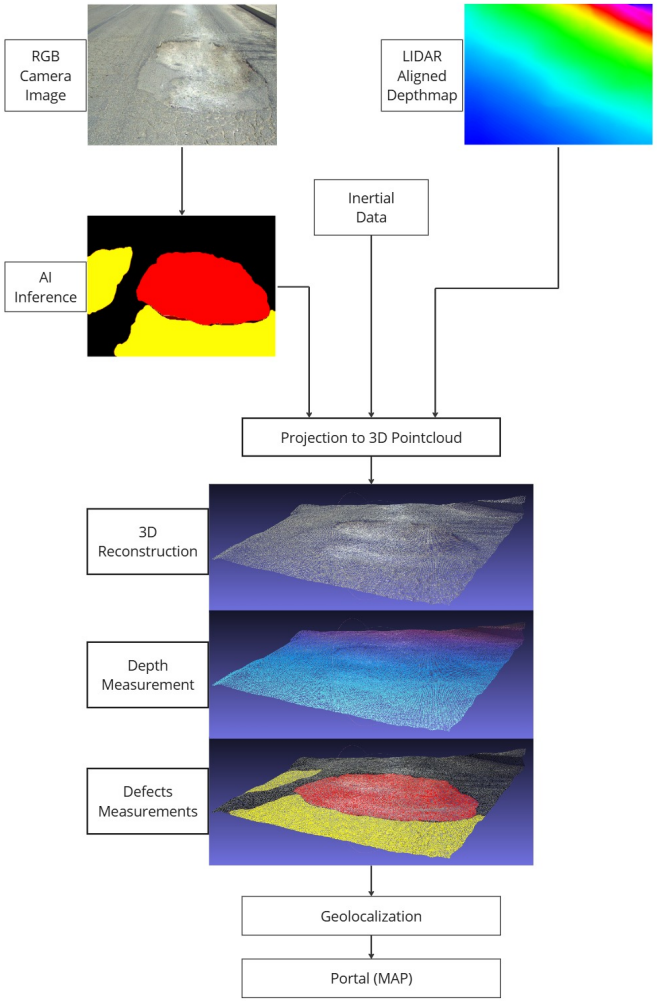

Figure 3. Software architecture components in ROS framework.

V. AI Models

The system's ability to accurately identify road defects depends on an AI pipeline composed of two distinct deep learning models. The development process involved careful model selection, extensive dataset preparation, rigorous training, and detailed validation.

A. Model Selection

Two primary architectures were chosen for their proven performance. For defect detection, a You Only Look Once (YOLO) model [8] version "v8small-seg" was selected for its powerful instance segmentation capabilities. This task extends beyond simple object detection by not only providing a bounding box for each detected object but also generating a pixel-perfect mask that outlines its exact shape. This capability is crucial for our application, as the generated masks are later projected into the 3D domain to enable precise geometric measurements of defects, such as their area and volume. The "s" variant (YOLOv8s-seg) was specifically chosen as it offers an excellent trade-off between accuracy and computational efficiency, making it highly suitable for real-time processing on our edge computing platform.

In parallel, a SegFormer variant B1 model [9] is used for scene understanding. This model is based on a Vision Transformer (ViT) architecture, which, unlike traditional CNNs, excels at capturing long-range dependencies and global context within an image. This makes SegFormer particularly effective for semantic segmentation, the task of assigning a class label to every pixel. Its primary role in our pipeline is to generate a highly accurate and reliable "road mask". This mask serves as a critical contextual filter: by intersecting the defect detections from YOLO with this road mask, we can effectively eliminate false positives that may occur on sidewalks, vegetation, or other non-road surfaces, thereby significantly increasing the overall reliability of the system.

B. Dataset Preparation and Training

The performance of the AI models relies on carefully prepared training data. For the scene understanding model, a transfer learning approach was used, employing a SegFormerB1 model pre-trained on the well-known Cityscapes dataset [10]. The dense annotations of urban scenes in this large-scale dataset enabled high-performance road segmentation without the need to create a new dataset from scratch.

For the defect detection model, however, an initial analysis of existing public datasets revealed that none fully met the project's requirements in terms of camera perspective, labeling quality, and class definitions. Consequently, a significant effort was dedicated to creating a custom, high-quality dataset. The process began by curating a base set of images from the public Road Damage Detection (RDD) dataset [11], selecting only those from geographical regions with road conditions and perspectives relevant to the target operational environment.

This base set was strategically composed to include a substantial number of images without any defects to train the model to minimize false positives on well-maintained road surfaces. A critical challenge identified was the underrepresentation of certain scenarios, particularly images containing both potholes and manholes, which often led to misclassifications. To address this, the dataset was enriched with hundreds of additional images sourced from various other public repositories, specifically chosen to increase the variety of potholes and provide negative examples of manholes.

The final curated dataset consisted of 6,504 images, split into training (5,199), validation (652), and test (653) sets. A meticulous manual re-labeling process was undertaken with the help of Segment Anything Model (SAM) [12] to create precise instance segmentation masks for two target classes: pothole and crack, which consolidated various types of fissures like alligator and linear cracks into a single category. To further enhance the model's robustness and its ability to generalize, the training set was expanded through extensive data augmentation. Techniques such as flipping, rotation, and adjustments to saturation, brightness, and noise were applied, resulting in a final training dataset of 25,995 images.

The training process itself employed a fine-tuning strategy, starting from the Common Objects in Context (COCO) pre-trained model. This approach leverages the generalized features learned on a large-scale dataset and adapts them to the specific task of defect detection. The default training configurations were used, which include additional data augmentation techniques like mosaicing to improve the model's performance on objects at various scales. The training was configured to run for 100 epochs with a standardized input image size of 640x640 pixels, using the custom dataset described above.

C. Model Validation and Performance Metrics

After training, the models were rigorously evaluated on their respective test sets, to provide an unbiased assessment of their generalization capabilities. This validation involved both a quantitative analysis through standard computer vision metrics and a qualitative visual inspection of the model's predictions. The quantitative evaluation is based on the confusion matrix, which categorizes predictions into True Positives (TP), False Positives (FP), and False Negatives (FN). Key metrics include Precision, which measures the model's ability to avoid false positives, and Recall, which measures the model's ability to find all relevant instances in an image. Additionally, also the F1-score is taken into account, providing the balance between Precision and Recall in one formula:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

where P is the Precision and R is the Recall value.

For tasks involving spatial localization, such as segmentation and detection, accuracy is quantified by the Intersection over Union (IoU). This metric measures the overlap between the predicted region (mask or bounding box) and the ground-truth region, providing a score for spatial accuracy. From this, the Average Precision (AP) is calculated for each class by averaging the precision values over the Precision-Recall curve. The primary summary metric for object detection models is the mean Average Precision (mAP), which is the mean of the AP values across all classes and, often, across a range of IoU thresholds, e.g., mAP@0.5:0.95. The mAP is calculated with the following equation:

$$\text{mAP} = \frac{1}{N_c} \sum_{c=1}^{N_c} \frac{1}{N_{\text{IoU}}} \sum_{i=1}^{N_{\text{IoU}}} \text{AP}_c^{(i)} \tag{2}$$

where N_c is the total number of classes, N_{IoU} is the number of IoU thresholds, and $\text{AP}_c^{(i)}$ is the average precision for class c at IoU threshold i.

The YOLOv8s-seg defect detection model achieved an mAP of 0.564. While the performance for the pothole class was strong, the instance segmentation of cracks proved more challenging due to their ambiguous and continuous nature, making it difficult for the model to distinguish separate instances. However, when evaluating the crack class from a semantic segmentation perspective (merging all predicted crack masks), the model's ability to correctly identify crack pixels versus background was excellent, confirming its effectiveness for the project's use case. The F1-Confidence curve indicated an optimal confidence threshold of 0.343, at which the model reached a balanced F1-score of 0.57.

The SegFormerB1 model demonstrated outstanding performance for road segmentation. On the Cityscapes test set, it achieved an F1-score of 0.98, specifically for the main road class, with a correct pixel classification rate of 99%. On the other hand, semantic segmentation models are often evaluated with meanIoU metrics, calculated with the following formulas:

$$\text{IoU}_c = \frac{|A_c \cap B_c|}{|A_c \cup B_c|} \tag{3}$$

$$\text{meanIoU} = \frac{1}{N_c} \sum_{c=1}^{N_c} \text{IoU}_c \tag{4}$$

where A_c is the predicted segmentation for class c, B_c is the ground truth segmentation for class c, and N_c is the number of classes.

The resulting meanIoU was actually 0.43, not as excellent as other metrics, but this is expected as the metric heavily penalizes minor shape deviations, which are common in complex road scenes. For the primary task of creating a reliable road mask, the performance was deemed excellent. Following validation, both models were optimized using NVIDIA TensorRT with FP16 precision, which more than doubled their inference speed on the edge device with negligible impact on accuracy. A summary of the resulting metrics is shown in Table I. A qualitative visual analysis further confirmed the two models' performance and robustness across various scenarios as shown in Figure 4.

TABLE I. SUMMARY OF AI MODEL PERFORMANCE.

Model	Main Task	Key Metric	Value
Yolov8s-seg	Road Defect Detection	mAP@0.5 (all classes)	0.564
		F1-Score (optimal)	0.57
SegFormerB1	Road Segmentation	F1-Score ("Road" class)	$\approx$ 0.98
		MeanIOU ("Road" class)	0.43

VI. EXPERIMENTAL SYSTEM VALIDATION

The system's validation followed a two-stage process. First, controlled laboratory tests were conducted to calibrate the sensors and benchmark core functionalities. Then, extensive on-site tests were run to evaluate its performance and robustness in real-world scenarios.

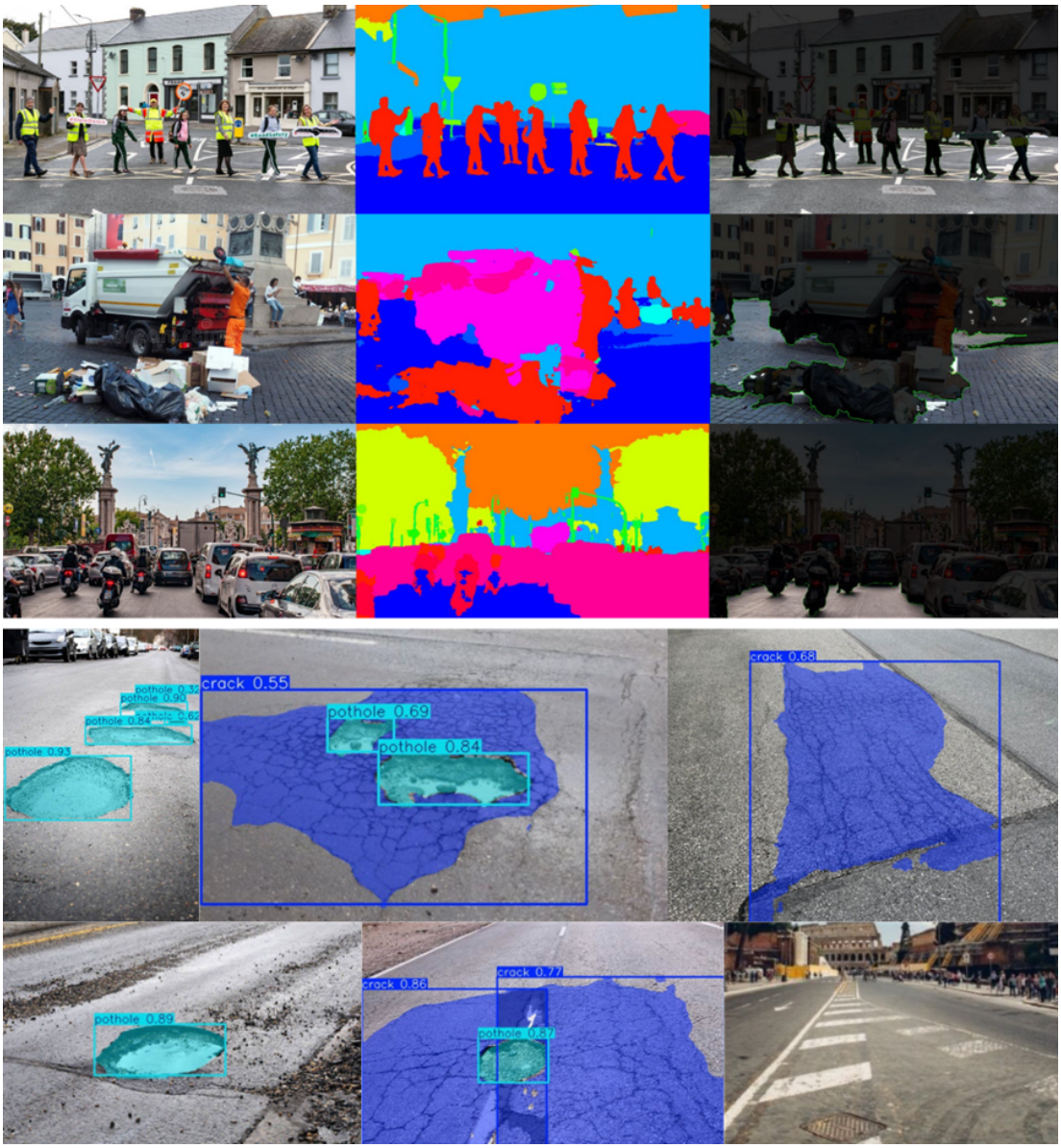

Figure 4. Qualitative inspection of model outputs. The top rows illustrate the SegFormerB1 model's road segmentation, displaying the input image, the complete segmentation map colored by CityScapes class, and the activation map for the "road" class. The bottom rows present a mosaic of detections from the YoloV8s-seg model, which identifies cracks (blue) and potholes (cyan) using bounding boxes and masks, each annotated with a confidence score.

A. Laboratory Calibration and Testing

Before on-site deployment, the integrated prototype was subjected to a series of crucial tests in a controlled laboratory environment. This preparatory phase was fundamental to ensuring the system's reliability and accuracy.

The most critical step was the multi-sensor calibration. Since the system relies on fusing data from a 2D camera and a 3D LiDAR, it was essential to precisely determine the geometric relationship between them. This process was divided into two parts. First, an intrinsic camera calibration was performed using a standard chessboard pattern viewed from multiple angles, as shown in Figure 5.

This allowed us to calculate the camera's internal parameters, i.e., the K matrix shown in Table II, and to generate correction maps to remove lens distortion. As shown in Figure 6, this undistortion process transforms the raw, warped image into a geometrically accurate one, which is a prerequisite for any precise measurement.

TABLE II. INTRINSIC MATRIX (K).

$$K = \begin{bmatrix} 304.6712 & 0.0 & 313.5861 \\ 0.0 & 380.5664 & 165.4646 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Next, an extrinsic LiDAR-camera calibration was conducted. This procedure establishes the rigid transformation, i.e., rotation and translation, between the LiDAR's and the

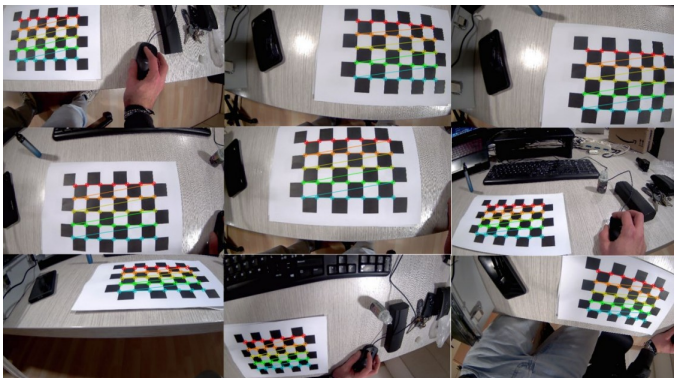

Figure 5. Intrinsics calibration process: using multiple images of chessboard from different points of view to calculate intrinsics matrix.

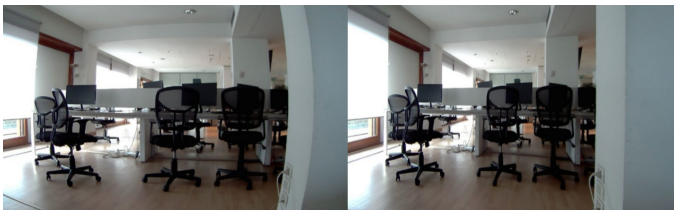

Figure 6. Undistortion process, after intrinsics calibration. Left: original image. Right: Undistorted image.

camera's coordinate systems. By manually identifying a set of corresponding points in both the 3D LiDAR point cloud and the 2D camera image, we used a Perspective-n-Point (PnP) algorithm to compute the 4x4 roto-translation matrix, shown in Table III. The accuracy of this calibration was verified by reprojecting the 3D LiDAR points onto the 2D image using the calculated matrix; the near-perfect alignment confirmed the success of the calibration. This matrix is the key that enables the accurate fusion of data from the two sensors. Final results of reprojection are shown in Figure 7.

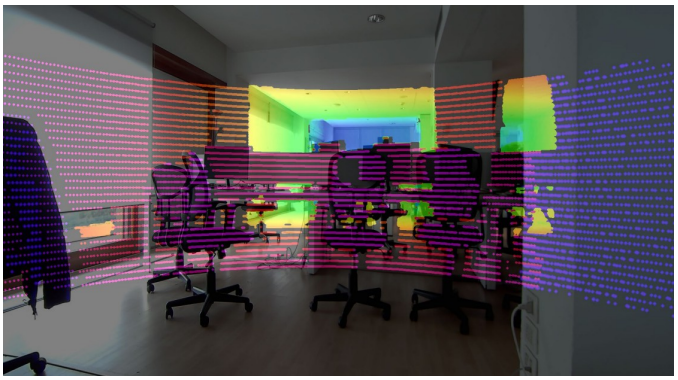

Figure 7. Point cloud reprojection onto RGB image, after intrinsics and estrinsics calibration.

With the sensors calibrated, we proceeded to benchmark the system's core functionalities. Throughput tests on the ROS AI node were conducted to measure its computational performance. The system demonstrated a stable processing average of 11.12 Frames Per Second (FPS), confirming its capability for real-time processing at typical urban driving speeds.

TABLE III. ESTRINSINCS ROTO-TRANSLATION MATRIX (RT).

$$RT = \begin{bmatrix} 0.998 & -0.008 & -0.059 & -0.111 \\ 0.008 & 1.000 & 0.005 & 0.014 \\ 0.059 & -0.005 & 0.998 & 0.093 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

The system's telemetry accuracy was evaluated using targets of known size - specifically, two manholes of different shapes - to simulate road defects. The results were highly encouraging, showing a low average relative error of just 6% for surface area measurements. The more challenging depth and volume estimations also yielded respectable average relative errors of 24% and 19%, respectively, demonstrating a good approximation capability.

Finally, static geolocation tests were performed to assess positioning accuracy. The system reported the coordinates of known points, which were then compared against high-precision ground truth data obtained from a topographic survey. The tests revealed a mean horizontal error of 2.13 meters, an accuracy level that is well within acceptable limits for the primary goal of dispatching maintenance crews to the correct location. A summary of the resulting metrics is shown in Table IV.

TABLE IV. SUMMARY OF THE PROTOTYPE'S MEASUREMENT AND POSITIONING PERFORMANCE.

Category	Metric	Calculated Value
Defect Dimensional Estimation (Telemetry)		
Mean Relative Error	Area	6%
Mean Relative Error	Depth	24%
Mean Relative Error	Volume	19%
Geographic Defect Localization		
Mean Error	Horizontal	2.131 meters
Root Mean Squared Error	Horizontal RMSE	2.780 meters

B. On-Site Testing

The most comprehensive and conclusive validation of the system was an extensive on-site testing campaign. This involved deploying the fully calibrated prototype on a vehicle and conducting multiple data acquisition sessions across approximately 70 km of public roads in Rome, Italy. The system has been positioned on a vehicle, pointing forward as shown in Figure 9. The routes were strategically selected to cover a wide spectrum of real-world conditions, including different road types (from high-speed arteries to narrow residential streets), varying traffic densities, and diverse lighting environments. This campaign was structured as an iterative process of testing, analysis, and refinement, allowing to systematically address challenges encountered in the field.

A primary challenge identified during the initial test runs was the misclassification of manholes. The AI model fre-

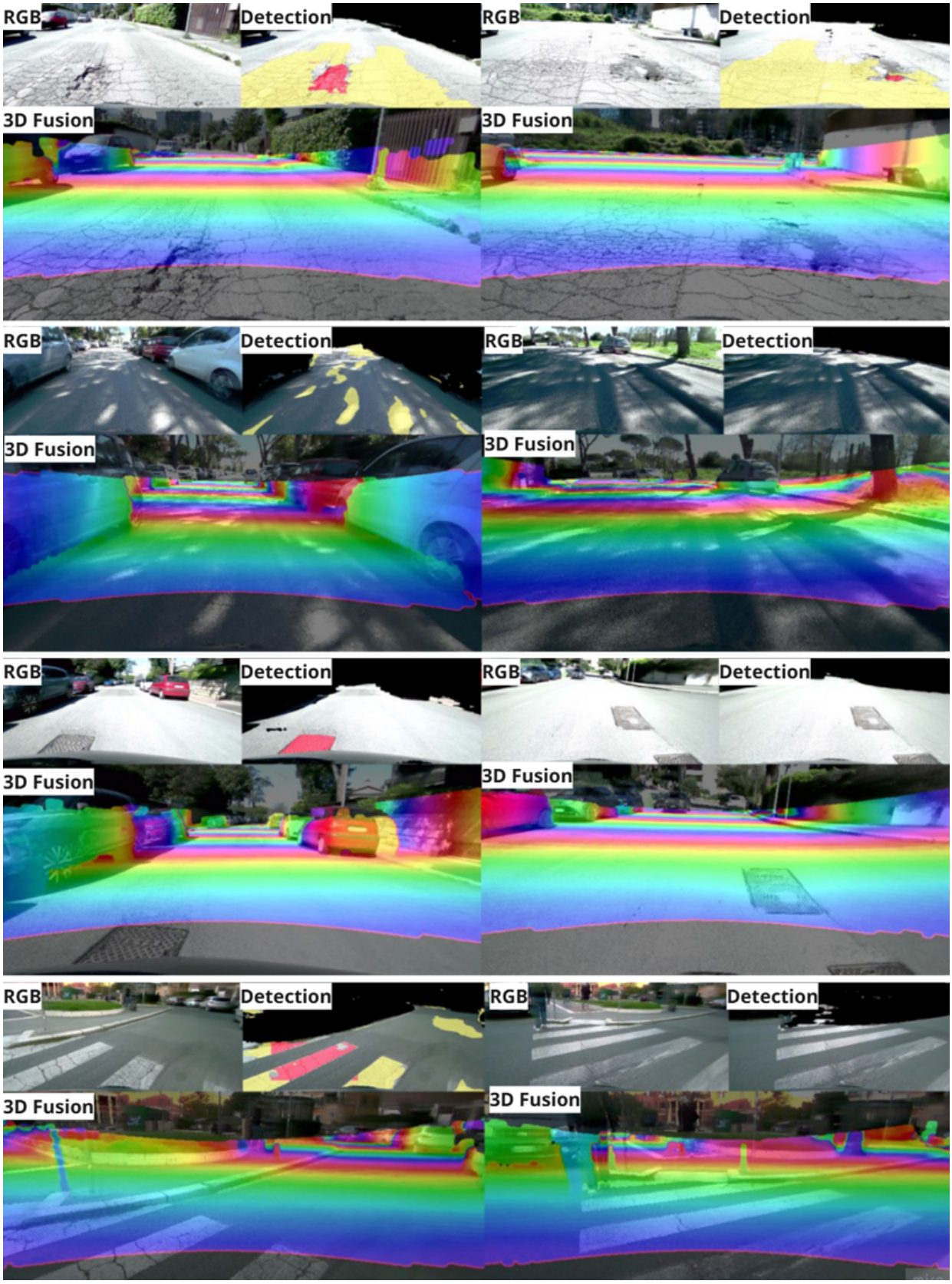

Figure 8. Iterative tuning of on-site tests: each row represents different situations. On the left: the system output before correction. On the right: output after tuning. The main issues were related to: false positives related to manholes, sharp shadows producing fake crack output, dark environment and road marks sensitivity.

Figure 9. A picture of the system prototype.

Figure 10. A screenshot of the web application for the visualization of results.

quently identified manholes, particularly those slightly recessed or with damaged edges, as "potholes". Through a careful analysis of these false positives, it has been determined that the confidence scores assigned to manholes were consistently lower than those for genuine potholes. Based on this observation, we iteratively adjusted the detection threshold, raising the minimum score threshold value for potholes to 0.30. This seemingly simple change proved highly effective, making the model more selective and drastically reducing manhole-related false positives without compromising its sensitivity to actual defects.

Another significant issue was caused by environmental factors, specifically the strong, hard-edged shadows cast by buildings and trees on sunny days. The high contrast along these shadow lines was often misinterpreted by the model as "cracks". This led to another round of data-driven tuning. After experimenting with different values, the confidence threshold for the minimum score threshold value for cracks was also set to 0.30. This found an optimal balance, successfully eliminating the vast majority of shadow-induced artifacts while still reliably detecting significant cracks, and also helped in correctly ignoring most worn-out road markings. Examples of challenges overcome through iterative tuning are shown in Figure 8.

The campaign culminated in a final, long-duration test session of approximately 35 km, which served as a validation run for the fully tuned system. This test confirmed the system's operational stability and thermal resistance over an extended period. Throughout this iterative process, the web portal as shown in Figure 10, proved to be a valid tool, allowing for rapid qualitative inspection of results and enabling the data-driven refinements that led to a robust and reliable final configuration for automated road condition assessment.

VII. Conclusion and Future Work

This paper has presented an extended and in-depth analysis of a multi-sensor system for automated road defect assessment, building upon previous work [1]. By detailing the system's architecture, its AI pipeline, and the results of extensive validation, this work has demonstrated a robust and viable solution for enhancing Smart City infrastructure management. The rigorous testing campaign confirmed the system's high performance. The AI-driven pipeline achieves reliable detection of critical defects suitable for practical applications. Furthermore, the fusion of LiDAR and camera data enables quantitative measurements and geolocalization with a precision that provides actionable data for maintenance planning. The iterative, data-driven tuning of the AI models was essential for overcoming real-world challenges and achieving a stable final configuration. The deployment of the entire processing pipeline on an edge computing device ensures real-time capabilities and operational autonomy. The final system represents a significant step forward from conventional inspection methods, offering a scalable, cost-effective, and data-rich approach to proactive road maintenance. Future work will specifically focus on improving the segmentation performance for complex crack patterns by exploring advanced model architectures and expanding the dataset with more diverse crack topologies, alongside further refining depth and volume estimation algorithms. Ultimately, this work validates the power of integrating advanced sensing and artificial intelligence to create smarter and more efficient urban environments.

Acknowledgement

This project has been co-financed by the European Union through the PR FESR 2021–2027 RSI program of Regione Lazio, managed by LazioInnova. The authors would like to thank the European Union and Regione Lazio for their support in enabling this research. Additionally, we extend our gratitude to all partners and collaborators who contributed to the successful implementation and validation of the proposed system.

References

[1] G. Nardini et al., "Smart city road maintenance: A lidar and ai-driven approach for detecting and mapping road defects", in *The Fourteenth International Conference on Smart Cities, Systems, Devices and Technologies (SMART 2025)*, Valencia, Spain, Apr. 2025.

[2] S. Girisan, T. V. Sreelakshmi, N. V. Swetha, V. Suresh, and K. M. Vipin, "Pothole detection based on accelerometer method", *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 5, May 2020.

[3] A. Ahmadi, S. Khalesi, and M. R. Bagheri, "Automatic road crack detection and classification using image processing techniques, machine learning and integrated models in urban areas: A novel image binarization technique", *Journal of Industrial and Systems Engineering*, vol. 11, pp. 85–97, 2018.

[4] K. Li, B. Wang, Y. Tian, and Z. Qi, "Fast and accurate road crack detection based on adaptive cost-sensitive loss function", *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 1051–1062, Feb. 2023, ISSN: 2168-2275.

[5] Y. Li, C. Yin, Y. Lei, J. Zhang, and Y. Yan, "Rdd-yolo: Road damage detection algorithm based on improved you only look once version 8", *Applied Sciences*, vol. 14, p. 3360, Apr. 2024.

[6] E. M. Thompson et al., "Shrec 2022: Pothole and crack detection in the road pavement using images and rgb-d data", in *SHREC2022 3D Shape Retrieval Challenge*, 2022.

[7] A. Talha, M. Karasneh, D. Manasreh, A. Oide, and M. Nazzal, "A lidar-camera fusion approach for automated detection and assessment of potholes using an autonomous vehicle platform", *Innovative Infrastructure Solutions*, vol. 8, Sep. 2023.

[8] J. Zeng, H. Ouyang, M. Liu, L. U. Leng, and X. Fu, "Multi-scale yolact for instance segmentation", *Journal of King Saud University - Computer and Information Sciences*, vol. 34, Oct. 2022.

[9] E. Xie et al., "Segformer: Simple and efficient design for semantic segmentation with transformers", in *Neural Information Processing Systems (NeurIPS)*, 2021.

[10] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[11] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "Rdd2022: A multi-national image dataset for automatic road damage detection", in *Crowdsensing-based Road Damage Detection Challenge (CRDDC)*, 2022.

[12] A. Kirillov et al., "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.