Low-Cost, Multi-Modal Road Defect Detection & Assessment

Deepesh Singh
University of Washington
dsingh01@uw.edu

Aatish Parson
University of Washington
aatishp@uw.edu

Ishan Gupta
University of Washington
doma9235@uw.edu

Abstract

Urban road maintenance still depends on labor-intensive inspections, ad-hoc citizen reports, or infrequent and costly mobile surveys. These conventional methods are often slow, inaccurate, or prohibitively expensive, resulting in incomplete data and delayed maintenance actions by municipalities. In response, we propose a cost-effective, fine-grained road defect detection system that fuses top-down camera imagery with LiDAR-derived depth data to generate pixel-level RGBD representations of road surfaces. These fused inputs are processed using deep learning models to detect, classify, and grade road defects—specifically cracks and potholes—under varying lighting and weather conditions. To identify an effective detection system, we auto-annotate the dataset using a pretrained YOLOv8 and SAM-2 model and finetune and evaluate three distinct models: Swin-Small, PVT-Small, and CoAtNet-0, to determine a more nuanced classification and assessment. System validation is conducted on road networks surrounding the neighborhoods of the University of Washington and the City of Seattle where we anticipate precise defect identification and geolocation at significantly reduced costs compared to current surveying methodologies.

1. Introduction

Road infrastructure maintenance is vital for public safety and economic activity, yet existing road surveying methods are limited. Manual inspection by city workers is slow, labor-intensive, and costly, requiring inspectors to cover vast road networks on foot or by vehicle while documenting defects through subjective visual assessments that can vary significantly between inspectors. Crowd-sourced reporting, where residents call or email about road damage, often lacks visuals and precise localization, hampering prioritization and repair efforts while creating inconsistent data quality that makes it difficult to assess severity or track changes over time. Meanwhile, private contractors offer high-fidelity 3D road surveys using specialized vehicles equipped with advanced sensors, but charge upwards of $2500 per mile, leading to infrequent data collection that leaves municipalities with outdated information about rapidly deteriorating road conditions. In cities like Seattle, comprehensive surveys happen once every two years and are restricted to arterial roads, despite over 25,000 pothole repairs reported annually [2]. This reactive approach to maintenance results in higher long-term costs as minor defects evolve into major infrastructure failures, while creating safety hazards and vehicle damage that could have been prevented through timely intervention.

Road defects principally manifest as cracks and potholes, each requiring different repair strategies based on their dimensional characteristics. Cracks typically include longitudinal cracks running parallel to traffic flow, transverse cracks perpendicular to the roadway, and alligator cracks that form interconnected patterns resembling reptilian skin. Potholes represent the most severe form of pavement failure, ranging from shallow depressions to deep cavities that expose the underlying base layer. Repair decisions are critically dependent on defect severity measurements: hairline cracks less than 3mm wide may only require crack sealing, while cracks exceeding 12mm in width necessitate more extensive patching or resurfacing. Similarly, pothole depth dictates repair methodology—shallow potholes under 25mm deep can be addressed with cold-mix asphalt, whereas deeper cavities require hot-mix asphalt application and compaction, meaning repairs become exponentially more expensive as dimensions increase. Accurate measurement of these dimensional parameters is therefore essential for cost-effective maintenance planning and resource allocation.

Recent research has centered on applying cutting-edge, open-source object-detection models and vision transformers to standard, publicly available road-damage datasets. Although this work has advanced the field, three important gaps remain: (1) existing datasets and model outputs lack the fine-grained labels required to capture damage severity, surface area, and specific defect types; (2) no published pipeline demonstrates how to integrate low-density LiDAR scans effectively within deep-learning frameworks; and (3) current approaches have been validated almost exclusively on benchmark datasets, leaving their ability to generalize across varied pavement materials, climates, and lighting conditions largely untested. This paper aims to introduce a system that tackles the first two issues.

Naturally, this methodology is subject to the vulnerability that the low-cost hardware is far too lacking in key information to allow for the discernment of road from potholes and its severity relative to more expensive options. However, with any level of success, developing a road defect detection system for under $150 has the potential to democratize infrastructure monitoring by enabling small municipalities, developing regions, and grassroots organizations to collect relatively high-quality road condition data without relying on expensive commercial surveys. Widespread deployment of such low-cost systems could lead to more frequent and equitable maintenance, reducing vehicle damage, improving safety, and optimizing public spending on infrastructure repairs.

2. Related Works

Pavement management and assessment capabilities have evolved drastically over the past decade, yet historical methods of manual defect detection and maintenance endure. On a municipality level, visual surveys, photographic documentation, and physical measurements based on standardized protocols such as ASTM D6433 still are the subjective, ruling method of judgment, which is evidently a labor-intensive and time-consuming process [5]. Thanks to the remarkable progress in machine learning and computer vision these tasks can overtaken by automated detection systems; although verification and more intricate tasks such as grading curvature remain [7].

2.1. Computer Vision

Early research in road defect detection relied on classical image processing techniques that exploited textural and geometric properties of pavement distresses through edge detection, morphological operations, and thresholding-based segmentation methods. While these approaches established foundational principles for automated defect identification, their reliance on hand-crafted features limited robustness across diverse imaging conditions and pavement types, which often times produced frequent misses and false alarms [12]. Additionally, although LiDAR-based approaches demonstrated effectiveness for precise depth measurement, image-based methods dominated literature due to lower costs and simpler processing requirements.

2.2. Deep Learning

Deep learning has emerged as the dominant paradigm in road defect detection research, though the field grapples with fundamental trade-offs between accuracy and computational efficiency. While 3D point-cloud networks achieve exceptional precision, their high labeling and computational demands favor 2D image-based approaches for practical deployment.

First there were two-stage detectors that showed promise but faced significant limitations. Gou *et al*. suggested an enhanced Faster R-CNN that improved robustness at slower inference speeds [6] and Shen *et al*. improved upon a Cascade R-CNN that broadened classification and increased accuracy relative to traditional vision methods.

Due to their demonstrably superior performance, streamlined architecture, and increased computational efficiency, single-stage object detection algorithms then grabbed the attention of academics. YOLO variants prioritized deployment speed through innovations like YOLOv5-CBoT (Bottleneck-Transformer and C2f blocks for crack patterns) [14], BL-YOLOv8 (BiFPN fusion with SimSPPF pooling) [13], and mobile-focused approaches like YOLO9tr and GASYOLO that reduced model size while maintaining precision [9] [15].

The latest models to reshape detection have been Vision Transformers, most notablely DETR. The original DETR enabled end-to-end prediction but suffered from slow convergence and poor small-object detection [3]. Improvements included Deformable DETR (sparse attention, multi-scale features), DAB-DETR (anchor-like priors) [8], and DINO (contrastive denoising) [17], though quadratic attention costs limited real-time use. RT-DETR addressed latency through hybrid encoders, IoU-aware query selection, and early-exit decoders, achieving DETR-level accuracy at video frame rates—making it compelling for pavement detection requiring robustness, speed, and deployability [18].

This past month Pavement-DETR was released, which builds on RT-DETR by inserting Channel-Spatial Shuffle attention into ResNet layers for sharper road-feature focus, replacing the neck with Conv3XC fusion blocks for better foreground-background separation, and adopting composite loss combining Powerful-IoU v2 and Normalized Wasserstein Distance for improved small-object precision [19].

Leveraging the advancements discussed, this paper outlines a robust, low-cost, end-to-end, road-defect system for expanded classifications, severity assessment, and detailed measurements for a practical use by government and industry applications.

3. Methodology

3.1. Sensor Platform and Mounting Configuration

A video camera and light detection and ranging (LiDAR) sensor were setup on a custom three-dimensional (3D) printed bracket. The bracket was rigidly affixed to a vehicle's (Acura MDX 2009) grill at a 30° upwards angle and a vertical height of 1.0m above the pavement surface. Compared with a previously evaluated 0° (parallel-to-ground)

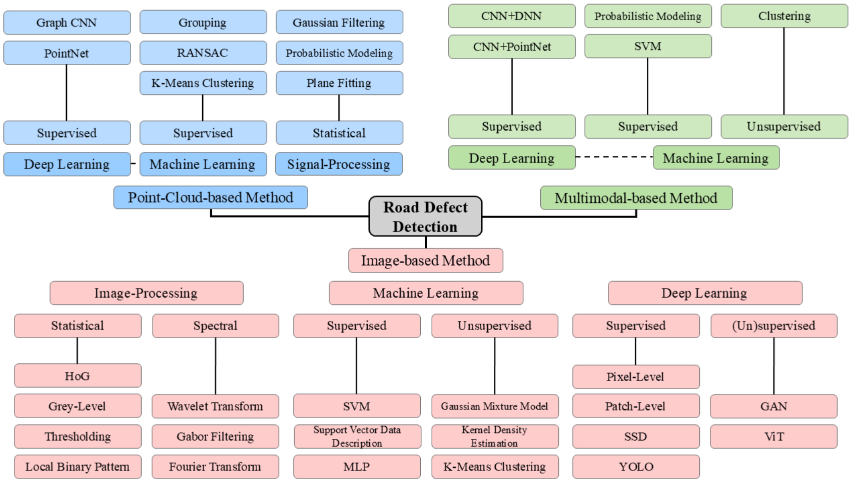


Figure 1. Taxonomy of Road Defect Detection Models [16]

orientation at 0.95m, the inclined configuration increased the horizontal field of view for both modalities, enabling wider lane coverage and improved contextualisation of disturbed locations relative to lane boundaries.

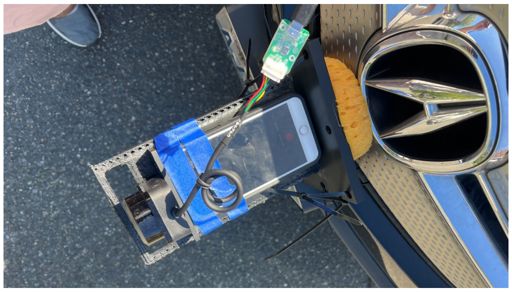


Figure 2. Data Collection Setup on Vehicle

3.1.1 Sensor Specifications

- Camera: Apple iPhone 13, ultra-wide lens (0.5× optical zoom), 30 fps
- LiDAR: RPLIDAR C1, 360° field of view, 10Hz rotational frequency (one revolution per 100ms), 12m measuring distance, 0.05m blind range

3.2. Data Collection

Field campaigns were performed on urban asphalt roadways using the aforementioned vehicle operated at 10–15 mph (4.5-6.7 m/s). Due to scan overlap and temporal stitching over a 1-second window (15 frames pre/post), each image benefits from multiple LiDAR scans contributing spatial context. Accounting for this temporal fusion, the system effectively capturef 2-3 LiDAR measurements per 1-meter segment of roadway, enabling sufficient spatial sampling for generating coarse depth maps. Two videos amounting to approximately 20-min worth of driving sequence were recorded, yielding 36,000 RGB frames (30 fps). Unfortunately, due to variety of reasons, the usable amount of frames was far lower ($\approx$ 2000). Before pre-processing, frames were indexed and time-stamped to maintain precise synchrony with LiDAR scans.

3.3. Data Pre-Processing

3.3.1 LiDAR Projection and Depth Map Generation

The process began by projecting LiDAR points (320° - 40°) onto image frames using a fitted equation that mapped measurement angles to image coordinates. Since LiDAR points are along a given y in the frame, we calculated the x-coordinates using:

$$angle_{new} = \begin{cases} angle_{old} - 320, & \text{if } angle_{old} > 100 \\ angle_{old} + 40, & \text{if } otherwsie \end{cases}$$

$$x = 9.97354644 \times angle_{new} - 18.97857$$

For each frame, the closest LiDAR points (by timestamp) were selected to create a single rotation overlay, with each point represented as a small bounding box—establishing a mapping between the image and LiDAR data. All image pixels within a given LiDAR point bounding-box were assigned a depth D(u,v) based on the value of the LiDAR point.

To enhance depth coverage, a temporal frame window of 1.0s (i-15...i+15) was considered. The 10 LiDAR overlays from these frames were aligned to the current image via planar homographies whose correspondences were derived by Harris corner detection and HOG descriptors and smoothed using linear blending. The LiDAR bounding boxes were then translated accordingly and aggregated across frames.

At the end, overlapping depth estimates were averaged to produce a dense, per-pixel depth map for each image. Local surface elevation anomalies were defined as pixels whose depth deviated by $> 5\%$ from the local median. Solely positive deviations (potholes) were considered given that it was hard to assess the causality of negative deviations (bumps). Due to the RPLIDAR's ranging noise, depth for narrow ($<$ 10mm) cracks was deemed unreliable and excluded from quantitative analysis.

3.3.2 Auto-Annotation

Each raw frame was also processed with the Ultralytics Auto-Annotate pipeline using a YOLOv8 detector (pre-trained on the Crowd-sensing Road Damage Detection Challenge 2022) with the Segment-Anything Model v2 (SAM-2) to obtain pixel-level masks. YOLOv8 was preferred over newer versions YOLOv9/YOLOv10/YOLOv11 owing to its superior performance on small-scale, near-field targets such as pavement cracks [10].

Detected defects were assigned to one of four mutually exclusive classes: **(1)** Longitudinal Crack **(2)** Transverse Crack **(3)** Alligator Crack **(4)** Pothole

3.3.3 Severity Grading and Further Classification

Even with auto-annotation, raw frames were re-classified using an extended set of categories that included: Good Road, Block Cracking, Slippage Cracking, Corner Break, and Raveling [1]. Severity was qualitatively graded following the Pavement Condition Index (PCI):

- Level 0 (None): Dark Green (negligible)
- Level 1 (Low): Light Green/Yellow (minor)
- Level 2 (Medium): Light Red/Medium Red (moderate)
- Level 3 (High): Dark Red/Grey (severe)

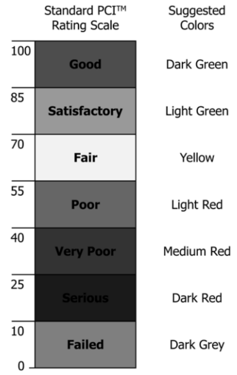


Figure 3. Pavement Condition Index (PCI), Rating Scale, and Suggested Colors [1]

3.4. Model

3.4.1 Architecture

For efficient training, incoming 720p (1280 x 720 pixel) frames were center-cropped, down-scaled to a network-friendly square, then gently compressed in height before reaching the model (450 x 360). Cropping, resizing, augmentation and FP16 casting all happened in the data loader, so only the current batch resided in memory. GPU-side augmentations were also ran in place while the CPU prefetched the next batch, leading to minimal memory footprints and negligible latency. Upon memory-optimizing the setup, a lightweight, dual-head Convolution Neural Network (CNN) was used to operate on 4-channel RGB-D frames (RGB + per-pixel depth deviation). This model consisted of three main blocks:

1. Convolution Backbone: A short stack of progressively down-sampling convolution blocks converted the 4-channel input into a low-resolution, high-depth feature map whose receptive field covered the entire image. Batch normalization, ReLU activations, and occasional max-pooling promoted fast convergence while suppressing memory use. Global average pooling then condensed this map to a single feature vector that retained the salient edge, texture and depth cues associated with road defects.

2. Shared fully-connected block: A small two-layer MLP refined the pooled descriptor, mixing coarse cues into richer abstractions (e.g. clustered micro-cracks or broad depressions). Heavy dropout was applied between layers to combat over-fitting on the limited dataset.

3. Task-specific heads: The network finally split into two lightweight branches: one for classifying defect type and the other for estimating severity. Each branch added a shallow dense layer and a focal or soft-max output, respectively. Sharing almost all parameters across tasks allowed the model to learn common visual signals once while still tailoring its final decision boundaries, enabling efficient single-GPU training.

ViTs, specifically swin-small and maxvit-small, were briefly tested; however, results were not separable from the original model.

3.4.2 Loss Function

For each sample i, the total loss combined classification and severity objectives:

$$L_i = \lambda_{cls} L_i^{Focal} + \lambda_{sev} L_i^{CSE}$$

λ_{cls} was the weight attached to the focal loss for defect classification and λ_{sev} was the weight attached to cross-entropy loss for severity.

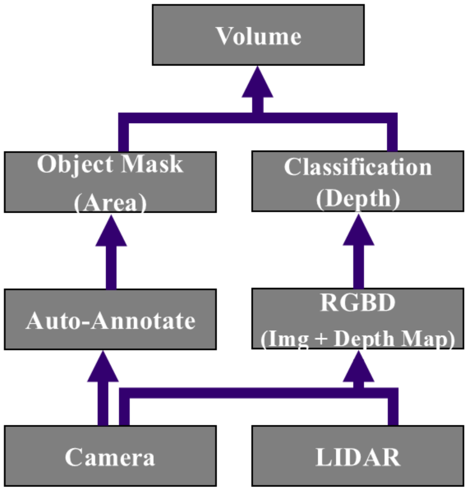


Figure 4. Full Experiment Pipeline from Input to Output

4. Experiments

4.1. Data Partitioning

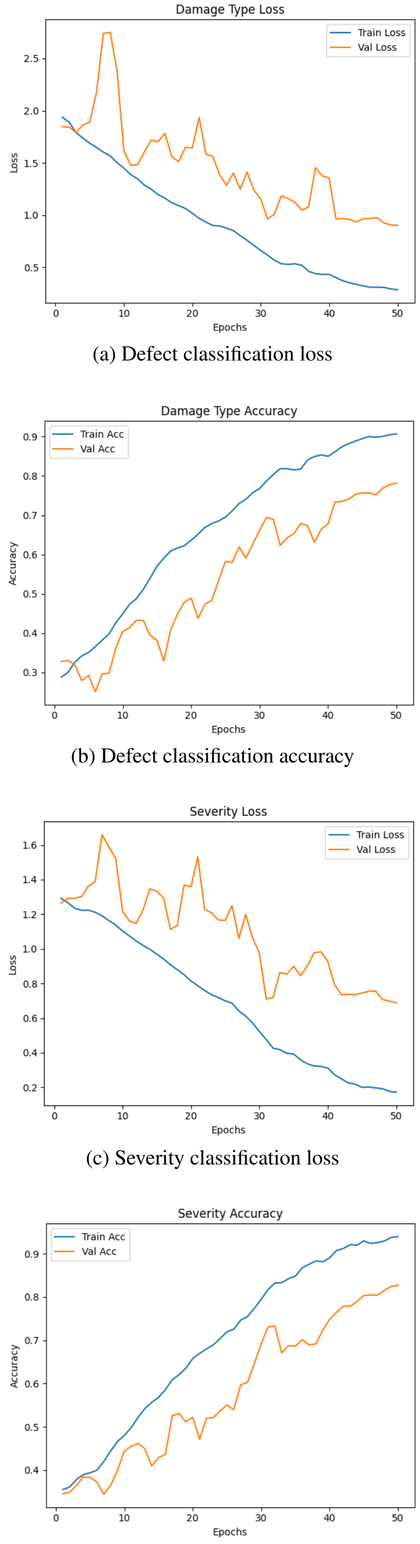
All annotated defect regions (≈2,000) were stratified at the *trip* level to avoid spatial correlation leakage and split **60% / 20% / 20%** into training, validation, and test subsets, respectively. Standard geometric augmentations (random horizontal flip, perspective warp, and cropping) were applied to increase dataset size and increase robustness to viewpoint changes.

Each sample comprised of (i) a raw image, (ii) a scalar severity ($s \in \{1, 2, 3\}$) and classification label, and (iii) a per-pixel depth map $D(u, v)$.

4.2. Evaluation

Each sample was assessed based on the classification of the defect and severity individually.

During training, the network's total validation loss was also monitored, so that the run terminated when this loss had not improved for 10 consecutive epochs. This was to ensure that the checkpoint giving the lowest validation loss was the one later evaluated.



(a) Defect classification loss

(b) Defect classification accuracy

(c) Severity classification loss

(d) Severity classification accuracy

Figure 5. Training and Validation Curves for Defect and Severity-Level Classifiers

Metric	Training	Validation	Test
Defect accuracy	0.9109	0.7785	0.7883
Severity accuracy	0.9247	0.8339	0.8241
Defect loss	0.2710	0.8814	–
Severity loss	0.2069	0.6782	–
Total loss	0.4780	1.5848	–

Table 1. Overall Accuracy and Loss Metrics

Jumps in validation loss for both defect and severity classification can in large part be attributed to the lack of diversity in our dataset even with the addition of synthetic data.

4.3. Other Metrics

Alongside accuracy and loss, precision, recall, and F1-score metrics were tracked for each type of defect and severity level. Relative to other defect types, the lower scores for longitudinal cracks and potholes were again attributable to the limitations of the curated dataset. With that said, raveling is a closely related defect to potholes, and it achieved high performance metrics across the board, so there may be some redeemable qualities.

Defect type	Precision	Recall	F1-score
good road	0.94	0.85	0.89
alligator crack	0.68	0.91	0.78
longitudinal crack	0.14	0.32	0.26
transverse crack	0.76	0.79	0.77
slippage crack	0.64	0.60	0.62
pothole	0.12	0.22	0.15
block crack	0.67	0.48	0.56
corner break	0.83	0.73	0.77
raveling	0.72	1.00	0.84

(a) Damage-type classification

Severity level	Precision	Recall	F1-score
none	0.93	0.83	0.88
low	0.83	0.86	0.84
medium	0.76	0.83	0.79
high	0.42	0.38	0.40

(b) Severity-level classification

Table 2. Metrics for Defect and Severity

4.4. Post-Processing

Outside of classification, geometric descriptors were computed externally for the select auto-annotated defects, namely the volume:

$$V = \Sigma_{(u,v)\in mask} D(u,v)$$

where (u,v) represents the coordinates of the pixel, mask is the SAM-derived binary mask, and D(u,v) the pixel's respective depth value.

5. Discussion

Based on the results, the proposed multi-modal CNN attains encouraging results (91.1% defect accuracy and 92.5% severity accuracy) on the training split, yet there is an understandably marked drop on the validation and test sets (Table 1), which likely reveals a generalization gap. Validation defect accuracy settles at 77.9% and severity accuracy at 83.4%, with a total validation loss almost 3.3 × higher than the training loss. Even though the data was stratified at the *trip* level to minimize spatial correlation leakage, the network still probably over-fits to the limited visual variability present in the curated dataset. The early-stopping criterion prevented runaway over-fitting, but Figure 5 shows that the validation curves plateau notably earlier than the training curves, indicating that additional regularization or data diversity is required.

5.1. Class-Wise Behavior

Defect Types. Performance across defect categories is highly non-uniform (Table 2a). Classes with abundant and visually distinctive examples (i.e. *good road*, *alligator crack*, and *raveling*) achieve F1-scores above 0.75. While, in contrast, *longitudinal cracks* and *potholes* yield F1-scores of only 0.26 and 0.15, respectively. Longitudinal cracks are slender and often occluded by shadows or lane markings, making them visually ambiguous; meanwhile, potholes suffer from a limited number of annotated instances in the current dataset. Interestingly, raveling (a defect visually and mechanically related to potholes) scores an F1 of 0.84, suggesting that the network has learned useful texture cues but fails once the damage crosses a geometric threshold that distinguishes raveling from an actual hole. A targeted re-labeling effort that balances minor and severe manifestations of potholes (and longitudinal cracks) should therefore result in significant improvements.

Severity Levels. Severity classification mirrors the aforementioned trend. The model is reliable for the *none*, *low*, and *medium* classes (F1 $\geq$ 0.79) but struggles with the *high* category (F1 = 0.40). This is probably be due to the simple fact high-severity examples are naturally rarer on public roads since they are usually high priority and resolved quickly, leading to class imbalance. Augmentations that explicitly distort depth maps or employ photorealistic generative models could help bridge this gap.

6. Further Work

Building on the limitations and opportunities highlighted above, several research directions are outlined below that could potentially enhance low-cost pavement-defect detection and assessment.

Extra Sensors and Robust. As previously stated, vision alone struggles with severity judgment. *(1) Vibrational cues* such as mounting an inexpensive tri-axial accelerometer, as in Singh *et al.* [11], would add a robust, lighting-independent signal that adds to the ability of the vision model to distinguish between varying levels of defects. *(2) LiDAR-driven shape estimation*—Faisal *et al.* [4] show that a curvature-based detector plus boundary delineation and voxelization quantifies pothole geometry with <10 % error even after thinning to 205 ppsm. Adding this to the pipeline could provide more precise width, depth, and volume estimates that go beyond the present labels and annotations.

Stronger Models. An apparent improvement with GPU headroom is scaling the current model to a larger backbone like a heavier vision transformer or a dual-path CNN–Transformer hybrid. Implementing stochastic depth and layer-wise LR decay can also be used to curb overfitting while keeping training feasible.

Richer data. Another clear improvement could be made by expanding the current dataset to different asphalt mixes, lighting regimes, and weather states (snow, standing water, oil staining). This will diversify the feature distribution and shrink the domain gap observed in the current results. Prioritizing the lesser frequent defects such as longitudinal cracks and severe potholes will redress the heavy-tail label distribution and raise the worst-case F1 scores. Furthermore, generative models can fabricate high-severity defects rarely seen in real driving footage.

With refinement, this low-cost road defect detection and assessment system will hopefully lend itself to be one of the superior choices for government pavement management practices. In the future, next steps would include prioritizing, mapping, and planning to resolve defects across the city or other municipality.

References

[1] Standard practice for roads and parking lots pavement condition index surveys, 2018. Approved 1 Jan 2018. 4

[2] Ethan Bancroft. We filled 25,000 potholes in 2023. our crews are working hard to address more potholes this winter, Jan. 2024. Accessed 21-May-2025. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020, LNCS 12346*, pages 213–229, 2020. 2

[4] Ali Faisal and Suliman Gargoum. Cost-effective lidar for pothole detection and quantification using a low-point-density approach. *Automation in Construction*, 172:106006, 2025. 7

[5] Rui Fan, Jianhao Jiao, Jie Pan, Huaiyang Huang, Shaojie Shen, and Ming Liu. Real-time dense stereo embedded in a uav for road inspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 535–543, Long Beach, CA, USA, June 2019. 3rd Int. Workshop on Computer Vision for UAVs (UAVision). 2

[6] Cong Gou, Bo Peng, Tianrui Li, and Ziping Gao. Pavement crack detection based on the improved faster-RCNN. In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 962–967, Dalian, China, 2019. 2

[7] Marco Leo, Antonino Furnari, Gerard G. Medioni, Mohan Trivedi, and Giovanni M. Farinella. Deep learning for assistive computer vision. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, volume 11134 of *Lecture Notes in Computer Science*, pages 3–14. Springer, Sept. 2018. 2

[8] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. *arXiv preprint arXiv:2201.12329*, 2022. 2

[9] Jianxi Ou, Jianqin Zhang, Haoyu Li, and Bin Duan. An improved YOLOv10-based lightweight multi-scale feature fusion model for road defect detection and its applications. SSRN preprint, Mar. 2025. Accessed 25-May-2025. 2

[10] Vung Pham, Lan Dong Thi Ngoc, and Duy-Linh Bui. Optimizing YOLO architectures for optimal road damage detection and classification: A comparative study from YOLOv7 to YOLOv10. In *Optimized Road Damage Detection Challenge (ORDDC'24) at IEEE BigData 2024*, Oct. 2024. arXiv:2410.08409. 4

[11] Premjeet Singh, Rashinda Wijethunga, Ayan Sadhu, and Jagath Samarabandu. Expert evaluation system for pothole defect detection. *Expert Systems with Applications*, 277:127280, 2025. 7

[12] Bin Wan, Xiaofei Zhou, Yaoqi Sun, Tingyu Wang, Shuai Wang, Haibing Yin, and Chenggang Yan. ADNet: Anti-noise dual-branch network for road defect detection. *Engineering Applications of Artificial Intelligence*, 132:107963, 2024. 2

[13] Xueqiu Wang, Huanbing Gao, Zemeng Jia, and Zijian Li. BL-YOLOv8: An improved road defect detection model based on YOLOv8. *Sensors*, 23(20):8361, 2023. 2

[14] Xuezhi Xiang, Zhiyuan Wang, and Yulong Qiao. An improved YOLOv5 crack detection method combined with transformer. *IEEE Sensors Journal*, 22(14):14328–14335, 2022. 2

[15] Sompote Youwai, Achitaphon Chaiyaphat, and Pawarotorn Chaipetch. YOLO9tr: A lightweight model for pavement damage detection utilizing a generalized efficient layer aggregation network and attention mechanism. *Journal of Real-Time Image Processing*, 21(5):163, 2024. 2

[16] Jongmin Yu, Jiaqi Jiang, Sebastiano Fichera, Paolo Paoletti, Lisa Layzell, Devansh Mehta, and Shan Luo. Road surface defect detection – from image-based to non-image-based: A survey. *arXiv preprint arXiv:2402.04297*, Feb. 2024. 3

[17] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[18] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16965–16974, 2024. 2

[19] Cuihua Zuo, Nengxin Huang, Cao Yuan, and Yaqin Li. Pavement-detr: A high-precision real-time detection transformer for pavement defect detection. *Sensors*, 25(8):2426. 2