End-to-End Pothole Detection and Semantic 3D City Model Enrichment Using LoRA-Adapted Foundation Models and Open Street-Level Imagery

Badred-Edine Ouzougarh[1], Jannik Matijevic[1], Huynh Duc An Son Nguyen[1], Lukas Arzoumanidis[1], Youness Dehbi[1]

[1] Computational Methods Lab, HafenCity University Hamburg, Germany –
(badred-edine.ouzougarh, jannik.matijevic, son.nguyen, lukas.arzoumanidis, youness.dehbi)@hcu-hamburg.de

Keywords: Pothole Detection, SAM 3, Low-Rank Adaptation, CityGML 3.0, Semantic 3D City Model.

Abstract

Regular road inspection is essential for maintaining pavement service life and ensuring traffic safety. Conventional approaches require specialized survey vehicles and sensor payloads, while damage assessment relies predominantly on supervised computer vision models trained on large annotated datasets. Recent advances in foundation models, notably the Segment Anything Model 3 (SAM 3), have demonstrated strong zero-shot and few-shot segmentation capabilities guided by natural language prompts, offering a promising alternative for domain-specific defect d etection. However, most existing studies address detection or segmentation in isolation without integrating results into standardized spatial data infrastructures suitable for infrastructure management. This paper presents a Location-Based Support Decision System (LBSDS) that bridges this gap: from openly available Mapillary street-level imagery acquisition, through pothole detection using a parameter-efficient Low-Rank Adaptation-adapted SAM 3 model, to monoplotting-based geolocation, and semantic enrichment of CityGML 3.0 road network models via direct 3DCityDB insertion. Comparative benchmarking against a YOLO-based instance segmentation model and fully fine-tuned SAM 3 shows that Low-Rank Adaptation detects 2.5 times more potholes than the supervised baseline at competitive precision, while training only 4.96% of model parameters; the supervised detector retains an advantage in pixel-level mask quality on the defects it does detect. The resulting pipeline produces schema-validated features within the CityGML Transportation Module, enabling interoperable road condition representation within semantic 3D city models. We demonstrate the practical feasibility of our pipeline through a real-world case study in Hamburg, Germany.

1. Introduction

Road health assessment is a resource-intensive process that demands significant human involvement and substantial financial investment. It plays a critical role in ensuring rider comfort and safety. In Germany alone, over 27,000 traffic accidents in 2024 were attributed to deteriorated road surface conditions (Statistisches Bundesamt, 2024). This growing concern has prompted regional authorities to take action; the Hamburg Senate, for instance, recently announced a dedicated initiative to address road surface defects across the city (Norddeutscher Rundfunk, 2024), underscoring the urgent need for efficient and scalable road inspection solutions. Traditional inspection methods rely heavily on manual visual assessment conducted by trained personnel, which is inherently subjective, labor-intensive, and difficult to scale. This has motivated an increasing interest in automated assessment systems capable of providing objective and reproducible evaluations of road surface conditions. Early automated approaches employed rule-based image processing techniques, such as edge detection and texture analysis, to identify surface defects (Ouma and Hahn, 2017; Ma et al., 2022). While effective under controlled conditions, these classical methods lack the robustness required for diverse real-world scenarios.

More recent approaches leverage advanced computer vision methods, including image classification, object detection, and semantic segmentation, to achieve more reliable defect identification. Among these, specialized models such as You Only Look Once (YOLO) (Redmon et al., 2016) have demonstrated strong performance on road defect tasks. However, such models require fine-tuning on large, accurately labeled datasets, which remains both costly and time-consuming. Furthermore, models trained on a single dataset frequently struggle to generalize across diverse environmental conditions, camera configurations, or annotation inconsistencies. To address these limitations, generalist foundation models capable of zero-shot segmentation have recently emerged, such as Segment Anything Model 3 (SAM 3) (Carion et al., 2025). These models demonstrate remarkable capabilities in instance segmentation without requiring task-specific training data, offering a promising direction for scalable and generalizable road surface assessment. Beyond routine maintenance, road surface inspection serves a broader range of applications, including traffic safety analysis, autonomous vehicle navigation, and long-term infrastructure monitoring. Despite this relevance, existing approaches predominantly focus on pothole detection in isolation, often relying on curated high-resolution datasets without addressing real-world data constraints. Moreover, these methods typically treat detection as the final objective, overlooking the potential for further integration of the results into spatial data models.

To address these gaps, this paper presents a Location-Based Support Decision System (LBSDS) that couples pothole detection with semantic 3D city modeling. The proposed approach leverages openly available street-level imagery from Mapillary[1], whose diverse environmental conditions, variable image quality, and heterogeneous camera configurations more closely reflect operational deployment than curated benchmark datasets. Detected potholes are geolocated from single posed images via monoplotting and integrated as semantically enriched `tran:Hole` features within the CityGML 3.0 (Kutzner et al., 2020) Transportation Module of a semantic 3D city model, bridging computer vision-based detection and standardized urban data infrastructures (Figure 1).

[1] `https://www.mapillary.com/`

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-267-2026

Figure 1. End-to-end pothole enrichment on a road segment in Hamburg, Germany. Left: enriched semantic 3D city model showing the inserted `tran:Hole` feature (yellow) on the road surface. Middle: corresponding Mapillary street-level image with the detected pothole outlined in red. Right: top-down view with a zoomed inset of the geolocated `tran:HoleSurface` geometry.

The core contribution is a parameter-efficient adaptation of SAM 3 using Low-Rank Adaptation (LoRA) (Hu et al., 2022; AI Research Group, KMUTT, 2025), training only 4.96% of the model parameters while outperforming fully fine-tuned SAM 3 across all metrics and achieving substantially higher detection coverage than a supervised YOLOv11m baseline, which in turn produces more precise masks on the subset of potholes it detects. A case study in Hamburg, Germany, demonstrates the complete pipeline from imagery acquisition to schema-validated CityGML output, confirming its applicability under real-world operational conditions and serving as location-based service for the detection and georeferenced mapping of potholes[2].

To that end the remainder of this paper is organized as follows. Section 2 reviews related work on pothole detection. Section 3 describes the four-stage pipeline covering data acquisition, detection and segmentation, monoplotting-based geolocation, and semantic 3D city model enrichment. Section 4 presents the comparative benchmarking, geolocation accuracy assessment, and enrichment results. Section 5 concludes with limitations and directions for future work.

2. Related work

The current state of the art reveals that previous research in automated road defect detection can be broadly categorized into location sensor-based approaches, LiDAR and 3D sensing methods, and image-based computer vision techniques.

Regarding location-based sensors, the earliest efforts in automated pothole detection relied on non-visual sensor modalities. Mednis et al. (2011) pioneered the use of smartphone accelerometers for real-time, crowd-sourced detection without camera input. Subsequent work by Wu et al. (2020) extended this concept by combining accelerometer and GPS data with classical machine learning classifiers, enabling low-cost distributed monitoring. In parallel, LiDAR and 3D sensing technologies have been explored for more precise geometric assessment. Tsai and Chatterjee (2018) combined 3D point cloud data with the watershed segmentation method to classify potholes by severity, while more recently, Li et al. (2025) integrated laser scanning with vision transformers for aggregate-level pavement deterioration analysis. Qiu et al. (2026) introduced SUD-ROAD, a high-resolution LiDAR benchmark dataset accompanied by a cross-dimensional semantic segmentation pipeline. Although sensor-based approaches offer practical advantages in terms of cost and deployment simplicity, they are inherently limited in their ability to characterize defect geometry, severity, or spatial extent, however, LiDAR-based methods provide high geometric accuracy, they require specialized and costly acquisition hardware, limiting their applicability for large-scale network-level monitoring.

Subsequently, image-based pothole detection has attracted research interest due to the widespread availability of cameras, and the introduction of deep learning has significantly advanced detection performance. Radopoulou and Brilakis (2017) presented an early automated system capable of detecting multiple pavement defect types simultaneously. Ye et al. (2021) developed a Convolutional Neural Network (CNN) model specifically for pothole localization in asphalt pavement images, while Ahmed (2021) proposed dilated convolutions to increase the receptive

[2] `https://github.com/badreddineouzougarh/Pothole_project`

field without losing spatial resolution. Chen et al. (2020) introduced location-aware CNNs that incorporate Global Positioning System (GPS) context to improve localization accuracy. Jakubec et al. (2023) provided a systematic benchmark of CNN-based models under real-world adverse conditions, offering practical guidance for deployment, and Vedula et al. (2025) extended the paradigm to Unmanned Aerial Vehicle (UAV) based imagery for evaluating pavement quality during construction.

Complementary techniques have been investigated to address data quality and representation challenges. Salaudeen and Çelebi (2022) combined GAN-based image enhancement with object detection to improve performance on low-quality images. Fan et al. (2021) integrated graph attention layers into semantic segmentation networks, contributing both a dedicated benchmark dataset and a novel architectural component. Lu et al. (2026) employed regional graph models to assess pavement condition data quality for digital twin applications. A substantial body of research demonstrates the dominance of YOLO family of detectors for real-time pothole detection. Ukhwah et al. (2019) established an early baseline using YOLO for asphalt pavement detection. In addition to Ping et al. (2020) bench-marked YOLO approaches against CNN and Support Vector Machine (SVM) baselines. Park et al. (2021) systematically compared YOLOv3, YOLOv4, and YOLOv5, while Sathvik et al. (2022) and Reddy and V (2022) demonstrated further improvements with YOLOv7. Moreover B and K.C (2022) adapted the anchor-free YOLOX algorithm, showing improved speed-accuracy tradeoffs. More recently, Kong et al. (2025) improved YOLOv8 for complex driving environments including night, rain, and occlusion conditions. Despite the strong performance of YOLO-based approaches, several persistent limitations remain for instance they require large, accurately labeled datasets for training, which is both costly and time-consuming to produce.

To overcome the data dependency and generalization limitations of task-specific models, recent research has turned to foundation models. Generalist vision models capable of zero-shot or few-shot segmentation, such as SAM, offer the ability to segment previously unseen object categories without task-specific training data. A parallel line of work explores vision-language models (VLMs) and large language models (LLMs) for road monitoring tasks. Zhang and Liu (2025) proposed a hybrid framework that fuses the detection capabilities of YOLO with the semantic understanding of Contrastive Language-Image Pretraining (CLIP) based vision-language models, specifically Qwen2-VL, to assess multiple categories of road damage within a unified pipeline. Xu et al. (2025) developed a robust prompt engineering strategy tailored for road evaluation, employing large language models to assess pavement surface conditions relative to the Pavement Surface Condition Index (PSCI), demonstrating that carefully designed prompting can yield competitive results without task-specific fine-tuning. Advancing further into multi-agent architectures, Yang et al. (2025) introduced a visual-language reasoning framework that leverages multiple specialized agents to address diverse highway scene understanding tasks, including pavement wetness assessment and surface condition classification. These emerging approaches demonstrate that foundation models can provide complementary capabilities to traditional detection pipelines, particularly in handling novel defect categories and providing richer semantic understanding of road conditions. Finally, street-level imagery platforms themselves increasingly provide model-derived semantic layers. Mapillary exposes per-image detections through its Graph API, including an `object--pothole` category produced by platform-internal segmentation models (Mapillary, n.d.). However, our evaluation of this layer over the Hamburg study region (Section 3.1) found it too sparse and unreliable to serve as a detection source or as annotation ground truth, motivating the dedicated detection stage pursued in this work.

3. Methodology

The proposed methodology comprises four sequential steps: (1) street-level imagery acquisition from the Mapillary platform, (2) pothole detection and instance segmentation using deep learning models selected through comparative benchmarking, (3) spatial filtering and monoplotting-based geolocation to establish real-world locations and elevations coordinates, and (4) enrichment of a semantic 3D city model with geolocated defect geometry following the CityGML 3.0 standard as shown in Figure 2. This work focuses exclusively on potholes as the target defect class.

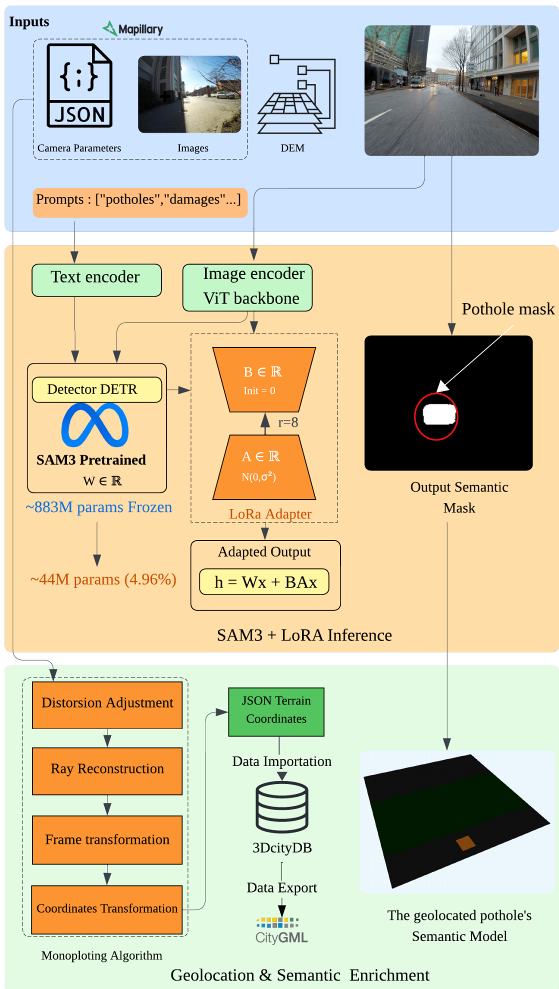

Figure 2. Overview of the methodology pipeline.

3.1 Data Acquisition

Street-level imagery is retrieved from the Mapillary platform via a custom interface supporting image-key, sequence-key, and bounding-box queries. For the Hamburg study area, imagery was retrieved by sequence keys covering road segments within the extent of the target CityGML transportation network, ensuring spatial consistency between imagery and the infrastructure to be enriched. Mapillary's platform-provided `object--pothole` detections were assessed as a potential data source: of $\sim$4,270 probed images in Hamburg, including a 500-image sequence with numerous visible potholes, only two carried pothole detections (five instances, two of them false positives on the recording equipment). As the class is furthermore absent from the spatially queryable map-features index, the layer was excluded, and all Mapillary training samples were manually annotated (Section 3.2.1).

3.2 Pothole Detection and Segmentation

3.2.1 Data Preparation The training dataset was compiled from three complementary sources to maximize sample diversity and model generalisation. Table 1 summarizes the data sources and their characteristics.

Source	Images	Size (px)	Reference
Dataset 1	2,235	1280 × 729	Passos et al. (2020)
Dataset 2	50	720 × 720	Rahman and Patel (2020)
Dataset 3	386	2048 × 1536	Mapillary[3]

Table 1. Overview of datasets used for pothole detection.

Dataset 1 (Passos et al., 2020) provides separately annotated binary masks for cracks and potholes; only the pothole class was retained for this work. Dataset 2 (Rahman and Patel, 2020) consists of pothole images from a controlled acquisition environment; a subset of 50 images was selected, as the full dataset does not reflect the heterogeneous conditions encountered in crowd sourced street-level imagery. The Mapillary subset comprises 386 images (216 positive, 179 negative) manually labeled within different cities in Germany, capturing diverse lighting, road surface, and camera viewpoint conditions representative of operational deployment.

Each source was filtered and harmonized into a single-class (pothole) annotation schema. All annotations include bounding boxes and instance segmentation polygons in Common Objects in Context (COCO) format, subsequently converted to YOLO format for the YOLOv11 experiments. Data augmentation, including horizontal flipping, brightness adjustment, and rotation, was applied exclusively to the Mapillary subset to increase its effective contribution while preserving the original distribution of the external datasets. The consolidated dataset was split into training, validation, and test sets as detailed in Table 2.

Split	Images	Annotations	Classes	Format
Train	1,320	1,923	1	COCO/YOLO
Validation	170	340	1	COCO/YOLO
Test	134	169	1	COCO/YOLO

Table 2. Dataset splits for training and evaluation. Negative images are included as background samples in all splits.

3.2.2 Model Selection and Benchmarking Three detection paradigms were evaluated to identify the most effective approach for pothole detection on heterogeneous street-level imagery.

1. **Supervised detection (YOLOv11m):** A YOLOv11 medium model was trained from pretrained weights on the pothole dataset for both detection and instance segmentation tasks, representing the conventional supervised approach requiring task-specific labeled data.

2. **Zero-shot segmentation (SAM 3 base):** We conducted a preliminary investigation into the sensitivity of SAM,3's zero-shot detection to prompt phrasing by testing five text prompts of varying complexity on representative test images. The concise spatially grounded prompt *pothole on asphalt road surface"* yielded the highest detection confidence (0.888), while generic terms such as *crack"* and overly verbose descriptions produced no detections. The single-word prompt *"pothole"* used throughout this study represents a conservative baseline. These observations suggest that moderate prompt specificity improves detection confidence, though a systematic evaluation across the full test set is needed to draw definitive conclusions and is left for future work.

3. **Parameter-efficient fine-tuning (SAM 3 + LoRA):** SAM 3 was adapted to the pothole domain using LoRA, which freezes the pretrained weights and injects small trainable low-rank matrices into selected layers, updating only 4.96% of the total parameters while preserving the foundation model's generalization capacity.

LoRA was applied to the vision backbone, the DETR-style encoder and decoder, and the mask prediction head ($r = 8$, $\alpha = 32$). Training used AdamW with cosine scheduling, bfloat16 mixed precision, and early stopping. Full hyper parameters are provided in Table 3. All models were evaluated on identical validation and test sets, with segmentation quality additionally assessed via the Dice coefficient and Intersection over Union (IoU).

Parameter	Value
Base model	SAM 3 (facebook/sam3)
Adaptation method	LoRA
LoRA rank	8
LoRA alpha	32
LoRA dropout	0.1
Trainable parameters	$\sim$44 M (4.96% of total)
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
LR scheduler	Cosine
Warm-up steps	200
Batch size	4
Gradient accumulation	8 (eff. batch = 32)
Epochs	30
Early stopping	Patience = 10
Mixed precision	bfloat16
LoRA target modules	Vision backbone (qkv, proj, fc1, fc2), DETR encoder/decoder, mask decoder

Table 3. SAM 3 LoRA fine-tuning configuration.

3.2.3 Inference Pipeline The model performs inference directly on each retrieved image. Detection results are represented as instance segmentation polygons delineating pothole boundaries and encoded as GeoJSON features for subsequent geolocation processing.

3.3 Pothole Geolocation via Monoplotting

Knowing where a pothole is located on the road surface is essential for routing maintenance crews and tracking infrastructure

[3] https://www.mapillary.com/

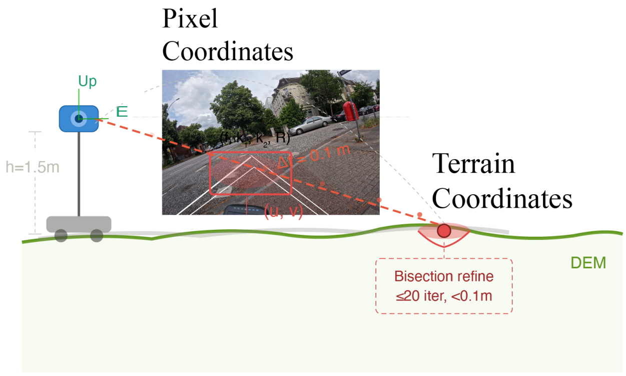

Figure 3. Monoplotting-based geolocation. A viewing ray is constructed from the detected pothole pixel coordinates through the camera model, corrected for radial lens distortion, and intersected with a DEM to estimate ground coordinates.

condition over time. Because crowdsourced street-level imagery rarely provides stereo overlap, we adopt monoplotting, a single-image photogrammetric technique that estimates the ground position of an image feature by intersecting a camera viewing ray with a Digital Elevation Model (DEM) (Mikhail et al., 2001; Jauregui et al., 2002; Sheng, 2005, 2008; Bozzini et al., 2012). As potholes are surface-level defects, the assumption that the target lies on the DEM surface is appropriate (Jauregui et al., 2002; Bozzini et al., 2012).

The pipeline supports two output modes. In *point mode*, a single representative ground coordinate is estimated from the mask centroid. In *polygon mode*, each vertex of the segmentation mask is projected individually, yielding a georeferenced polygon that captures the spatial extent of the defect. Both modes export results as GeoJSON for direct GIS ingestion.

3.3.1 Method Overview For each detected pothole the pipeline requires three inputs: (i) the segmentation polygon produced by the detection model, (ii) Mapillary image metadata comprising the SfM-refined camera position (`computed_geometry`) and exterior orientation (`computed_rotation`, a Rodrigues angle-axis vector), both derived from Mapillary's sequence-level OpenSfM reconstruction rather than the raw GNSS capture pose, together with interior calibration parameters (focal ratio, radial distortion coefficients), and (iii) a local DEM subset around the camera location.

Pixel-to-ray conversion. The target pixel, either the mask centroid or an individual contour vertex, is converted to normalized image coordinates using the camera focal length and principal point. Radial lens distortion is then removed via iterative inversion of the OpenSfM perspective camera model (OpenSfM, 2025), and the corrected coordinates are assembled into a unit viewing ray in the camera frame.

Coordinate transformation. The camera-frame ray is rotated into a local East-North-Up (ENU) frame using the inverse of the Rodrigues rotation and subsequently transformed into Earth-Centered, Earth-Fixed (ECEF) coordinates, following standard photogrammetric practice (Mikhail et al., 2001).

Ray-DEM intersection. Starting from the camera position (DEM elevation plus an assumed mounting height of 1.5 m), the ray is marched forward in small increments. Whenever a sample point falls below the DEM surface, a bisection search refines the intersection to sub-decimetre accuracy (Sheng, 2005, 2008). The resulting geodetic coordinates are reported together with the camera-to-target distance.

3.3.2 Accuracy Assessment The geolocation pipeline was evaluated using sewer manholes as ground-truth reference points. Manholes serve as suitable proxies for potholes: they are flat, surface-level features identifiable in both street-level imagery and digital orthophotos, whose planimetric position can be approximated by the DEM surface. Reference coordinates for 50 manholes were digitised in QGIS from leaf-off orthophotos (GSD 0.2 m) (Landesbetrieb Geoinformation und Vermessung Hamburg, 2025). Each manhole was annotated in corresponding Mapillary images and matched to its ground-truth identifier. Eight cases were excluded due to unfavourable viewing geometry or unreliable metadata, yielding $n = 42$ valid predictions. Quantitative results are reported in Section 4.

3.4 Semantic 3D City Model Enrichment

Geolocated potholes are integrated into a semantic 3D city model conforming to the CityGML 3.0 standard, transforming isolated detection results into standardised spatial objects within the transportation network. This enables downstream applications including infrastructure condition assessment, safety analysis, and route planning.

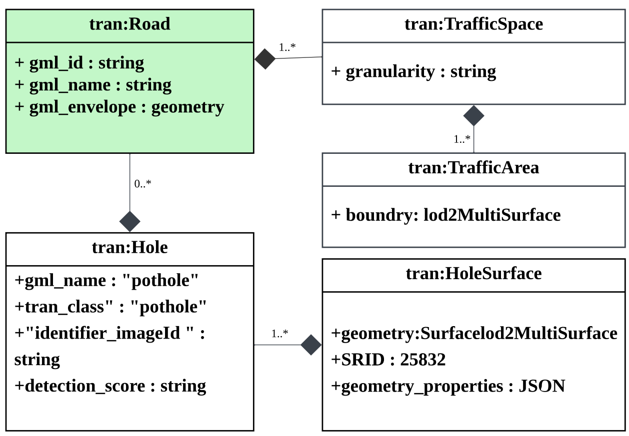

Figure 4. UML class diagram of the CityGML 3.0 Transportation Module showing the feature hierarchy used for road network enrichment.

3.4.1 CityGML Feature Hierarchy The CityGML 3.0 Transportation Module represents road infrastructure hierarchically. A `tran:Road` contains `tran:TrafficSpace` and `tran:AuxiliaryTrafficSpace` elements, each bounded by surface features carrying LoD2 geometry. The module defines `tran:Hole` as a dedicated feature type for surface discontinuities, linked to its parent road via a composition property and containing a `tran:HoleSurface` element that holds the defect geometry. Figure 4 illustrates this hierarchy and the enrichment scope. Each inserted feature carries a `gml:name` attribute set to *pothole*, a `tran:class` for semantic classification, and generic attributes storing the source image identifier and detection confidence score.

3.4.2 Database Insertion and Road Matching The 3D City Database (3DCityDB) v5.1.1 (Yao et al., 2025), deployed on PostgreSQL 18 with PostGIS 3.6, serves as the persistence layer. The database stores CityGML features across three principal

Metric	SAM 3 Base	SAM 3 LoRA	YOLOv11m
mAP@50	0.337	**0.753**	0.300
mAP@75	0.219	**0.302**	0.199
mAP@50:95	0.199	**0.351**	0.189
Precision	0.197	0.416	**0.507**
Recall	0.375	**0.677**	0.267
F1	0.258	**0.515**	0.350

Table 4. Detection performance on the pothole test set. Bold indicates the best value per metric.

tables: one for feature identity and spatial envelope, one for typed semantic properties and inter-feature relationships, and one for PostGIS geometries with associated metadata. GeoJSON polygons in WGS 84 are re-projected to EPSG:25832 and each pothole is spatially matched to the nearest `tran:Road` feature by computing the minimum distance between the pothole centroid and all road envelopes in the database.

3.4.3 Validation The enriched model is exported via `citydb-tool`[4] and validated against the CityGML 3.0 XML Schema using `citygml-tools`[5] validate, confirming that inserted features conform to the Transportation Module schema with correctly embedded geometry and properly assigned attributes.

4. Results and Discussion

This section presents results across three dimensions of the proposed workflow: model selection for pothole detection and segmentation, geolocation accuracy of detected defects, and semantic integration into the CityGML-based road network model. Each component is evaluated independently and discussed in the context of the end-to-end pipeline.

4.1 Model Selection and Comparative Evaluation

Three model configurations were bench-marked on the test set (134 images, 73 of which contain no potholes): the pretrained SAM 3 in zero-shot mode, SAM 3 adapted via (LoRA, rank 8, $\alpha = 32$), and YOLOv11m trained for instance segmentation. An additional SAM 3 variant with full parameter fine-tuning was included as a reference to illustrate the effect of unconstrained adaptation on a small dataset. Results are reported across three evaluation perspectives: detection performance (Table 4), segmentation mask quality (Table 5), and an instance-level breakdown of the SAM 3 model family (Table 6).

The SAM 3 LoRA model achieved the strongest overall detection performance with mAP@50 of 0.753 and F1 of 0.515, substantially outperforming both the zero-shot baseline (mAP@50 = 0.337) and YOLOv11m (mAP@50 = 0.300).

Parameter-efficient versus full adaptation. The LoRA-adapted model outperformed full fine-tuning across every metric despite training only 4.96% of parameters. Table 6 reveals the underlying cause: unconstrained optimization of all 883 million parameters on approximately 1,300 training images led to severe overfitting, as evidenced by the low precision (0.057) and F1 (0.100) of the fully fine-tuned variant. The LoRA rank-8 decomposition constrains adaptation to a low-dimensional subspace, preserving the pretrained representations while allowing targeted domain adaptation. This confirms that parameter-efficient fine-tuning serves as an effective regulariser when training data is limited.

Metric	SAM 3 Base	SAM 3 LoRA	YOLOv11m
Mean Dice	0.379	0.436	**0.525**
Mean IoU	0.326	0.375	**0.500**
Median Dice	0.190	0.507	**0.760**
Median IoU	0.106	0.340	**0.613**

Table 5. Pixel-level segmentation quality on the pothole test set. Metrics are computed over combined instance masks per image.

Foundation model versus supervised detector. YOLOv11m achieved the highest pixel-level median Dice (0.760) and IoU (0.613), indicating high mask quality when a pothole is successfully detected. However, this comes at the cost of recall: YOLOv11m detected only 26.7% of potholes, missing nearly three quarters of all defects. The gap between median and mean Dice (0.760 versus 0.525) reflects this quantitatively, as background images and missed detections contribute zero-score entries that lower the mean. SAM 3 LoRA detected 2.5 times more potholes (recall 0.677 versus 0.267) while maintaining reasonable precision of 0.416, an advantage attributable to the foundation model's broad pretrained visual representations. Which model is preferable therefore depends on the application priority: for network-level condition screening, where missing three quarters of all defects is unacceptable, detection coverage dominates and SAM 3 LoRA is the appropriate choice; where boundary fidelity of individual detections is paramount, the supervised detector's higher pixel-level mask quality may be preferred. As our target application is network-scale inventory, we select SAM 3 LoRA for the downstream pipeline.

Metric	Base (zero-shot)	LoRA (r=8)	Full FT
Instance Dice	0.499	**0.775**	0.511
Instance IoU	0.454	**0.692**	0.430
Precision	0.197	**0.416**	0.057
Recall	0.375	**0.677**	0.449
F1	0.258	**0.515**	0.100

Table 6. Instance-level comparison of SAM 3 adaptation strategies (IoU matching threshold = 0.5, confidence threshold = 0.3).

Zero-shot baseline. The base SAM 3 model without domain adaptation achieved mAP@50 of 0.337, notably surpassing YOLOv11m (0.300) without any task-specific training. This demonstrates that foundation models possess transferable knowledge about surface-level anomalies. However, the low precision (0.197) confirms that zero-shot performance alone is insufficient for operational deployment. The improvement from base to LoRA-adapted SAM 3, an increase of 30 percentage points in recall and 22 in precision, quantifies the value of targeted adaptation even with minimal parameter modification.

Segmentation quality. At the instance level, SAM 3 LoRA achieved a mean Dice of 0.775 and IoU of 0.692 for correctly matched detections, confirming that when the model identifies a pothole the segmentation boundary closely matches the ground truth.

Data source considerations. The use of Mapillary as the primary imagery source was motivated by its open accessib-

[4] https://github.com/3dcitydb/citydb-tool
[5] https://github.com/citygml4j/citygml-tools

ility, global coverage, and diverse contributor community. However, crowd-sourced imagery introduces challenges absent from curated datasets: variable resolution, inconsistent camera calibration, and uncontrolled acquisition conditions. The performance levels reported here, while lower than controlled benchmarks, reflect operational reality more faithfully, reinforcing the case for evaluating detection models on deployment-representative data. Based on these results, SAM 3 LoRA was selected as the detection backbone for the subsequent geolocation and enrichment pipeline.

4.2 Geolocation Accuracy

We evaluated the monoplotting pipeline on 42 sewer manholes with known ground-truth coordinates (Section 3.3.2). Monoplotting produced a more accurate location estimate than the raw camera position in 33 of the 42 cases (79%). The mean positional error decreased from 14.86 m (camera baseline) to 9.36 m, and the median error from 9.85 m to 5.04 m. To assess whether this improvement is statistically significant, we applied both a paired t-test and a Wilcoxon signed-rank test. Both tests confirm that monoplotting yields coordinates significantly closer to the ground truth than the camera position alone ($p < 0.001$ in both cases).

Distance-dependent performance. We examined whether the distance between the camera and the target influences prediction accuracy. A Pearson correlation test reveals a strong positive relationship ($r = 0.715$, $p < 0.001$): defects farther from the camera are geolocated less accurately, consistent with the geometric error amplification inherent in ray-based projection methods. Within 20 m of the camera ($n = 32$), the mean error is 6.06 m (median 4.19 m), and the distance–error correlation becomes non-significant ($p = 0.144$), indicating that, at close range, camera metadata quality rather than distance is the dominant source of error.

Practical implications. Without reaching sub-metre precision, monoplotting still reduces the median positional error from 9.85 m (camera position) to 5.04 m (4.19 m within 20 m), requiring only a single image and the metadata already embedded in platforms such as Mapillary. Nine predictions (21%) were less accurate than the camera baseline, and long-range targets (> 30 m, $n = 6$) yielded a mean error of 26.52 m; the latter is mitigated by restricting predictions to close range. The remaining close-range error is dominated by camera-pose uncertainty rather than ray–DEM intersection: although the pipeline uses Mapillary's SfM-refined poses (`computed_geometry`, `computed_rotation`) rather than raw GNSS positions, a single-view estimate absorbs the full per-frame pose error and, in particular, converts camera-height uncertainty into along-ray range error. As potholes are typically visible in several consecutive frames, multi-view triangulation of per-frame rays offers the most direct improvement: it exploits the more accurate relative SfM poses, eliminates the camera-height term, and leaves only the sequence-level georegistration bias, with the DEM retained as a vertical consistency check. Together with correction of the radial distortion reported in `camera_parameters`, sub-metre accuracy appears attainable.

4.3 Semantic 3D City Model Enrichment

In the study area, 20 street-level images were collected from Mapillary, yielding 18 confirmed pothole detections after visual inspection. During the semantic enrichment phase, 12 potholes were successfully matched to road segments in the CityGML transportation network (matching distance $\leq$ 20 m), while 6 remained unmatched due to geolocation inaccuracies that placed them outside road boundaries. All geolocated potholes were inserted into the 3DCityDB instance as `tran:Hole` and `tran:HoleSurface` features conforming to the CityGML 3.0 Transportation Module. The enriched model was exported and validated against the CityGML 3.0 XML Schema without errors (see Figure 1).

Integration challenges. The insertion process revealed several practical barriers to CityGML-based infrastructure enrichment. First, the 3DCityDB v5 relational schema distributes a single CityGML feature across multiple tables (`feature`, `property`, `geometry_data`), each requiring correctly typed entries with specific `datatype_id`, `namespace_id`, and `val_relation_type` values. A single incorrect value for example, a missing `namespace_id` or an absent `geometry_properties` JSON, results in either silent data loss during export or schema validation failure. This complexity presents a significant barrier for practitioners unfamiliar with the internal data model, as the mapping between CityGML XML elements and relational table entries is not self-evident.

Second, the CityGML 3.0 Transportation Module introduces a rich but deeply nested feature hierarchy: a road pothole requires instantiating a `tran:Hole` feature linked to its parent `tran:Road` via a typed composition property, a child `tran:HoleSurface` linked via a boundary property, and a geometry record with associated metadata, amounting to two feature records, one geometry record, and at least five property records per pothole.

5. Conclusion

This paper presented an end-to-end workflow for automated pothole detection and semantic 3D city model enrichment using openly available Mapillary street-level imagery. The pipeline comprises four stages: data acquisition, deep learning-based detection and segmentation, monoplotting-based geolocation, and CityGML 3.0 Transportation Module enrichment via direct 3DCityDB insertion. A comparative evaluation of four model configurations demonstrated that SAM 3 with LoRA (rank 8, = 32) achieved the best detection performance (mAP@50 = 75.3%, F1 = 0.515) while training only 4.96% of the model parameters, outperforming both full fine-tuning and YOLOv11m on heterogeneous street-level imagery.

A monoplotting geolocation module improved positional accuracy in 79% of evaluated cases, reducing the median error to 5.04 m. Detected potholes were successfully integrated as tran:Hole features within the CityGML road network, validated against the CityGML 3.0 schema. Several limitations were identified. The geolocation accuracy of approximately 5–10 m remains insufficient for lane-level maintenance planning and is sensitive to camera metadata quality. The current scope is restricted to pothole detection as a single defect class, whereas operational road condition assessment requires a comprehensive taxonomy of pavement distresses. Furthermore, the CityGML enrichment process demands detailed knowledge of the 3DCityDB relational schema, presenting a barrier to wider adoption.

Future work will address these limitations along three directions. First, we plan and currently working to extend the dataset corpus to a multi-class segmentation framework capable of identifying and classifying diverse pavement distresses including cracking, rutting, raveling, and surface deformation aligned with standardized road condition rating systems such as the Pavement Surface Condition Index (PSCI). This would enable the pipeline to produce not only defect geometries but also quantitative condition scores per road segment. Second, we aim to develop a large language model (LLM)-based interface that allows non-expert users to interact with the pipeline through natural language: a user would provide a street name or address together with the corresponding CityGML file, and the system would autonomously retrieve imagery, perform defect detection and classification, execute geolocation, and return a fully enriched CityGML model with embedded road condition assessment. Such an LLM-driven workflow would encapsulate the technical complexity of coordinate transformation, database schema mapping, and feature insertion behind a conversational interface, substantially lowering the barrier to operational use. Third, multi-view triangulation across overlapping Mapillary sequence frames will be implemented to exploit the higher accuracy of SfM relative poses, with the DEM retained as a vertical consistency check; together with full radial-distortion correction, this is expected to bring geolocation accuracy toward the sub-metre level required for lane-level maintenance planning.

Acknowledgment

This research was supported by the project "Next Generation City Networking" (Grant No. 19DZ24004) at the Hanseatic Wireless Innovation Competence Center (HAWICC), funded by the Federal Ministry of Transport through the German Center for Future Mobility (DZM). The authors further thank Mapillary and its contributor community for openly providing the crowd-sourced street-level imagery used in this work.

References

Ahmed, K. R., 2021. Smart Pothole Detection Using Deep Learning Based on Dilated Convolution. *Sensors*, 21(24).

AI Research Group, KMUTT, 2025. SAM3-LoRA: Low-rank adaptation for fine-tuning. GitHub repository, URL: `https://github.com/Sompote/SAM3_LoRA`. Accessed: 2026-04-07.

B, M. P., K.C, S., 2022. Enhanced Pothole Detection System Using YOLOX Algorithm. *Autonomous Intelligent Systems*, 2(1), 22.

Bozzini, C., Conedera, M., Krebs, P., 2012. A New Monoplotting Tool to Extract Georeferenced Vector Data and Orthorectified Raster Data from Oblique Non-Metric Photographs. *International Journal of Heritage in the Digital Era*, 1(3), 499–518.

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.-H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H. K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C., 2025. SAM 3: Segment anything with concepts. *International Conference on Learning Representations (ICLR)*, 138846–138923. Accessed: 2026-04-07.

Chen, H., Yao, M., Gu, Q., 2020. Pothole Detection Using Location-Aware Convolutional Neural Networks. *International Journal of Machine Learning and Cybernetics*, 11(4), 899–911.

Fan, R., Wang, H., Wang, Y., Liu, M., Pitas, I., 2021. Graph Attention Layer Evolves Semantic Segmentation for Road Pothole Detection: A Benchmark and Algorithms. *IEEE Transactions on Image Processing*, 30, 8144–8154.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*. Accessed: 2026-04-07.

Jakubec, M., Lieskovská, E., Bučko, B., Zábovská, K., 2023. Comparison of CNN-Based Models for Pothole Detection in Real-World Adverse Conditions: Overview and Evaluation. *Applied Sciences*, 13(9).

Jauregui, M., Vílchez, J., Galán, L., 2002. A Procedure for Map Updating Using Digital Mono-Plotting. *Computers & Geosciences*, 28(4), 513–523.

Kong, S., Meng, Q., Li, X., Wang, Z., Liu, X., Li, B., 2025. Road Perception for Autonomous Driving: Pothole Detection in Complex Environments Based on Improved YOLOv8. *IEEE Access*, 13, 80914–80929.

Kutzner, T., Chaturvedi, K., Kolbe, T. H., 2020. CityGML 3.0: New Functions Open Up New Applications. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(1), 43–61.

Landesbetrieb Geoinformation und Vermessung Hamburg, 2025. Luftbilder hamburg – DOP zeitreihe unbelaubt. Dataset series, URL: `https://metaver.de/trefferanzeige?docuuid=86bbb89e-df36-4436-b394-913998629ba5&q=`. Publication date: 2025-01-12. Accessed: 2026-04-07.

Li, Y., Liu, C., Weng, Z., Wu, D., Du, Y., 2025. Aggregate-Level 3D Analysis of Asphalt Pavement Deterioration Using Laser Scanning and Vision Transformer. *Automation in Construction*, 178, 106380.

Lu, L., Pan, Y., Sheil, B., Luo, P., Brilakis, I., 2026. Regional Graph-Based Pavement Condition Data Quality Assessment in Support of Trustworthy Highway Infrastructure Digital Twins. *Automation in Construction*, 181, 106603.

Ma, N., Fan, J., Wang, W., Wu, J., Jiang, Y., Xie, L., Fan, R., 2022. Computer Vision for Road Imaging and Pothole Detection: A State-of-the-Art Review of Systems and Algorithms. *Transportation Safety and Environment*, 4(4), tdac026.

Mapillary, n.d. Detections – API documentation. URL: `https://www.mapillary.com/developer/api-documentation/detections`. Accessed: 2026-07-12.

Mednis, A., Strazdins, G., Zviedris, R., Kanonirs, G., Selavo, L., 2011. Real time pothole detection using android smartphones with accelerometers. *2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, IEEE, Barcelona, Spain, 1–6.

Mikhail, E. M., Bethel, J. S., McGlone, J. C., 2001. *Introduction to Modern Photogrammetry*. Wiley, New York.

Norddeutscher Rundfunk, 2024. Hamburger senat kündigt schlagloch-offensive an. URL: `https://www.ndr.de/nachrichten/hamburg/hamburger-senat-kuendigt-schlagloch-offensive-an,schlagloecher-114.html`. Accessed: 2026-04-07.

OpenSfM, 2025. Geometric models. OpenSfM 0.5.2 documentation, URL: https://opensfm.org/docs/geometry.html. Accessed: 2026-03-06.

Ouma, Y. O., Hahn, M., 2017. Pothole Detection on Asphalt Pavements from 2D-Colour Pothole Images Using Fuzzy C-Means Clustering and Morphological Reconstruction. *Automation in Construction*, 83, 196–211.

Park, S.-S., Tran, V.-T., Lee, D.-E., 2021. Application of Various YOLO Models for Computer Vision-Based Real-Time Pothole Detection. *Applied Sciences*, 11(23).

Passos, B. T., Cassaniga, M. J., Fernandes, A. M. R., Medeiros, K. B., Comunello, E., 2020. Cracks and Potholes in Road Images. Mendeley Data, URL: https://doi.org/10.17632/t576ydh9v8.4. Accessed: 2026-04-07.

Ping, P., Yang, X., Gao, Z., 2020. A deep learning approach for street pothole detection. *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, Oxford, UK, 198–204.

Qiu, Z., Ghasemlou, A., Martínez-Sánchez, J., Arias, P., 2026. AI-Driven Road Inspection with SUD-ROAD: High-Resolution LiDAR Benchmark and a Novel Cross-Dimensional Semantic Segmentation Pipeline. *Computers and Electrical Engineering*, 132, 110993.

Radopoulou, S. C., Brilakis, I., 2017. Automated Detection of Multiple Pavement Defects. *Journal of Computing in Civil Engineering*, 31(2), 04016057.

Rahman, A., Patel, S., 2020. Annotated Potholes Image Dataset. Kaggle, URL: https://doi.org/10.34740/KAGGLE/DSV/973710. Accessed: 2026-04-07.

Reddy, E. S. T. K., V, R., 2022. Pothole detection using CNN and YOLO v7 algorithm. *2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECAA)*, IEEE, Coimbatore, India, 1255–1260.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, 779–788.

Salaudeen, H., Çelebi, E., 2022. Pothole Detection Using Image Enhancement GAN and Object Detection Network. *Electronics*, 11(12).

Sathvik, M., Saranya, G., Karpagaselvi, S., 2022. An intelligent convolutional neural network based potholes detection using YOLO-V7. *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, India, 813–819.

Sheng, Y., 2005. Theoretical Analysis of the Iterative Photogrammetric Single-Ray Backprojection Algorithm. *Photogrammetric Engineering & Remote Sensing*, 71(7), 863–871.

Sheng, Y., 2008. Modeling Algorithm-Induced Errors in Iterative Mono-Plotting Process. *Photogrammetric Engineering & Remote Sensing*, 74(12), 1529–1537.

Statistisches Bundesamt, 2024. General causes of accidents involving personal injury. URL: https://www.destatis.de/EN/Themes/Society-Environment/Traffic-Accidents/Tables/general-causes-of-accidents-involving-personal-injury.html. Accessed: 2026-04-07.

Tsai, Y.-C. J., Chatterjee, A., 2018. Pothole Detection and Classification Using 3D Technology and Watershed Method. *Journal of Computing in Civil Engineering*, 32(2), 04017078.

Ukhwah, E. N., Yuniarno, E. M., Suprapto, Y. K., 2019. Asphalt pavement pothole detection using deep learning method based on YOLO neural network. *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, IEEE, Surabaya, Indonesia, 35–40.

Vedula, N., Beheshti, M., Madasu, S., Ozer, H., 2025. Automated Framework for Evaluating Asphalt Pavement Construction Using UAV Imagery. *Automation in Construction*, 180, 106498.

Wu, C., Wang, Z., Hu, S., Lepine, J., Na, X., Ainalis, D., Stettler, M., 2020. An Automated Machine-Learning Approach for Road Pothole Detection Using Smartphone Sensor Data. *Sensors*, 20(19).

Xu, S., Zhao, K., Loney, J., Li, Z., Visentin, A., 2025. Image-Based Large Language Model Approach to Road Pavement Monitoring. *Computer-Aided Civil and Infrastructure Engineering*, 40(26), 4448–4464.

Yang, Y., Xu, N., Yang, J. J., 2025. Multi-Agent Visual-Language Reasoning for Comprehensive Highway Scene Understanding. *arXiv preprint arXiv:2508.17205*. Accessed: 2026-04-07.

Yao, Z., Nagel, C., Kendir, M., Willenborg, B., Kolbe, T. H., 2025. The New 3D City Database 5.0 – Advancing 3D City Data Management Based on CityGML 3.0. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 241–248.

Ye, W., Jiang, W., Tong, Z., Yuan, D., Xiao, J., 2021. Convolutional Neural Network for Pothole Detection in Asphalt Pavement. *Road Materials and Pavement Design*, 22(1), 42–58.

Zhang, Y., Liu, C., 2025. Vision-Enhanced Multi-Modal Learning Framework for Non-Destructive Pavement Damage Detection. *Automation in Construction*, 177, 106389.