Unified Scene Segmentation, Reconstruction, and Conflation for Road Closure Mapping

ABM Musa
musaabm@amazon.com
Amazon
USA

Rohit Keshari
rokeshar@amazon.com
Amazon
USA

Amr Abdelraouf
amrhra@amazon.com
Amazon
USA

Vaibhav Kango
kangov@amazon.com
Amazon
USA

Izzeddin Teeti
izzteeti@amazon.com
Amazon
USA

Mohamed Moustafa
mmoustm@amazon.com
Amazon, Last Mile
USA

Abstract

We present a unified scene understanding system for automated road closure mapping at scale. Our approach detects, segments, and geo-localizes road objects and surfaces from fleet camera image sequences, and conflates them with map segments—all without human intervention. For detecting road closure objects, we adopt a two-stage strategy using single-stage detection and segmentation models: a lightweight detection pass first identifies scenes likely containing road closures, enabling rapid filtering of the large-scale image stream; selected scenes then undergo detailed instance segmentation. To address training data scarcity, we introduce a scalable annotation pipeline combining SAM2-assisted labeling, third-party road closure data for targeted collection, and few-shot vision-language models for bootstrapping rare classes. For 3D scene reconstruction, we apply Structure-from-Motion (SfM) to image sequences to obtain geo-referenced point clouds and camera poses. We then precisely localize detected objects in world coordinates and generate road surface polygons via ray casting, providing surrounding road context for closure objects. An efficient object grouping method leverages feature correspondences to aggregate detections of the same physical object across multiple views into a single 3D instance. Map conflation associates the localized closure objects (like Barricades, Drums) to the affected road segment through selection of most relevant candidates in the vicinity. Road closure segmentation exceeds 96% precision on the most frequent closure objects, and our system localizes them to 1.4 m (p50) and 5 m (p90) accuracy. We associate about 80% of the closure locations with high confidence without a human in the loop.

CCS Concepts

• **Information systems** → **Geographic information systems**; **Location based services**; • **Computing methodologies** → **Reconstruction**; **Object identification**; **Scene understanding**.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
SIGSPATIAL '26, Riverside, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2950-8/2026/11
https://doi.org/10.1145/3841645.3843273

Keywords

Road closure mapping, Scene understanding, Instance segmentation, Structure-from-Motion, 3D reconstruction, Object geo-localization, Map conflation, Multi-object tracking, HD maps

ACM Reference Format:
ABM Musa, Rohit Keshari, Amr Abdelraouf, Vaibhav Kango, Izzeddin Teeti, and Mohamed Moustafa. 2026. Unified Scene Segmentation, Reconstruction, and Conflation for Road Closure Mapping. In *The 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26), November 03–06, 2026, Riverside, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3841645.3843273

1 Introduction

Accurate and up-to-date maps are essential for routing, navigation, and safe driving. The road network topology alone is insufficient; higher-level road attributes such as traffic signs, signals, and dynamic road conditions like closures are critical for efficient and safe delivery.

To address this, we developed a mapping platform that equips fleet vehicles with multiple cameras, a sub-meter accurate Global Navigation Satellite Systems (GNSS) receiver with Real-Time Kinematic (RTK) correction and Inertial Navigation System (INS) [15, 21, 23], and on-board edge compute. This platform captures geo-referenced imagery at strategic intervals (every 1–3 meters) across thousands of routes. We built mapping applications on this platform for detecting and localizing traffic signs and signals at scale, and for near real-time mapping of dynamic road closures using object segmentation and ground surface localization. Together, these systems can process billions of images and map millions of road objects. Our detection pipeline spans 11 traffic sign classes, 7 lane and median classes, and 15 road closure and surface classes, using single-stage detector architectures optimized for both edge and cloud deployment [16, 30, 31, 33, 34, 41].

In this paper, we present a unified system that advances the core techniques from these efforts. We detect and segment road attribute objects (e.g., traffic signs, signals), temporary road objects (e.g., barricades, cones), and road surfaces (e.g., drivable area, sidewalk) in a sequence of images, group them into unique objects, and reconstruct the scene in 3D for accurate localization. We use Structure-from-Motion (SfM) to obtain precise geo-referenced point clouds and camera poses, generate road surface polygons

through ray casting, and perform object grouping using feature correspondences across images. Our primary contributions are:

- A multi-task detection and segmentation framework spanning traffic signs, lane and median attributes, and road closure objects, with a novel SAM2-assisted annotation pipeline and multi-source data collection strategy for closure detection.
- A robust method for high-precision localization of road objects (traffic signs, signals, closure objects) using 3D scene reconstruction.
- A ground surface localization method using ray casting from camera poses to generate accurate polygons of ground surfaces (e.g., road, sidewalk) for enhanced scene and context understanding.
- An object tracking algorithm that groups multiple detections of the same object across images, achieving over 90% on MOTA, MOTP, and IDF1 metrics.
- An effective map conflation method to associate road closures to affected road segments.

2 Related Work and Background

Traffic sign detection. Traffic sign detection has been well studied using both traditional computer vision and deep learning approaches. Modern methods predominantly use CNN-based object detectors such as Faster R-CNN [26], SSD [17], and the YOLO family [33, 34]. The German Traffic Sign Detection Benchmark (GTSDB) [12] has been a standard evaluation dataset, though real-world fleet-sourced imagery presents additional challenges including motion blur, occlusion, and scale variation. Most prior work targets a limited set of traffic sign classes at small scale. In contrast, our system handles a wide variety of road objects including traffic signs, traffic lights, and road closure objects.

Road scene segmentation. Semantic and instance segmentation of road scenes has advanced significantly with architectures such as PSPNet [39], SegFormer [37], and YOLO-based segmentation variants [33, 34]. Lane detection methods range from traditional Hough transform approaches to recent deep learning methods including LATR [20] and DV-3DLane [19]. Multi-task models like YOLOP [36] jointly perform object detection, drivable area segmentation, and lane detection. Our approach differs in that we perform instance segmentation of lane markings and medians as distinct classes, enabling direct inference of lane count and median type at the road-segment level through polygon counting.

Work zone and road closure detection. Automated detection of temporary traffic control (TTC) devices and work zones has received limited attention compared to permanent road objects. Existing approaches typically rely on fixed infrastructure sensors or third-party (3P) data providers (e.g., INRIX [1], Waze [2]) that offer real-time closure reports but with limited coverage and location accuracy. Recent work on few-shot vision-language models [10] has shown promise for detecting rare on-road objects with minimal labeled data. Our work formulates road closure detection as a 15-class instance segmentation problem and introduces a practical pipeline combining 3P-guided data collection, VLM-based candidate retrieval, and SAM2-assisted annotation [25] to address the challenges of training data scarcity and annotation quality for this task.

3D reconstruction for mapping. Structure-from-Motion (SfM) [3, 27, 28] has been widely used for 3D scene reconstruction from image sequences. However, SfM produces an up-to-a-scale 3D model that is not aligned with the real world. To recover the actual scale and orientation, an external sensor such as GNSS [14] or an Inertial Measurement Unit (IMU) [29] is required. While SfM is primarily an offline process, Simultaneous Localization and Mapping (SLAM) [22, 24] operates in real time. However, SLAM suffers from drift and accumulates errors due to dead-reckoning. Our system builds on these techniques, using SfM in combination with sub-meter accurate RTK GNSS and INS [15, 21, 23] for high accuracy localization. Additionally, we use an enhanced Inverse Perspective Mapping (IPM) [7] approach based on ray casting for ground surface polygon generation.

3 System Overview

Figure 1 outlines our pipeline for detecting, reconstructing, and geo-referencing road objects and surfaces from vehicle imagery. The system ingests synchronized sequences of images and GNSS-INS location data, processed in the cloud to produce a unified, geo-referenced scene with precisely localized objects and surfaces.

As a first stage, a lightweight closure candidate detection model identifies image sequences likely to contain road closure scenes, filtering out non-closure image sequences before triggering the full pipeline.

Next, detection models segment road closure objects and ground surfaces (e.g., road, sidewalk), generating pixel-level masks for each frame. Concurrently, an SfM pipeline reconstructs a metric 3D point cloud through feature matching [5, 13, 18], triangulation [11], and bundle adjustment [32], which is then geo-aligned using GNSS-INS data to ensure metric accuracy. The semantic labels from detection guide the extraction of object positions directly from the segmented point cloud, while ray casting intersects camera rays with the ground plane to localize ground surfaces as precise polygons, creating a unified and geo-referenced scene with localized objects and ground surface classes.

The resulting unified geo-referenced scene is then fed into an association model that identifies the relationship of the localized detections and road network graph data to select the affected road segment. The final output is a complete map update combining semantic understanding, spatially accurate geometry, and road-segment-level attribution.

4 Road Closure Candidate Generation

The vast majority of image sequences captured by our fleet do not contain road closure scenes. Running the full detection, reconstruction, and localization pipeline on every sequence would incur unnecessary computational cost and network bandwidth for uploading imagery to the cloud. To address this, we employ a lightweight filter stage that identifies sequences likely to contain closures before triggering downstream processing.

We use a single-stage object detection model trained to detect six closure object classes: barricades, drums, cones, Temporary Traffic Condition (TTC) signs, arrow boards, and vertical bars. Given a

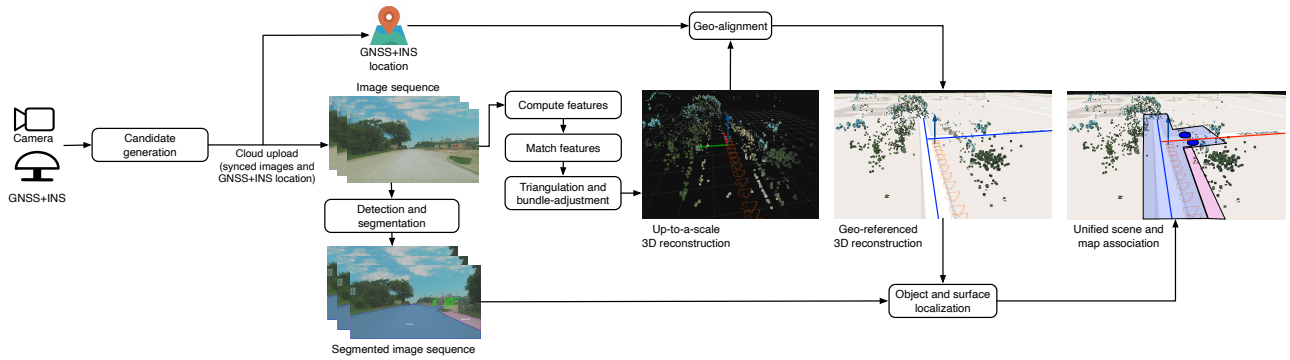


Figure 1: System overview.

sequence of images captured along a route, the model flags the sequence as a *closure candidate* if N consecutive frames contain at least one closure object detection. This temporal consistency requirement suppresses isolated false positives while preserving sensitivity to true closure scenes. We select a single-stage model for its low computational cost and suitability for deployment on both vehicle-edge hardware and in the cloud.

5 Detection and Segmentation

Our system employs a suite of detection and segmentation models to identify road objects and surfaces from our imagery. We address three distinct detection tasks: (1) road closure attribute detection for dynamic, temporary road conditions, (2) traffic sign detection, and (3) lane and median segmentation for road geometry. We present our road closure detection pipeline in this section; the remaining two tasks are detailed in Appendix D and Appendix E, respectively. Each task presents unique challenges in terms of object scale, class distribution, annotation quality, and data availability, which we address through task-specific model selection, training strategies, and data pipelines.

5.1 Road Closure Attribute Detection

Detecting temporary road closures in near real-time is a distinct challenge from permanent road attribute mapping. Closure objects—such as barricades, cones, drums, and arrow boards—are inherently temporary, visually diverse, and appear in complex, cluttered work-zone scenes. Our primary contribution here is the formulation and end-to-end solution of this problem rather than achieving state-of-the-art detection performance on any single benchmark.

Problem formulation. We define road closure detection as a scene-level segmentation task encompassing 15 classes: 11 closure object classes (Cone, Drum, Vertical Bar, Work Vehicle, Arrow Board, Barricade, TTC Sign, TTC, Message Board, Barrier, Fence, Work Equipment) and 4 road surface classes (road, sidewalk, earth, dirt-track). The inclusion of road surface classes is critical because determining whether a closure object is *on the road* (indicating an actual closure) versus on the sidewalk or off-road requires spatial context that pure object detection cannot provide.

Data collection. Our fleet captures imagery across large geographic areas, but identifying images that contain road closures from this massive corpus is impractical through manual inspection. We address this through a multi-source candidate selection strategy: (1) we leverage 3P providers of real-time road closure data, which provides closure event reports with approximate latitude-longitude coordinates—though this data is neither exhaustive nor highly accurate, it serves as a useful signal to identify our fleet images captured near reported closures; (2) we apply a few-shot vision-language model (VLM) [10] to retrieve candidate closure images from our images at selected stations; and (3) we incorporate images identified through divergent-route analysis by human annotators. This combined approach yields a diverse training set spanning different closure types, road geometries, lighting conditions, and geographic regions.

SAM-assisted annotation. A key challenge in training segmentation models for closure objects is annotation quality. Human annotators produce inconsistent mask boundaries, particularly for irregularly shaped objects such as barricades and fences, leading to noisy labels that degrade model generalization. To address this, we leverage SAM2 [25] for mask generation. Given human-annotated bounding boxes, we prompt SAM2 with each bounding box along with five points sampled within the box to generate instance masks. SAM2 produces more consistent and higher-quality boundaries than human annotators, particularly along object edges. However, SAM2 masks only capture the visible portion of objects, which can be problematic for occluded instances. To handle this, we additionally run SAM2's automatic mask generator on the full image to obtain class-agnostic masks for all visible regions. These are combined with the prompted masks, and a BEiT-based semantic segmentation model is used to fill in surface-level labels, producing the final training annotations. The complete label generation pipeline is detailed in Appendix B.

Road surface pseudo-labels. Since ground-truth mask annotations for road surfaces are not available, we generate pseudo-label masks for the 4 road surface classes using a pre-trained BEiT [4] semantic segmentation model trained on the ADE20K dataset [40]. We evaluated multiple architectures—BEiT, ConvNeXt, PSPNet, SegFormer [37], and SegNeXt—through qualitative comparison on our imagery and selected BEiT for its superior boundary quality on road and sidewalk regions. The BEiT-generated masks are used directly as training labels for the road surface classes in our segmentation model, enabling joint closure object and surface segmentation without requiring manual surface annotation.

Model and augmentation. We train a single-stage instance segmentation model on the 15-class dataset. Given the severe class imbalance—ranging from 12 training images for work-equipment to 6,673 for the most frequent class—we apply class-balanced augmentation with ratios from 20% to 1000% depending on class frequency. The augmentation pipeline includes (a) Speed blur, (b) rain, (c) fog, (d) copy-paste with the pasted region highlighted by a bounding box (placing closure objects onto negative images, restricted to the bottom half of the image to maintain spatial realism), and (e) darkness. Figure 2 shows examples of these.

6 3D Reconstruction and Localization

The detection and segmentation models described in Section 5 produce per-image 2D predictions—bounding boxes and pixel-level masks—but do not provide information about where these objects exist in the physical world. To localize detected objects and surfaces with metric accuracy, we now describe our 3D reconstruction pipeline that provides the geometric foundation for geo-referencing.

6.1 Scene 3D reconstruction

To localize various road objects and surfaces, we use Structure-from-Motion (SfM) [3, 27, 28] to reconstruct a scene in 3D from a sequence of images. We upload a sequence of 30 images at a 1-meter sampling interval from the vehicle's front camera with precisely time-synced GNSS-INS location and other metadata. This ensures good view overlap between adjacent images and similar lighting and environmental conditions, which helps to find good correspondences among the images. Next, we find correspondences across images by computing 2D corner features [5, 18] and feature matching [13]. Next, we run triangulation [11] to obtain 3D scene points and bundle adjustment (joint optimization of camera pose and scene 3D points) [32] to obtain an accurate 3D point cloud model of the scene. We also obtain precise 6-DoF (x, y, z, roll, pitch, yaw) camera poses for the images. The 3D model from SfM has an arbitrary coordinate frame (scale and orientation) and does not correspond to the real world, which is not suitable for localizing objects absolutely. We transform this arbitrary coordinate frame according to the GNSS-INS locations of all the frames, which aligns the cameras accurately to the real world. Since the whole 3D model is rigid, this transformation also scales and orients all 3D points representing the scene to the real world.

6.2 Object grouping

A single physical object in the scene — such as a stop sign or a barricade — is typically visible in multiple consecutive images captured as the vehicle drives past it. Each image produces an independent detection of that object, resulting in multiple detections corresponding to the same physical entity. Without grouping, these detections would be treated as separate objects, leading to duplicate localizations and inflated object counts. We therefore need to associate and group detections of the same object across images into a single unified representation before localization.

While there has been extensive work in multi-object tracking [8, 35, 38], most methods operate on video or continuous sequences of images with high frame rates. Since our system processes sparse images that may be captured at different times, we instead leverage the 3D reconstruction from SfM, which provides transitive feature correspondences across multiple images, enabling us to establish associations between the same physical object observed in different views.

We construct a bipartite graph $G = (V, E)$ where vertices represent detected objects in specific images, and edges connect detections that likely correspond to the same physical object. Two detections are connected if they satisfy two conditions: (1) they share the same semantic label (e.g., both are barricades), and (2) they share common 2D feature points within their detection regions that produced 3D points, ensuring both visual and geometric consistency. To maintain computational efficiency, we only compare detections within a temporal window of nearby images rather than all possible pairs. We then extract connected components from this graph, where each component represents an object tracklet — a set of detections corresponding to the same physical object across multiple images. This approach effectively handles occlusions and viewpoint changes while remaining computationally efficient. Additionally, this process also filters out false detections, as it is unlikely for a spurious detection to consistently appear in the same region across multiple images with valid feature correspondences.

Figure 3 shows a visual example of object grouping. Here, same-colored lines connect detections of the same physical object across images, with different colors representing distinct objects. Note that false detections can be filtered out during object association, as shown by the blue detection — it is unlikely for a false detection to appear in the same region across two images while also having feature correspondences between them.

6.3 Object localization

Once we obtain the geo-referenced 3D point cloud of a scene, we can localize an object from its detection bounding box or segmentation mask. Since the 3D model contains known 2D feature to 3D scene point correspondences, we identify the 3D points that fall within an object's bounding box or mask. These points form a tight cluster, and we take the centroid of this cluster as the location of the object. Figure 4 demonstrates this process, showing the segmentation mask of a barricade, its corresponding point cloud, and the inferred location. We use segmentation masks when available to avoid including background 3D points, which can introduce noise in localization.

6.4 Ground surface localization

Beyond localizing discrete road objects, understanding the layout of ground surfaces (e.g., roads, sidewalks) is essential for comprehensive scene understanding. For example, knowing whether a detected barricade is on the road, on the sidewalk, or far from the road entirely is critical for accurately inferring road closure status. To achieve this, we localize ground surfaces and generate geo-referenced polygons representing different surface types.

We leverage the precise 6-DoF camera poses and ground-plane information obtained during 3D reconstruction to estimate geo-referenced coordinates of ground points using an enhanced Inverse Perspective Mapping (IPM) approach. Our method projects rays from the camera center through image points and intersects them with the ground plane to compute geo-coordinates via ray casting.



Figure 2: Examples of data augmentation techniques applied to training images. (a) Speed blur, (b) rain, (c) fog, (d) copy-paste with the pasted region highlighted by a bounding box (inset shows a cropped and zoomed view of the pasted area), and (e) darkness.



Figure 3: Object association across adjacent images using feature correspondences. Same-colored lines connect detections of the same physical object across images, with different colors representing distinct objects.

This formulation exhibits superior numerical stability compared to traditional homography-based IPM — which transforms perspective views to bird's-eye views using four-point correspondences [7] — particularly for camera orientations nearly parallel to the ground plane, a common scenario for vehicle-mounted cameras.

For ground-surface mapping, we use segmentation masks from our detection model to identify different surface types (e.g., roads, sidewalks) in each image. By applying ray casting to all points within these segmentation masks, we generate geo-referenced polygons representing ground surfaces, as illustrated in Figure 4. We then merge these per-image polygons to construct comprehensive surface maps for the entire scene. To suppress artifacts, we enforce a 100-meter distance threshold that removes spurious intersections caused by near-horizontal rays, which would otherwise project to unreliable distant locations.

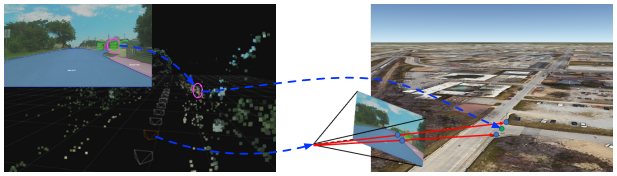


Figure 4: Localization of road objects and ground surfaces.

7 Unified scene representation

The object detection and segmentation, 3D reconstruction, object grouping, object localization, and ground surface localization together produce a unified scene representation that jointly represents the precise location of each unique road object and the spatial extent of ground surfaces in a common geo-referenced coordinate frame.

This unified representation provides rich contextual information essential for downstream applications. For instance, the spatial relationship between closure objects (e.g., barricades, cones) and road/sidewalk polygons can determine whether a road is fully closed, partially closed, or whether the closure object is not on the road at all. Figure 5 demonstrates this through three scenarios: a full closure where the object lies entirely within the road polygon, a partial closure where the object partially overlaps the road, and a case where the closure object is located off the road on an adjacent surface. We show additional examples in Appendix A.

This unified scene representation is a generic framework applicable to a range of mapping tasks beyond any single use case, such as mapping road attributes, detecting road closures and hazards, and enriching map data with fine-grained surface information.

7.1 Map Association

The final step in the pipeline is associating each localized closure object with the affected road segment in the digital map. This step bridges the gap between object location estimation and actionable map updates by determining which specific road segment is impacted in the scene. Once closure objects in the scene are geo-referenced, we select a candidate set of nearby road segments from the map data. For each candidate road segment, the model evaluates a rich set of geometry and road network graph features, such as road class of the segment, the angle between vehicle heading and road segment, proximity of vehicle and object location to the road segment and network topology features such as one-way/two-way or degree of intersection. We feed these features into an XGBoost model [9], which is trained in a supervised manner with labeled examples of object to segment associations. High-confidence associations are automatically applied to the map. This design helps us associate closure locations without human validation at a large scale.

8 Results

In this section, we present quantitative evaluations of the detection, localization, and object grouping components of our system. We evaluate on real-world data spanning diverse road environments with varying lighting, occlusion, and viewpoint conditions.

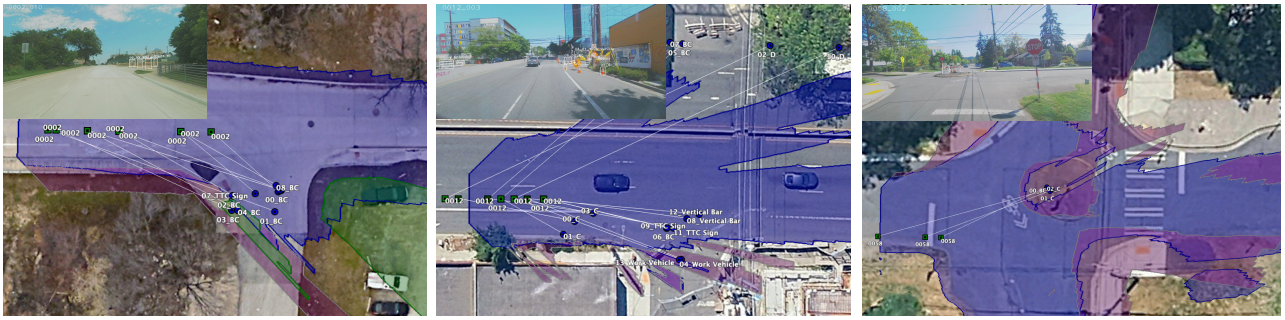


Figure 5: Unified localization of road objects and ground surfaces. Left: full closure, middle: partial closure, right: closure object not on road (green square: image location, blue circle: object location, blue polygon: road, purple polygon: sidewalk, green polygon: ground).

8.1 Road Closure Candidate Generation Accuracy

We evaluate the road closure scene detection module on a dataset of 1,863 image sequences containing 521 closure scenes (*42,612* total images), curated using the same data collection procedure described in Section 5.1. We set $N = 3$ consecutive frames as the detection threshold.

Table 1 presents performance across detection confidence thresholds. At a confidence threshold of 0.4, the module achieves 95.6% precision and 94.6% recall on closure candidate identification. Of the 1,342 non-closure sequences in the dataset, only 23 are incorrectly flagged as candidates at this operating point, filtering out **98.3%** of non-closure sequences from the remainder of the pipeline. Higher confidence thresholds improve precision at the cost of recall, as shown in the table.

Table 1: Road closure scene detection performance across confidence thresholds.

Confidence Threshold	Precision	Recall
0.4	0.956	0.946
0.5	0.972	0.928
0.6	0.983	0.886
0.7	0.987	0.800
0.8	0.989	0.559
0.9	1.000	0.023

8.2 Detection Accuracy

Lane and median segmentation. Our segmentation model achieves 95% precision and 85% recall for lane and median segmentation across 7 classes. Data augmentation (brightness, darkness, rain, blur, snow, pothole, fog) improves recall by approximately 5% on underrepresented conditions.

Road closure attribute detection. Table 2 presents the test set performance of our 15-class segmentation model (8,889 images). For the 11 closure object classes (11,184 instances with human-annotated ground truth), the model achieves an overall box mAP@0.5 of 0.842 and mask mAP@0.5 of 0.776. The 4 road surface classes are trained using pseudo-labels generated by a pre-trained BEiT model [4] (see Section 5.1); their test set metrics reflect agreement with these pseudo-labels rather than human ground truth. Notably, the dirt-track class was absent from the test set, resulting in zero performance for that class. Additional experimental results are provided in Appendix C.

8.3 Localization Accuracy

Our system achieves high absolute localization accuracy through two key factors: precise 3D scene reconstruction and accurate georeferencing using RTK GNSS and INS. As shown in Table 3, our localization of road closure objects achieves a median (p50) error of 1.38 meters and a 90th percentile (p90) error of 5.04 meters, demonstrating robust performance across diverse real-world conditions.

8.4 Grouping Accuracy

We evaluate our object grouping method using standard multi-object tracking metrics [6], as our system must consistently track and group multiple objects across image sequences. Table 4 presents our performance across three key metrics: Multi-Object Tracking Accuracy (MOTA) for overall tracking correctness, Multi-Object Tracking Precision (MOTP) for bounding box localization consistency, and IDF1 for identity preservation across frames. Our method achieves over 90% on all three metrics, demonstrating robust and consistent object grouping performance.

8.5 Map Association Accuracy

To evaluate whether the affected road segments are correctly identified by the model, we noted the precision and recall metrics as presented in Table 5. We observe that the most frequent incorrect associations occur in challenging scenarios such as complex intersections or scenes with multiple affected road segments. To enable large-scale conflation, we filter and apply high-confidence associations for about 80% of the upstream cases in the map.

9 Conclusion

We presented a unified system for detecting, segmenting, and localizing road objects and surfaces from vehicle imagery at scale. Our system addresses traffic sign detection (11 classes), lane and median segmentation (7 classes), and road closure detection (15 classes)

Table 2: Road closure detection: 15-class model on held-out test set. Road surface metrics reflect agreement with BEiT pseudo-labels, not human ground truth. Dirt-track was absent from the test set.

Class	Inst.	Box P	Box R	Box mAP@.5	Mask P	Mask R	Mask mAP@.5
Barricade	2327	0.850	0.885	0.889	0.692	0.729	0.657
Drum	2476	0.883	0.851	0.912	0.872	0.853	0.903
Cone	1704	0.808	0.764	0.805	0.787	0.758	0.788
TTC Sign	1885	0.798	0.825	0.829	0.791	0.830	0.827
Arrow Board	462	0.817	0.825	0.845	0.810	0.827	0.851
Vertical Bar	1515	0.773	0.723	0.768	0.731	0.703	0.723
Barrier	364	0.714	0.703	0.674	0.648	0.648	0.600
Work Vehicle	331	0.796	0.870	0.850	0.734	0.807	0.783
TTC Msg Board	65	0.694	0.803	0.739	0.695	0.805	0.759
Fence	44	0.786	0.795	0.818	0.522	0.545	0.465
Work Equipment	11	0.683	0.909	0.835	0.536	0.727	0.588
Road†	13223	0.953	0.817	0.915	0.950	0.822	0.916
Sidewalk†	13594	0.838	0.492	0.661	0.837	0.505	0.668
Earth†	10698	0.814	0.414	0.585	0.801	0.421	0.582
Dirt Track†	34	0	0	0	0	0	0

† Road surface classes trained with BEiT [4] pseudo-labels (no human GT).

Table 3: Localization accuracy for road closure objects.

Percentile	**p50**	**p90**	**p95**
Localization error (meters)	1.38	5.04	9.61

Table 4: Object grouping accuracy.

Metric	**MOTA**	**MOTP**	**IDF1**
Value	94.75%	93.24%	90.68%

Table 5: Map Association Accuracy.

Metric	**P**	**R**	**F1**
Value	96.1%	80.1%	87.3%

using single-stage architectures, with a practical data pipeline combining third-party-guided collection, VLM-assisted retrieval, SAM2-based annotation, and class-balanced augmentation for road closure detection. We reconstruct scenes in 3D using SfM along with object grouping and precise localization of objects and surfaces, achieving localization accuracy of 1.4m (p50) and 5m (p90), and over 90% on MOTA, MOTP, and IDF1 tracking metrics. The resulting unified scene representation provides a generic framework for a range of mapping tasks including road attribute mapping, road closure and hazard detection, and fine-grained surface enrichment. Future work includes extending detection to additional road attributes, improving closure detection recall for rare object classes, and replacing the current rule-based nearest-segment association with multi-modal learning-based methods that can handle ambiguous cases at complex intersections.

References

[1] 2026. INRIX: Intelligent Mobility Solutions and Traffic Analytics. https://inrix.com/.

[2] 2026. Waze: Real-Time Traffic and Navigation App. https://www.waze.com/.

[3] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. 2009. Building Rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*. 72–79. doi:10.1109/ICCV.2009.5459148

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I* (Graz, Austria) *(ECCV'06)*. Springer-Verlag, Berlin, Heidelberg, 404–417. doi:10.1007/11744023_32

[6] Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance: the CLEAR MOT metrics. In *EURASIP Journal on Image and Video Processing*, Vol. 2008. 1–10.

[7] M. Bertozzi and A. Broggi. 1998. GOLD: a parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Transactions on Image Processing* 7, 1 (1998), 62–81. doi:10.1109/83.650851

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. 3464–3468. doi:10.1109/ICIP.2016.7533003

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[10] Minsoo Choi, Ravi Garg, and Moustafa Mohamed. 2025. Few-shot Vision-language Prompt Tuning of VLMs for On-road Object Detection. In *Proceedings of the 33rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

[11] Richard Hartley and Andrew Zisserman. 2003. *Multiple View Geometry in Computer Vision* (second ed.). Cambridge University Press, Cambridge, UK.

[12] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. 2013. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*. 1–8.

[13] Shiguo Huang, Guobing Sun, and Minglun Li. 2021. FAST and FLANN for feature matching based on SURF. In *2021 33rd Chinese Control and Decision Conference (CCDC)*. 1584–1589. doi:10.1109/CCDC52312.2021.9601366

[14] M.R. James, S. Robson, S. d'Oleire Oltmanns, and U. Niethammer. 2017. Optimising UAV topographic surveys processed with structure-from-motion: Ground control quality, quantity and bundle adjustment. *Geomorphology* 280 (2017), 51 – 66. doi:10.1016/j.geomorph.2016.11.021

[15] Elliott Kaplan and Christopher Hegarty. 2017. *Understanding GPS/GNSS: Principles and Applications, Third Edition.*

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot multibox detector. In *European Conference on Computer Vision*. 21–37.

[18] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110. doi:10.1023/B:VISI.0000029664.99615.94

[19] Yueru Luo et al. 2024. DV-3DLane: End-to-end Multi-modal 3D Lane Detection with Dual-view Representation. In *International Conference on Learning Representations*.

[20] Yueru Luo, Chaoda Zheng, Jianhua Yan, Bingbing Chen, Wei Xu, Zhaoxiang Zhang, and Shuicheng Zhao. 2023. LATR: 3D Lane Detection from Monocular Images with Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[21] Hongyang Ma, Qile Zhao, Sandra Verhagen, Dimitrios Psychas, and Xianglin Liu. 2020. Assessing the Performance of Multi-GNSS PPP-RTK in the Local Area. *Remote Sensing* 12, 20 (2020). doi:10.3390/rs12203343

[22] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics* 31, 5 (2015), 1147–1163.

[23] ABM Musa, Chris Baker, Emre Eftelioglu, and Amber Roy Chowdhury. 2022. High-accuracy GNSS localization with low-cost. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems* (Seattle, Washington) *(SIGSPATIAL '22)*. Association for Computing Machinery, New York, NY, USA, Article 95, 4 pages. doi:10.1145/3557915.3561031

[24] K. Pirker, M. Ruther, and H. Bischof. 2011. CD SLAM - continuous localization and mapping in a dynamic world. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3990–3997. doi:10.1109/IROS.2011.6094588

[25] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolber, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024).

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, Vol. 28.

[27] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. doi:10.1109/CVPR.2016.445

[28] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.* 25, 3 (July 2006), 835–846. doi:10.1145/1141911.1141964

[29] S. Song, M. Chandraker, and C. C. Guest. 2016. High Accuracy Monocular SFM and Scale Correction for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 4 (2016), 730–743. doi:10.1109/TPAMI.2015.2469274

[30] Mingxing Tan, Ruoming Pang, and Quoc V. Le. 2020. EfficientDet: Scalable and Efficient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[32] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. 2000. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, Bill Triggs, Andrew Zisserman, and Richard Szeliski (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 298–372.

[33] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7464–7475.

[34] Chien-Yao Wang and Hong-Yuan Mark Liao. 2024. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv preprint arXiv:2402.13616* (2024).

[35] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing, China). IEEE Press, 3645–3649. doi:10.1109/ICIP.2017.8296962

[36] Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. 2022. YOLOP: You Only Look Once for Panoptic Driving Perception. In *Machine Intelligence Research*, Vol. 19. 550–562.

[37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, Vol. 34.

[38] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 1–21. doi:10.1007/978-3-031-20047-2_1

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2881–2890.

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 633–641.

[41] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).

A Qualitative examples

Figure 6 demonstrates our system's performance for road closure objects, showing qualitative examples of accurately localized road closure objects and their associated ground surfaces. Here, the top-left example shows a full road closure scenario, the top-right example shows a non-closure scenario with two barricades on the road but still drivable by vehicles, the bottom-left example shows non-closure with objects on the sidewalk, and the bottom-right example shows non-closure with objects far from the actual road. This demonstrates the importance of accurate scene reconstruction of road closure objects and ground surfaces to understand road closure scenarios accurately.

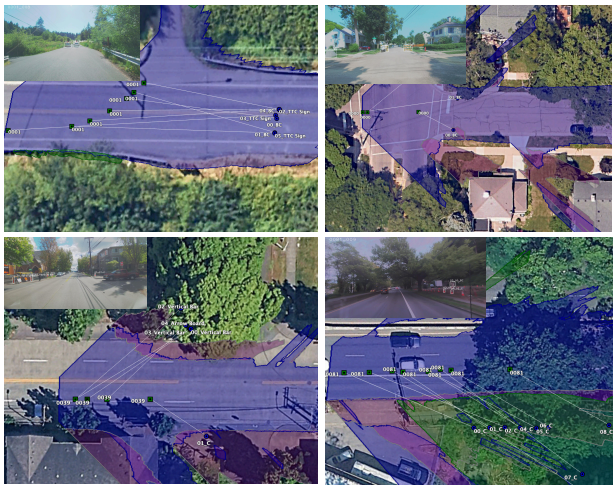


Figure 6: Road closure object and ground surface localization examples (green square: image location, blue circle: object location; blue polygon: road, purple polygon: sidewalk, green polygon: ground).

Figure 7 presents representative examples for traffic signs, illustrating our system's ability to accurately map traffic signs and signals in real-world scenarios. Here, the top-left example shows localization of a stop sign, the top-right example shows localization of multiple traffic signs and lights in a complex intersection, the bottom-left example shows localization of a speed-limit sign, and the bottom-right example shows another complex intersection with localization of multiple traffic signs and traffic lights.

B Ground-truth Mask generation from foundational and Specialized model

The foundational label generation process operates in the following steps as shown in Figure 8. First, for each object with a human-annotated bounding box, we obtain a mask prediction from SAM using the box as a prompt (box-to-mask). Simultaneously, we run SAM's automatic mask generator to produce class-agnostic mask proposals across the entire image. For each bounding box, we select the largest mask from either the box-to-mask or auto-generated predictions, constrained to remain within the bounding box boundaries. Next, we compute the coverage of each finalized mask relative



Figure 7: Traffic signs and signals localization examples (green square: image location, blue circle: object location, green circle: ground truth location).

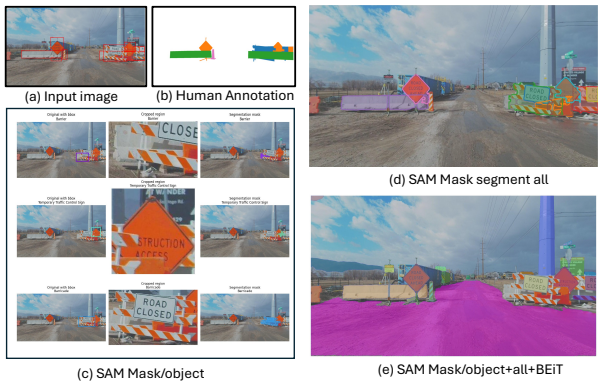


Figure 8: Overview of the foundational label generation pipeline. (a) Input image with detected objects, (b) human annotation with class-level masks, (c) SAM-based per-object segmentation showing original crops, cropped regions, and resulting segmentation masks, (d) SAM segment-all output producing class-agnostic masks over the full image, and (e) combined segmentation result by merging SAM per-object masks, SAM segment-all masks, and BEiT [4] surface predictions.

to its bounding box and calculate the permissible area for the remaining overlapping objects. Since SAM predicts masks only for the visible portion of an object, when multiple objects overlap, the finalized mask of one object is subtracted from the human-annotated mask of the overlapping object to avoid double-counting occluded regions. This ensures that each pixel is assigned to the object that is visually dominant in that region. Finally, a BEiT [4] semantic segmentation model is applied to generate surface-level labels (e.g.,

road, sidewalk, sky), which are merged with the object-level masks to produce the complete pixel-wise annotation for each image.

C Additional Segmentation Results

Table 6: Road closure detection: Additional 15-class model results

Class	Precision	Recall	F1	TP	FP	FN
Cone	0.963	0.424	0.588	2721	106	3701
Drum	0.986	0.543	0.700	6616	97	5576
Vertical Bar	0.978	0.331	0.495	1075	24	2169
Barricade	0.823	0.648	0.725	3763	807	2044
Work Vehicle	0.883	0.311	0.460	639	85	1413
Arrow Board	0.838	0.353	0.496	134	26	246
TTC Sign	0.630	0.397	0.487	415	244	631
TTC Msg Board	0.424	0.608	0.500	104	141	67
Barrier	0.120	0.008	0.016	3	22	351
Fence	0	0	0	0	0	304
Work Equipment	0	0	0	0	0	13
Overall (Macro)	**0.604**	**0.329**	**0.406**	15470	1552	16515

Table 6 presents the additional evaluation of the 15-class segmentation model. The model achieves high precision on the most frequent closure classes: 96.3% for cones, 98.6% for drums, 97.8% for vertical bars, and 82.3% for barricades. Recall is lower due to the high diversity of real-world closure scenes, with barricade achieving the highest recall at 64.8%. Notably, the Fence and Work Equipment classes show zero detection performance. Fences are inherently transparent and lack distinctive texture, making it challenging for the model to learn discriminative features against varied backgrounds. For Work Equipment, the evaluation set contains only 13 instances, and the appearance of work equipment in this additional evaluation differs significantly from the training distribution. These classes would benefit from targeted data augmentation and additional training samples to improve generalization.

D Traffic Sign Detection

Traffic sign detection is the foundational detection task in our system, enabling the mapping of permanent road attributes such as stop signs, speed limits, turn restrictions, and one-way indicators. We detect 11 categories of traffic signs chosen based on their impact on route planning and navigation: No-Left, No-Right, One-Way, Do-Not-Enter, Stop, Max-Speed-Limit, Left-Only-Turn, Traffic-Light, No-U-Turn, Height-Restriction, and Weight-Restriction. We employ a single-stage object detection model that achieves up to 0.98 precision and 0.97 F1-score on high-frequency classes such as stop signs and speed limits. The model processes high-resolution images, enabling detection of smaller and more distant signs. Performance varies across the 11 sign categories, with rarer classes such as height and weight restrictions showing lower recall due to limited training data.

E Lane and Median Segmentation

Accurate lane count and median classification are essential road attributes for route planning and navigation. We formulate this as an instance segmentation task with 7 classes: drivable-area, drivable-area-reverse, median-physical, median-double-yellow-line, lane-line-white, lane-line-double-white, and lane-line-yellow.

We train our instance segmentation models using Amazon SageMaker Ground Truth. The segmentation masks are post-processed to infer lane count (by counting drivable-area polygons in the forward direction) and median presence/type (physical vs. double-yellow-line) at the road-segment level. Our best model achieves 95% precision and 85% recall on the evaluation dataset.

To improve robustness across diverse conditions, we apply the same augmentation pipeline used for traffic signs: brightness, darkness, rain, blur, snow, pothole, and fog transformations. The augmentation ratios are calibrated per class based on training set size, with rare classes receiving up to 10× augmentation.