The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

Object localization and change detection in urban environments using dashcam videos

Aziza Zhanabatyrova[1], Yu Xiao[1], Ahmad Elalailyi[2,3], Fabio Remondino[2]

[1] Dept. of Information and Communications Engineering, Aalto University, Espoo, Finland – <zhanabatyrova.aziza, yu.xiao>@aalto.fi
[2] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy – <aelalailyi, remondino>@fbk.eu
[3] Dept. of Architecture, Built environment and Construction engineering (ABC), Politecnico di Milano, Italy

Keywords: 3D mapping, semantic segmentation, object detection, monocular, change detection, MMT

Abstract
The rapid evolution of urban landscapes necessitates efficient mapping solutions. Traditional high-accuracy semantic maps generated using expensive sensors and mobile mapping vehicles provide precise spatial data, but face challenges related to cost and scalability. Crowdsourced dashcam videos present a practical alternative for acquiring urban visual data, leveraging widely available and low-cost camera technology. Recent advances in photogrammetry and computer vision - such as Structure from Motion (SfM), Simultaneous Localization and Mapping (SLAM), semantic segmentation and object detection - enable the extraction of both 3D and semantic information from monocular images. Building upon previous research, we propose a pipeline for constructing and updating semantic 3D maps using crowdsourced low-cost dashcam footages, with a particular emphasis on automatic change detection. Our approach compares metadata related to urban landmarks (e.g., traffic signs) to identify modifications in cityscapes. We evaluate the robustness of the proposed approach with various sequences captured under challenging conditions, including rain, darkness and fog, comparing the performance of SfM-based and SLAM-based 3D reconstruction methods. Results show the effectiveness of the proposed low-cost methodology in localizing urban objects and changes, although accuracy needs to be improved with better georeferencing procedures.

1. Introduction

Rapid urban development and evolving cityscapes demand continuous change detection, efficient mapping and update strategies to effectively monitor dynamic environments and keep map databases up-to-date (Tran et al., 2018; Lee and Hsu, 2021; Stilla and Xu, 2023; Kharroubi et al., 2025). Typical change detection solutions include remote sensing technologies, such as optical satellites, LiDAR and optical cameras on aircraft or drones, as well as close-range or terrestrial methods that utilize videos, images or point clouds. With respect to images, 3D point clouds provide a valuable alternative solution offering different modalities and enabling highly detailed 3D geometric analyses with attribute enrichment. While high-accuracy, expensive Mobile Mapping Technology (MMT) vehicles (Elhashash et al., 2022) equipped with multiple cameras, LiDAR and GNSS/IMU sensors offer centimeter-level precision and redundant data for semantic map generation, their widespread deployment is constrained by cost and scalability. In contrast, crowdsourced dashcam videos provide a cost-effective and widely available alternative for urban visual data collection (Zhanabatyrova, 2025). Moreover, recent advances in photogrammetry and computer vision, including Structure from Motion (SfM) (Schoenberger and Frahm, 2016; Pan et al., 2024), Simultaneous Localization and Mapping (SLAM) (Kazeroui et al., 2022), semantic segmentation (Mo et al., 2022) and object detection (Kaur and Singh, 2023), enable the extraction of 3D and semantic information from monocular images, also for change detection purposes (Lin et al., 2022). Due to the high computational complexity - and sometimes challenges - of image-based point cloud generation, effective 3D reconstructions from crowdsourced visual data is still an open task. Consequently, it is essential to develop efficient methods for detecting and localizing environmental changes, enabling the selective re-mapping of only those regions where modifications have occurred.

1.1 Paper's Aim

The goal of this work is to further investigate how low-cost dashcam videos acquired from vehicles moving in urban environments could support the creation and updating of semantic 3D maps. Our previous work (Zhanabatyrova et al., 2023) introduced a pipeline that reconstructs SfM-based semantic 3D maps from dashcam videos and automatically detects urban changes based on metadata comparison (Zhanabatyrova et al., 2023). In this study, we further evaluate the robustness of our approach using additional visual data sequences captured under challenging conditions (such as rain, darkness, fog). Additionally, we compare a SfM-based methodology with a SLAM pipeline (Campos et al., 2021) assessing how point cloud quality impacts the system performance. In summary, the paper's aim is:

- to assess the robustness of the SfM-based pipeline across diverse visual conditions (fog vs. clear summer);
- to compare the performance of SfM-based and SLAM-based pipelines on the same data sequence, while investigating resolution-induced sensitivities in object detection, as well as trade-offs in accuracy and computational cost within the pipeline;
- to demonstrate how crowdsourced dashcam data can complement or update - with certain accuracy limitations - pre-existing maps, reducing the need for costly MMT sensors.

The proposed methodology distinguishes itself by leveraging recent advances in SfM- and SLAM-based 3D reconstruction while operating under the constraints of low-cost, crowdsourced data acquired with dashcams. Unlike methods that rely on dense LiDAR point clouds or pre-calibrated multi-camera systems, our approach, although not achieving very high accuracy results, is tailored for scenarios where only single-view RGB data is available and where crowdsourced videos are the only source of data. Given these constraints, the system prioritizes efficiency and scalability over centimeter-level accuracy.

2. Related works

Recent literature has focused on several aspects related to image-based semantic 3D map generation and change detection in urban environments. In monocular video processing, several self-supervised approaches have demonstrated impressive depth estimation and object

detection from dashcam videos (Godard et al., 2019; Lee et al., 2019; Shabestari et al., 2023). Similarly, several methods have been developed to enhance object detection and semantic mapping from car videos, utilizing deep learning-based segmentation techniques to generate semantically enriched maps and point clouds (McCormac et al., 2017; Li et al., 2019; Qin et al., 2020; Roddick and Cipolla, 2020; Cheng et al., 2022; Zhang et al., 2023).

Change detection in urban environments has also attracted attention, with works addressing the challenge based on image or point cloud data, from ground or aerial perspective (Shirowzhan et al., 2019; Zhang et al., 2021; de Gélis et al., 2021; de Gélis et al., 2023; Xiao et al., 2023). More recently, change detection is accomplished exploiting visual language models (Lin et al., 2025), 3D Gaussian Splatting (Lu et al., 2025), vision transformers (Alpherts et al., 2025) or a SAM-based zero-shot framework (Kim and Kim, 2025).

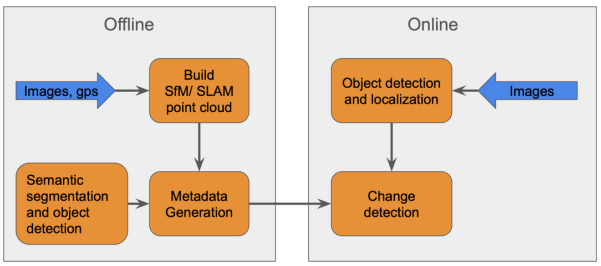

Figure 1. Overview of the pipeline, which comprises both offline and online stages.

3. Methodology

Figure 1 provides a high-level overview of the system architecture, outlining a dual-stage process. The first stage involves offline map creation using either SfM or SLAM, combined with semantic segmentation and object detection to extract object information and generate metadata. The second stage focuses on online change detection, leveraging newly acquired dashcam data to identify updates and refine the existing map. The individual components of the pipeline will be validated for robustness and adaptability, especially when handling the newly introduced challenging scenarios:

- *Offline Map Generation*: dashcam videos are processed using COLMAP (Schonberger et al, 2016) to retrieve camera poses and build an initial sparse 3D point cloud of the environment. In addition to a SfM-based approach, this study also includes a comparative evaluation using a SLAM-based approach (Campos et al., 2021) on selected sequences to assess performance variations. To enhance semantic understanding of the surveyed scene, semantic segmentation networks - based on architectures similar to Seamseg (Porzi et al., 2019) and object detectors - derived from the SSDResNet framework (Lu et al., 2019) are used to associate 3D points of interest with class labels (e.g., traffic sign types). SfM and SLAM identify images that observe each 3D point and record its corresponding pixel coordinates, known as image keypoints. These keypoints are then classified using the aforementioned image processing techniques, ultimately generating a semantic map that represents the initial state of the urban environment. The semantic map contains metadata detailing object types, such as traffic signs, locations, and characteristics.
- *Online Change Detection and Map Update*: new visual data sequences, collected successively to the creation of a semantic map, are fed into the pipeline. For each frame, camera poses are estimated and pixelwise 3D object localization is performed using a modified deep learning network built upon BTS (Lee et al., 2019). The camera pose is estimated by registering the image to an existing point cloud utilizing custom matching in COLMAP. In this custom matching, the incoming image is matched with the nearest image in the point cloud in terms of Euclidean distance. The pixelwise 3D localization method assigns precise 3D coordinates to every pixel in the image. To classify the pixels of interest, an object detection algorithm is applied, ensuring correct identification of traffic signs. By combining the estimated 3D coordinates with camera pose information, objects can be precisely localized within the map. The extracted objects are then matched against the offline map exploiting georeferencing information. Differences trigger a candidate change detection mechanism. A thresholding algorithm is applied to minimize false positives and detected changes can eventually trigger an automatic map update.

In this work, visual sequences (including rainy, dark, and "foggy" conditions – Figure 2) are considered to test the robustness of the proposed pipeline. Through quantitative metrics and qualitative visual comparisons, we investigate both the performance under standard conditions and the edge cases imposed by harsh weather. The original object localization method presented in Zhanabatyrova et al. (2023) is tested with SfM as well as with a SLAM pipeline.

4. Experiments

4.1 Datasets

To evaluate the proposed methodology, two image sequences are recorded (Table 1, Figure 2 and Figure 3).

Figure 2: Example frames from (a) Sequence 1 and (b) Sequence 2

	Sequence 1	*Sequence 2*
Date	30.10.2020	30.06.2022
Camera	Garmin	iPhone 12
Weather	rain, dark, fog	sun, summer
FPS	10	10
Resolution [px]	1920 x 1080	3840×2160
Camera set-up	(a)	(b)
Frames	1202	3136
Length [m]	617	970
Speed	17 km/h	10 km/h

Table 1: Specifications of the sequences used in the evaluation.

The first sequence consists of 3600 frames along a 617 m road within a university campus. The sequence was collected using a (calibrated) Garmin dashcam, similar to the single view setup employed in the previous studies (Zhanabatyrova et al., 2023;

2024). The dashcam was positioned in one corner of the car's front window, slightly convergent to the side to observe both the middle and the opposite side of the road, allowing for a broader field of view (Figure 4a). The frames were recorded at a resolution of 1920x1080 pixels and a frame rate of 30 FPS. This sequence includes data captured in urban environments under challenging conditions, such as rainy, dark and foggy weather (Figure 2a). During the data collection, the camera set-up imaged a portion of the car in every frame, negatively affecting the 3D mapping algorithm's performance. Therefore, during the undistortion process (i.e. mapping of original distorted pixel to undistorted locations) part of the scene is eliminated (Figure 5), improving the 3D mapping process. To maintain consistency with previous experiments and optimize performance, we adjusted the frame rate to 10 fps to enhance the reliability of the results.

Figure 3: Trajectories of data collection in the urban area: green - sequence 1, pink - sequence 2.

Figure 4. Camera set up: a) sequence 1; b) sequence 2.

Figure 5: Example of sequence 1 images before (above) and after (bottom) undistortion.

The second sequence was recorded at a slow driving pace of 10 km/h with the dashcam in the center and facing forward (Figure 4b), using an iPhone 12 Pro Max. The initial frame rate of 30 fps is later reduced to 10 fps, resulting in a dataset of 3136 images. Captured at a resolution of 3840x2160 pixels, the sequence benefits from clear summer weather conditions, providing excellent visibility and optimal lighting for feature extraction (Figure 2b). This sequence was collected in the same region, covering a total distance of 970 m, allowing for direct comparisons with previously recorded sequences under different environmental conditions. The two data sequences, recorded nearly two years apart in different seasons, exhibit substantial variations in traffic signs, weather conditions, resolution and lighting, making them suitable for our change detection experiment.

4.2 Evaluation on foggy and low-visibility data

Despite the accumulation of noticeable drift, the COLMAP sequential matching demonstrated superior performance on sequence 1 compared to vocabulary-tree matching (Figure 6a,c), largely due to challenging visibility conditions, hence low feature recognition in the images. The sequential SfM reconstruction generated 104,132 3D points and 1202 poses out of 1202 undistorted images, demonstrating the effectiveness of the pre-processing steps in improving the overall quality of the reconstruction. Due to a lack of onboard GNSS observations, data geo-registration is performed using manually selected points on a map (Helmert transformation), resulting in an alignment error of 8.18 meters (mean) and 8.14 meters (median), reflecting the need of good geographic points for proper georeferencing purposes. Despite these challenges, the pipeline remains functional, demonstrating its ability to process challenging visual data. Camera poses and sparse point cloud of the entire sequence 1 are shown in Figure 6d.

Due to low visibility conditions, the successive object detection algorithm successfully identified 11 out of 21 traffic signs using a threshold 0.15 (Figure 7a), but struggled to detect the remaining ones, even when located in close proximity. Setting the confidence threshold at 0.15 helps reduce false positives and improve detection accuracy. Lowering the threshold to 0.1 increases sensitivity, allowing the model to detect more objects, including the third traffic sign in the background (Figure 7b). However, this heightened sensitivity also introduces misclassification. The algorithm exhibited misclassifications, often confusing visually similar signs, such as "No Parking" and "No Stopping," or other traffic signs with similar shapes and appearances (Figure 7c,d). Furthermore, traffic signs located at greater distances proved challenges in the detection, as diminished resolution and atmospheric interference reduced recognition accuracy. The overall object detection process operates efficiently, with an average processing time of 0.229 seconds per image, without significantly sacrificing performance.

The successive semantic segmentation component correctly detects and outlines the borders of traffic signs (Figure 8a).

Despite the extreme challenges posed by the dark and foggy dataset - particularly for object detection and SfM - the proposed framework performed well, detecting a significant number of traffic signs. This demonstrates the robustness of the approach, even under suboptimal conditions and highlights its potential for real-world applications. Enhancing image quality and refining detection algorithms could further improve results, making this pipeline an effective tool for urban mapping in complex environments. Georeferencing could be improved with an onboard GNSS.

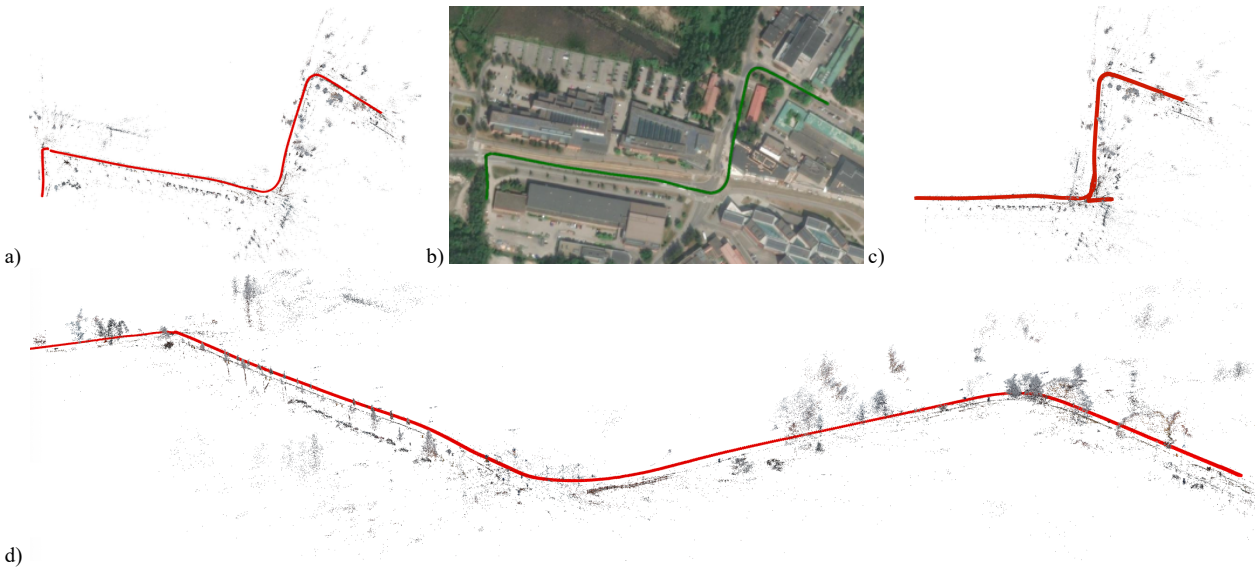

Figure 6: 3D reconstruction using sequential matching (a). Recovered trajectory mapped on a 2D map (b). 3D reconstruction results using vocabulary-tree matching (c). Overview of the entire recovered trajectory and sparse point cloud for sequence 1 (d).

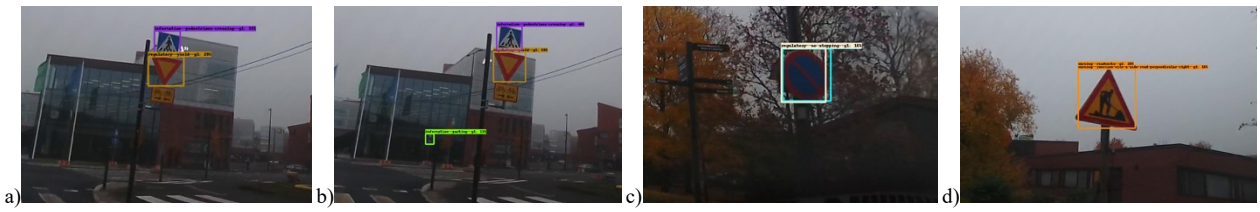

Figure 7. Examples of object detection results with threshold=0.15 (a) and threshold=0.1 (b); misclassification examples (c,d).

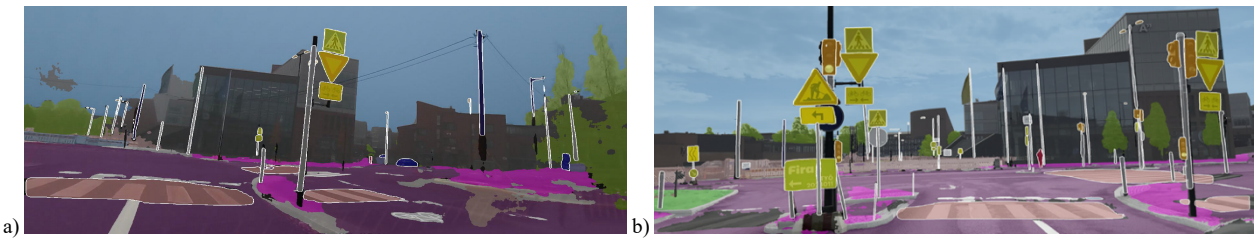

Figure 8. Semantic segmentation results on an image of sequence 1 (a) and 2 (b).

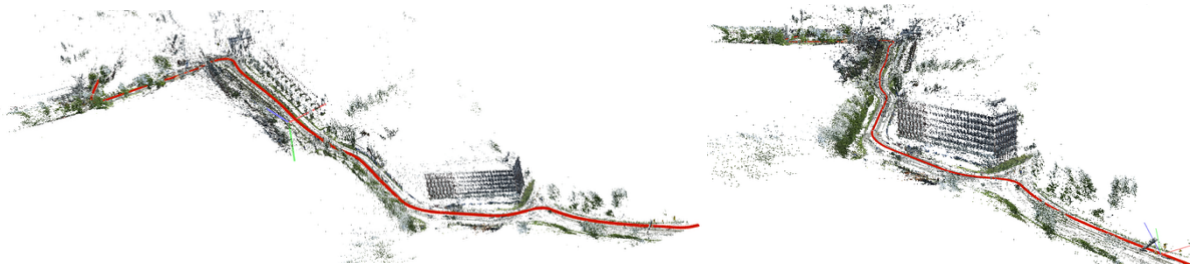

Figure 9. Views of SfM-based recovered trajectory (red signs) and sparse 3D reconstruction of sequence 2.

4.3 Evaluation on clear high-resolution data

For the second test, with images acquired a high-quality iPhone camera and under clear weather conditions, the visual differences between sequential and vocabulary-based matching are not immediately apparent. However, after geo-registration (5 reference points), vocabulary matching demonstrated superior performance, achieving an alignment error of 2.7 meters (mean) and 1.9 meters (median), compared to 3.1 meters (mean) and 2.2 meters (median) for sequential matching. Additionally, vocabulary-based matching yielded a higher number of 3D points, with 654,969 3D points compared to 611,565 3D points from sequential matching. Therefore, the vocabulary-based 3D results are used for the further steps of the framework. The reconstruction in Figure 9 clearly reveals buildings and trees, highlighting a significantly higher point cloud density compared to the foggy dataset. Several factors contributed to this improvement, including the slower speed at which the dataset is collected, enabling more stable and detailed image capture. Additionally, while both datasets had a reduced frame rate of 10 fps, the enhanced visibility, superior camera quality, and the feature matching method played a crucial role in achieving a denser and more accurate reconstruction.

Despite the improved resolution and visibility of the data, object detection (Figure 10a) struggles with classification inconsistencies. Adjusting detection thresholds reduces false positives but inadvertently increases false negatives, causing

lower-confidence traffic signs to be ignored despite being correctly detected. Future work should focus on improving detection reliability by expanding the training dataset for underrepresented traffic sign classes, ensuring better feature recognition across varying conditions and image resolutions.

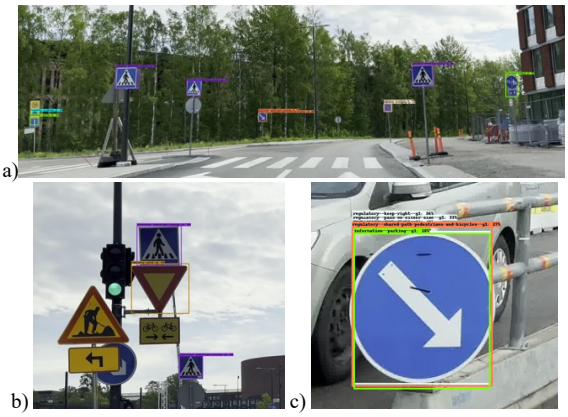

Figure 10. Object detection for sequence 2 (a) and close-up examples (b, c).

While most signs are successfully detected, misclassifications persist. Expanding the training dataset with more diverse and representative samples could significantly enhance classification accuracy and detection reliability. For instance, in our model the construction work class (Figure 10b) is particularly affected by limited training data. To address the issue of multiple classes being assigned to the same bounding box (Figure 10c), an additional filtering step is added. If bounding boxes overlap by more than 80%, only the class with the highest confidence score is retained. Additionally, all classifications with a confidence score below 24% are removed. The higher resolution and improved clarity in this dataset significantly benefit semantic segmentation, allowing for much cleaner boundary delineation, leading to more reliable semantic segmentation results compared to foggy data (Figure 8b). However, this improvement comes at the cost of processing speed - higher-resolution images require increased computational resources, which significantly slows inference time.

Figure 11. Object localization on sequence 2 using SfM (a), and SLAM (b). Red: ground truth; Blue: predictions

The evaluation results of the object localization algorithm on the second data sequence in clear weather conditions highlight the good performance of the detection system in identifying traffic signs, with 32 true positives out of 35 total traffic signs (Table 2, Figure 11a), indicating that most ground truth signs are successfully detected. However, the presence of 16 false positives and 3 false negatives (due to misclassification of object detection model) suggests that the object detection algorithm is overly sensitive, generating detections for non-sign objects and repeatedly classifying the same object multiple times. A median error of 4.22 m and a standard deviation of 1.06 m are observed for the detected traffic signs, indicating a consistent offset in localization (Table 3). This aligns with the observed pattern of predictions being systematically shifted in the same direction, likely due to geo-registration misalignment, as shown in Figure 12. The presence of this directional bias suggests a need for improved calibration of geospatial mapping methods to ensure more accurate localization.

	GT	*True positive*	*False positive*	*False negative*
SfM	35	32	16	3
SLAM+DF(20m)	35	28	7	7
SLAM+DF(30m)	35	28	9	7

Table 2. Object localization results (distance filtering - DF).

	Mean	*Std*	*Median*
SfM error [m]	4.22	1.06	4.22
SLAM error + DF (20 m)	6.58	4.25	5.43

Table 3. Offsets of localized objects (distance filtering - DF).

Figure 12. Geo-registration shifts in the SfM results, affecting the localization performance. Blue-ground truth, red-predictions.

Sequence 2 is also processed with the SLAM-based pipeline: out of the 3136 total images, 1417 are successfully oriented as key-frames, creating a final point cloud of 57,319 points (Figure 13).

Figure 13. Georeferenced SLAM trajectory (a) and point cloud (b) for sequence 2.

In the object detection step, the evaluation results indicate 28 true positives, 7 false positives and 7 false negatives out of 35 total traffic signs in the ground truth (Figure 11b, Table 3). With respect to the SfM-based results, the number of false positives is reduced due to lower number of 3D points in the SLAM-generated point cloud - since predictions for missing 3D points are simply discarded. This demonstrates that when the object detection model is overly sensitive, lower number of 3D points appeared to be beneficial. To improve noisy SLAM point cloud results, we applied additional distance filtering (DF) to 3D points, removing noisy points beyond 20 meters from the camera pose. This refinement reduced the total number of points to 26,439 but improved localization quality and optimized processing efficiency. A comparative analysis of different distance thresholds showed that filtering 3D points at 30 m distance introduced 4 additional false positives, while filtering at 15 meters resulted in 5 extra false negatives. Therefore, a 20-meter threshold proved to be the optimal balance, minimizing false negatives while maintaining a high true positive rate. The higher detection performance of the SfM-based approach, as reflected in the number of true positives, can largely be attributed to its denser point cloud.
The average localization error is measured at 6.58 meters, with a standard deviation of 4.25 meters, while the median error is 5.43 meters (Table 3).
The results based on the SLAM-derived point cloud demonstrated strong localization/identification performance while maintaining a more lightweight computational cost compared to SFM-based approach. Distance filtering was applied exclusively to the SLAM-based experiment due to the higher noise levels observed in SLAM-generated point clouds. While SfM could potentially benefit from similar filtering, its point cloud remained more stable and less affected by excessive outliers, making additional filtering unnecessary for this evaluation.

4.4 Change detection

The two sequences (Table 1) We conducted change detection tests, using as ground truth results from the first data sequence, representing the initial state of the environment. Based on this reference, change detection is performed on the second sequence in the overlapping part of the trajectory. The change detection results show that out of 21 total traffic signs in the aligned trajectory, and 13 documented changes (Figure 15 and 16) in the ground truth dataset (6 newly appeared signs and 7 disappeared ones), the proposed approach identified 8 traffic signs as appearing (Table 4). Among these, 6 are correctly classified as true positives (85.7% precision, Table 5), aligning with the ground truth, while 2 are false positives - resulting from the object detection algorithm due to a lack of sufficient training data. Additionally, the algorithm detected 6 out of 7 disappeared signs, resulting in 1 false negative. The model correctly identified 8 matched signs as true negatives. Figure 14 shows the confusion matrix summarizing change detection performance. Taking these into account, the overall recall for change detection is 92.3% (Table 5), confirming the system's strong capability in recognizing both newly introduced and removed traffic signs. This demonstrates strong performance in finding both newly introduced and missing objects. Further refinement to reduce false positives can be obtained through training the object detection algorithm on a larger dataset.

4.5 Discussion

Sequence 1 and 2 feature two very different environmental conditions and acquisition sensor and camera set-up. Figure 16 show how such conditions can affect the derived 3D scene.

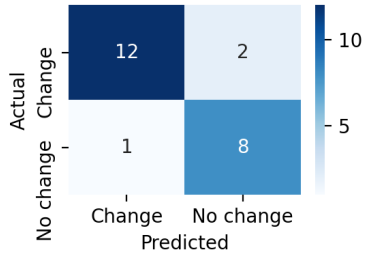

Figure 14. Confusion matrix for change detection results.

Total	*GT*	*True positives*	*True negatives*	*False positives*	*False negatives*
21	13	12	8	2	1

Table 4. Detected changes between the sequences 1 and 2.

Metric	*Value*
Precision: TP / (TP + FP)	85.7%
Recall: TP / (TP +FN)	92.3%

Table 5. Precision and recall values for the change detection process.

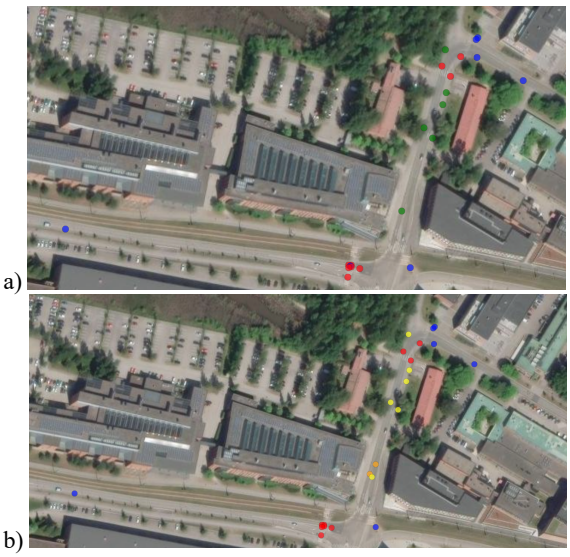

Figure 15. Change detection ground truth (a) - Green: appeared, Blue: disappeared, Red: matched. Predictions results (b) - Blue: disappeared, Yellow: appeared, Orange: false positives, Red: matched/ true negatives.

The sparsity of the point cloud stems from several factors affecting the 3D reconstruction (Figure 16a). Low feature density in uniform areas, poor visibility (darkness and fog), limited camera perspective and sequential image matching reduced the number of valid correspondences. Additionally, the 10 fps frame rate limited fine-detail capture between frames. The imposed constraints resulted in fewer matched features and a sparser reconstruction for sequence 1. Within the purple dotted rectangle, the reconstructed traffic signs on the right side of the road are visible. This indicates that certain images successfully captured the traffic sign features, allowing spatial reconstruction within the point cloud. Figure 16b illustrates a denser point cloud from sequence 2 (sunny weather, better camera), where the traffic sign reconstruction is more detailed. The corresponding detected feature example is presented in Figure 16d. However, if the SLAM-based approach is not selected a particular keyframe, objects might

not be detected and the system might fail to correctly localize signs, preventing their integration into an updated map.

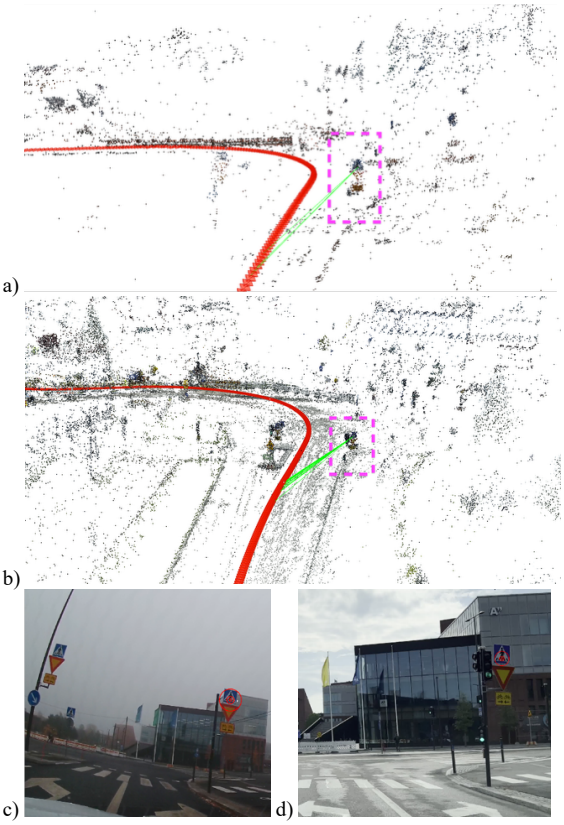

Figure 16. Reconstructed scene for sequence 1 (a) and 2 (b). Detected features in sequence 1 (c) and 2 (d) corresponding to the same location in town.

5. Conclusions

This work has introduced a low-cost crowdsourcing methodology for urban environment 3D mapping using crowdsourced dashcam videos, with a focus on automatic change detection. Leveraging techniques such as SfM, SLAM, semantic segmentation and object detection, the proposed approach enables scalable and cost-effective updates of urban maps in the meter-accuracy range. The system's robustness is evaluated with complex sequences captured under challenging environmental conditions, specifically darkness and fog. Additionally, performance is assessed using both SfM-based point clouds and lightweight SLAM-based point clouds. The results demonstrate that despite the sparser SLAM-based 3D point cloud, SLAM can achieve comparable object detection and localization performance to SfM, while significantly reducing processing time. Furthermore, the change detection method successfully identified most environmental modifications between two states of the environment (2-years apart), achieving a recall of 92.3% and a precision of 85.7%, despite the challenging weather conditions of sequence 1. While SfM offers slightly better numerical accuracy, SLAM significantly improves processing efficiency, making it a viable alternative for real-time or resource-constrained environments without significantly compromising reliability.
The results highlight the potential for (real-time) cityscape monitoring, robust mapping under diverse environmental conditions and cost-effective map updates. Future improvements could further refine system components, particularly the object detection module, enhancing robustness and contributing to more reliable and adaptable urban mapping workflows. Additionally, separating the 3D model into multiple sections may improve geo-registration accuracy, enabling better alignment with road geometry and reducing localization errors. On-board GNSS receiver could be used to support geolocalization, with the known issues of satellite signals in urban environments.

References

Alpherts, T., Ghebreab, S., van Noord, N., 2025. EMPLACE: Self-supervised urban scene change detection. Proc. *AAAI*.

Campos, C., Elvira, R., Gómez Rodríguez, J.J., Montiel, J.J.M., Tardós, J.D., 2021. ORB-SLAM3: An accurate open-source library for visual, visual-Inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874-1890.

Cheng, Q., Zeller, N., Cremers, D., 2022. Vision-based large-scale 3D semantic mapping for autonomous driving application. Proc. *ICRA*, pp. 9235-9242.

Elhashash, M., Albanwan, H., Qin, R., 2022. A Review of mobile mapping systems: from sensors to applications. *Sensors*, 22, 4262.

de Gélis, I., Lefèvre, S., Corpetti, T., 2023. Siamese KPConv: 3D multiple change detection from raw point clouds using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing,* Vol. 197, pp. 274-291.

de Gélis, I., Lefèvre, S., Corpetti, T., 2021. Change Detection in Urban Point Clouds: An Experimental Comparison with Simulated 3D Datasets. *Remote Sensing*, 13, 2629.

Godard, C., Aodha, O. M., Firman, M., Brostow, G., 2019. Digging into self-supervised monocular depth estimation. Proc. *CVPR*, pp. 3827–3837.

Kaur, R., Singh, S., 2023. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, Vol. 132, 103812.

Kharroubi, A., Remondino, F., Ballouch, Z., Hajji, R., Billen, R., 2025. Semantic and geometric fusion for object-based 3d change detection in LiDAR point clouds. *Remote Sensing*, 17, 1311.

Kazerouni, I.A., Fitzgerald, L., Dooly, G., Toal, D., 2022. A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications*, Vol. 205(1), 117734.

Kim, J.W., Kim, U.H., 2025. Towards Generalizable Scene Change Detection. Proc. *CVPR*.

Lee, J.H., Han, M.K., Ko, D.W., Suh, H., 2019. From big to small: multi-scale local planar guidance for monocular depth estimation. *arXiv:1907.10326*

Lee, M.J.L., Hsu, L.T., 2021. Semantic 3D map change detection and update based on smartphone visual positioning system. *arXiv:2103.11311.*

Li, X., Wang, D., Ao, H., Belaroussi, R., Gruyer, D., 2019. Fast 3D semantic mapping in road scenes. *Applied Science*, 9, 631.

Lin, Z., Yu, J., Zhou, L., Zhang, X., Wang, J., Wang, Y., 2022. Point cloud change detection with Stereo V-SLAM: dataset, metrics and baseline. *IEEE Robotics and Automation Letters*, Vol. 7(4,) pp. 12443-12450.

Lin, C.J., Garg, S., Chin, T.J., Dayoub, F., 2025. Robust Scene Change Detection Using Visual Foundation Models and Cross-Attention Mechanisms. Proc. *ICRA*.

Lu, X. et al., 2019. Object Detection Based on SSD-ResNet. Proc. *ICCIS*.

Lu, Z., Ye, J., Leonard, J., 2025. 3DGS-CD: 3D Gaussian Splatting-based Change Detection for Physical Object Rearrangement. IEEE *Robotics and Automation Letters*.

McCormac, J., Handa, A., Davison, A., Leutenegger, S., 2017. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. Proc. *ICRA*, pp. 4628-4635

Mo, Y., Wu, Y., Yang, X., Liu, F., Liao, Y., 2022. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, Vol. 493(7), pp. 626-646.

Pan, L., Baráth, D., Pollefeys, M., Schoenberger, J.L., 2024. Global structure-from-motion revisited. Proc. *ECCV*.

Porzi, L., Rota Bulo', S., Colovic, A., Kontschieder, P., 2019. Seamless scene segmentation. Proc. *CVPR.*

Qin, T., Chen, T., Chen, Y., Su, Q., 2020. AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot. Proc. *IROS*, pp. 5939-5945.

Roddick and Cipolla, 2020. Predicting semantic map representations from images using pyramid occupancy networks. Proc. *CVPR*.

Schoenberger, J.L., Frahm, JM., 2016. Structure-from-motion revisited. Proc. *CVPR*.

Shabestari, Z.B., Hosseininaveh, A., Remondino, F., 2023. Motorcycle detection and collision warning using monocular images from a vehicle. *Remote Sensing*, 15, 5548.

Shirowzhan, S., Sepasgozar, S., Li, H., Trinder, J., Tang, P., 2019. Comparative analysis of machine learning and point-based algorithms for detecting 3D changes in buildings over time using bi-temporal LiDAR data. *Automation in Construction*, 105, 102841.

Stilla, U., Xu, Y., 2023. Change detection of urban objects using 3D point clouds: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 228–255.

Tran, T.H.G., Ressl, C., Pfeifer, N., 2018. Integrated change detection and classification in urban areas based on airborne laser scanning point clouds. *Sensors*, *18*, 448.

Xiao, W., Cao, H., Tang, M., Zhang, Z., Chen, N., 2023. 3D urban object change detection from aerial and terrestrial point clouds: A review. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 118, 103258.

Zhanabatyrova, A., Leite, C., Xiao, Y., 2023. Automatic map update using dashcam videos. *IEEE Internet of Things Journal*.

Zhanabatyrova, A, Leite, C., Xiao, Y., 2024. Structure from motion-based mapping for autonomous driving: Practice and experience. *ACM Transactions on Internet of Things*.

Zhanabatyrova, A., 2025. *Crowdsourced 3D semantic mapping and change detection in urban driving environments.* PhD thesis, Aalto University, Finland. Available at https://urn.fi/URN:ISBN:978-952-64-2412-5

Zhang, P., Zhang, M., Liu, J., 2021. Real-time HD map change detection for crowdsourcing update based on mid-to-high-end sensors. *Sensors*, Vol. 21, no. 7.

Zhang, H., Venkatramani, S., Paz, D., Li, Q., Xiang, H., Christensen, H.I., 2023. Probabilistic semantic mapping for autonomous driving in urban environments. *Sensors*, 23, 6504.