Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 11, No. 04, April, 2025

Manholes Detecting and Mapping Using Open-World Object Detection and GIS Integration

Ibrahim F. Ahmed [1*], Mohammed Alheyf [2], Ahmed Ali [1], Mohamed S. Yamany [3, 4]

[1] *Construction Eng. & Utilities Department, Faculty of Engineering, Zagazig University, Zagazig 44159, Egypt.*

[2] *Department of Civil Engineering, College of Engineering, King Saud University, Riyadh 12372, Saudi Arabia.*

[3] *Department of Construction Engineering, Faculty of Engineering, Zagazig University, Zagazig 44159, Egypt.*

[4] *Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, WY 82071, United States.*

Received 19 January 2025; Revised 08 February 2025; Accepted 02 March 2025; Published 01 April 2025

Abstract

Accurate detection and mapping of manholes are essential for urban infrastructure management, facilitating efficient maintenance and safety. This paper introduces a novel methodology that integrates the open-world object detection model, Grounding DINO, with geographic information systems (GIS) to detect and geolocate manholes in urban environments. Unlike traditional object detection approaches that rely on extensive labelled datasets and predefined object categories, Grounding DINO, a transformer-based model, leverages natural language processing for adaptable, scalable detection. Grounding DINO processes natural language descriptions to detect the manholes in an open-world context, overcoming the limitations of predefined object categories. Detected manholes are localized using multi-view triangulation, which refines their 3D positions by leveraging redundant camera viewpoints and intrinsic calibration parameters, which ensures accurate geometric mapping of manhole centers. The resulting geospatial coordinates are transformed into the WGS84 system using a global navigation satellite system/inertial navigation system (GNSS/INS) for compatibility with GIS platforms. The proposed approach achieved sub-meter precision, with mean localization errors of 0.36 meters in easting and 0.34 meters in northing, evaluated on KITTI dataset sequences under various urban conditions. The seamless integration of object detection and geospatial mapping demonstrates the potential of this approach for efficient and scalable urban infrastructure management.

Keywords: Manhole Detection; Open-World Object Detection; Geographic Information Systems (GIS); Grounding DINO.

1. Introduction

The management of urban infrastructure is important for the safety, effectiveness, and sustainability of public utility systems, such as roadways and sewage infrastructures. Manhole detection and maintenance are two crucial tasks in such management. Traditional methods for manhole detection and mapping rely mainly on manual fieldwork and practices, which are often laborious and inefficient for large-scale urban settings [1, 2]. Besides, increased urban landscape complexity raises the need for more intelligent automated solutions. To cope with such challenges, technologies such as geographic information systems, computer vision, and deep learning have become more prominent. Previous studies pointed out that these technologies integrated into the operation flow significantly contribute to acquiring higher accuracy, lower labor intensiveness, and real-time data availability [3-5].

* Corresponding author: Ibrahim_fouad_ghonim@yahoo.com; ifahmd@eng.zu.edu.eg

http://dx.doi.org/10.28991/CEJ-2025-011-04-07

The traditional object detection methods, such as Haar cascades and HOG-SVM, were effective only in controlled environments and were based on hand-crafted features [6]. With the introduction of convolutional neural networks (CNNs), it was possible to have an improved detection accuracy in more complex scenarios [7]. Modern object detection architectures, such as YOLO (You Only Look Once) and Faster R-CNN, have gained widespread adoption, given their state-of-the-art accuracy and speed in detecting objects across a wide range of applications [8, 9]. However, these approaches are constrained by their dependence on predefined object classes and extensive labeled datasets. In areas like urban infrastructure, where manholes and other objects vary greatly in size, shape, and appearance, this dependence leads to bottlenecks [10, 11].

A novel approach that overcomes all these drawbacks is open-world object detection. This does not require any pre-training on certain classes of objects, in contrast to all the aforementioned models. Instead, this employs natural language processing (NLP) to enable the performance of descriptive text searches for detecting never-seen objects [12, 13]. This feature is especially beneficial in dynamic and diverse urban environments, where detecting objects, such as manholes, will often be in complicated setups. This avoids the need to extensively pre-train the model, improving adaptability and scalability while drastically cutting down the preparation time [14, 15]. In addition, open-world object detection offers a novel approach to urban applications, which allows systems to adjust to a variety of unpredictable situations [16, 17]. This is particularly important when it comes to managing urban infrastructure since the physical attributes of items like drainage gates, utility covers, and manholes can differ greatly between cities and regions [18].

This paper introduces a novel approach for manhole detection and mapping, which integrates open-world object detection with a geographic information system for city infrastructure management. The developed method leverages the Grounding DINO model, which can perform well without extensive pre-training for most urban scenes and identifies objects based on natural language queries. Triangulation methods are utilized to estimate the manhole position. The estimated coordinates are then converted into a geographical data frame that is compatible with the geographic information system software. This ensures the absolute accuracy of manhole position mapping in practical applications. The KITTI dataset has been used to prove that the proposed methodology can operate well in terms of robustness and scalability in various urban scenarios. This paper is organized as follows: Section 2 presents the related work, including a review of the latest developments regarding manhole detection using deep learning and an investigation of applications related to Grounding DINO in civil engineering. The detection and mapping approach proposed in this article is discussed in full detail in Section 3; the data used in the evaluation are described in Section 4; experimental results are discussed in Section 5; and conclusions are presented in Section 6.

2. Related Works

Manhole detection and localization in an urban environment have been widely investigated with different deep learning-based approaches. Previous research work has significantly improved the accuracy of detection, precision in localization, and computational efficiency. However, many of these prior studies still suffer from shortcomings regarding either too-extensive training datasets or not being directly able to integrate their results into geospatial databases for practical purposes.

Several research works have used deep learning models for real-time manhole detection. In Kumar et al. [19], a CNN-based predictive model was proposed for real-time manhole hazard detection, providing authorities with a tool to promptly address hazardous conditions and enhance public safety in urban environments. Pang et al. [20] evaluated deep-learning-based object detection algorithms that resulted in the detection of pavement manhole covers with an accuracy of over 90%, but the detection methodology was dependent on predefined datasets for training and did not consider the precision of localization. Zhang et al. [21] proposed a method of data augmentation in order to improve the detection of abnormal manhole covers by enhancing the robustness of the methodology in variable conditions; however, it did not integrate with geospatial mapping to visualize the results. In Yang et al. [9], an improved YOLOX-based model for manhole detection was proposed that achieved high precision and recall in real-time applications. Leni et al. [22] presented a YOLOv8-based real-time pothole detection system designed for smart city infrastructure monitoring. The study introduced a multi-stage optimization framework that enhanced detection accuracy while maintaining real-time performance. Notably, the approach supported multi-class detection, identifying potholes, manholes, and road surfaces simultaneously, achieving 79% precision and 74.6% recall for pothole detection. However, all the previous methods focused on detection and did not associate the location of manholes with geospatial coordinates for creating the database.

Researchers have also employed more advanced architectures. An improved Faster R-CNN architecture has been proposed in Zhang et al. [23], which successfully detected pavement manhole covers with a location accuracy as low as 0.5 meters. Their approach requires heavy training on labeled datasets and hence limits its scalability to new urban environments. The work of Lin et al. [24] proposed using a structured light camera to detect structural anomalies with manhole covers. They suggested using point cloud data processing techniques such as RANSAC and DBSCAN to accurately measure subsidence and improve road maintenance efficiency. Similarly, Qing et al. [25] introduced a

combination of mobile LiDAR and deep learning for offering rapid detection at sub-meter localization precision. Their approach leverages high-cost sensor data and has never integrated easily with a full-fledged GIS platform. Commandre et al. [16] conducted a similar manhole detection using aerial imagery coupled with deep learning. While this approach offers scalability, it is an expensive solution due to the costs associated with acquiring and processing aerial images. Moreover, manholes could be covered or occluded by trees, cars, or other objects, further complicating the detection process and reducing its reliability. Accuracy in localization has been one of the prime evaluation points in these works. Methods such as those in Zhang et al. [23] and Qing et al. [25] achieve sub-meter accuracy but have dependencies on fixed datasets and no integration of geospatial databases for the maintenance of urban infrastructure; thus, both are practically inoperable. In Khare et al. [26], researchers utilized the YOLOv8 model to detect more general road hazards, including manhole detection. Again, this work does not focus on how to produce actionable maps for maintenance teams.

Being a transformer-based open-set object detection model, Grounding DINO has attracted more attention for its flexibility and generalization capabilities. In contrast to traditional models, which are trained using large-scale training datasets, this model enables natural language-based detection and generalizes to new object categories. Its application in civil engineering has been proposed in several works. Its application in civil engineering has been explored in several works. Ma et al. [27] proposed a marker recognition method for substation engineering progress monitoring using Grounding DINO, demonstrating its adaptability to infrastructure-related tasks. Cai et al. [28] proposed using Grounding DINO, which combines image and text modalities, to improve safety monitoring in construction sites by enabling the detection of both predefined and unknown construction elements. In de Moraes Vestena et al. [29], they employed Grounding DINO in conjunction with SAM algorithms to classify pavement types, thereby demonstrating an open-vocabulary detection capability. Nevertheless, these studies only demonstrate the ability for the detection tasks, not for the localization and geospatial mapping of the detected entities.

While these approaches have very promising results for detection accuracy, they all suffer from a number of limitations that include large labeled training datasets, not being integrated into GIS platforms, and almost complete disregard for efforts toward the creation of comprehensive geospatial databases. These limit their applicability for practical use in maintaining urban infrastructure. For example, no method has related the detections back to the actual geographic locations of manholes in a manner that supports making real decisions, such as scheduling maintenance. The research work in this paper attempts to fill those knowledge and practice related gaps by integrating the open-world object detector Grounding DINO with GIS platforms.

3. Research Methodology

This study presents a novel approach for manhole detection and mapping that combines a cutting-edge open-world object detection algorithm with GIS for urban infrastructure management. The methodology integrates an advanced object detection model, Grounding DINO, with triangulation and GIS technologies to achieve accurate and scalable detection and mapping. The overall methodology flowchart is shown in Figure 1, which summarizes the processes and steps of the proposed approach.

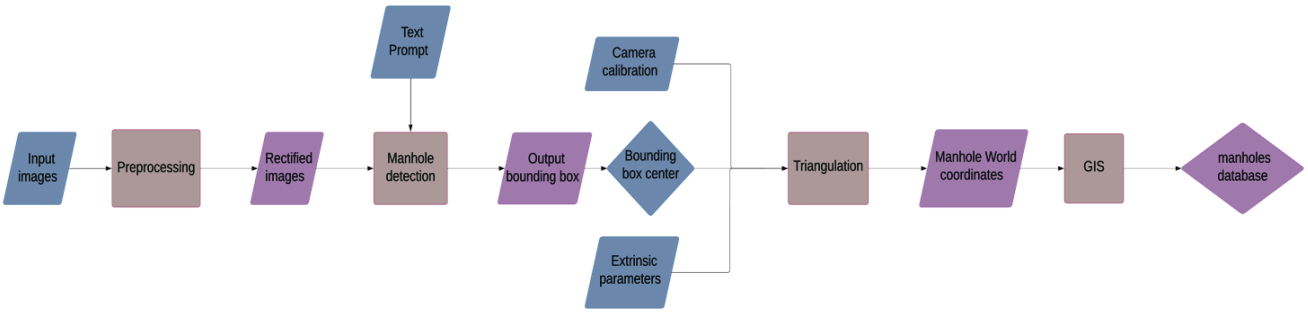


Figure 1. Workflow of the proposed manhole detection methodology

The detection pipeline's foundation is grounding DINO. This model uses natural language queries to detect objects in an open-world environment by integrating a transformer-based architecture with grounded pre-training. Grounding DINO performs exceptionally well at recognizing new items described by textual inputs, in contrast to conventional object detection models that depend on preset categories and large labelled datasets. Several essential elements compose its architecture, as illustrated in Figure 2. While a BERT-based text encoder creates features from natural language inputs, a Swin Transformer serves as the image backbone, extracting multi-scale image features. The model's feature enhancer aligns the modalities so that accurate detection can happen. It does this by fusing text and visual features through self-attention and cross-attention methods. A language-guided query selection module further refines the detection procedure by initiating object detection queries that align descriptive text with picture features. A cross-modality decoder processes these queries, assigning labels and fine-tuning the object-bounding boxes to ensure

correctness even in intricate or untested situations [30]. For example, when provided with the natural language query *"circular metal manhole cover on the road,"* the model extracts semantic features from the text and searches for visual patterns that correspond to the description. The text features guide the query selection module, ensuring the detection focuses on objects with circular metallic structures embedded in road surfaces. Unlike traditional models that rely on fixed object categories, Grounding DINO dynamically matches text prompts with real-world features, allowing it to detect manholes of varying shapes, colors, and surface conditions without predefined labels.

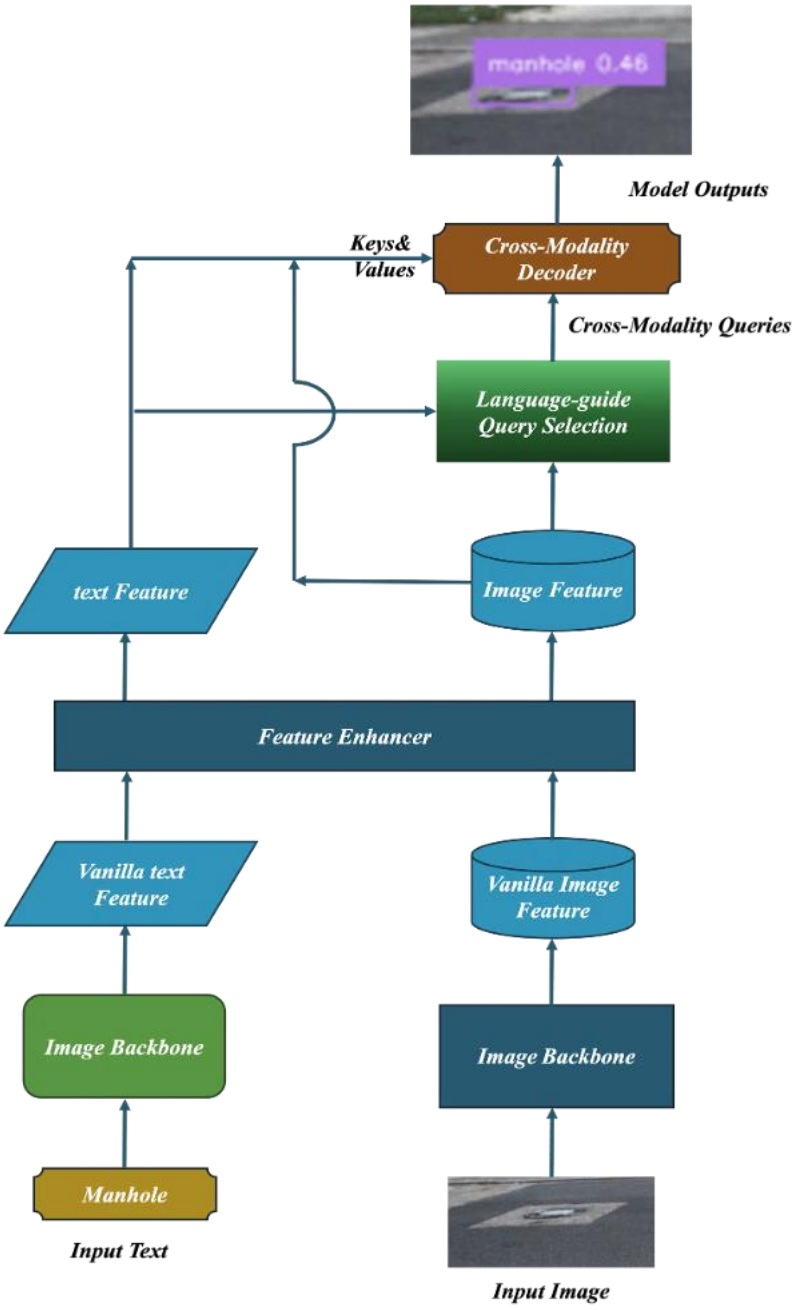


Figure 2. Grounding DINO Model architecture

Grounding DINO was trained on a wide range of datasets to guarantee reliable performance under a variety of circumstances. To improve detection capabilities, datasets including COCO, Objects365, and OpenImages were used. Grounding data like GoldG and RefC, offered more annotations that connected text descriptions to visual regions. In order to enhance the generalization capability of the model to new item categories, extensive caption data and image-caption pairs were utilized. The model's performance on benchmarks demonstrated how effective it is. With a remarkable Average Precision (AP) of 52.5 on the COCO zero-shot detection benchmark without training on COCO images, Grounding DINO significantly outperformed earlier models. The model proved its adaptability and precision by achieving an accuracy of 90.56% in fine-tuned settings on referring expression problems, such as the Ref-COCO/+/g datasets. Despite using a single-word prompt "manhole", Grounding DINO effectively detected manholes by leveraging its open-set object recognition and cross-modal feature alignment. This enabled the model to generalize across varying

appearances, lighting conditions, and occlusions, ensuring robust detection in diverse urban environments. However, challenges such as occlusions, varying lighting conditions, and diverse road textures might occasionally impact detection confidence. To mitigate these issues, the model's ability to generalize across different environments was enhanced through multi-view detection, ensuring more robust localization and minimizing false positives. Once manholes are detected, their precise localization is estimated using triangulation. This process comprises detecting the same manhole from multiple camera viewpoints, leveraging redundancy to mitigate errors caused by occlusions or environmental noise. Camera calibration ensures that intrinsic and extrinsic parameters are accounted for, enabling accurate geometric relationships between image pixels and real-world coordinates. The triangulation process employs epipolar geometry to calculate the 3D coordinates of detected manhole centers in the camera frame, transforming image detections into actionable geospatial data. The redundancy in camera views ensures that each detected manhole is observed from multiple angles, allowing for error correction and improved depth estimation. By analyzing overlapping viewpoints, inconsistencies in depth measurements can be minimized, leading to higher localization accuracy even in occluded or cluttered environments.

After triangulating the manhole centers in the camera coordinate frame using multiple views, the coordinates of a manhole center (Xc, Yc, Zc) are transformed into the world coordinate system84 (WGS84). This transformation involves a series of sequential rotations and translations using the extrinsic calibration parameters between the camera, LiDAR, IMU (body frame), and global navigation satellite system (GNSS) reviver. The final result maps the detected manhole locations into the Earth-Centered Earth-Fixed (ECEF) coordinate system and subsequently into WGS84 geodetic coordinates.

The coordinates in the camera frame are first transformed to the LiDAR frame using a rotation matrix R_{velo_cam} and a translation vector T_{velo_cam}. The transformation is represented as:

$$\begin{bmatrix} X_{\text{velo}} \\ Y_{\text{velo}} \\ Z_{\text{velo}} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{\text{velo\_cam}} & T_{\text{velo\_cam}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{1}$$

Next, the coordinates are transferred to the IMU (body frame) using the rotation matrix R_{imu_velo} and translation vector T_{imu_velo}:

$$\begin{bmatrix} X_{\text{imu}} \\ Y_{\text{imu}} \\ Z_{\text{imu}} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{\text{imu\_velo}} & T_{\text{imu\_velo}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{velo} \\ Y_{velo} \\ Z_{velo} \\ 1 \end{bmatrix} \tag{2}$$

The IMU coordinates are then transformed into the local East-North-Up (ENU) frame. The orientation of the IMU in the ENU frame is represented by the rotation matrix R_{imu_ENU}, which is constructed using roll, pitch, and yaw angles. The local ENU frame is then rotated and translated into the global ECEF frame. This transformation uses the rotation matrix R_{ENU_ECEF}, which is computed from the latitude ϕ and longitude λ of the ENU origin. The transformation is given by:

$$R_{\text{ENU\_ECEF}} = \begin{bmatrix} -\sin(\lambda) & -\sin(\phi)\cos(\lambda) & \cos(\phi)\cos(\lambda) \\ \cos(\lambda) & -\sin(\phi)\sin(\lambda) & \cos(\phi)\sin(\lambda) \\ 0 & \cos(\phi) & \sin(\phi) \end{bmatrix} \tag{3}$$

The translation vector T_{ENU_ECEF} is computed by converting the GNSS-derived latitude, longitude, and altitude into Cartesian coordinates. The overall transformation from the camera frame to the WGS84 frame can be summarized as:

$$RT_{\text{cam\_ECEF}} = RT_{\text{ENU\_ECEF}} \cdot RT_{\text{imu\_ENU}} \cdot RT_{\text{velo\_imu}} \cdot RT_{\text{cam\_velo}} \tag{4}$$

The final step involves converting the ECEF coordinates (XECEF, YECEF, ZECEF) into geodetic latitude, longitude, and altitude on the WGS84 ellipsoid using standard geodetic formulas.

The final stage of the methodology integrates these geospatial coordinates into GIS platforms. Detected manhole coordinates are exported in standard geospatial formats and then visualized and analyzed within GIS software. This integration allows for the creation of comprehensive maps that detail manhole locations across urban environments. These maps provide a valuable resource for urban planners, maintenance teams, and safety inspectors, offering a centralized database for infrastructure management and facilitating data-driven decision-making.

4. Dataset

The KITTI dataset provides a wide range of real-world data collected in Karlsruhe, Germany, shown in Figure 3, with a cutting-edge sensor suite mounted on a car. It is widely recognized for its contributions to research on autonomous driving and mobile robotics. This platform consists of an OXTS GNSS/inertial navigation system (INS), a Velodyne HDL-64E 3D laser scanner, and two calibrated stereo cameras (color and grayscale). Accurate depth estimation is made possible by the stereo camera arrangement, which produces synchronized left and right images with a resolution of 1242×375 pixels. Accurate data fusion and trajectory mapping are ensured by precisely calibrating intrinsic and extrinsic parameters across all sensors. In this study, sequences 2, 9, 14, and 17 were selected to evaluate the proposed methodology for detecting and localizing manholes representing a variety of driving conditions.

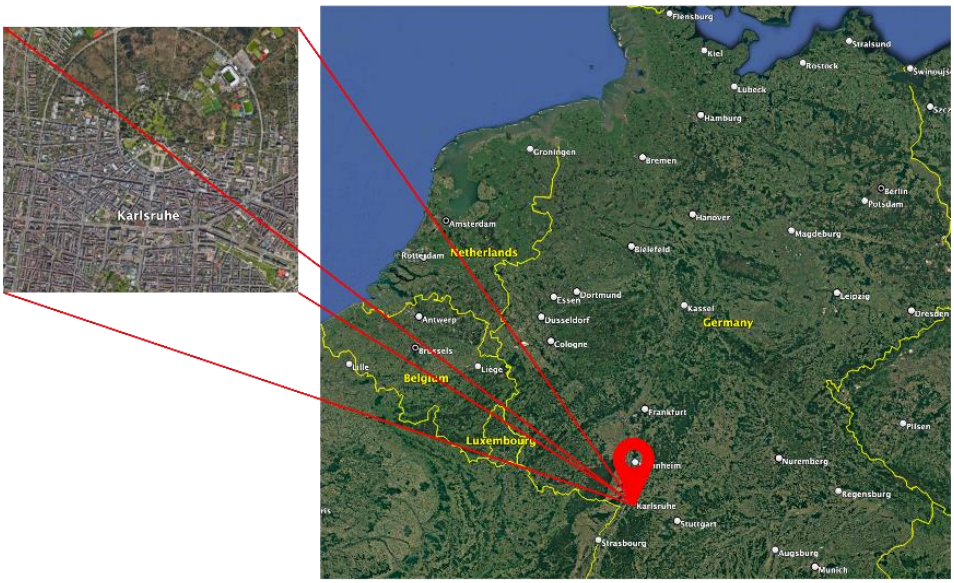
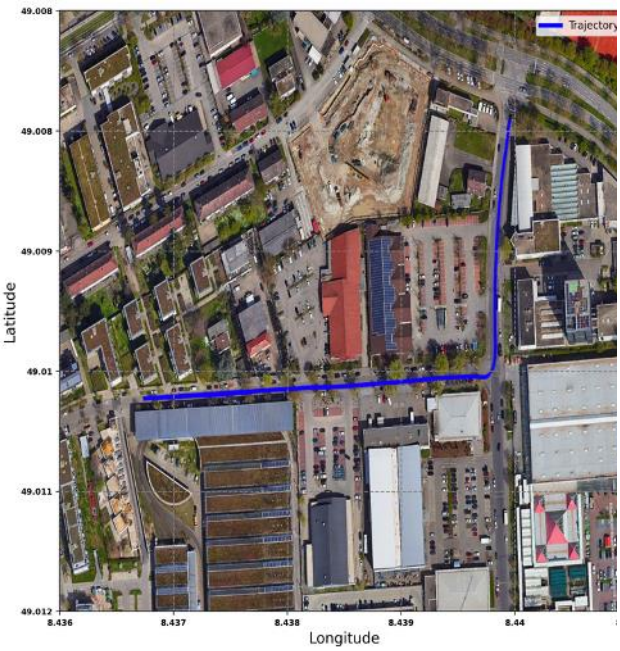


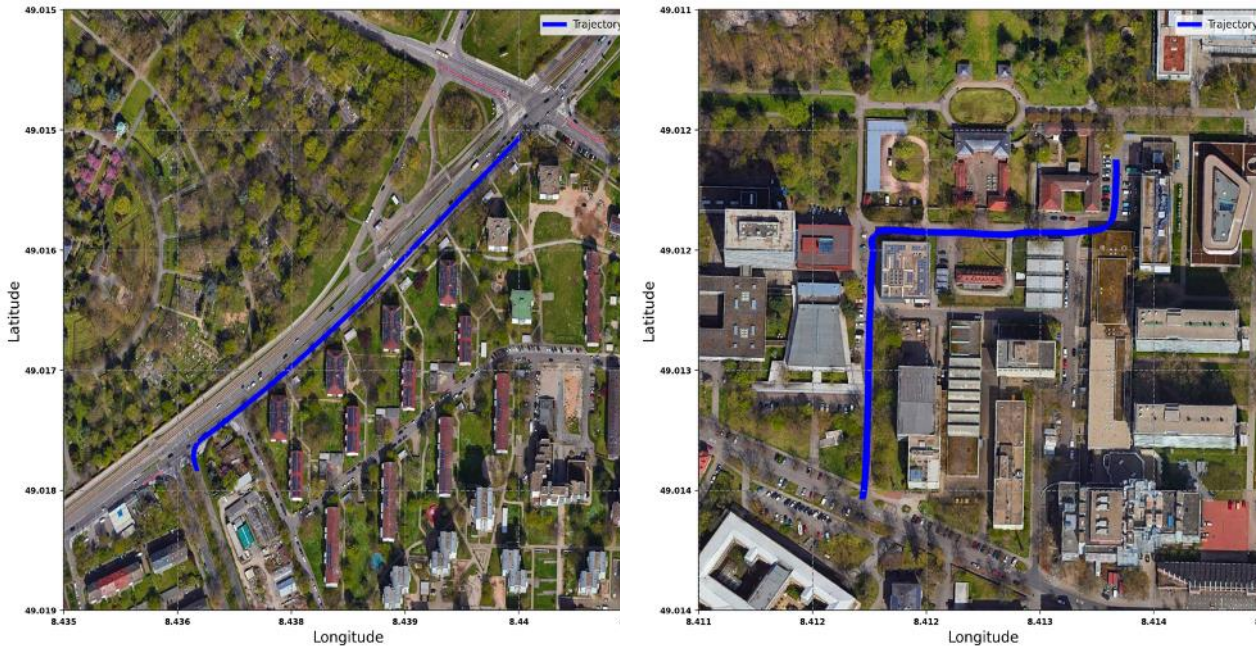
Figure 3. Study Area of the KITTI Dataset: The figure illustrates the location of Karlsruhe, Germany, where the KITTI dataset was captured

The purpose of this work is to leverage the KITTI dataset to detect manholes in urban environments and reconstruct their 3D geospatial positions using a combination of deep learning, multi-view geometry, and the captured GNSS/INS data. By selecting these specific sequences, the study aims to evaluate the robustness and scalability of the methodology in diverse settings. The trajectories of the selected sequences are illustrated in Figure 4, which highlights the paths traversed during data collection, enabling visualization of the environments where manhole detection and localization were conducted.

Trajectory for sequence 2



Trajectory for sequence 9



Trajectory for sequence 14 Trajectory for sequence 17

Figure 4. Trajectories of selected KITTI sequences (2, 9, 14, and 17) used in the study

5. Results and Discussion

The proposed methodology showed considerable efficiency in manhole detection and localization on different urban scenes by using KITTI dataset sequences. To design a reliable ground truth for validation, a high-resolution image was georeferenced using ArcGIS Pro based on the method of image-to-image registration over a base map. The image registration needed the identification of 13 well-spread ground control points (GCPs) to ensure good alignment. Figure 5 presents an overlay of high-resolution imagery on the base maps with an emphasis on the spatial distribution of GCPs. The residual statistics of GCPs obtained from this georeferencing, as shown in Table 1, confirm the accuracy of the registration process, and the total RMS errors for the forward, inverse, and forward-inverse residuals were all recorded at 0.16, 0.57, and 0.09, respectively. The forward residual measures errors in spatial reference units, while the inverse residual measures errors in pixel units; the forward-inverse residual, on the other hand, represents the overall alignment accuracy. These statistics indicate the high accuracy achieved in the georeferencing process. The accuracy of the georeferencing process was critical in establishing a reliable ground truth for evaluation. The use of 13 well-spread GCPs ensured minimal transformation distortions, preserving the spatial integrity of manhole positions. This methodological rigor enhances confidence in the validation process, minimizing systematic errors that could otherwise influence detection performance assessment. Thus, the manhole centers' coordinates were extracted from this high-resolution georeferenced image and used as ground truth for the evaluation of the proposed detection and localization method.

Figure 5. High-resolution image overlayed on base map

Table 1. Image to image registration residuals statistics

ID	Source X	Source Y	Map X	Map Y	Residual	Residual X	Residual Y
1	2,009.926	-1,577.075	939,057.919	6,276,475.083	0.157	0.153	0.037
2	7,353.536	-4,899.152	939,543.509	6,278,131.184	0.040	-0.027	0.029
3	7,455.363	-562.487	938,374.607	6,277,858.392	0.026	-0.003	0.026
4	247.350	-346.633	938,865.671	6,275,885.529	0.037	-0.034	0.013
5	715.314	-3,791.218	939,747.878	6,276,295.166	0.067	0.005	-0.066
6	6,907.970	-2,794.387	939,013.669	6,277,875.548	0.105	0.026	-0.101
7	5,351.342	-561.774	938,534.091	6,277,295.155	0.072	0.026	0.067
8	5,175.661	-2,995.341	939,197.800	6,277,430.553	0.312	0.129	0.283
9	3,970.591	-4,481.688	939,686.559	6,277,216.754	0.225	0.059	-0.217
10	6,061.638	-1,327.831	938,683.130	6,277,545.746	0.102	-0.041	-0.093
11	3,729.199	-2,565.429	939,191.566	6,277,014.339	0.273	-0.267	-0.056
12	3,545.473	-526.027	938,663.118	6,276,806.057	0.073	-0.001	-0.073
13	2,186.339	-4,650.678	939,866.570	6,276,758.255	0.154	-0.023	0.152

Manhole boundaries and centers were successfully detected using the Grounding DINO model, which also showed remarkable performance with confidence scores that indicate how accurate each detection was. Grounding DINO and other object detection models employ the confidence score as a metric to quantify the certainty that an object they have spotted belongs to a given category (in this case, a "manhole"). Greater confidence in the detection is indicated by higher scores, which range from 0 to 1. In order to maintain a balance between recall and precision, a threshold of 0.4 was selected. Because of the model's inherent flexibility, a confidence score of 0.4 is deemed adequate and dependable for Grounding DINO, which is intended for open-world object detection. By using natural language descriptions and a transformer-based architecture, Grounding DINO is able to generalize to new object categories with less reliance on labelled training data than traditional object detection models, which require extensive pre-training with predefined object categories. Although this adaptability is a significant advantage, it also implies that models trained on a fixed dataset with well-defined categories typically have slightly higher confidence scores. A threshold that is set too high may prevent legitimate detections, particularly in difficult situations such as partially obscured or distant objects. On the other hand, a lower threshold could permit an excessive number of false positives, which would impair the results' dependability.

Confidence scores above the 0.4 level were retained, as seen in Figures 6 to 9, guaranteeing accurate detections while reducing false positives. A confidence score analysis revealed that most valid detections exceeded 0.5, with a relatively small proportion falling between 0.4 and 0.5. The chosen threshold of 0.4 ensured that potentially valid detections were not mistakenly discarded, particularly in cases where partial occlusions or varying illumination affected feature visibility. This balance between recall and precision aligns with the open-world object detection framework, allowing for reliable detection while minimizing false positives. For instance, detections with scores higher than this threshold, such as 0.64 in Figure 8's first image, suggest a strong probability that the object detected is, in fact, a manhole. On the other hand, detections with a score of 0.38, such as the red detection in Sequence 9's second image in Figure 7, is discarded because they are below the predetermined threshold, preventing false positives from influencing the mapping outcomes.

Detection on the first image

Detection on the second image

Manhole boundary and center extraction on the first image

Manhole boundary and center extraction on the second image

Figure 6. Sequence 2 manhole detection example

Detection on the first image

Detection on the second image

Manhole boundary and center extraction on the first image

Manhole boundary and center extraction on the second image

Figure 7. Sequence 9 manhole detection example

Detection on the first image

Detection on the second image

Manhole boundary and center extraction on the first image

Manhole boundary and center extraction on the second image

Figure 8. Sequence 14 manhole detection example

Detection on the first image

Detection on the second image

Manhole boundary and center extraction on the first image

Manhole boundary and center extraction on the second image

Figure 9. Sequence 17 manhole detection example

While the methodology demonstrated promising results, several factors, most notably the camera's proximity to the manhole and the angle at which the images were taken, influenced the detection confidence scores. First, closer proximity to the camera significantly improves detection confidence due to clearer feature details. For instance, as seen in Figure 8, the first image's confidence score was 0.64, which increased to 0.7 in the second image as the manhole was closer to the camera. This pattern demonstrates how a shorter distance improves the model's precision in item detection and classification, thereby enhancing Grounding DINO's dependability in a variety of urban environments.

Second, the angle at which the image is taken significantly impacts the confidence score. Images taken from an angle closer to congruence with the surface of the manhole give better visibility and clarity of features compared to images taken from oblique angles. For example, detection instances from Sequence 9, as shown in Figure 7, were found to have lower confidence scores compared to instances from other sequences. This can be attributed to the fact that this image

was taken perpendicular to the road; hence, the shape of the manhole was somewhat distorted, while the other sequences required taking pictures as the car moved directly parallel to the road, resulting in clearer and relatively uniform views. These observations accentuate the issue of how enhancing data acquisition methods can effectively improve detection capability and imply the potential for improving image acquisition angles or supplementing with additional sensing modalities under difficult conditions in the future. Moreover, road surface textures and illumination played a role in detection performance. Brightly lit urban areas provided more distinct edge features, aiding detection, whereas shadowed environments or high-glare surfaces introduced challenges. Future work could explore adaptive exposure settings or polarization filters to counteract these effects, ensuring more uniform detection reliability across varying environmental conditions.

This offered further validation of the accuracy of localization by triangulation, which in most instances achieved sub-meter precision. Table 2 summarizes positional errors with a mean of 0.36 m in easting and 0.34 m in northing coordinates. The RMSE for the position was 0.53 m, which demonstrated the reliability of the approach. Figures 10-13 include graphical representations of mapped manhole locations overlaid on Google Earth for practical utility of the approach and a visual correspondence with Google Earth imagery. The small discrepancies noted between the triangulated and actual ground truth positions emphasize the methodological rigor in place, particularly for sequences 2 and 14, where average positional inaccuracies were less than 0.4 m. The integration of localized manholes into the GIS platform finally made possible the generation of detailed geospatial maps, as demonstrated in Figures 14-17. The maps provide many important tools in managing urban infrastructures.

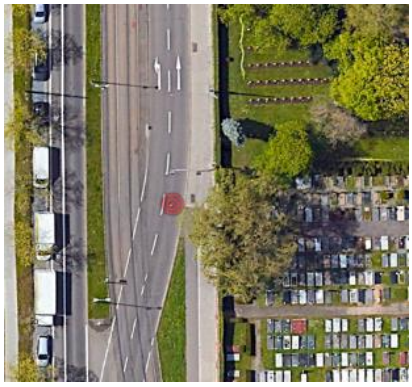


Figure 10. Sequence 2 triangulated manhole projected on Google Earth

Figure 11. Sequence 9 triangulated manhole projected on Google Earth

Figure 12. Sequence 14 triangulated manhole projected on Google Earth

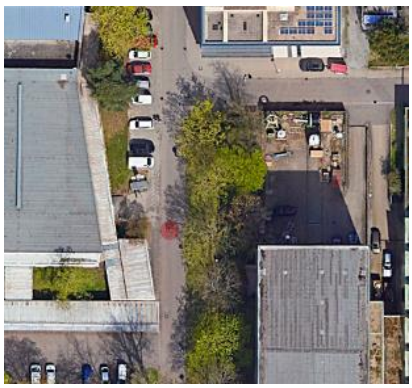


Figure 13. Sequence 17 triangulated manhole projected on Google Earth

Figure 14. Sequence 2 manhole database on ArcGIS

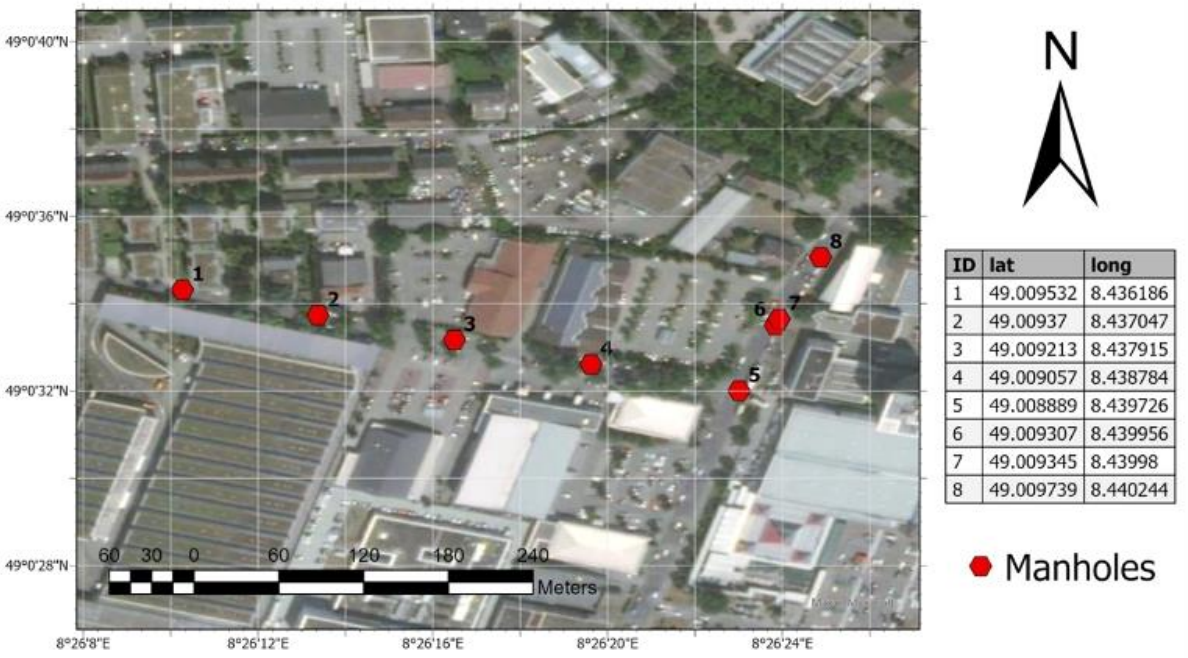
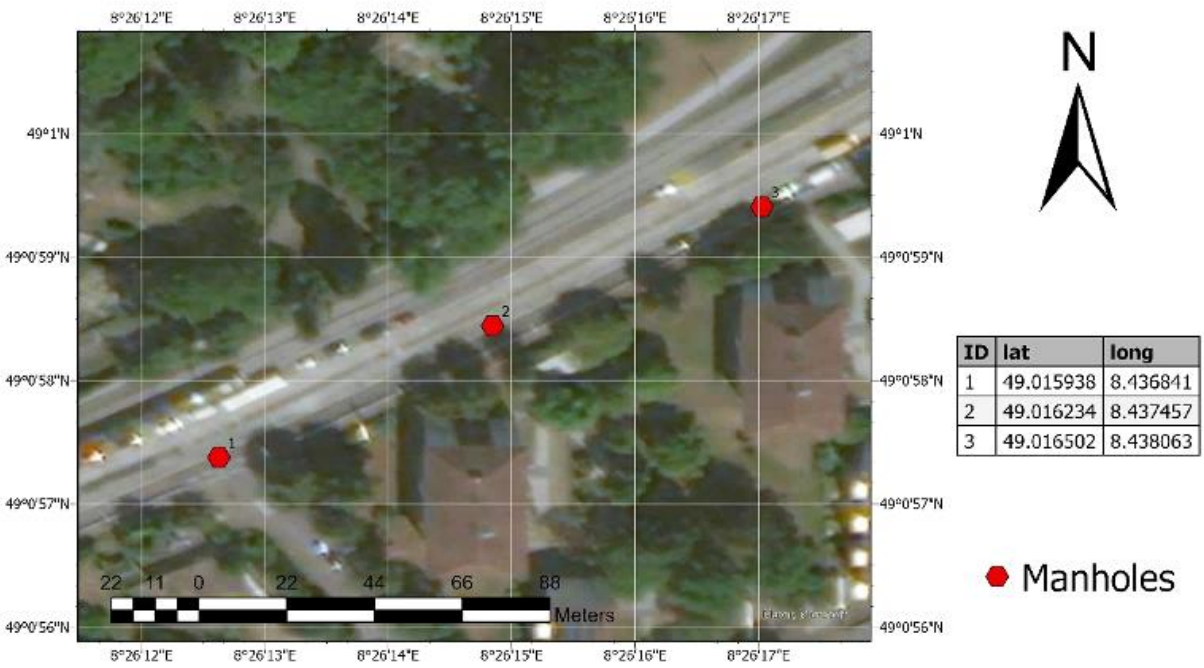


Figure 15. Sequence 9 manhole database on ArcGIS



ID	lat	long
1	49.015938	8.436841
2	49.016234	8.437457
3	49.016502	8.438063

Figure 16. Sequence 14 manhole database on ArcGIS

ID	lat	long
1	49.012193	8.411894
2	49.012345	8.411912
3	49.012789	8.411957
4	49.013008	8.411986
5	49.01306	8.411889
6	49.012978	8.413593
6	49.013021	8.413588
8	49.013038	8.41377

Figure 17. Sequence 17 manhole database on ArcGIS

Error analysis in Figures 18 to 20, in addition to Table 2, shows that Sequences 2 and 14 consistently outperform Sequences 9 and 17 with respect to positional accuracy. Figure 18 shows the distribution of errors in all sequences. From this figure, it is evident that Sequences 2 and 14 have clusters of errors with tight bounds, centered at lower values; hence, these sequences are more consistent and more reliable with regard to the performance of positional accuracy. On the other hand, Sequences 9 and 17 give a larger range of errors, including some higher deviations. In this regard, Sequence 2 reached a mean positional error of 0.36 m, while Sequence 14 followed at 0.40 m. In addition, Sequence 9 logged a mean error of 0.49 m, while Sequence 17 obtained the largest mean error of 0.51 m, emphasizing the challenges posed by environmental factors in those sequences.

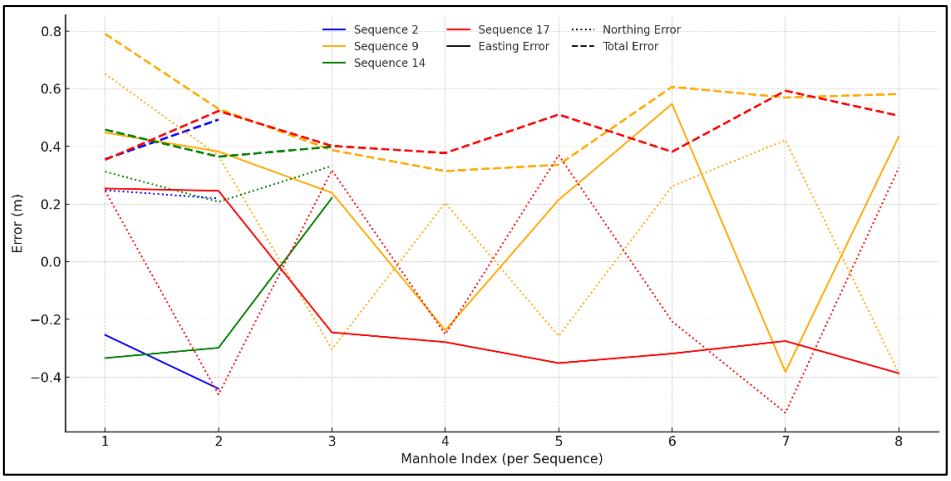


Figure 18. Error variations: Distribution of positional errors for all detected manholes across the four sequences

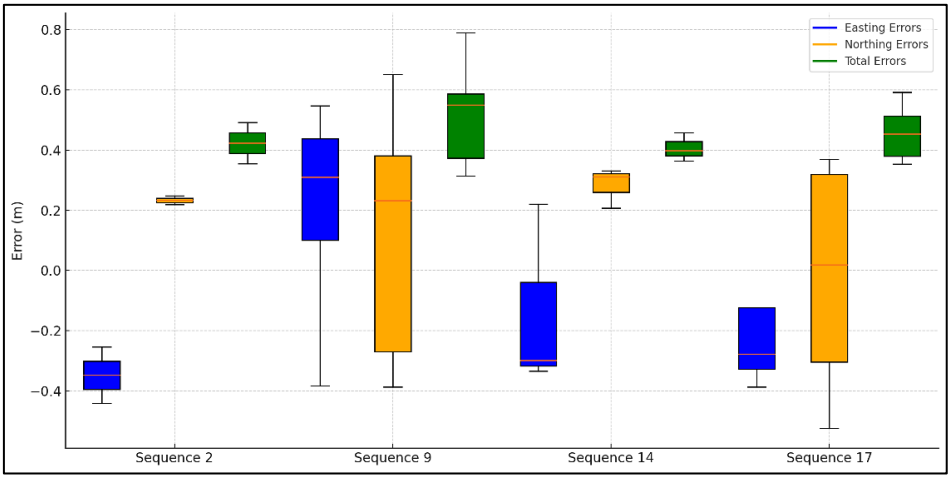


Figure 19. Error component: Breakdown of easting and northing error components for all manholes

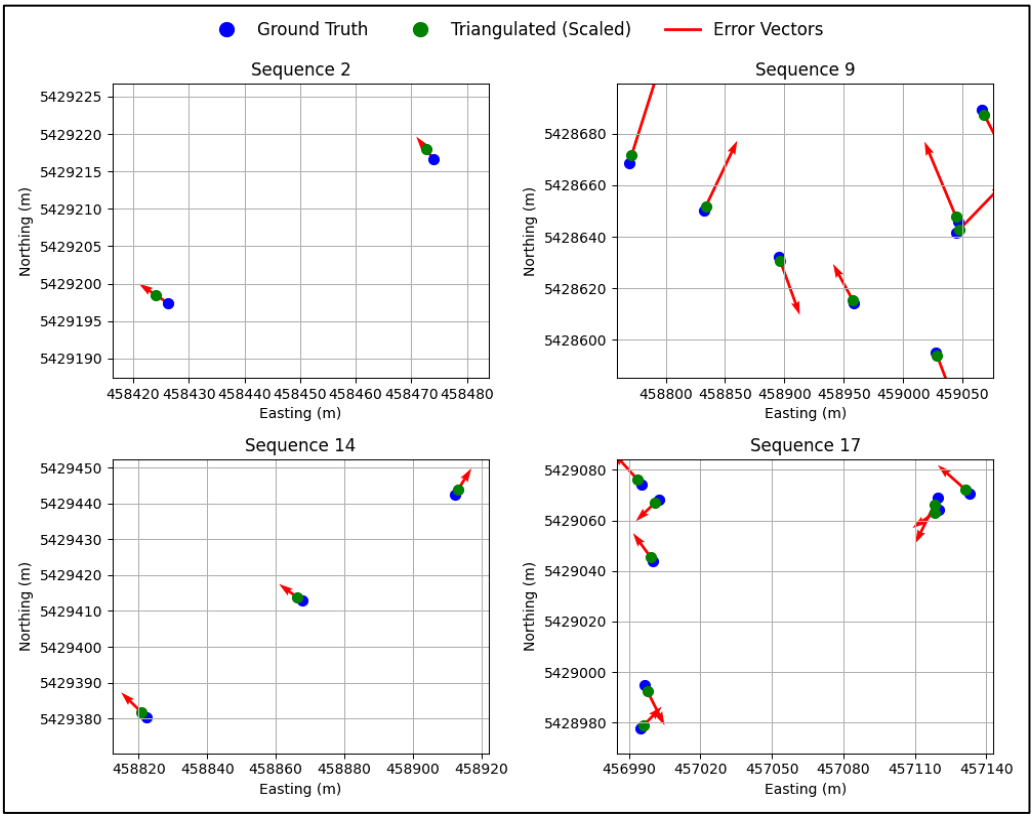


Figure 20. Error vectors: Visualization of error vectors for all detected manholes

These discrepancies are further investigated in Figures 19 and 20. Figure 19 decomposes the errors into their easting and northing components. Sequences 2 and 14 possess smaller error components that show less deviation from each other compared to Sequences 9 and 17, whose error components are far higher in magnitude. Figure 20 shows that the error vectors of Sequences 2 and 14 are much smaller in magnitude and distributed closer together compared to Sequences 9 and 17, whose error vectors are larger and distributed over a larger area. This is explained by the fact that Sequences 9 and 17 suffered from significant degradation due to occlusions caused by tree cover and other obstacles, as can be seen in Figure 9 from Sequence 17. The difference arises due to the fact that Sequences 2 and 14 passed through open-sky areas, being recorded on main streets, and suffered from minimal occlusion, allowing for better GNSS data. Sequences 9 and 17 passed through a lot of tree cover and were recorded over a side road, reflecting a higher positional error. This deep breakdown of error components reveals that while the average error remains within sub-meter precision, small inconsistencies in depth estimation and GNSS signal drift contributed to slight variations. The error distribution suggests that redundant multi-view detections played a key role in maintaining accuracy, particularly in open-sky sequences. These results outline how sensitive the performance of localization is to good environmental conditions and GNSS signals. It is likely that localization stability could be improved even more in environments with obstacles by further integrating sensor fusion techniques, such as combining GNSS with visual-inertial odometry. In addition, future improvement will be devoted to integrating different data sources, such as LiDAR or depth sensors, in order to overcome these limits.

6. Conclusion

This study presents a transformative approach in manhole detection and mapping by integrating the Grounding DINO-based open-world object detection with GIS platforms. The proposed methodology overcomes some of the critical limitations associated with traditional detection techniques, namely, reliance on pre-defined object categories and extensive labeled datasets. It does this by using natural language processing to make object detection adaptable and scalable. Triangulation for precise localization and transformation of detected coordinates to geospatial formats contributes to high accuracy with practical applicability in urban environments. The performance of the approach was thoroughly evaluated employing sequences from the KITTI benchmark, which is well known for the diversity encountered in urban scenarios. Thus, the system reached a sub-meter localization precision, with mean errors in easting and northing of 0.36 and 0.34 meters, respectively. Considering each sequence individually, Sequences 2 and 14 were the most accurate ones, where the mean positional deviations and tightly grouped errors were below 0.4 meters. This was attributed to better conditions, such as open-sky conditions and fewer obstructions to receive more accurate data from the GNSS. By contrast, Sequences 9 and 17 had more dispersed error distributions with mean positional errors of 0.49 and 0.51 meters, respectively. These were considered larger as a result of environmental influences in the form of tree cover and obstacles that hampered the quality of the GNSS signals. The seamless integration of localized manhole coordinates into GIS platforms made it possible to generate accurate geospatial maps. This demonstrates the robustness of the methodology under varied urban conditions and its ability to respond to challenging scenarios.

7. Declarations

7.1. Author Contributions

Conceptualization, I.F.A. and A.A.; methodology, I.F.A., A.A., M.A., and M.S.Y.; software, I.F.A., A.A., and M.S.Y.; Validation, I.F.A., A.A., and M.S.Y.; formal analysis, I.F.A., A.A., M.A. and M.S.Y.; investigation, I.F.A., A.A., and M.A.; resources, M.A.; data curation, I.F.A., A.A., and M.S.Y.; writing–original draft preparation, I.F.A. and A.A.; writing–review & editing, M.S.Y. and M.A.; visualization, M.S.Y. and M.A.; supervision, M.S.Y.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Specific sequences from the KITTI dataset utilized in this study are publicly accessible at *http://www.cvlibs.net/datasets/kitti/*. Additional data supporting the findings, including processed geospatial coordinates and model outputs, can be shared subject to institutional and ethical guidelines.

7.3. Funding

This research study is funded by the Researchers Supporting Project number (RSPD2025R902), King Saud University, Riyadh, Saudi Arabia.

7.4. Acknowledgements

The authors extend their appreciation to the Researchers Supporting Project number (RSPD2025R902), King Saud University, Riyadh, Saudi Arabia, for funding this work.

7.5. Conflicts of Interest

The authors declare no conflict of interest.

8. References

[1] Rasheed, W. M., Abdulla, R., & San, L. Y. (2021). Manhole cover monitoring system over IOT. Journal of Applied Technology and Innovation, 5(3), 1-6.

[2] Alshaiba, O., Núñez-Andrés, M. A., & Lantada, N. (2020). Automatic manhole extraction from MMS data to update basemaps. Automation in Construction, 113, 103110. doi:10.1016/j.autcon.2020.103110.

[3] Oulahyane, A., & Kodad, M. (2024). Advancing Urban Infrastructure Safety: Modern Research in Deep Learning for Manhole Situation Supervision Through Drone Imaging and Geographic Information System Integration. International Journal of Advanced Computer Science and Applications, 15(7), 211–219. doi:10.14569/IJACSA.2024.0150721.

[4] Vishnani, V., Adhya, A., Bajpai, C., Chimurkar, P., & Khandagle, K. (2020). Manhole Detection using Image Processing on Google Street View imagery. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 684–688. doi:10.1109/icssit48917.2020.9214219.

[5] Wei, Z., Yang, M., Wang, L., Ma, H., Chen, X., & Zhong, R. (2019). Customized Mobile LiDAR System for Manhole Cover Detection and Identification. Sensors, 19(10), 2422. doi:10.3390/s19102422.

[6] Boller, D., Moy de Vitry, M., D. Wegner, J., & Leitão, J. P. (2019). Automated localization of urban drainage infrastructure from public-access street-level images. Urban Water Journal, 16(7), 480–493. doi:10.1080/1573062X.2019.1687743.

[7] Mishra, R., Patil, C., Kasture, S., Dhake, D. (2022). Manhole Quality Management and Sensing Using IOT. International Journal of Advanced Research in Science, Communication and Technology, 512–515. doi:10.48175/ijarsct-4533.

[8] Wang, D., & Huang, Y. (2024). Manhole Cover Classification Based on Super-Resolution Reconstruction of Unmanned Aerial Vehicle Aerial Imagery. Applied Sciences (Switzerland), 14(7), 2769. doi:10.3390/app14072769.

[9] Yang, L., Hao, Z., Hu, B., Shan, C., Wei, D., & He, D. (2024). Improved YOLOX-based detection of condition of road manhole covers. Frontiers in Built Environment, 10, 10. doi:10.3389/fbuil.2024.1337984.

[10] M, A., S, K., & T, S. (2022). Smart Manhole Managing and Monitoring System using IoT. International Journal for Research in Applied Science and Engineering Technology, 11(12), 998–1001. doi:10.4108/eai.16-4-2022.2318149.

[11] Xiao, Y., Li, S., Li, Z., & Qu, Z. (2023). Design of an Intelligent Manhole Cover System Based on BeiDou Navigation. 2023 13th International Conference on Information Technology in Medicine and Education (ITME), 505–509. doi:10.1109/itme60234.2023.00106.

[12] Kumar, S. V. S., Padmaja, J. N., Mattaparty, S. H., Ismail, S., Varma, N. M. K., & Vaishnavi, P. (2024). Data-Driven Urban Safety: A CNN-Based Predictive Model for Manhole Hazard Detection. 2024 IEEE Students Conference on Engineering and Systems (SCES), 1–5. doi:10.1109/sces61914.2024.10652445.

[13] Yadav, M., Lohani, B., & Goel, S. (2022). Geometric and radiometric constraints-based extraction of urban road manhole covers and their maintenance-related information using mobile laser scanning data. Geocarto International, 37(27), 16716–16735. doi:10.1080/10106049.2022.2115151.

[14] Rajasekar, A., Aditya, J. J., Sundar, K. S., Suman, M., & Danush, S. (2024, April). Drain Block Detection and Controlling System. In 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT), 1-4. doi:10.1109/IC3IoT60841.2024.10550345.

[15] Liu, Y., Du, M., Jing, C., & Bai, Y. (2013). Design of supervision and management system for ownerless manhole covers based on RFID. 2013 21st International Conference on Geoinformatics, 1–4. doi:10.1109/geoinformatics.2013.6626149.

[16] Commandre, B., En-Nejjary, D., Pibre, L., Chaumont, M., Delenne, C., & Chahinian, N. (2017). Manhole Cover Localization in Aerial Images with a Deep Learning Approach. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1/W1, 333–338. doi:10.5194/isprs-archives-xlii-1-w1-333-2017.

[17] Nan, T., Xiangyang, M., Lin, L., Liu, C., & Xiuhan, J. (2003). Manhole detection and location for urban pavement. Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems, 2, 1552–1555. doi:10.1109/itsc.2003.1252744.

[18] Moccardi, A. (2023). AI Driven Potholes Detection for Equitable Repair Prioritization: Human centred AI-driven methodology as support of road management system. Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice, 56–56. doi:10.1145/3633083.3633224.

[19] Kumar, S. V. S., Padmaja, J. N., Mattaparty, S. H., Ismail, S., Varma, N. M. K., & Vaishnavi, P. (2024). Data-Driven Urban Safety: A CNN-Based Predictive Model for Manhole Hazard Detection. 2024 IEEE Students Conference on Engineering and Systems: Interdisciplinary Technologies for Sustainable Future, SCES 2024. doi:10.1109/SCES61914.2024.10652445.

[20] Pang, D., Guan, Z., Luo, T., Su, W., & Dou, R. (2023). Real-time detection of road manhole covers with a deep learning model. Scientific Reports, 13(1), 16479–16479. doi:10.1038/s41598-023-43173-z.

[21] Zhang, D., Yu, X., Yang, L., Quan, D., Mi, H., & Yan, K. (2023). Data-Augmented Deep Learning Models for Abnormal Road Manhole Cover Detection. Sensors, 23(5), 2676. doi:10.3390/s23052676.

[22] Leni, E. S., Akey, S., Vaishnovi Mane, P. K., R, R. S., & Adere, K. (2025). Yolov8-Based Real-Time Pothole Detection System for Smart Cities: A Multi-Stage Optimization Approach. doi:10.2139/ssrn.5088949.

[23] Zhang, H., Dong, Z., He, A., Zhang, A. A., Wang, K. C. P., Liu, Y., Xu, J., Shang, J., & Ai, C. (2022). Efficient approach to automated pavement manhole cover detection with modified faster R-CNN. Intelligent Transportation Infrastructure, 1, 1. doi:10.1093/iti/liac006.

[24] Lin, G., Zhang, H., Xie, S., Luo, J., Li, Z., & Wang, Y. (2024). Research on Point Cloud Structure Detection of Manhole Cover Based on Structured Light Camera. Electronics, 13(7), 1226. doi:10.3390/electronics13071226.

[25] Qing, L., Yang, K., Tan, W., & Li, J. (2020). Automated Detection of Manhole Covers in MLS Point Clouds Using a Deep Learning Approach. IEEE International Geoscience and Remote Sensing Symposium (IGARSS (2020)), 1580–1583. doi:10.1109/igarss39084.2020.9324137.

[26] Khare, O., Gandhi, S., Rahalkar, A., & Mane, S. (2023). YOLOv8-Based Visual Detection of Road Hazards: Potholes, Sewer Covers, and Manholes. 2023 IEEE Pune Section International Conference (PuneCon), Pune, India, 1–6. doi:10.1109/punecon58714.2023.10449999.

[27] Ma, L., Zhou, M., Wu, Q., Zhang, T., Zhang, H., & Cai, J. (2024). Research on Marker Recognition Method for Substation Engineering Progress Monitoring Based on Grounding DINO. 2024 The 9th International Conference on Power and Renewable Energy (ICPRE), 776–780. doi:10.1109/icpre62586.2024.10768306.

[28] Cai, R., Guo, Z., Chen, X., Li, J., Tan, Y., & Tang, J. (2025). Automatic identification of integrated construction elements using open-set object detection based on image and text modality fusion. Advanced Engineering Informatics, 64, 103075. doi:10.1016/j.aei.2024.103075.

[29] de Moraes Vestena, K., Phillipi Camboim, S., Brovelli, M. A., & Rodrigues dos Santos, D. (2024). Investigating the Performance of Open-Vocabulary Classification Algorithms for Pathway and Surface Material Detection in Urban Environments. ISPRS International Journal of Geo-Information, 13(12), 422. doi:10.3390/ijgi13120422.

[30] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. Computer Vision – ECCV 2024, ECCV 2024, Lecture Notes in Computer Science, 15105, Springer, Cham, Switzerland. doi:10.1007/978-3-031-72970-6_3.