Automatic Detection, 3D Localization, and Semantic Enrichment of Commercial Signboards Using 360° Mobile Mapping Imagery: A Case Study in Temara, Morocco

Hiba Doi[1], Rafika Hajji[1],Imane Jeddoub[2], Roland Billen[2]

[1] College of Geomatic Sciences and Surveying Engineering, Agronomy and Veterinary Institute Hassan II, Rabat 10101, Morocco (doihiba,r.hajji)@iav.ac.ma
[2] GeoScITY, Spheres Research Unit, University of Liège, 4000 Liège, Belgium (i.jeddoub,rbillen)@uliege.be

Keywords: Signboard Detection, YOLO, 3D Geolocation, MMS Imagery, Line of Bearing , GPT-4o

Abstract

The regulation of urban advertising signage is critical for preserving visual harmony and ensuring regulatory compliance in modern cities. This study presents a novel pipeline for the automatic detection, tracking, geolocation, and textual identification of storefront signboards from 360° Panoramic imagery acquired via Mobile Mapping Systems (MMS). We first fine-tune a YOLOv11 object detection model on a custom-labeled dataset of urban scenes, enabling robust identification of signboards across varied viewing angles. To associate detections across consecutive frames, we leverage the integrated YOLOv11 tracking mode, which assigns consistent object IDs based on motion and appearance features. Each tracked instance is then localized in 3D space using a photogrammetric Line of Bearing (LoB) method, relying on known camera poses and pixel coordinates. In parallel, we extract the textual content from each detected sign using advanced GPT-4o Vision, which has demonstrated improved performance in complex visual environment. The proposed pipeline offers a scalable alternative to manual inspection, providing precise spatial and semantic information about urban signage. The pixel-wise projection precision, quantified by an average RMSE of 7.75 pixels (median 7.17px, std dev 2.80px) derived from LoB intersection consistency, confirms the pipeline's reliability for automated urban inventory systems and smart city applications.

1. Introduction

Commercial Signboards represent a widespread form of publicity used by businesses to promote their services directly on the street. However, when these advertising panels are not properly monitored, their chaotic distribution can lead to visual pollution and reduce the aesthetic quality of urban environments. Traditionally, inspecting the presence, compliance, and content of these panels requires human auditors to physically inspect each storefront. This process is not only time-consuming but also resource-intensive. In this work, we propose an automated approach to detect, localize, and identify the textual content of storefront advertising panels using 360° Panoramic imagery acquired by Mobile Mapping Systems (MMS). Our goal is to provide municipalities with a scalable alternative to manual field inspections for monitoring urban signage. This work integrates state-of-the-art computer vision techniques including object detection and tracking, with photogrammetric principles to accurately project detected objects into real-world coordinates. This allows not only identifying the panels in the image but also geolocating them accurately in the urban space. While previous research has mainly focused on the detection and segmentation of buildings or large urban structures in street-level imagery, few studies have addressed the detection of storefront advertising panels. To our knowledge, no prior study, has tackled the extraction of their attributes (such as textual content and business name) and precise spatial location. Our approach is the first to address this gap in the literature through an innovative method by simultaneously identifying, segmenting, reading, and georeferencing these urban objects.

2. Related Work

2.1 Urban Object and Signboard Detection in Computer Vision

Over the past decade, urban object detection has been predominantly addressed using Convolutional Neural Networks (CNNs). Traditional approaches, such as Faster R-CNN (Ren et al., 2016), rely on a two-stage detection process. In the first stage, a Region Proposal Network (RPN) identifies regions of interest, and in the second stage, these proposals are refined by a classifier to deliver precise bounding boxes and object labels. This decoupled process typically results in high detection accuracy, which is essential for analyzing complex urban scenes with diverse elements like vehicles, pedestrians, and building components.

In contrast, the YOLO series of detectors employs a single-stage pipeline that directly predicts bounding boxes and class probabilities from full images (Khanam and Hussain, 2024). While early versions of YOLO had some limitations in precision compared to two-stage methods, the YOLO family has seen significant improvements over time, achieving competitive accuracy along with enhanced efficiency and streamlined design. These advancements in the YOLO approach have made it an attractive option for applications where scalability and performance in detection precision, rather than strict real-time capability, are prioritized.

This general object detection framework has also been adapted to more specific façade□level detection, while extensive research has focused on the segmentation of building features such as windows, doors, and arches (Sezen et al., 2022), relatively few studies have targeted the precise detection of storefront advertising boards. One notable investigation in this area is the study conducted by (Bochkarev and Smirnov, 2019); The

authors proposed a fast CNN-based method for detecting illegal advertising on building façades. They leveraged a curated training dataset composed of rectified images to ensure that building façades were pre-aligned to a frontal view, thereby simplifying the detection task and enhancing performance. However, such an approach inherently restricts applicability to scenarios where strict image rectification is feasible.

2.2 3D Localization of Urban Objects

The georeferencing of detected objects in three-dimensional space has traditionally been performed by segmenting and classifying elements directly within LiDAR point clouds or dense 3D reconstructions (Sun et al., 2018). While this approach excels in detecting large and geometrically distinct urban structures, such as poles, trees, or building edges, it often falls short when applied to fine-grained or visually heterogeneous objects like signboards. These elements tend to be thin, planar, and mounted flush against façades, resulting in sparse or incomplete point cloud representations, especially when the LiDAR sensor is distant or at a steep angle. Additionally, the lack of distinct 3D geometric features makes it difficult for even advanced deep learning models to differentiate signboards from the surrounding façade textures or noise(Sun et al., 2018).

As an alternative, image-based localization techniques have gained attention due to their broader applicability in such cases. Among them, methods relying on monocular or stereo depth estimation, such as MiDaS (Mixed Datasets for Monocular Depth Estimation), estimate per-pixel depth maps directly from RGB images. When combined with calibrated camera poses, these depth cues enable approximate 3D positioning of objects from imagery alone. This strategy has proven particularly useful in urban navigation and semantic mapping tasks, where LiDAR coverage is partial or the objects of interest (like signage) are too small or geometrically subtle to be consistently captured in point cloud data (Ranftl et al., 2020). However, despite its flexibility, monocular depth estimation remains sensitive to occlusions, reflectivity, and scene complexity, requiring post-processing for scale refinement and spatial consistency.

Another category of methods leverages geometric constraints and photogrammetric principles, such as Line of Bearing (LoB) localization. LoB methods infer object positions from known camera poses and pixel coordinates in spherical or panoramic imagery. They are particularly effective in MMS setups, where high-resolution and geo-referenced imagery is available, even in the absence of dense LiDAR. LoB-based approaches are also better suited for localizing signage in urban settings, where precise façade alignment and angular coverage are critical (Doi et al., 2024).

Despite progress across all fronts, current methods rarely combine detection, tracking, and geospatial localization into a unified pipeline. Our approach bridges this gap by integrating YOLO-based object detection, multi-frame tracking, and LoB-based spatial localization, enriched with textual attribution extracted from each signboard.

2.3 Text Extraction in Urban Imagery

Optical Character Recognition (OCR) remains a key component for extracting textual information from storefront signs. While classical OCR systems perform well on clean, well-aligned documents, they tend to struggle in real-world urban scenes, where text may appear in varying fonts, orientations, lighting conditions, and levels of occlusion. Recent advances in deep learning, such as the Convolutional Recurrent Neural Network (CRNN) architecture, have substantially improved robustness to such variability by combining visual feature encoding with sequence modeling. More recently, the emergence of vision-language models (VLMs) such as GPT-4o has opened new possibilities for text recognition in complex visual settings. In a recent benchmark by (Nagaonkar et al., 2025), GPT-4o outperformed conventional OCR engines across multiple domains, including code overlays, advertisements, and broadcast video, achieving lower Word and Character Error Rates in highly dynamic scenes. These results highlight GPT-4o's potential for accurate text extraction in visually challenging urban environments like storefront signage.

3. Methodology

3.1 Overview of the Pipeline

The proposed pipeline is designed to automate the detection, tracking, spatial localization, and textual interpretation of storefront advertising signboards from 360° Panoramic imagery acquired by MMS. This approach leverages deep learning methods and photogrammetry techniques in a modular architecture, as illustrated in Figure 1.

The pipeline consists of six main components:

1. **360° MMS image acquisition:** Panoramic street-level images are collected using a calibrated Mobile Mapping System. Each frame is associated with accurate exterior orientation parameters (EOPs), including camera position and heading, enabling photogrammetric projection of image-space observations into georeferenced coordinates.
2. **Object Detection using YOLOv11:** A custom-trained YOLOv11 object detection model is employed to identify advertising signboards in each panoramic frame. This single-stage detector offers a favorable balance between detection accuracy and computational efficiency across wide field-of-view images.
3. **Multi-frame Tracking via YOLOv11 Tracking Model:** Detected instances are associated across sequential frames using the built-in tracking mode in YOLOv11. This mechanism assigns persistent object IDs based on spatiotemporal consistency, allowing each signboard to be tracked over time, even as viewpoint and appearance vary.
4. **Pixel Refinement via Segment Anything Model (SAM):** To improve the accuracy of pixel-level localization, each detected bounding box is refined using the Segment Anything Model (SAM). The segmentation masks generated by SAM allow precise delineation of signboard contours, reducing background noise and enhancing the reliability of downstream geolocation. The refined mask centroids are used as updated keypoints for 3D localization.
5. **3D Geolocation via Line of Bearing (LOB):** A photogrammetric multi-view LoB method is employed to estimate the 3D position of each signboard. For each tracked instance, bearing rays are constructed from multiple camera positions using the known exterior orientation and the detected pixel coordinates. The signboard's spatial location is estimated by computing the optimal intersection point of these rays in 3D space.

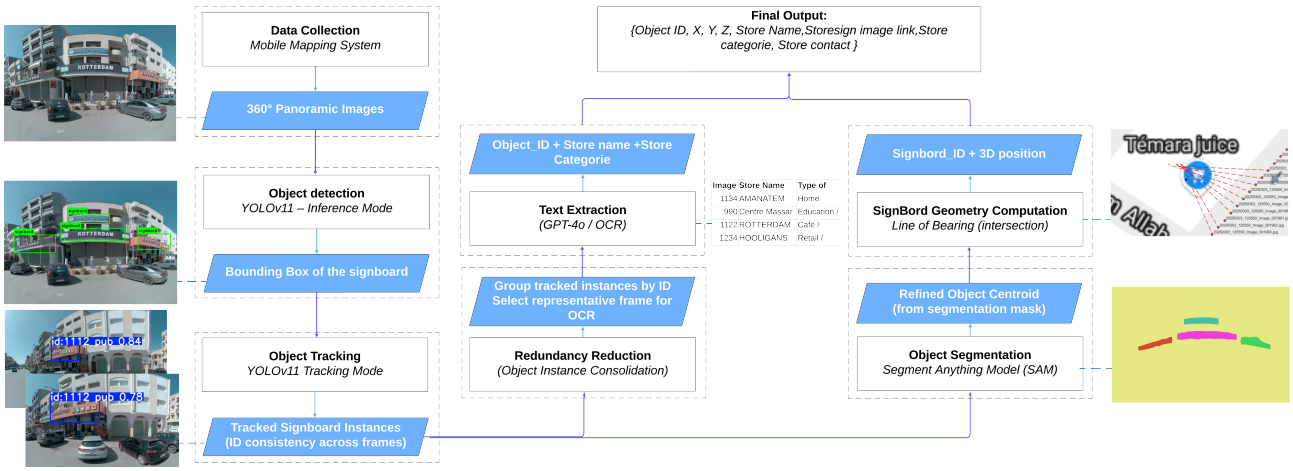


Figure 1. Methodology workflow

6. **Text Extraction using Vision-Language OCR:** Cropped image patches corresponding to each signboard are processed using GPT□4o, a vision-language model, to extract textual content. This approach leverages the model's advanced multimodal capabilities to improve recognition accuracy in complex, multilingual, and unstructured urban environments.

3.2 Dataset Preparation and Annotation

The dataset used in this study was acquired during a structured urban data collection campaign in Temara, a city in the Rabat province of Morocco. The survey was conducted using a Viametris MS-96 Mobile Mapping System (MMS) mounted on a ground vehicle. The system integrates a panoramic camera, a GNSS receiver, and an Inertial Measurement Unit (IMU), providing accurate exterior orientation parameters (EOPs) for each frame, including the 3D position (X, Y, Z) and orientation angles (pitch, yaw, roll) of the camera.

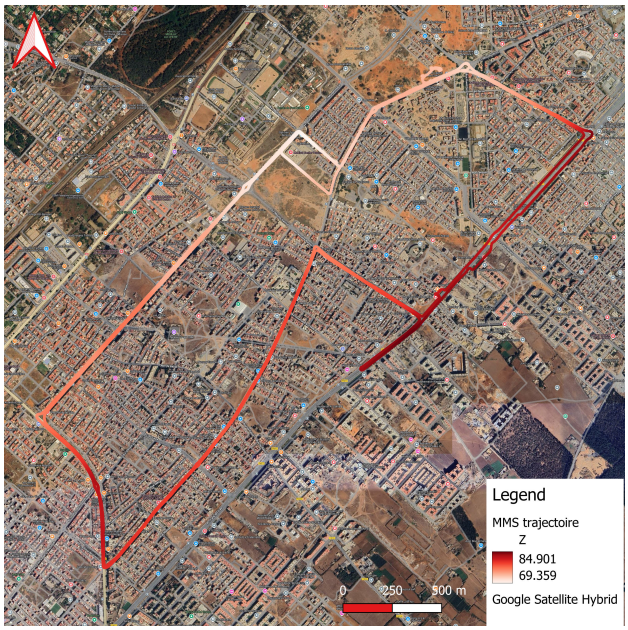


Figure 2. MMS Trajectory – Temara Center | 03 March 2025

A total of 16,072 equirectangular panoramic images were collected during the mission (see Figure 2). Each with a native resolution of 12,600 × 6,400 pixels, later downsampled to 2,560 × 1,260 pixels during the detection phase to optimize training efficiency and GPU memory usage. The imagery is georeferenced using the WGS 84 / Pseudo-Mercator coordinate system (EPSG:3857). Images were extracted from a continuous video stream captured at regular 0.5-meter intervals, resulting in dense spatial coverage along approximately 18 kilometers of urban road network, corresponding to 9 kilometers of street surveyed in both directions (round trip). The spatial trajectory of the survey, color-coded by height (Z), is illustrated in Figure 2, demonstrating the extent and vertical variability of the collected route.

A total of 708 storefront signboards were manually annotated using the YoloLabel tool, following the standard YOLO bounding box format. Labels were applied to each signboard instance with high attention to occlusion, angle, and visibility variations. This annotation process served as the ground truth for training and evaluating the object detection model. All labeled images were split into training (80%) and validation (20%) subsets, ensuring a balanced representation of signage density, orientation, and urban variability.

Annotating the dataset presented challenges specific to urban environments and the nature of equirectangular 360° Panoramic imagery. Although not all scenes were severely impacted, distortions near the top and bottom poles of the panorama complicated the consistent placement of rectangular bounding boxes. In some cases, such as figure 3, signs appeared visibly skewed due to projection effects. This phenomenon tends to be more noticeable when signboards are positioned close to the camera, where local perspective effects are more pronounced in the equirectangular projection. Despite these annotation challenges, the training process benefits from the robustness of YOLOv11, whose architecture is well-suited to handle geometric distortions, partial occlusions, and non-linear signboard shapes common in equirectangular 360° Panoramic imagery.

3.3 Signboard Detection

To automatically detect advertising signboards on storefronts, we fine-tuned a YOLOv11 object detection model a fast, single-stage detector belonging to the "You Only Look Once" family of networks. This model was trained on a custom-labeled dataset.

Training was conducted locally on a workstation equipped with an NVIDIA GeForce RTX 2080 SUPER GPU (8GB VRAM),

Figure 3. Skewed Signboard

using the YOLOv11 framework. The model was trained for 500 epochs with an input image resolution of 640×640 pixels.

The performance of the trained model was evaluated on the validation set using standard object detection metrics. The model achieved a precision of 0.795, recall of 0.724, and a mean Average Precision (mAP) of 0.797 at IoU=0.5. Additionally, the F1 score, which provides a harmonic mean of precision and recall, was calculated to be 0.758, indicating a strong balance between detection accuracy and completeness.

3.4 Implementation of YOLOv11 Tracking

For multi-frame tracking, we utilize the ByteTrack algorithm integrated within YOLOv11's tracking framework (Zhang et al., 2022). ByteTrack associates detections across frames using IoU-based matching combined with Kalman filter motion prediction. Low-confidence detections are recovered through a secondary association step, which is particularly beneficial for partially occluded signboards in 360° imagery. Object identities are maintained using appearance similarity and spatial proximity, enabling robust tracking despite viewpoint variations inherent in panoramic sequences. This module integrates appearance-based similarity and spatial proximity to associate detections over time, ensuring that the same physical object is reliably identified throughout the image sequence. An example of this process is illustrated in Figure 4, where the same signboard is consistently tracked across multiple frames, maintaining a unique ID throughout the sequence.

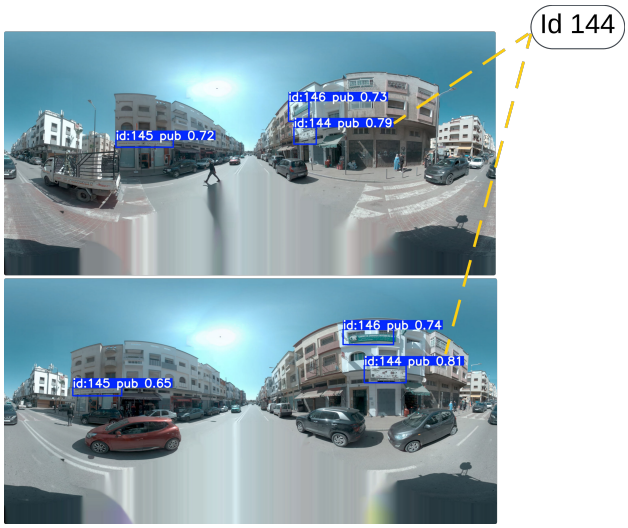


Figure 4. Tracking ID consistency

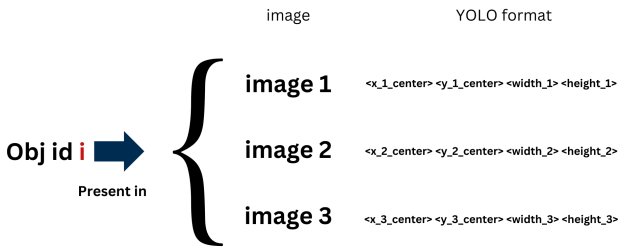


Figure 5. Object tracking with YOLO bounding boxes across multiple images

Tracking plays a critical role in our localization pipeline. By maintaining identity continuity for each detected signboard, we are able to group multiple observations of the same object from different viewpoints/Images, as illustrated in Figure 5. These grouped instances form the basis for our Line of Bearing (LoB) triangulation method, enabling accurate estimation of the object's position in 3D space.

Thus, the tracking module serves as the critical bridge between 2D detection and spatial refinement. By consistently grouping detections for each signboard across multiple frames, we establish a reliable dataset for subsequent processing. In the next step, we employ the Segment Anything Model (SAM) to refine the pixel coordinates of each detection, calculating a more accurate visual centroid that truly represents the object's geometry, before proceeding to 3D localization using our Line of Bearing (LoB) triangulation strategy.

3.5 Object Segmentation and Centroid Estimation

To improve the spatial accuracy of 3D geolocation, we introduced an intermediate segmentation refinement step. Rather than relying solely on the geometric center of the bounding box produced by the object detector, we leverage the Segment Anything Model (SAM) to extract a pixel-level binary mask of the signboard, allowing for a more accurate estimation of its visual footprint. The SAM model is applied to the cropped region defined by the detected bounding box , as shown in Figure 6, using a box-guided prompt strategy to generate a binary segmentation mask. This process captures the actual shape of the signboard, including cases where the object is elongated, tilted, or partially occluded, scenarios where the bounding box center may deviate significantly from the visual centroid.

After computing the refined pixel centroid for each signboard using the segmentation mask, we merged this output with the corresponding tracking results. This operation allowed us to associate each centroid with its respective object ID, forming a consolidated dataset in which each tracked instance is grouped across multiple panoramic frames.

3.6 Text extraction

We initially explored traditional OCR tools, namely Tesseract, EasyOCR, and PaddleOCR, to extract textual content from storefront signs, in an effort to evaluate whether free, open-source alternatives could meet the needs of our application. Although we were aware of the superior performance of GPT-4o, we sought to empirically assess the viability of these tools. While they

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-73-2025 | © Author(s) 2025. CC BY 4.0 License.

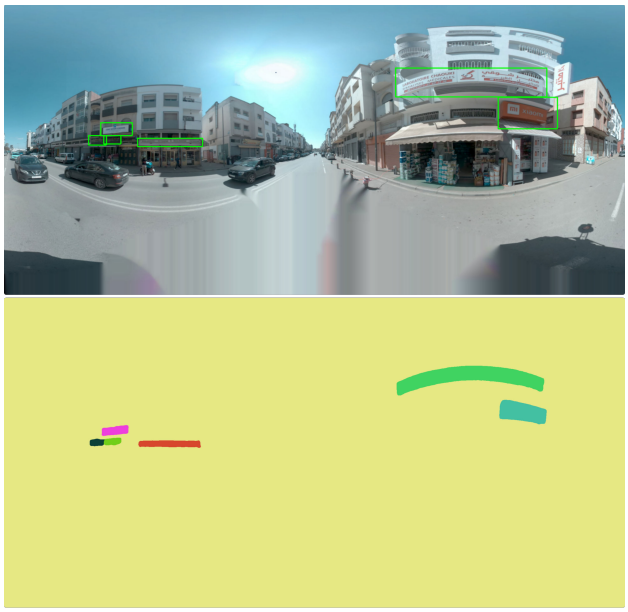


Figure 6. SAM model is applied to the cropped region defined by the detected bounding box

performed reasonably well on clean, front-facing signs, they showed significant limitations when applied to the multilingual, angled, and visually complex signage typical of urban environments.

To address these challenges, we adopted GPT-4o, a vision-language model, to directly interpret the content of cropped signboard images. This approach significantly improved recognition accuracy. GPT-4o was able to return structured semantic fields, including the store name, business category, and contact information, even when the text appeared in Arabic, French, or English, or when stylized fonts and partial occlusion were present.

Figure 7 presents visual sequences of five tracked storefront signboards, where each row corresponds to a distinct detected instance (e.g., ID_14, ID_63, etc.). For each sequence, the median image from the chronological tracking sequence was selected and submitted to the GPT-4o vision API. This median selection strategy automatically identifies the most central viewpoint, typically corresponding to the most frontal view with optimal text visibility and minimal perspective distortion, as the mobile mapping system passes the signboard location. This selective submission approach optimizes semantic extraction accuracy while reducing the number of API calls required. The extracted information, including normalized store names, categorized business types, and structured contact details, is summarized in Table 1.

object_id	store_name	store_category	contact_info
14	MICHWAT BOUAFOUD	Food specializing in roast chicken	
63	Laboratoire Chaouki	D'Analyses Médicales	{'phone': '+212 (0) 5 37 60 72 36', 'email': 'labochaouki@gmail.com'}
136	Banque Populaire	Bank	
146	Cabinet de Chirurgie Dentaire Al Kadi	Dental Surgery	{'phone': '05.37.58.07.63'}

Table 1. Structured text extraction results using GPT-4o

3.7 Tracking-Informed LoB Triangulation: Motivation and Design

As illustrated in Figure 9, our method employs multi-frame tracking to associate object identities prior to Line of Bearing (LoB)

Figure 7. Tracked visual sequences of storefront signboards for five detected instances (ID_14, ID_63, ID_136, ID_146, and ID_432), used for semantic enrichment and spatial mapping

triangulation. This enables us to compute LoBs solely for the frames where an object is consistently tracked and segmented. In contrast, classical LOB-based methods such as (Doi et al., 2024, Li et al., 2022) generate LoBs for all detected features across frames and rely on post-hoc spatial filtering to remove ghost intersections (see Figure8). By focusing only on frames with verified object presence, our approach significantly reduces the search space and computational burden of constraint-based algorithms, eliminating the need for extensive thresholding or clustering steps and leading to more robust and efficient localization.

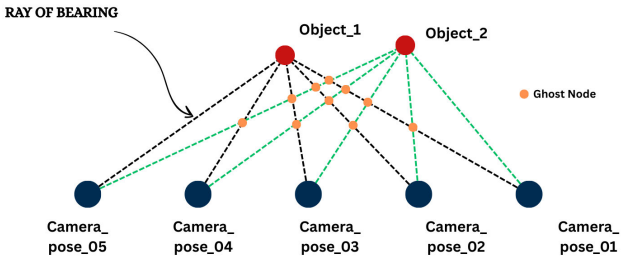
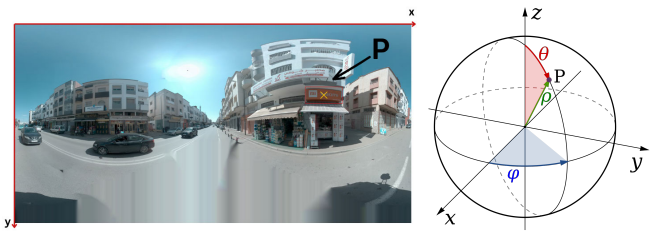


Figure 8. Illustration of classical LoB-based triangulation. Rays are generated for all detected features across frames, leading to multiple false intersections, known as "ghost nodes", due to the absence of object identity association

3.8 Line of Bearing Angle Calculation and Geospatial Coordinate Transformation

First, we convert the pixel coordinates (x,y) from the panoramic image into spherical coordinates (φ, λ), where φ represents the azimuth angle and λ the elevation angle. This transformation is illustrated in Figure 10 (left to right). A detailed explanation of the full pipeline, from pixel coordinates to spherical, then Cartesian, and finally world coordinates-can be found in (Doi et al., 2024).



((a)) ((b))

Figure 10. This figure illustrates the process of localizing a detected signboard point from image space to spherical coordinates

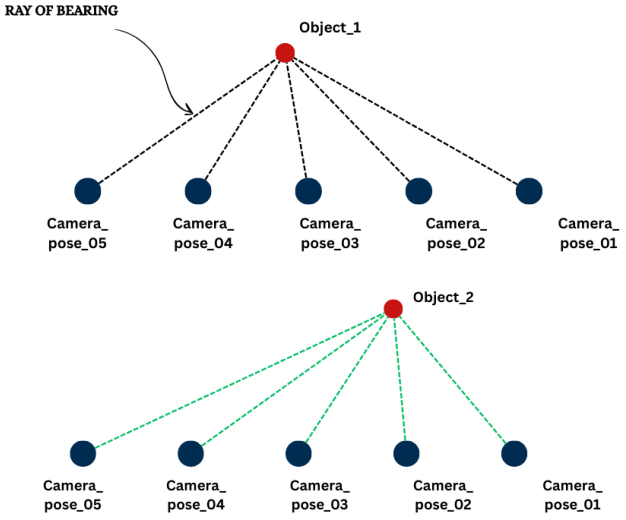


Figure 9. Illustration of our tracking-informed LoB triangulation. By leveraging consistent multi-frame object identities, LoBs are computed only for verified detections, resulting in accurate object localization without spurious ghost intersections

In Figure 10, subfigure (a) shows the point P identified in pixel coordinates (w,h) within a 360° panoramic image. Subfigure (b) illustrates the same point P represented in spherical coordinates by projecting it onto the unit sphere, yielding the azimuth and elevation angles that define its direction in 3D space.

The transformation from image space coordinates to world coordinates is given by:

$$\begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = s \cdot R \cdot \begin{bmatrix} x_c \\ y_c \\ -z_c \end{bmatrix} + \begin{bmatrix} x_{\text{cam}} \\ y_{\text{cam}} \\ z_{\text{cam}} \end{bmatrix} \tag{1}$$

$$R = Rz(yaw) \cdot Ry(pitch) \cdot Rx(roll)$$

where s is the depth coefficient, R is the rotation matrix, (x_c, y_c, z_c) are the image space coordinates, and $(x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}})$ are the camera position coordinates in the world frame.

Using Equation (1), we obtain the 3D projection rays corresponding to each detected point in image space. From these, we extract the line of bearing (LoB) using Equation (2), and the vertical angle using Equation (3). The LoB allows us to estimate the (x,y) coordinates by intersecting the bearing lines computed from multiple frames, as illustrated in Figure 9. The vertical angle is then used to estimate the height (z-coordinate) relative to the (x,y) position, resulting in a full 3D localization of the object.

- Line of Bearing

$$bearing = \arctan\left(\frac{y_c - y_{\text{cam}}}{x_c - x_{\text{cam}}}\right) \tag{2}$$

- Elevation angle

$$\text{V} = \arctan\left(\frac{(-z_c) - z_{\text{cam}}}{\sqrt{(x_c - x_{\text{cam}})^2 + (y_c - y_{\text{cam}})^2}}\right) \tag{3}$$

Line of Bearing is represented by l, as shown in Equation 4:

$$l = (x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}}, bearing) \tag{4}$$

This yields a 3D line (ray) originating from the known camera position and extending in the direction of the signboard's pixel centroid. For each object, we collect all such rays from the frames in which it was detected and tracked.

Because each set of rays corresponds to exactly one object, we compute the optimal intersection point among them using a least-squares method. The mean of all pairwise ray intersection points is used to estimate the final (X, Y, Z) coordinates of the signboard in real-world space.

3.9 Fusion of Spatial and Textual Signboard Data

In parallel, each localized signboard is assigned precise 3D coordinates (x,y,z), and linked, via a unique object ID, to semantic attributes such as store name, business category, and contact information, extracted using the GPT-4o Vision API. This integration results in a spatially anchored semantic dataset that combines geometric and textual data for each detected object. The final output of the pipeline is shown in Figure 11,and Figure 12, where each signboard instance is both geolocated in 3D and enriched with its semantic content.

Figure 11. Final 3D positioning and semantic enrichment of signboards:Covers food□service businesses (cafés and small shops)

Figure 12. Final 3D positioning and semantic enrichment of signboards:Concerns the medical / pharmaceutical sector (pharmacies, dental practices)

4. Results and Discussion

4.1 Object Detection Performance

We evaluated the detection performance of the YOLOv11 model on a validation set comprising 20% of our annotated images. The model achieved a mean Average Precision mAP of 0.797, with a recall of 0.724 and a precision of 0.795, resulting in an F1 score of 0.758. These results demonstrate the model's robustness in detecting signboards under varied illumination, occlusion, and 360° Panoramic imagery distortions. A related study by Bochkarev et al. (Bochkarev and Smirnov, 2019), which also focused on signboard detection, reported a best mAP of 0.59 using Faster R-CNN with Inception v2 on rectified façade images in controlled conditions. In contrast, our method operates directly on unrectified panoramic street-view imagery, indicating a significant performance gain under more challenging and realistic acquisition settings.Our method aims to robustly detect signboards across diverse urban settings, ultimately supporting a more scalable and realistic framework for monitoring urban advertising. This broader applicability is essential for real-world applications where strict image alignment cannot be assured.

4.2 Tracking Performance Evaluation

To enable rigorous tracking evaluation, we developed a custom manual annotation platform designed specifically for 360° panoramic imagery tracking tasks. This platform facilitates frame-by-frame annotation of signboard instances with persistent identity assignment across temporal sequences, accommodating the unique challenges of panoramic distortion.

We evaluated tracking performance using standard MOT metrics on a manually-annotated subset of 400 frames containing 662 ground truth signboard detections from our 1,610 frame dataset.

MOT Evaluation Results:

- MOTA (Multiple Object Tracking Accuracy): 0.809
- MOTP (Multiple Object Tracking Precision): 0.693
- IDF1 (Identity F1 Score): 0.804

Trajectory Analysis:

- Mostly Tracked (MT): 38/47 tracks (80.9%)
- Partially Tracked (PT): 9/47 tracks (19.1%)
- Identity Switches: 22 across 662 detections (3.3% switch rate)

The MOTA score of 0.809 indicates robust tracking accuracy despite panoramic distortions, while IDF1 of 0.804 demonstrates effective identity preservation across viewpoint variations. The MOTP score of 0.693 reflects the inherent challenges of precise localization in 360° imagery but remains within acceptable ranges for our application context.

We assessed the tracking performance also through visual inspection. As shown in Figure 4, the majority of signboard instances were consistently tracked across frames, maintaining stable identities despite variations in viewpoint and occlusion. In a few cases, the tracker assigned multiple IDs to the same object across different frames Figure13 or failed to maintain tracking altogether. However, these instances were infrequent and did not significantly affect the downstream 3D localization or semantic linking steps. Overall, the tracking module demonstrated reliable performance under challenging 360° panoramic conditions.

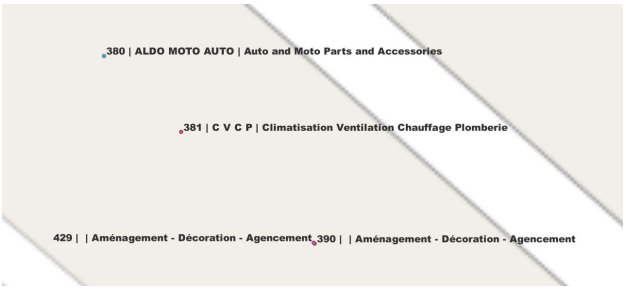


Figure 13. Exemple of tracker assigned multiple IDs to the same object

4.3 Localization Accuracy

To quantify the precision of the targeted LoB triangulation, we computed complementary metrics across multiple evaluation dimensions (Table 2). Internal precision analysis demonstrated substantial improvement when using SAM-derived centroids compared to YOLO bounding box centers, representing a 20% precision enhancement in average RMSE performance. SAM maintains superior performance across the entire range, from best-case to challenging conditions.

Spatial agreement between methods indicates systematic positioning differences, with performance ranging from near-perfect alignment to significant divergence across varying signboard geometries. The pixel-level accuracy demonstrates reliable sub-pixel localization capabilities. These comprehensive metrics confirm that the system achieves facade-mounted signboard localization with precision levels suitable for automated urban inventory applications.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-73-2025 | © Author(s) 2025. CC BY 4.0 License.

Table 2. YOLO Bounding□Box Centres vs. SAM□Derived Centroids for 3D Position Estimation

Metric	Average	Median	Best	Worst	Std. Dev.
Internal Precision (RMSE, m)					
YOLO BBOX centres	0.499	0.516	0.042	0.784	—
SAM□derived centroids	0.398	0.406	0.106	0.750	—
Improvement	**20.2%**	**21.3%**	—	**4.3%**	—
Spatial Agreement Between Methods					
BBOX vs. SAM (RMSE)	0.881	0.869	0.004	2.332	0.431
Pixel□Level Localisation Accuracy					
SAM pixel accuracy	7.75	7.17	1.88	14.58	2.80

4.4 Semantic Enrichment Results

From each localized instance, we extracted semantic content using GPT-4o. Table 1 presents a subset of the semantic extraction results, showcasing successfully parsed store names, business categories, and where available contact information. The extraction was performed using GPT-4o Vision, which demonstrated robust performance even in challenging conditions such as angled views, low resolution, and multilingual signage. In our evaluation, GPT-4o consistently outperformed traditional OCR tools such as Tesseract and EasyOCR on our test set, particularly in terms of accuracy, contextual understanding, and completeness of extracted information.

5. Conclusion

In this paper, we presented a comprehensive pipeline for the automatic detection, tracking, 3D localization, and semantic enrichment of commercial signboards using 360° MMS imagery. Our approach combines modern object detection (YOLOv11), multi-frame tracking, geometry-aware centroid refinement via SAM segmentation, and a targeted LoB triangulation strategy to accurately estimate the spatial coordinates of signboards in real-world coordinates.

Beyond localization, we introduced a semantic layer by leveraging GPT-4o for robust textual content extraction, enabling each spatial instance to be enriched with business-specific attributes such as store name and category. The results demonstrate strong detection performance, reliable tracking consistency, and the ability to generate structured, georeferenced inventories of urban signage.

This pipeline is designed to be modular, scalable, and adaptable to diverse urban environments. The integration of detection, tracking, localization, and semantic extraction supports applications in urban inventory management, signage compliance monitoring, and large-scale geospatial annotation.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Geoptima, especially Ms. Khaoula Kilani for conducting the data collection mission, and Viametris, especially Mr. Omar Motaib for providing the software tools and technical support used in this research.

References

Bochkarev, K., Smirnov, E., 2019. Detecting advertising on building façades with computer vision. *Procedia Computer Science*, 156, 338-346. 8th International Young Scientists Conference on Computational Science, YSC2019, 24-28 June 2019, Heraklion, Greece.

Doi, H., Yarroudh, A., Jeddoub, I., Hajji, R., Billen, R., 2024. Automatic Detection and 3D Modeling of City Furniture Objects using LiDAR and Imagery Mobile Mapping Data. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 125–132.

Khanam, R., Hussain, M., 2024. Yolov11: An overview of the key architectural enhancements.

Li, G., Lu, X., Lin, B., Zhou, L., Lv, G., 2022. Automatic positioning of street objects based on self-adaptive constrained line of bearing from street-view images. *ISPRS International Journal of Geo-Information*, 11(4), 253.

Nagaonkar, S., Sharma, A., Choithani, A., Trivedi, A., 2025. Benchmarking vision-language models on optical character recognition in dynamic video environments.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Sezen, G., Cakir, M., Atik, M., Duran, Z., 2022. Deep learning-based door and window detection from building façade. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 315–320.

Sun, Z., Xu, Y., Hoegner, L., Stilla, U., 2018. Classification of MLS point clouds in urban scenes using detrended geometric features from supervoxel-based local contexts. *ISPRS Annals of Photogrammetry, Remote Sensing Spatial Information Sciences*, IV-2.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. Bytetrack: Multi-object tracking by associating every detection box.