ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XII-4/W1-2026
21st International 3D GeoInfo Conference 2026, 28–29 September 2026, Sofia, Bulgaria

Data-Aware 3D Model Enrichment: A Multimodal Pipeline Combining LiDAR Point Clouds and Geotagged Imagery

Rashid Haffadi[1], Roland Billen[1]

[1]GeoScITY, Spheres Research Unit, University of Liège, 4000 Liège, Belgium - (rhaffadi,rbillen)@uliege.be

Keywords: 3D City Model Enrichment, LiDAR and CityJSON, Object Detection, 3D Ray Casting, Semantic Segmentation, Urban Data Integration.

Abstract

Enhancing 3D city models with facade-level semantics is essential for urban analyses such as flood damage estimation, energy performance modelling, and accessibility assessment. Many existing enrichment methods rely on a single data source and do not systematically produce standardised CityJSON geometry. This study presents a data-aware pipeline that enriches LoD2 building models with windows, doors, and floor levels through two interchangeable workflows, one operating on Mobile Laser Scanning (MLS) point clouds and one on geotagged street-level imagery, that converge to a shared floor estimation and CityJSON enrichment backend. Both workflows employ the Grounding DINO open vocabulary detector in zero shot mode, eliminating the need for task specific training data. Evaluated on data from the Vesdre valley (Belgium), the LiDAR workflow processed 206 buildings and achieved 95.5% and 97.0% within one count accuracy for window and door counts, with 90.8% exact match floor count accuracy. The image workflow processed 338 buildings from a separate area, reaching 93.5% and 95.0% within one count accuracy for windows and doors, with comparable floor estimation, the lower opening accuracy reflecting street-level occlusions. The resulting enriched CityJSON models contain Window, Door, and FloorSurface semantics at LoD3, demonstrating a scalable approach to urban model enrichment that adapts to locally available data.

1. Introduction

Three-dimensional city models play an increasingly important role in urban applications such as flood risk assessment, energy demand estimation, and urban planning (Biljecki et al., 2015). While LoD2 building models, which represent roofs and wall surfaces but generally omit facade details, are increasingly available through national and regional mapping initiatives such as 3DBAG in the Netherlands (Stoter et al., 2020) and large scale reconstruction pipelines like Optim3D (Yarroudh et al., 2023), they often lack the semantic richness needed for many downstream analyses. In particular, the absence of openings (windows, doors) and floor level information limits their use in applications where water ingress points or storey-by-storey building characteristics are required (Dottori et al., 2016, Nouvel et al., 2015).

Enriching these models toward LoD3 requires exploiting complementary data sources. Mobile Laser Scanning (MLS) point clouds provide dense 3D coverage of building facades, enabling the direct geometric extraction of openings (Zolanvari et al., 2019). Street level geotagged imagery from mobile mapping systems offers broad coverage along road networks and provides imagery well suited to deep learning based object detection (Doi et al., 2025). Existing CityJSON models provide the structured geometric context, including wall and roof surfaces, onto which detected features are mapped (Ledoux et al., 2019). Each source alone gives only a partial picture; integrating them within a unified framework is essential for consistent enrichment across heterogeneous urban environments.

However, current methods for facade feature extraction and 3D model enrichment remain limited in several respects. Many point cloud based approaches focus on point level semantic segmentation and do not explicitly produce opening level geometries in standard 3D city model formats (Thomas et al., 2019, Hu et al., 2020), while image based pipelines often stop at 2D detections and only partially address reprojection onto existing buildings (Müller et al., 2007, Hensel et al., 2019, Doi et al., 2025). Most systems also rely on a single data modality and task specific training data, which reduces transferability across study areas.

This paper presents a data-aware pipeline that addresses these limitations by offering two interchangeable processing workflows, one for LiDAR data and one for geotagged imagery, that converge to a shared backend for floor level estimation and CityJSON enrichment. The pipeline is termed data-aware because it explicitly selects the processing workflow according to the locally available input data, while preserving a common downstream enrichment procedure. Both workflows rely on the Grounding DINO open vocabulary detector to identify facade elements without dedicated training data. The main contributions are:

1. A modular, data-aware pipeline that selects a LiDAR or image workflow based on data availability, converging to a shared floor estimation and CityJSON enrichment backend.
2. A constrained 3D ray casting approach that reprojects 2D detections from geotagged images onto CityJSON wall surfaces using frustum based spatial filtering.
3. An end-to-end CityJSON enrichment process that integrates detected openings and estimated floor levels into existing LoD2 models as standardised semantic surfaces (Window, Door, FloorSurface).

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents the experimental setup and results. Section 5 discusses findings and limitations. Section 6 concludes.

2. Related Work

2.1 Opening Detection from Point Clouds

Detecting building openings from point clouds has been approached through both geometric and learning based methods. Traditional approaches extract facade planes using RANSAC or region-growing and identify openings as void regions with reduced point density (Truong-Hong and Laefer, 2014, Zolanvari and Laefer, 2017), which works well on dense Terrestrial Laser Scanning (TLS) data but is sensitive to the sparser and noisier acquisitions typical of Mobile Laser Scanning (MLS). Deep learning has significantly improved point cloud segmentation, with networks such as KPConv (Thomas et al., 2019) and RandLA-Net (Hu et al., 2020) classifying points into semantic categories on urban benchmarks. (Zolanvari et al., 2019) provide an annotated urban LiDAR dataset for facade-level segmentation, and (Wysocki et al., 2023) combine segmentation with parametric modelling to reconstruct LoD3 buildings from MLS data. More recently, (Yarroudh et al., 2024) demonstrated an automated LoD2.2 to LoD3 upgrade by projecting facade point clouds to 2D images and applying Grounding DINO for zero shot opening detection. Despite these advances, many methods still focus primarily on point level outputs and do not explicitly produce structured opening level geometries in standardised CityJSON form.

2.2 Opening Detection from Imagery

Image based facade analysis ranges from classical grammar based parsing (Müller et al., 2007) to modern convolutional detectors. Methods based on Faster R-CNN and YOLO have been applied to window and door detection in street-level photographs (Ren et al., 2015, Redmon et al., 2016, Neuhausen et al., 2018), achieving robust 2D localisation. The recent emergence of open vocabulary models, in particular Grounding DINO (Liu et al., 2024), enables zero shot detection via text prompts and removes the need for task specific annotated datasets. Bridging these 2D detections with 3D localisation remains challenging: monocular depth estimation and Structure-from-Motion provide approximate geometry (Bhat et al., 2023, Schönberger and Frahm, 2016), but relatively few works explicitly reproject their outputs onto existing 3D building models such as CityJSON.

2.3 3D City Model Enrichment

Large scale LoD2 building models are now produced through both public mapping initiatives, such as 3DBAG (Stoter et al., 2020), and automated reconstruction pipelines, such as Optim3D (Yarroudh et al., 2023). CityJSON (Ledoux et al., 2019) provides the standard data model for representing building semantics at increasing levels of detail. Enriching LoD2 models with facade features has been identified as a key research need (Biljecki et al., 2016), and several approaches have been proposed, for instance LoD3 reconstruction from MLS data (Wysocki et al., 2023) or automated LoD2.2 generation using Roofer (Peters et al., 2022). However, these pipelines rely on a single data modality and follow a fixed processing chain. To the best of our knowledge, few existing frameworks adapt their workflow depending on which data sources are available.

2.4 Floor Level Estimation

Floor count and interstorey height are critical inputs for flood damage models such as INSYDE (Dottori et al., 2016) and for energy demand estimation. Common estimation strategies rely on simple heuristics, such as dividing wall height by a fixed constant of about 3 m (Biljecki et al., 2016), or on census based records. Clustering approaches based on elevation histograms of facade points have also been explored (Biljecki et al., 2016), but they typically assume uniform floor heights. Deriving floor levels directly from the elevations of detected openings, where windows at the same storey share a common height band, has not, to our knowledge, been attempted in a multimodal pipeline that combines both point cloud and image based detections.

Table 1 summarises how the proposed approach compares with existing methods.

Method	Point cloud	Imagery	Zero shot	3D enrichment	Data-aware
Zolanvari et al. (2019)	✓				
Wang et al. (2022)	✓				
Yarroudh et al. (2024)	✓		✓	✓	
Wysocki et al. (2023)	✓			✓	
Hensel et al. (2019)		✓		✓	
Proposed	✓	✓	✓	✓	✓

Table 1. Comparison of the proposed pipeline with existing methods.

3. Methodology

The pipeline uses as input an existing LoD2.2 CityJSON building model generated beforehand with Roofer (Peters et al., 2022). This model provides the wall and roof surfaces that are subsequently enriched with openings and floor semantics.

Figure 1 presents the overall architecture of the proposed pipeline. For each building in the study area, the system determines whether MLS point cloud data or geotagged street-level imagery is available and routes the building to the corresponding workflow. The LiDAR workflow (Section 3.1) segments the point cloud, isolates building facades, and detects openings on projected 2D images. The image workflow (Section 3.2) detects openings in photographs and reprojects them onto CityJSON wall surfaces through ray casting. Both workflows produce a common output, a set of 3D bounding boxes per building, that feeds into the shared backend (Section 3.3) for floor level estimation and CityJSON enrichment.

3.1 LiDAR Workflow

3.1.1 Semantic Segmentation. The raw MLS point cloud is first classified into semantic categories using KPConv (Thomas et al., 2019), a deep learning network that operates directly on unstructured 3D points. A pretrained model, trained on the Paris-Lille-3D urban benchmark (Roynard et al., 2018), is applied in direct inference mode without fine tuning on the target domain. The network distinguishes nine semantic classes typical of urban street scenes, including ground, vegetation, buildings, and vehicles. Only points labelled as buildings are retained for further processing.

3.1.2 Building Instance Extraction. The classified building points are then assigned to individual buildings using cadastral footprints as reference geometry. For each point, the nearest footprint polygon is identified through a spatial index (STRtree), and the point is assigned to that building if the horizontal distance falls below a configurable threshold. Points that

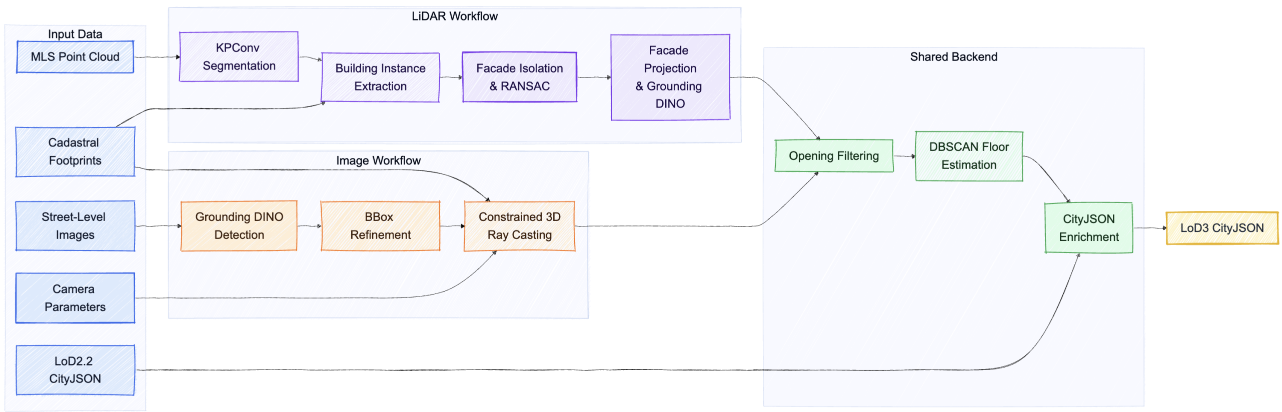

Figure 1. Data-aware pipeline architecture. Buildings are routed to the LiDAR or Image workflow based on data availability; both produce 3D bounding boxes that feed into the shared backend for floor estimation and CityJSON enrichment.

do not lie within the threshold of any footprint are discarded. The result is a set of per building point clouds, each containing only the points that belong to a single structure (Figure 2).

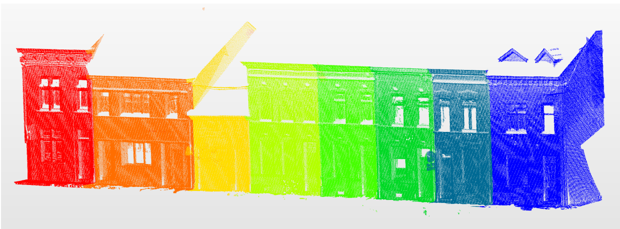

Figure 2. Building instance extraction from MLS data, where points are coloured by building identifier after assignment using cadastral footprints.

3.1.3 Facade Isolation and Plane Fitting. For each building, facade points are isolated by selecting points within a buffer zone around the footprint boundary edges. This buffer retains points close to the building envelope while discarding interior and roof points. The retained point cloud is then decomposed into individual planar segments using RANSAC (Fischler and Bolles, 1981). At each iteration, a plane is fitted and its normal vector is evaluated: planes whose normal is close to the vertical axis, indicating horizontal surfaces such as roofs or ground, are rejected based on a verticality threshold, so that only planes with near horizontal normals, corresponding to facade surfaces, are kept. The process repeats on the remaining points until no significant planar segment is left, and each extracted plane represents one facade face of the building (Figure 3).

3.1.4 Facade Projection and Opening Detection. Each planar facade segment is projected onto a 2D image to enable the use of an image based object detector. Two orthogonal basis vectors are computed from the plane normal, defining a local coordinate frame on the facade surface. The 3D points are projected onto this frame and mapped to pixel coordinates at a fixed resolution, producing a coloured raster image of the facade (Figure 4). This projection follows the same principle as (Yarroudh et al., 2024), where each pixel corresponds to one or more facade points.

The resulting image is passed to Grounding DINO (Liu et al., 2024), an open vocabulary object detector that identifies windows and doors from text prompts without task specific training. The detected 2D bounding boxes are backprojected to 3D using the stored projection parameters, specifically the plane origin, basis vectors, and coordinate ranges, yielding 3D bounding boxes for each opening on the facade surface. Postprocessing removes contained and overlapping detections by comparing bounding box areas and confidence scores.

Figure 3. Facade isolation and plane segmentation in the LiDAR workflow, with each colour representing a distinct RANSAC fitted wall plane.

3.2 Image Workflow

3.2.1 Street Level Detection. Geotagged street-level photographs are processed using Grounding DINO (Liu et al., 2024). Unlike the LiDAR workflow, which uses a large variant of the model on projected facade rasters where the lower image quality demands higher model capacity, the image workflow applies the base model directly to full resolution photographs, which already contain rich visual detail. For each image, the detector identifies windows and doors using text prompts, producing a set of 2D bounding boxes with associated confidence scores and semantic labels. No task specific fine tuning is performed; the model operates in zero shot mode, which allows it to generalise across different building styles and study areas.

3.2.2 Bounding Box Refinement. The axis aligned bounding boxes produced by the detector do not account for perspect-

Figure 4. Facade projection and opening detection. Left: RGB raster image from MLS facade points projected onto the RANSAC fitted plane. Right: Raw Grounding DINO detections with confidence scores, before postprocessing.

ive distortion and may not tightly fit the opening boundaries. A refinement step corrects this by recovering the actual quadrilateral shape of each opening using a Hough Transform that detects dominant line segments within each box, computes their intersections, and selects the four corners that best define the opening contour. In the experiments reported here, this Hough based refinement was used as the default, as it proved robust on the predominantly rectilinear facades of the study area, and its output replaces the original bounding boxes for the subsequent 3D reprojection step.

3.2.3 Constrained 3D Ray Casting.

Each refined 2D quadrilateral is reprojected onto the corresponding CityJSON wall surface through a constrained ray casting procedure. For each corner, a ray is constructed from the camera position through the undistorted pixel coordinate into the scene,

$$\mathbf{p} = \mathbf{o}_{\text{cam}} + t\,\mathbf{d}, \tag{1}$$

where $\mathbf{o}_{\text{cam}}$ is the camera position, $\mathbf{d}$ is the unit ray direction derived from the camera intrinsics and extrinsics, and $t \geq 0$ is the distance along the ray. To limit the search space, the camera's viewing frustum is used to select only buildings within the field of view (Figure 5), and, for each candidate building, wall surface polygons are extracted from the CityJSON model. Intersections between rays and wall polygons are then computed, and the closest valid intersection for each corner defines the 3D opening quad on that wall.

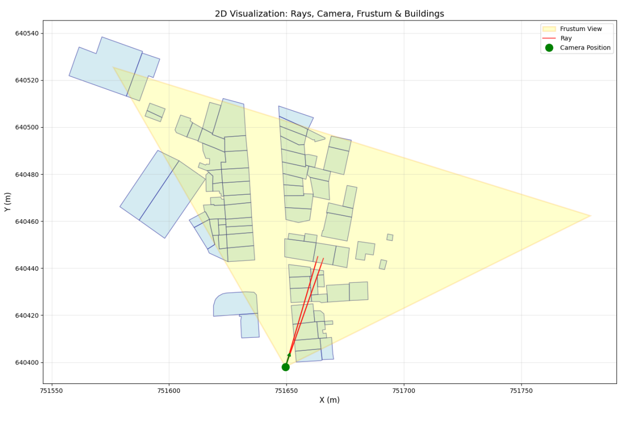

Figure 5. Camera position, viewing frustum, and sample rays intersecting building footprints used to constrain 3D ray casting.

3.3 Shared Backend

3.3.1 Opening Filtering and Validation.

The 3D openings produced by either workflow are filtered to remove low confidence and geometrically invalid detections. Openings below a confidence threshold or with too small an area, unrealistic edge lengths, or strongly nonrectangular proportions are discarded. Duplicate detections of the same physical opening, originating from multiple images or facade planes, are removed by comparing 3D bounding box overlap and retaining the highest confidence instance, and the surviving openings are regularised to axis aligned rectangles on the wall plane to produce CityJSON compliant geometry.

3.3.2 Floor Level Estimation.

Floor levels are estimated from the vertical distribution of validated openings. For each building, the upper edge elevation of every detected opening is extracted. These elevations are clustered using DBSCAN (Ester et al., 1996), where each cluster corresponds to a storey. Clusters that are closer than a minimum floor height are merged to avoid oversegmentation caused by vertically adjacent openings. The number of resulting clusters defines the floor count N_F. When $N_F \geq 2$, the interstorey height IH is derived as the vertical distance between the centres of the two uppermost clusters:

$$IH = \bar{Z}_{N_F} - \bar{Z}_{N_F - 1} \tag{2}$$

where $\bar{Z}_k$ denotes the mean elevation of cluster k. The ground reference elevation Z_G is obtained from the GroundSurface geometry already present in the input LoD2.2 CityJSON model. Floor levels are generated by stepping upward from Z_G at intervals of IH,

$$Z_f = Z_G + f \cdot IH, \qquad f = 1, 2, \ldots, N_F - 1. \tag{3}$$

Each computed floor elevation Z_f is shifted slightly above the upper edge of the corresponding opening cluster (by 0.15 m) (Figure 6) so that the floor surface clears the lintel line. The two uppermost clusters are used for the IH estimate because they are least affected by ground level irregularities such as elevated entrances or sloped terrain. For single storey buildings ($N_F = 1$), no interfloor division is generated.

3.3.3 CityJSON Enrichment.

The final step integrates the detected openings and floor levels into the existing CityJSON model (LoD2.2, as generated by Roofer). Each opening is assigned to a wall surface based on proximity, coplanarity, and 2D overlap with the wall polygon. Openings that do not satisfy all three criteria are discarded. The assigned openings are projected exactly onto the wall plane by removing the component perpendicular to the wall normal:

$$\mathbf{p}' = \mathbf{p} - \big((\mathbf{p} - \mathbf{c}) \cdot \hat{\mathbf{n}}\big)\,\hat{\mathbf{n}} \tag{4}$$

where $\mathbf{c}$ is the centroid of the wall polygon and $\hat{\mathbf{n}}$ is its outward unit normal. Coplanar wall surfaces that represent the same facade are merged prior to opening insertion. Each wall is then split at the computed floor heights to create per storey wall segments. Openings are cut from the wall polygon as holes, and the corresponding window or door geometries are added as child semantic surfaces. The resulting CityJSON model contains WallSurface, Window, Door, FloorSurface, and TINRelief (terrain) semantics, producing a building representation at LoD3 with ground level context.

4. Experiments and Results

The pipeline was evaluated in the Vesdre valley (Chaudfontaine, Belgium) using MLS point clouds acquired along the river corridor, airborne LiDAR for terrain modelling, PICC cadastral footprints projected in Belgian Lambert 72 (EPSG:31370), and LoD2.2 CityJSON models generated with Roofer (Peters et al., 2022). The two workflows were applied to nonoverlapping building sets: the LiDAR workflow processed 206 buildings for which MLS data was available, while the image workflow processed 338 buildings covered by geotagged street-level photographs. The Grounding DINO detection threshold was set to 0.23 for the LiDAR workflow and 0.35 for the image workflow. RANSAC plane fitting used a distance threshold of 0.25 m and a verticality threshold of 0.3 (dot product with the Z axis). Floor level clustering used DBSCAN with $\varepsilon = 1.0$ m, min_samples $= 1$, and a minimum interfloor distance of 2.0 m for cluster merging.

Across the Vesdre valley case study, both workflows achieve high agreement with manual reference counts. The LiDAR branch, evaluated on 206 buildings, reaches 64.1% and 73.3% exact match accuracy for windows and doors, and over 95% within one count accuracy for both, with 90.8% exact floor count accuracy. The image branch, applied to 338 buildings, attains slightly lower opening accuracies (60.9% and 68.0% exact match for windows and doors, with 93.5% and 95.0% within one count) but comparable floor count performance, confirming that floor levels are robustly recovered while opening accuracy depends on data modality and occlusion.

4.1 LiDAR Workflow Results

The LiDAR workflow was evaluated against manual ground truth established for all 206 processed buildings by counting windows, doors, and floor levels through inspection of the projected facade images and Google Street View.

Within one count accuracy is defined as the percentage of buildings for which the predicted count differs from the manual reference by at most one opening.

For window detection, the pipeline achieved an exact match accuracy of 64.1% (132 out of 206 buildings), with 95.5% within one count accuracy. The mean absolute error (MAE) was 0.40 windows per building. Across the study area, a total of 806 windows were predicted against 796 manually identified, indicating a slight overcount (+1.3%). For door detection, the exact match accuracy reached 73.3% (151/206), with 97.0% within one count accuracy and an MAE of 0.30 doors per building. However, the pipeline predicted only 181 doors against 218 in the ground truth, representing a 17.0% undercount, while the combined opening count (987 predicted vs. 1014 manual) remained close to the reference. This asymmetry between window overcount and door undercount indicates a systematic confusion between doors and windows, likely caused by the loss of contextual cues such as ground contact and depth information during 2D rasterisation of facade point clouds.

Floor count estimation reached 90.8% exact match accuracy (187/206), with all 206 buildings having a predicted floor count within one of the manual reference (MAE = 0.09). The 19 underestimated buildings were cases where a single detected opening per storey was insufficient for the DBSCAN clustering to identify separate floor levels, or where the uppermost storey had no visible openings. Considering all three metrics jointly, 117 out of 206 buildings (56.8%) had fully correct predictions for windows, doors, and floor count simultaneously, and the average per building completeness, defined as the mean of the three per building exact match indicators (window, door, floor), was 76.1%.

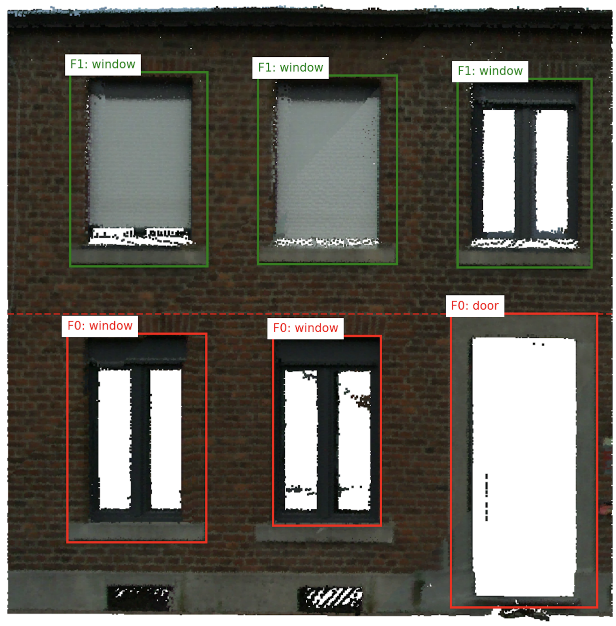

Figure 6. Refined detections with floor level assignment (same facade as Figure 4). Openings are colour coded by DBSCAN assigned storey: red = Floor 0, green = Floor 1.

4.2 Image Workflow Results

The image workflow was evaluated on 338 buildings covered by geotagged photographs from a separate sector of the Vesdre valley. Ground truth was established through a comparable manual annotation protocol, combining Google Street View inspection with site photographs.

For window detection, the pipeline achieved an exact match accuracy of 60.9% (206 out of 338 buildings), with 93.5% within one count accuracy (MAE = 0.46). For door detection, the exact match accuracy reached 68.0% (230/338), with 95.0% within one count accuracy (MAE = 0.35). The overall opening detection accuracy is slightly lower than that of the LiDAR workflow, which is attributable to street-level occlusions: parked vehicles, fences, and garden walls frequently obstruct ground floor openings, reducing both detection recall and the classifier's ability to distinguish doors from windows. The resulting window to door ratio of 6.4:1, compared with 4.4:1 in the LiDAR workflow, confirms that ground level doors are disproportionately affected by occlusion and perspective foreshortening.

Floor count estimation reached 90.8% exact match accuracy (307/338), with all 338 buildings having a predicted floor count within one of the manual reference (MAE = 0.09). The resulting floor count distribution (50.9% single storey, 34.6% two storey, 13.0% three storey, and 1.5% four storey) is consistent with the predominantly residential character of the study area. Considering all three metrics jointly, 50.0% of buildings (169/338) had fully correct predictions for windows, doors, and floor count simultaneously.

4.3 Cross Workflow Comparison

The two workflows were applied to nonoverlapping building sets, precluding a direct per building comparison on the same structures, but the accuracy metrics reveal consistent patterns. The LiDAR workflow achieves slightly higher exact match and within one count accuracies for openings, reflecting the more complete facade observation provided by MLS scanning, while the image workflow compensates with a larger spatial coverage (338 vs. 206 buildings). Floor count estimation performs comparably across both workflows, with 90.8% exact accuracy and 100% within one count accuracy, confirming that the DBSCAN based clustering is robust once a sufficient number of vertically distributed openings is detected. Despite the accuracy advantage of the LiDAR workflow, the consistent output schema across both branches confirms the interchangeability of the data-aware architecture.

4.4 Qualitative Results

The detected openings and estimated floor levels were integrated into the LoD2.2 CityJSON models following the enrichment procedure described in Section 3.3. The resulting models contain WallSurface, Window, Door, and FloorSurface semantics at LoD3, with each opening's surface geometry aligned to the wall surface normal vector. Figure 7 shows a single enriched building with its semantic surfaces: windows (light blue) on two storeys, a door (brown) at ground level, a floor division line separating the storeys, and the roof surface (red). Figure 8 presents an oblique view of the enriched zone, where the same semantic enrichment is applied across the full building stock along the Vesdre valley, rendered with the terrain model and road network. All enriched CityJSON files were validated with val3dity (Ledoux, 2018) to confirm geometric compliance.

Figure 7. Enriched LoD3 CityJSON building model with Window (light blue), Door (brown), FloorSurface, and RoofSurface (red) semantics.

5. Discussion

5.1 Workflow Comparison

The two workflows offer complementary trade-offs. The LiDAR workflow operates directly in 3D, producing geometrically accurate opening positions without requiring camera calibration or image–model alignment. However, it depends on MLS data, which remains expensive to acquire and limited in spatial coverage. The image workflow, by contrast, leverages widely available street-level photographs and can therefore be applied more broadly. Its accuracy is bounded by the quality of the underlying CityJSON model and the precision of the camera extrinsic parameters, which govern the ray casting step (Section 3.2). Because both workflows feed into the same floor estimation and enrichment backend (Section 3.3), they produce identical output schemas regardless of the input modality. This modularity validates the data-aware design principle: enrichment quality depends on the available data rather than on the choice of processing chain.

When both data sources are available for the same building, the architecture supports running both workflows in parallel and merging their outputs; this capability is implemented but has not yet been evaluated, as the current experiments use nonoverlapping building sets (Section 4). The LiDAR workflow is unaffected by lighting conditions and perspective distortion but suffers from point cloud sparsity on distant or oblique facades, whereas the image workflow captures fine visual details at range but is sensitive to camera pose errors. The potential for crossmodal fusion is discussed further in Section 5.4.

5.2 Comparison with Related Work

Table 1 (Section 2) positions the proposed pipeline relative to existing methods. Unlike (Zolanvari et al., 2019), (Wang et al., 2022), and (Wysocki et al., 2023), who rely exclusively on point clouds, the proposed system also supports image based input. While (Wysocki et al., 2023) reconstruct LoD3 CityGML models, the remaining methods do not produce standardised 3D city model outputs. The proposed pipeline generates CityJSON geometry with Window, Door, and FloorSurface semantics, combining multimodal input support with structured output. The closest comparable work is (Yarroudh et al., 2024), who also apply Grounding DINO to projected facade images from LiDAR. The present approach extends this in two directions: (i) a parallel image workflow with constrained ray casting for scenarios where MLS data is unavailable, and (ii) a floor level estimation stage that derives storey boundaries from detected opening elevations rather than relying on fixed height assumptions.

A key advantage of the open vocabulary detection strategy shared across both workflows is that no task specific annotated dataset is required. This zero shot property enables direct application to new study areas without retraining, which contrasts with supervised methods that require labelled facade images or annotated point clouds for each target domain.

Figure 8. Oblique view of the enriched study area with LoD3 buildings, terrain model, and road network.

5.3 Limitations

Several limitations affect the current pipeline. Facade occlusion by vegetation, parked vehicles, or neighbouring buildings prevents openings from being detected on obscured surfaces, directly reducing recall. MLS acquisition typically captures only street facing facades, leaving rear and courtyard elevations unobserved. The Grounding DINO detector can produce false positives on decorative facade elements, ventilation grilles, or wall textures that visually resemble openings, an effect observed in a small number of buildings in the image workflow, where repetitive facade patterns triggered dozens of spurious window detections.

A systematic door to window misclassification was observed in both workflows. This confusion arises because the 2D rasterisation of facade point clouds and the perspective geometry of street-level photographs both discard the contextual cues, including ground contact, handle hardware, and depth discontinuities, that distinguish doors from windows. Addressing this limitation through postclassification rules (e.g., assigning the lowest opening on a facade as a door) or fine tuning the detector on facade specific data would improve classification accuracy without affecting overall detection performance.

Floor count estimation is inherently limited to storeys with at least one visible opening; attic levels without windows and below grade basements are not captured by the current approach. Both workflows achieve over 90% exact match floor accuracy; occasional failures occur when an entire storey has no detected openings, due to occlusion or missing data, preventing the DBSCAN clustering step from identifying that floor level. Finally, the pipeline assumes that the input LoD2.2 model faithfully represents the building envelope; geometric errors in wall positions or orientations propagate into the opening assignment and floor splitting stages.

5.4 Future Work

Several directions could address the limitations identified above. First, crossmodal fusion of LiDAR and imagery on buildings where both data sources are available would increase facade coverage and enable crossvalidation of detections: openings missed in one modality could be recovered in the other, and detections confirmed by both would carry higher confidence. Second, a more detailed geometric evaluation is needed: annotating a subset of buildings with precise 3D opening polygons would allow Intersection-over-Union (IoU) assessment, and measuring actual interstorey heights from building plans or interior surveys would enable evaluation of estimated floor elevations against physical references. Third, extending the experiments to additional European urban datasets with different architectural styles and integrating the enriched CityJSON models into downstream applications such as flood damage estimation or building energy simulation would strengthen generalisability claims and demonstrate practical value beyond geometric accuracy.

6. Conclusion

This paper presented a data-aware pipeline for enriching LoD2 building models with facade openings and floor levels, producing standardised CityJSON geometry at LoD3. The pipeline offers two interchangeable workflows, one for MLS point clouds and one for geotagged imagery, that converge to a shared backend for floor estimation and CityJSON enrichment. A constrained 3D ray casting approach was introduced to reproject 2D detections from photographs onto CityJSON wall surfaces through frustum based spatial filtering. Both workflows rely on the Grounding DINO open vocabulary detector, enabling zero shot facade element recognition without task specific training data.

Evaluated on the Vesdre valley (Belgium), the LiDAR workflow achieved 95.5% and 97.0% within one count accuracy for window and door counts across 206 buildings, with 90.8% exact floor count accuracy. The image workflow, applied to 338 buildings, reached 93.5% and 95.0% within one count accuracy for windows and doors, with comparable floor estimation performance. Both workflows produce identical CityJSON output schemas with Window, Door, and FloorSurface semantics, validating the data-aware design principle.

The enriched models provide building level attributes, including opening counts, positions, and storey boundaries, that are directly usable in flood damage estimation and energy performance simulation. By adapting to locally available data sources, the proposed approach offers a scalable path toward systematic LoD3 enrichment across heterogeneous urban environments.

Acknowledgements

This work was carried out within the framework of the Interreg Meuse-Rhine project FloodWisdom.

References

Bhat, S., Birkl, R., Wofk, D., Wonka, P., Müller, M., 2023. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint*. arXiv:2302.12288.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59, 25–37.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A., 2015. Applications of 3D city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4), 2842–2889.

Doi, H., Hajji, R., Jeddoub, I., Billen, R., 2025. Automatic detection, 3D localization, and semantic enrichment of commercial signboards using 360° mobile mapping imagery: A case study in Temara, Morocco. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4/W6-2025, 73–80.

Dottori, F., Figueiredo, R., Martina, M., Molinari, D., Scorzini, A., 2016. INSYDE: A synthetic, probabilistic flood damage model based on explicit cost analysis. *Natural Hazards and Earth System Sciences*, 16(12), 2577–2591.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 226–231.

Fischler, M., Bolles, R., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Hensel, S., Goebbels, S., Kada, M., 2019. Facade reconstruction for textured LoD2 CityGML models based on deep learning and mixed integer linear programming. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-2/W5, 37–44.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11108–11117.

Ledoux, H., 2018. val3dity: Validation of 3D GIS primitives according to the international standards. *Open Geospatial Data, Software and Standards*, 3(1), 1.

Ledoux, H., Arroyo Ohori, K., Kumar, K., Dukai, B., Labetski, A., Vitalis, S., 2019. CityJSON: A compact and easy-to-use encoding of the CityGML data model. *Open Geospatial Data, Software and Standards*, 4(1), 4.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2024. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer.

Müller, P., Wonka, P., Haegler, S., Ulmer, A., Van Gool, L., 2007. Image-based procedural modeling of facades. *ACM Transactions on Graphics*, 26(3), 85.

Neuhausen, M., Herbers, P., König, M., 2018. Window detection in facade images for risk assessment in tunneling. *Automation in Construction*, 96, 325–338.

Nouvel, R., Brassel, K.-H., Bruse, M., Duminil, E., Coors, V., Eicker, U., Robinson, D., 2015. SimStadt, a new workflow-driven urban energy simulation platform for CityGML city models. *CISBAT 2015*, Lausanne, Switzerland.

Peters, R., Dukai, B., Vitalis, S., van Liempt, J., Stoter, J., 2022. Automated 3D reconstruction of LoD2 and LoD1 models for all 10 million buildings of the Netherlands. *Photogrammetric Engineering & Remote Sensing*, 88(3), 165–170.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28.

Roynard, X., Deschaud, J.-E., Goulette, F., 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research*, 37(6), 545–557.

Schönberger, J., Frahm, J.-M., 2016. Structure-from-Motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113.

Stoter, J., Peters, R., Commandeur, T., Dukai, B., Kumar, K., Ledoux, H., 2020. Automated reconstruction of 3D input data for noise simulation. *Computers, Environment and Urban Systems*, 80, 101424.

Thomas, H., Qi, C., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L., 2019. KPConv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6411–6420.

Truong-Hong, L., Laefer, D., 2014. Quantitative evaluation strategies for urban 3D model generation from remote sensing data. *Computers & Graphics*, 49, 82–91.

Wang, Y., Chen, Q., Zhu, Q., Liu, L., Li, C., Zheng, D., 2022. A survey of mobile laser scanning applications and key techniques over urban areas. *Remote Sensing*, 14(16), 3909.

Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U., 2023. Scan2LoD3: Reconstructing semantic 3D building models at LoD3 using ray casting and Bayesian networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6547–6557.

Yarroudh, A., Kharroubi, A., Billen, R., 2023. *Optim3D: Efficient and scalable generation of large-scale 3D building models*. Lecture Notes in Geoinformation and Cartography, Springer, Cham.

Yarroudh, A., Kharroubi, A., Billen, R., 2024. Automatic upgrade of 3D building models to LoD3 using 3D point clouds and Grounding DINO. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2/W8-2024, 471–476.

Zolanvari, S., Laefer, D., 2017. Slicing Method for curved facade and window extraction from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119, 334–346.

Zolanvari, S., Ruano, S., Rana, A., Cummins, A., da Silva, R., Rahber, M., Smolic, A., 2019. DublinCity: Annotated LiDAR point cloud and its applications. *Proceedings of the 30th British Machine Vision Conference (BMVC)*.