The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W6-2025
3rd International Workshop on Evaluation and BENCHmarking of Sensors, Systems and
GEOspatial Data in Photogrammetry and Remote Sensing (GEOBENCH), 20–21 November 2025, Wroclaw, Poland

CityJSON LOD3.3 Enrichment Using Zero-Shot Learning on Mobile Mapping Data

Eline Deblock[1], Suzanna Cuypers[1] and Maarten Bassier[1,*]

[1]Dept. of Civil Engineering, –Geomatics Section, KU Leuven - Faculty of Engineering Technology, Ghent, Belgium
eline.deblock1@student.kuleuven.be, suzanna.cuypers@kuleuven.be, maarten.bassier@kuleuven.be

Keywords: GIS, Semantic Segmentation, CityGML, Deep Learning, Point Clouds, Mobile Mapping.

Abstract

This paper presents an automated workflow for enriching existing LoD2 building models to LoD3 by integrating aerial LiDAR, mobile mapping imagery, and zero-shot vision–language models. The approach combines TU Delft's Geoflow for geometric reconstruction with Grounding DINO for façade element detection, followed by homography-based perspective correction and spatial reasoning filters to merge redundant detections. Parameter studies demonstrate that optimized Geoflow configurations achieve sub-decimeter accuracy, while the zero-shot detector reaches an average detection score of 83% with a false alarm rate below 10%. The final CityJSON models, validated through CJVal, show 95% geometric and semantic compliance with international standards. The proposed proof of concept demonstrates scalable, data-driven LoD3 reconstruction without retraining, bridging computer vision and geospatial modeling for large-scale urban digital twins.

1. Introduction

The generation of increasingly dense and detailed city models is pivotal for advancing urban informatics, supporting a variety of applications such as urban design studies, energy performance analysis, and Urban Building Energy Modeling (UBEM) (Rossknecht et al., 2020). Many cities have already been mapped at Levels of Detail (LoD) 0, 1, or even 2, leveraging high-resolution unmanned aerial vehicle (UAV) nadir imagery. However, enriching these models to LoD3, incorporating intricate details such as windows, doors, and other façade elements, remains an active area of research (Wysocki et al., 2024).

In this work, we propose a novel automated pipeline designed to upgrade existing LoD1 and LoD2 city models to LoD3 with mobile mapping data (Figure 1). Our approach combines aerial and terrestrial LiDAR datasets with imagery to identify and model missing details such as windows and doors. By aligning all datasets and their detections, we produce enriched CityJSON (Ledoux et al. 2019) outputs containing LoD3 models that capture detailed exterior features of urban structures.

2. Related Work

2.1 Reconstruction of City Models from Aerial and Terrestrial Datasets

Urban 3D reconstruction has progressed considerably with the proliferation of large-scale spatial datasets obtained from both aerial and terrestrial sensors. Aerial Laser Scanning (ALS) and photogrammetric techniques have long been employed to generate LoD1 and LoD2 models (Figure 2), effectively

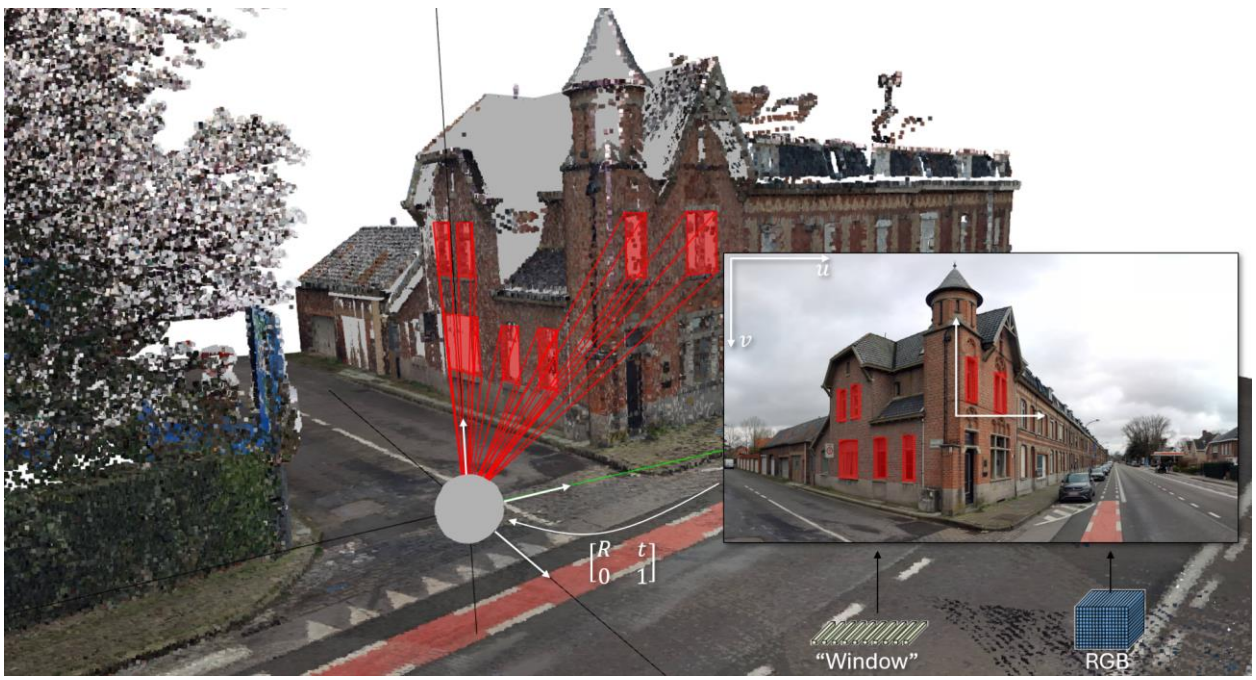
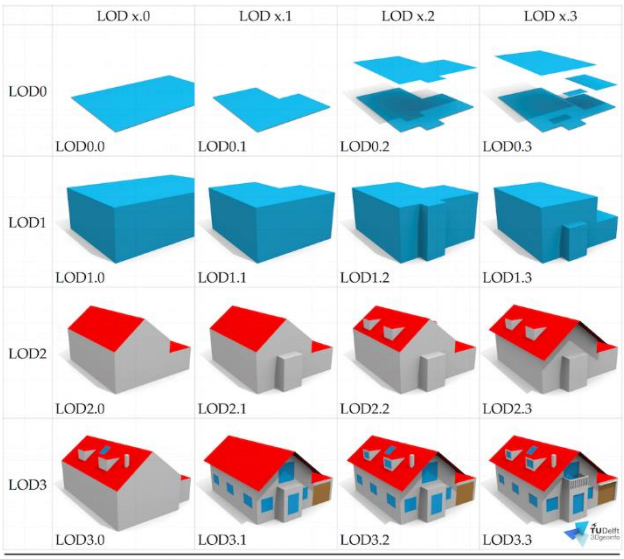


Figure 1. Overview of the proposed LoD3 enrichment framework using mobile mapping data to detect windows and doors improving existing city models.

capturing roof geometries and building footprints across extensive areas (Peters et al., 2022, Jayaraj et al., 2018). While these methods ensure broad coverage and positional accuracy, they often lack sufficient facade detail due to limited visibility from overhead perspectives. To address this, Mobile Laser Scanning (MLS) and mobile mapping imagery have become crucial for capturing high-resolution facade information (Zhang et al. 2024). MLS provides dense geometric data of vertical structures, whereas panoramic imagery contributes color and texture information for improved realism. Integrating aerial and terrestrial datasets enables a more complete representation of buildings, yet challenges persist regarding data registration, as even minor misalignments between datasets can cause geometric distortions in the fused models (Lewandowicz et al., 2022).



Requirements	Refined levels of detail															
	0.0	0.1	0.2	0.3	1.0	1.1	1.2	1.3	2.0	2.1	2.2	2.3	3.0	3.1	3.2	3.3
Individual buildings		•	•	•		•	•	•	•	•	•	•	•	•	•	•
Large building parts (>4 m, 10 m²)		•	•	•		•	•	•	•	•	•	•	•	•	•	•
Small building parts, recesses and extensions (>2 m, 2 m²)			•	•			•	•		•	•	•	•	•	•	•
Top surface(0)			S	M	S	S	S	M								
Explicit roof overhangs (if >0.2 m)												•		•	•	•
Roof superstructures(1) (larger than 2 m, 2 m²)											•	•	•		•	•
Other roof details (e.g. chimneys >1 m)													•		•	•
Openings(2) (>1 m, 1 m²)													R	W	•	•
Balconies (>1 m)														•	•	•
Embrasures, other façade and roof details, and smaller windows (>0.2 m)																•

(0) Applicable only to LOD0.y and LOD1.y: S—Single top surface; M—Multiple top surfaces if the difference in height of the extruded building elements is significant (larger than 2 m).
(1) It includes dormers and features of comparable size and importance (e.g. very large chimneys).
(2) R—only openings on roofs; W—only openings on walls. In R, openings on dormers are not required.

Figure 2. CityGML 2.0 LoDs industry standard for conveying the grade of 3D city models (Biljecki et al., 2016).

2.2 Joint Processing of LiDAR and Imagery in Enrichment Frameworks

The fusion of LiDAR and imagery has emerged as a robust approach for enriching existing 3D city models to higher levels of detail, particularly LoD3, which includes windows, doors, and other facade elements (Yang et al., 2015, Wen et al. 2019). LiDAR contributes accurate geometric structure, while imagery provides complementary spectral and semantic information (Behley et al., 2019). Hybrid frameworks typically integrate the two by performing ray-casting or visibility analyses on LiDAR data and object detection on images using convolutional neural networks (CNNs) (Wysocki et al., 2023, Pantoja-Rosero et al., 2022). These observations are then merged through probabilistic reasoning or graph-based optimization to obtain refined 3D representations (Froech et al., 2014). The joint processing of these modalities increases robustness against occlusion and lighting variation and improves the semantic segmentation of urban scenes. However, the accuracy of such enrichment frameworks depends strongly on sensor calibration and spatial alignment, as errors in image orientation or LiDAR registration can propagate through the entire reconstruction pipeline (Zhou & Neumann, 2013).

2.3 Detection Frameworks and Zero-Shot Learning for Scene Enrichment

The developments in computer vision have introduced flexible object detection paradigms that allow automated model enrichment without domain-specific training. Traditional supervised frameworks, such as Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017), achieve high detection accuracy when trained on extensive labeled datasets but often fail to generalize to new building styles or sensor modalities. To overcome these limitations, zero-shot vision–language models, including Grounding DINO (Liu et al., 2023) and CLIP-based models (Radford et al., 2021), have been adopted for facade enrichment tasks. These approaches leverage natural-language prompts to guide detection, enabling the identification of unseen objects without retraining. While zero-shot methods demonstrate exceptional adaptability, they can exhibit lower geometric precision, especially under occlusion or extreme perspective distortion, necessitating corrective measures such as homography rectification. Nevertheless, their scalability and minimal data requirements position them as promising tools for efficient and automated enrichment of LoD2 to LoD3 models in urban environments.

3. Methodology

The developed workflow automates the enrichment of existing city models from LoD2 to LoD3 by integrating LiDAR, panoramic imagery, and zero-shot detection. The process consists of four main stages (Figure 3): (1) preprocessing and semantic segmentation of source data, (2) reconstruction of LoD2 models using Geoflow, (3) zero-shot detection and 3D projection of facade elements, and (4) LoD3 model enrichment and validation in CityJSON.

3.1 Extraction of Building Footprints and Roof Structures

The workflow begins with parsing a building footprint dataset (e.g., a LoD0 City Database) to identify parcels of interest. In cases where no such database is available, semantic segmentation of LiDAR or image data is employed to extract building-related points, specifically roofs and vertical outlines. Let the point cloud be represented as:

$$P = \{p_i = (x_i, y_i, z_i, f_i) \mid i = 1, \dots, N\} \quad (1)$$

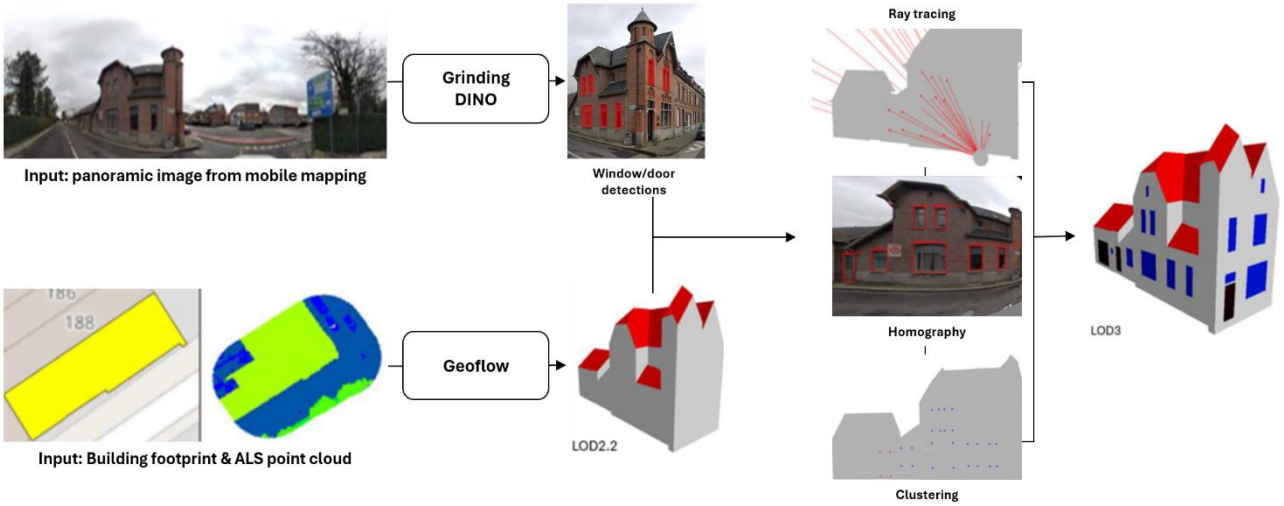


Figure 3. LOD3 enrichment methodology including the Geoflow LoD2 reconstruction, Zero-shot detection and LoD3 enrichment from mobile mapping data.

where f_i denotes the feature vector of point p_i. The goal of semantic segmentation is to assign each point a class label:

$$c_i = \underset{\{c \in C\}}{argmax}\, P(c \mid f_i, \theta) \quad (2)$$

with C = {roof, façade, ground, vegetation, others} and parameters θ learned by a deep learning segmentation model. Several pretrained machine learning frameworks were evaluated, including Point Transformer V3 (PTv3) (Wu et al., 2024) and UnetFormer (Wang et al., 2022). These networks were benchmarked on large-scale urban datasets such as ISPRS Potsdam, Hessigheim 3D (Kölle et al., 2021), Vaihingen (Gerke et al., 2014), Toronto-3D (Tan et al., 2020), and Paris-Lille 3D (Roynard et al., 2018).

3.2 LoD2 Reconstruction Using Geoflow

After the segmentation, the extracted building points are processed using TU Delft's Geoflow tool (Peters et al., 2022) to reconstruct 3D building models up to LoD2.3. Geoflow employs a data-driven roof reconstruction pipeline based on region growing and planar fitting. For each detected roof plane, the model minimizes the distance between observed points and fitted planes:

$$E(\pi_j) = \Sigma_{\{p_i \in \pi_j\}} ||n_j p_i + d_j||^2 \quad (3)$$

In our workflow, we empirically optimize the Geoflow parameters by assessing the reconstruction quality and occlusion rates, which are quantified for each building parcel. Occlusion O is defined as the proportion of missing facade points:

$$O = 1 - \frac{N_v}{N_t} \quad (4)$$

where N_v is the number of visible facade points and N_t is the total number of expected points based on footprint geometry. Concretely, Table 1 shows some of the key values that were iterated over to optimize the reconstruction quality of the LoD2 models.

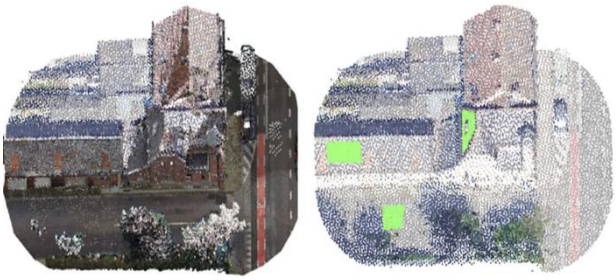


Figure 4. Point cloud density and relative accuracy: (Left) MLS $\sigma\ [2-3cm], > 1000\ pts/m^2$ and (right) MLS $\sigma\ [5-10cm], 25\ pts/m^2$, with co-registration accuracies of 4-6mm.

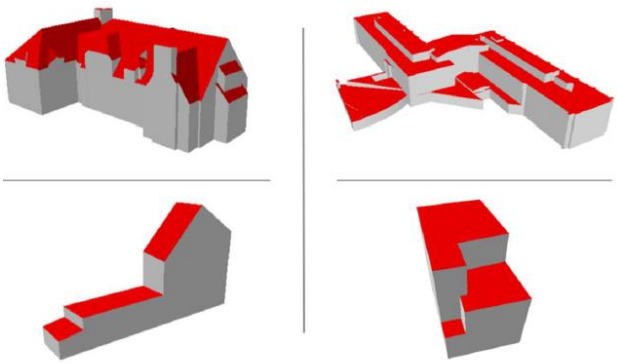


Figure 5. Geoflow reconstruction results from parcel and point cloud input.

Parameter Name	Abbr.	Default Value	Meaning / Description
Line epsilon	ε_l	0.4	Maximum distance between a detected line and its inliers
Normal k	k_n	5	Number of neighboring points used for surface normal estimation
Optimization data term	N_{ot}	10	Higher values yield more detailed models; lower values produce more simplified or generalized models

Plane epsilon	E_p	0.2	Maximum distance between a detected plane and its inliers
Plane k	k_p	15	Number of neighboring points used for region-growing plane detection
Plane minimum points	Min_p	15	Minimum number of inliers required for each detected plane

Table 1. Default Geoflow Values

3.3 Zero-Shot Detection and 3D Projection of Facade Elements

The third stage processes geolocated panoramic imagery from a mobile mapping system. Using Grounding DINO (Liu et al., 2023), the windows and doors are identified without any dataset-specific retraining. The objectness score S_b for each detected bounding box b is computed as:

$$S_b = \sigma(T_e \; I_f) \tag{5}$$

where T_e represents the embedded textual query (i.e., *window*, *door*), I_f the visual feature embedding from the panoramic imagery, and σ the sigmoid activation function indicating detection confidence. The result is a set of bounding boxes per image with a detection score. However, as these boxes are not in the same plane as the façade, and to combine multiple detections, we include a second processing step.

First, we apply a planar homography transformation to correct the perspective of the detected bounding boxes before projection. Each façade image is treated as a 2D observation of a planar surface in 3D space. The homography $H \in \mathbb{R}^{3\times3}$ relates the image coordinates $(u, v, 1)^\top$ to the corresponding façade coordinates $(x, y, 1)^\top$ via

$$H = K[R \mid t] \tag{6}$$

Where the intrinsic calibration K and are the extrinsic $[R \mid t]$ parameters of the cameras are used from the mobile mapping SLAM. This transformation rectifies the bounding boxes to a fronto-parallel orientation, ensuring that windows and doors detected in multiple oblique views are geometrically consistent when projected onto the 3D building surface.

Figure 6. (left) bounding boxes detected by Grounding DINO, (right) corrected detections using facade plane homography.

Next, the corrected detections are merged using spatial averaging to compute a single, reliable instance for each façade element.

$$\bar{p} = \frac{1}{k} \sum_{i=1-k} p_i \tag{7}$$

To this end, we perform a ray tracing to map the 2D box detections into 3D space using our group's GEOMAPI API (Bassier et al., 2024). A ray is cast from each camera center C through the pixel coordinates (u, v), which are converted to spherical coordinates $t(\theta, \phi)$.

$$p_i = C + td, \; t > 0 \tag{8}$$

where d is the direction vector in world coordinates. The intersection of the ray with the LoD2 mesh determines the corresponding 3D coordinates p_i of the facade element. At the same time, we filter out overlapping detections with deviating dimensions due to occlusions (Figure 7). The result is set of geometrically consistent façade elements.

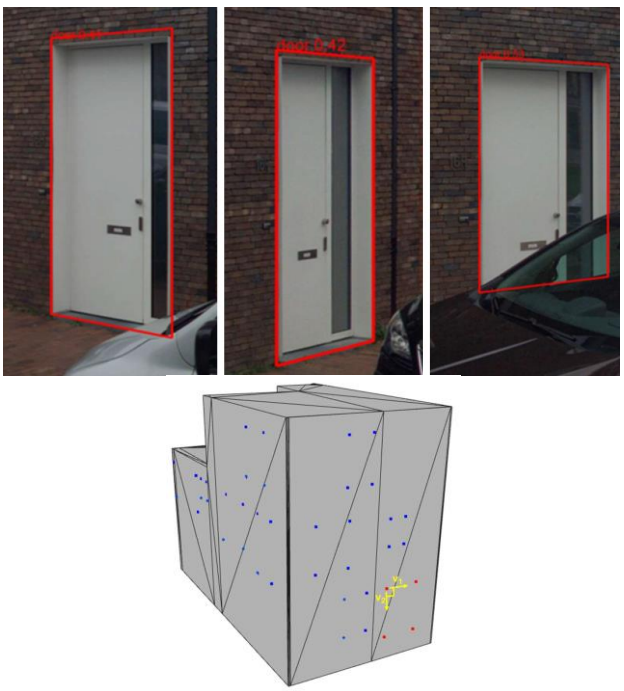


Figure 7. Filtering and region optimization using mean intersection from multiple detections.

3.4 Model Enrichment and Validation

The final step enriches the CityJSON representation with the newly detected facade features. Each feature is encoded as a rectangular face within the corresponding wall surface, and vertices are updated accordingly:

$$V' = V \cup \{v_1, v_2, v_3, v_4\} \tag{9}$$

As the LOD2 models from Geoflow are as Solids, we both adjust the façade's geometry and introduce new geometries for the windows and doors (Figure 8). The enriched models are validated using the CityJSON Ninja platform, which assesses geometric integrity, occlusion statistics, and semantic correctness. Quantitative evaluation metrics include Precision, Recall, and RMSE in planimetric and height dimensions. These indicators allow iterative refinement of the reconstruction pipeline and ensure the method's scalability for large-scale urban modeling.

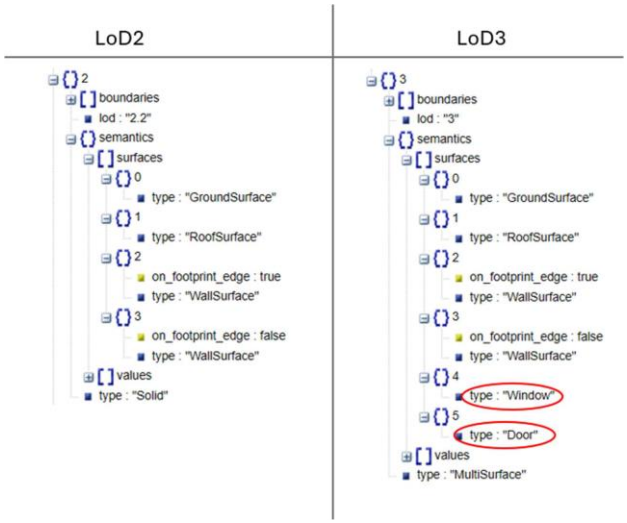


Figure 8. Enriched Geoflow LoD3 models with window and door elements.

4. Experiments

The experimental evaluation of the proposed workflow was conducted on multiple test areas in Flanders, representing typical mixed urban structures. The experiments aim to (1) quantify the reconstruction accuracy and parameter sensitivity of Geoflow, (2) evaluate the performance of zero-shot window and door detection using Grounding DINO, and (3) assess the reliability of model enrichment and filtering in the final CityJSON output.

4.1 Optimized Geoflow reconstruction accuracy

The parameter sensitivity was conducted for a wide range of buildings (Figure 9). Under default settings (Table 1), typical modelling errors occurred where roof or façade segments were represented by an insufficient number of points. To address this, the influence of individual parameters was systematically evaluated. The analysis revealed that lowering the line epsilon (ε_l) slightly improved façade line definition, while increasing both the plane neighborhood size (k_p) and the minimum plane inlier count (min_p) resulted in more stable region growing and fewer fragmented planes. The optimization data term (n_{opt}) mainly affected the model's level of detail: lower values simplified the geometry by removing smaller roof elements such as dormers, whereas higher values added fine-scale detail and local roof variations. Based on these findings, an optimal parameter configuration was derived.

$$\varepsilon_l = 0.35, \quad k_p = 20, \quad min_p = 20$$

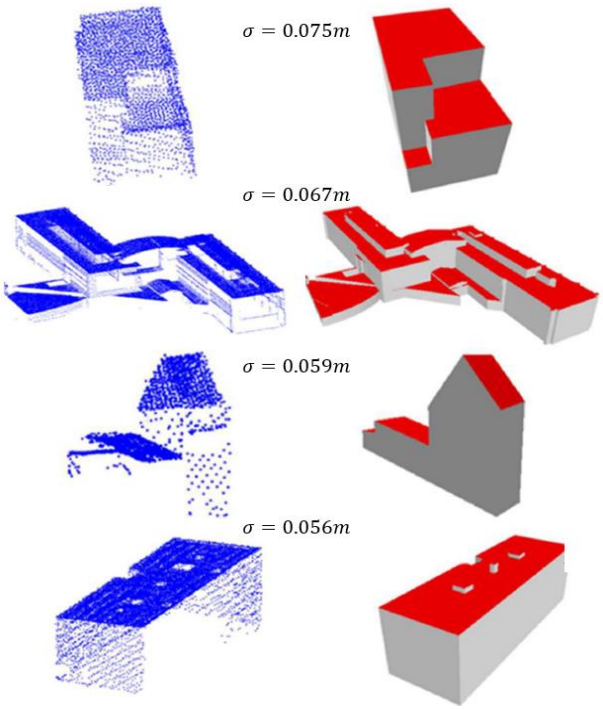


Figure 10. Example reconstruction results with Geoflow optimized parameters.

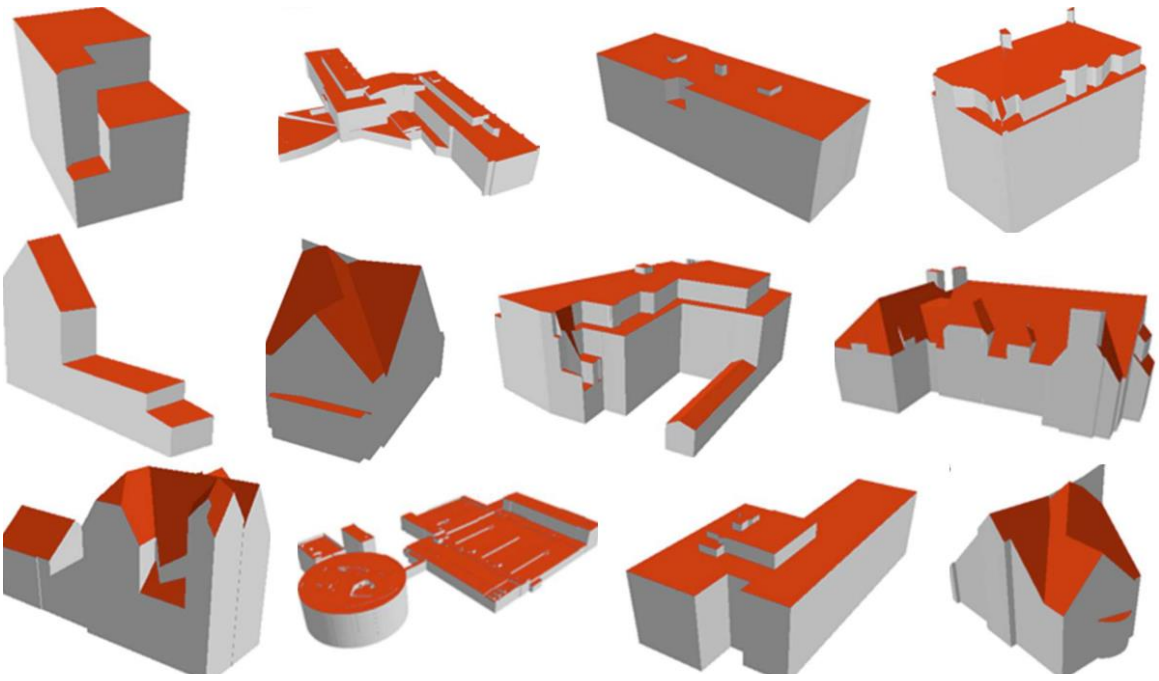


Figure 9. Buildings used for Geoflow parameter optimization.

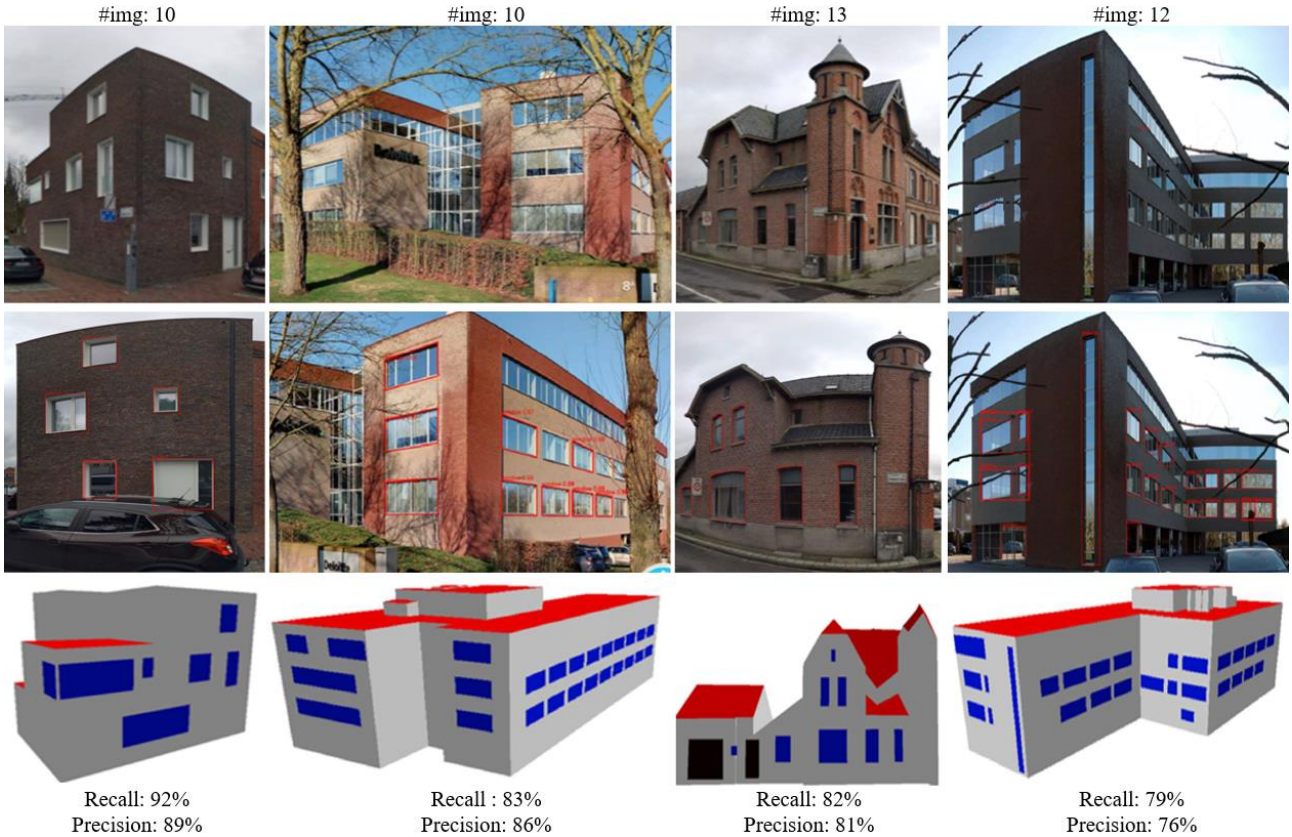


Figure 11. LOD3 reconstruction results including (row 1) geolocated panoramic imagery, (row 2) Grounding DINO detections and (row 3) final LOD3 reconstruction after reprojection, filtering and CityJson mutations.

The geometric accuracy of the reconstructed building models was evaluated on subset of these structures (Figure 10). The RMSE was computed for both roof and façade surfaces to capture variations in reconstruction performance between horizontal and vertical elements. Figure 10 reports mean values of $\sigma = 5.5$ cm and $2\sigma = 14.9$ cm, indicating that 68 % of all residual distances fall within 5.5 cm of the reconstructed surface, while 95 % remain within 15 cm. These results demonstrate that Geoflow achieves an accuracy that is well within the ALS accuracy.

4.2 Zero-shot detections

The detection accuracy of Grounding DINO, for the windows and doors from the mobile mapping data, was evaluated on 4 types of constructions (Figure 11) **.** The detector was operated in zero-shot mode, using natural-language prompts (“window” and “door”) without dataset-specific fine-tuning. A total of 45 façade images were processed, containing 71 windows (475 occurrences) and 20 doors (57 occurrences) manually annotated as ground truth. The evaluation was performed on a per-instance basis using standard precision and recall metrics.

Experiments were conducted using box confidence threshold between 0.4-0.6, applied to the same set of 45 façade images. The results indicate a clear trade-off between completeness and correctness. At a low threshold (0.4), the model detects nearly all window and door instances (DS > 0.88) but introduces many false positives, particularly reflections and repeated window frames. Increasing the threshold to 0.6 reduced the recall by on average 20% and increased the precision above 90%, but missed smaller or partially occluded objects. The best yielding setting for the box confidence threshold was 0.5, since we could still filter out false positives using the spatial filtering from multiple occurrences.

Using this setting, the zero-shot detector achieved an average precision of 84.2% and recall of 79.3%, indicating that the model generalizes well to unseen urban scenes. The results confirm that windows are more reliably detected than doors, largely due to their higher visual contrast, repetitive structure, and stronger contextual cues. Doors, in contrast, exhibited more variation in texture and color, particularly in shaded or recessed entrances, resulting in more missed detections. Additionally, doors suffered from a higher occlusion rate due to the terrestrial vantage point (i.e. by blocking cars).

Visual inspection of representative façades showed that detection accuracy is highly dependent on viewing angle, illumination, and occlusions from street furniture or vegetation. In oblique views, bounding boxes tended to overestimate the true object size due to perspective distortion.

4.3 Filtering & Model Enrichment

The detection merger and filtering was evaluated on the same four types of buildings. Concretely, the reconstructed corner points of the windows and doors were compared to ground truth created corner points. The maximum clustering distance for the corner points of the different detections was set to 0.5m. Overlapping occurrences that didn’t comply with the clustering distance threshold were removed. A minimum cluster size of two detections was required to confirm an object’s reliability, effectively filtering out single, uncertain detections from oblique or shadowed images.

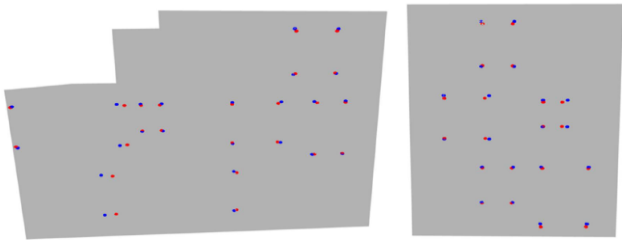


Figure 12. Clustering of detected windows/door corner points.

After clustering and filtering, each object cluster was reduced to a single representative instance by computing the mean centroid and averaged dimensions (Figure 12). The resulting enriched facades showed a substantial reduction in false positives and improved positional consistency. Concretely, the clustering reduced the number of false detections by approximately 36%, while maintaining geometric consistency with an average of 34mm difference between the center of the GT and the detected windows, with a standard deviation of 30mm.

The geometric and semantic validity of the final CityJSON models was verified using the CJVal web-based validation tool, developed as part of the CityJSON Ninja platform (Ledoux & Dukai, 2023). The validator checks conformance to the CityJSON v1.1 specification, assessing both syntactic correctness (JSON structure and attributes) and geometric validity (topology, vertex orientation, and manifoldness). The validation results confirm that all reconstructed buildings comply with the CityJSON standard.

5. Conclusion

The conducted research demonstrates the feasibility and robustness of an integrated workflow for automated LoD3 building model reconstruction, combining geometric reconstruction, zero-shot semantic enrichment, and standards-based validation. Building upon the Geoflow framework, the developed methodology successfully bridges the gap between conventional LoD2 representations and semantically enriched LoD3 models by integrating heterogeneous data sources including airborne LiDAR, mobile mapping imagery, and deep learning-based detection systems within a unified processing pipeline.

The experiments confirmed that Geoflow delivers accurate LoD2.3 reconstructions when optimized using the parameter configuration derived from the sensitivity study. The mean RMSE values of approximately 5–6 cm align with the expected accuracy range for airborne LiDAR data at 25 points/m². The influence of reconstruction parameters was significant: smaller region-growing radii and tighter planarity constraints improved geometric precision but occasionally reduced completeness in complex roof geometries. Hence, parameter selection must balance spatial detail and generalization according to the data density and building typology.

The integration of zero-shot detection using Grounding DINO introduced a flexible mechanism for façade enrichment without the need for dataset-specific training. The results illustrate the potential of large-scale vision–language models for urban scene understanding, achieving a mean detection score of 0.83 and a false alarm rate below 0.15 under optimal thresholding. Nevertheless, detection performance remained dependent on environmental conditions such as occlusion, lighting, and surface reflectivity. The homography-based rectification significantly improved the geometric consistency of projections but did not fully eliminate misalignments on highly reflective or irregular façades.

Subsequent spatial reasoning and clustering steps proved essential to ensuring semantic coherence across multiple viewpoints. The clustering parameters, i.e. a maximum centroid distance of 0.5m, successfully merged duplicate detections while maintaining the geometric fidelity of valid features. An average accuracy of 3.4cm and standard deviations of 3cm in position confirm the internal consistency of the enriched models.

The CJVal validation results further substantiate these findings, showing an overall conformity, which demonstrates the workflow's readiness for deployment within official 3D city model production pipelines.

While the framework performs robustly across varied urban contexts, several limitations remain. Detection reliability is affected by lighting conditions and occlusions, and the dependence on high-quality panoramic imagery constrains applicability in areas with limited street-level data. Additionally, the workflow assumes well-registered multimodal datasets; future work could address sensor misalignment through self-calibrating multimodal fusion methods.

Further improvements should focus on extending the semantic taxonomy beyond windows and doors to include balconies, roof structures, and vegetation, leveraging transformer-based multi-modal architectures. While the current pipeline focuses on planar façades and rectangular openings, future work should address more complex façade geometries. Integrating uncertainty quantification into the enrichment process would also enhance interpretability and quality control.

Acknowledgements

This work was made possible with Cyclorama data from Leiedal Intercommunal, EODas Open Lidar Vlaanderen, and internal KUL Geomatics Funding.

References

Bassier, M., J. Vermandere, S. De Geyter, and H. De Winter, 2024. GEOMAPI: Processing Close-Range Sensing Data of Construction Scenes with Semantic Web Technologies. *Automation in Construction* 164: 105454, https://doi.org/10.1016/j.autcon.2024.105454

Behley, J., et al., 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. *arXiv preprint* arXiv:1904.01416. https://doi.org/10.48550/arXiv.1904.01416.

Biljecki, F., H. Ledoux, and J. Stoter, 2016. An Improved LOD Specification for 3D Building Models. *Computers, Environment and Urban Systems* 59: 25–37. https://doi.org/10.1016/j.compenvurbsys.2016.04.005.

Buyuksalih, G., P. Baskaraca, S. Bayburt, I. Buyuksalih, and A. Abdul Rahman, 2019. 3D City Modelling of Istanbul Based on LiDAR Data and Panoramic Images – Issues and Challenges. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4/W12: 51–60. https://doi.org/10.5194/isprs-archives-XLII-4-W12-51-2019.

Froech, T., O. Wysocki, L. Hoegner, and U. Stilla, 2024.

Reconstructing Facade Details Using MLS Point Clouds and Bag-of-Words Approach. *arXiv preprint* arXiv:2402.06521. https://doi.org/10.48550/arXiv.2402.06521.

Gerke, M., and I.T.C, 2014. Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen). *ISPRS Benchmark on Urban Object Detection and 3D Building Reconstruction.*

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, 2017. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

Jayaraj, P., and A.M. Ramiya, 2018. 3D CityGML Building Modelling from LiDAR Point Cloud Data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5: 175–180. https://doi.org/10.5194/isprs-archives-XLII-5-175-2018.

Kölle, M., D. Laupheimer, S. Schmohl, N. Haala, F. Rottensteiner, J.D. Wegner, and H. Ledoux, 2021. The Hessigheim 3D (H3D) Benchmark on Semantic Segmentation of High-Resolution 3D Point Clouds and Textured Meshes from UAV LiDAR and Multi-View Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 1: 100001.

Ledoux, H., K. Arroyo Ohori, K. Kumar, B. Dukai, A. Labetski, and S. Vitalis, 2019. CityJSON: A Compact and Easy-to-Use Encoding of the CityGML Data Model. *Open Geospatial Data, Software and Standards* 4(1): 4. https://doi.org/10.1186/s40965-019-0064-0.

Lewandowicz, E., F. Tarsha Kurdi, and Z. Gharineiat, 2022. 3D LoD2 and LoD3 Modeling of Buildings with Ornamental Towers and Turrets Based on LiDAR Data. *Remote Sensing* 14(19): 4687. https://doi.org/10.3390/rs14194687.

Liu, S., et al, 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint* arXiv:2303.05499. https://doi.org/10.48550/arXiv.2303.05499.

Markus, R., and E. Airaksinen, 2020. Concept and Evaluation of Heating Demand Prediction Based on 3D City Models and the CityGML Energy ADE—Case Study Helsinki. *ISPRS International Journal of Geo-Information* 9(10): 602. https://doi.org/10.3390/ijgi9100602.

Pantoja-Rosero, B.G., R. Achanta, M. Kozinski, P. Fua, F. Perez-Cruz, and K. Beyer, 2022. Generating LOD3 Building Models from Structure-from-Motion and Semantic Segmentation. *Automation in Construction* 141: 104430. https://doi.org/10.1016/j.autcon.2022.104430.

Peters, R., B. Dukai, S. Vitalis, J. van Liempt, and J. Stoter, 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. https://doi.org/10.14358/PERS.21-00032R2.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, 2021. Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun, 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems (NeurIPS)* 28.

Roynard, X., J.-E. Deschaud, and F. Goulette, 2018. Paris-Lille-3D: A Large and High-Quality Ground Truth Urban Point Cloud Dataset for Automatic Segmentation and Classification. *arXiv preprint* arXiv:1712.00032. https://doi.org/10.48550/arXiv.1712.00032.

Tan, W., et al, 2020. Toronto-3D: A Large-Scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 797–806. https://doi.org/10.1109/CVPRW50498.2020.00109.

Wang, L., R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P.M. Atkinson, 2022. UNetFormer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 190: 196–214.

Wen, X., H. Xie, H. Liu, and L. Yan, 2019. Accurate Reconstruction of the LoD3 Building Model by Integrating Multi-Source Point Clouds and Oblique Remote Sensing Imagery. *ISPRS International Journal of Geo-Information* 8(3): 135. https://doi.org/10.3390/ijgi8030135.

Wu, X., L. Jiang, P.S. Wang, Z. Liu, X. Liu, Y. Qiao, and H. Zhao, 2024. Point Transformer V3: Simpler Faster Stronger. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4851.

Wysocki, O., B. Schwab, C. Beil, C. Holst, and T.H. Kolbe, 2024. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48: 493–500.

Wysocki, O., et al, 2023. Scan2LoD3: Reconstructing Semantic 3D Building Models at LoD3 Using Ray Casting and Bayesian Networks. *arXiv preprint* arXiv:2305.06314. https://doi.org/10.48550/arXiv.2305.06314.

Yang, B., and Q. Zhu, 2015. 3D Reconstruction of Building Facades with Fused Data of Terrestrial LiDAR and Images. *Optics and Lasers in Engineering* 66: 195–204.

Zhang, P., H. Ma, L. Wang, R. Zhong, M. Xu, and S. Chen, 2024. Automatic Registration of Panoramic Images and Point Clouds in Urban Large Scenes Based on Line Features. *Remote Sensing* 16(23): 4450.