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

SVI2LoD3: Agent-Driven Reconstruction of LoD3 Façade Openings in Semantic 3D City Models from Volunteered Street View Imagery using Large Language and Visual Models

Elmehdi Kanna*, Lukas Arzoumanidis*, Huynh Duc An Son Nguyen, Youness Dehbi

Computational Methods Lab, HafenCity University, Hamburg, Germany -
{elmehdi.kanna, lukas.arzoumanidis, son.nguyen, youness.dehbi}@hcu-hamburg.de

Keywords: vision foundation models, large language models, LoD3 reconstruction, façade openings, semantic 3D city models.

Abstract

This paper presents an end-to-end, agent-driven pipeline for the LoD3 reconstruction of façade openings in 3D city models, producing directly usable CityGML-conform outputs. In contrast to existing approaches that rely on supervised semantic segmentation and therefore require large amounts of manually annotated training data, the proposed method employs a zero-shot segmentation strategy. This substantially reduces the annotation effort while still achieving strong performance in our benchmark on the eTRIMS dataset. A further key contribution is the enforcement of correct partonomic hierarchies, thereby producing CityGML-conform LoD3 building models. Beyond the reconstruction pipeline itself, this work also introduces a novel evaluation metric for façade reconstruction, termed *Facade Feature Distance* (FFD). Unlike conventional metrics such as mIoU or FRDS, which assess similarity primarily through pixel-wise overlap, FFD measures distance in a high-level feature space derived from a vision transformer. In doing so, it captures both semantic correctness and architectural layout, providing a more suitable assessment of façade reconstruction quality. The proposed pipeline and evaluation strategy together offer a practical and scalable contribution toward the automated generation and analysis of semantically enriched 3D city models. The developed code is published at: `https://github.com/hcu-cml/citydb-SVI2LoD3-ai`.

1. Introduction

Advances in urban analysis and simulation tasks, such as modeling of urban temperature and wind flow, building energy demand estimation, and simulations of interactions between autonomous vehicles and Internet of Things (IoT) infrastructures, are increasingly dependent on highly detailed and semantically rich 3D city models (Kolbe and Donaubauer, 2021). These applications require not only accurate geometric representations of the urban environment but also rich semantic information.

Currently, the majority of publicly available semantic 3D city models remain at LoD2 (Wysocki et al., 2024; Biljecki et al., 2016a). While LoD2 models typically include differentiated roof structures and generalized building geometries, they remain limited in terms of detailed representations for façades (Gröger and Plümer, 2012). Consequently, fine-grained façade characteristics, such as material properties, façade composition, and the precise geometry and placement of openings, are missing. This limitation is primarily driven by the complexity of acquiring and reconstructing such detailed information at scale. Addressing this gap by enabling city-wide LoD3 reconstruction therefore requires advanced methods capable of capturing the heterogeneity of architectural styles and structural typologies (Pang and Biljecki, 2022).

Moreover, reconstruction processes typically rely on observations that are inherently noisy and incomplete, such as airborne or terrestrial LiDAR point clouds and street-view imagery. These observations are affected by measurement errors, occlusions, varying illumination conditions, and perspective distortions (Kolbe and Donaubauer, 2021). Since airborne and terrestrial LiDAR point cloud coverage is unavailable for most urban regions, volunteered street-view imagery (SVI) can serve as an alternative data source due to its broader availability. This wider coverage is largely attributable to the considerably less demanding acquisition and handling of camera images for contributors, compared to large point clouds from different vendors, which typically require additional processing.

As a result, this work proposes a novel agent-driven approach for the reconstruction of façade openings from SVI in semantic 3D city models using large-scale foundation models, as illustrated in Figure 1. Specifically, our approach leverages the high-level reasoning capabilities of Large Language Models (LLMs) and zero-shot image segmentation capabilities of Large Vision Models (LVMs) within an agent-driven framework to guide and coordinate the reconstruction process. To ensure syntactic as well as semantic, geometric, and topological correctness of the reconstructed 3D city models in accordance with the CityGML 2.0 LoD3 schema, we developed algorithms that support the agent in two distinct stages of the reconstruction process, which we refer to as *expert-designed programs*.

To the best of our knowledge, this approach is among the first to combine expert-designed programs into an agent-driven reconstruction approach. By coupling these complementary capabilities, the proposed approach is able to address the substantial architectural variability of building façades, including diverse arrangements and geometries of façade openings which has historically posed a significant challenge for automated reconstruction approaches and has limited the generalizability of many existing methods, as will be pointed out in Section 2. In addition, this work introduces a novel evaluation metric for façade openings in reconstructed LoD3 models that relies on a high-level feature distance instead of pixel-wise comparison strategies commonly adopted in previous approaches.

Our main contributions include:

* These authors contributed equally to this work.

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-XII-4-W1-2026-203-2026

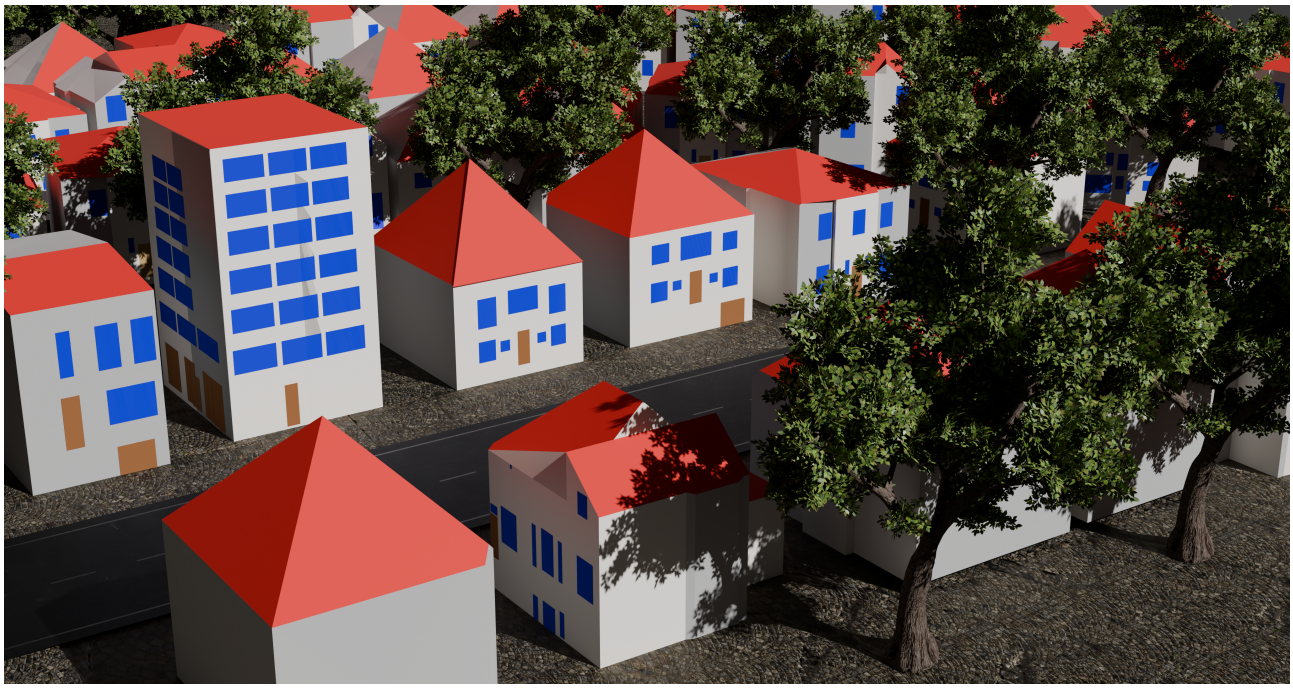


Figure 1. Reconstructed façade openings for LoD3 building models in a selected exemplary neighborhood in Hamburg, Germany, visualized in Blender©. Blue rectangles indicate reconstructed windows, while brown rectangles indicate reconstructed doors.

- an end-to-end approach for reconstructing CityGML-conformant LoD3 building models from volunteered SVI,
- the use of LVM for zero-shot segmentation of façade openings, eliminating the need for task-specific training data and enabling direct application to different cities and architecturally heterogeneous façades,
- an agent-driven framework that leverages the high-level reasoning capabilities of large language models for 2D-to-3D projection in combination with expert-designed algorithms, and
- a novel evaluation metric that focuses on semantic and topological accuracy, in contrast to state-of-the-art metrics that assess only semantic accuracy.

2. Related Work

In recent years, the reconstruction of façade openings has attracted increasing attention, with numerous deep learning-based approaches proposed that utilize observational data such as airborne or terrestrial LiDAR-derived point clouds, street-view imagery (SVI), or airborne RGB imagery. The following subsections review several of the most recent and influential approaches in this domain and highlight the key distinctions between these methods and the approach proposed in this work.

2.1 LiDAR-Based Façade Reconstruction

Over the past years, numerous machine learning-based approaches have been developed for the 3D reconstruction of buildings using terrestrial and airborne LiDAR point clouds. These methods typically exploit the geometric richness of point cloud data to infer building structures and façade elements.

Using terrestrial LiDAR point clouds, Dehbi et al. (2016) employed Support Vector Machines (SVMs) in combination with Statistical Relational Learning (SRL) through Markov Logic Networks (MLNs) to learn grammar-based rules for the reconstruction of 3D building structures. This approach integrates probabilistic reasoning with structural constraints to infer building components from point cloud observations. Similarly, a method proposed by Wang et al. (2016) combines multiple geometric algorithms within a rule-based framework to reconstruct building façades from terrestrial LiDAR point cloud data. In this framework, geometric feature extraction and predefined architectural rules are jointly utilized to derive façade structures and opening configurations. More recently, Wysocki et al. (2023) introduced an approach aimed at improving the accuracy of semantic LoD3 building reconstruction. Their method enhances façade-level semantic 3D segmentation by incorporating knowledge of laser scanning physics together with prior information derived from existing 3D building models to probabilistically identify model conflicts.

2.2 Image-Based Façade Reconstruction

Using Structure-from-Motion (SfM) and deep-learning-based segmentation techniques, Pantoja-Rosero et al. (2022) developed a pipeline for the automatic reconstruction of LoD3 building models of free-standing structures, with a particular focus on detecting and reconstructing façade openings. Pang and Biljecki (2022) proposed a method for reconstructing 3D building models from a single street-view image using image-to-mesh reconstruction techniques. Their approach leverages street-view imagery to enhance the level of detail of commonly available block models (LoD1). In the same context, (Wang et al., 2024) propose an automated framework for semantic building-model reconstruction at an urban scale. Specifically, they introduce a façade layout graph model to represent the geometric and topological relationships of semantic entities on building façades, thereby enabling the inference of structural completeness and the reconstruction of semantic façade models. Rather than using street-view imagery, their approach exploits mesh textures that are often available for LoD2 models

and uses deep learning-based computer vision methods to extract semantic information from these textures in order to enrich the existing LoD2 building models. Tang et al. (2025) present a method for reconstructing LoD3 building models by combining low-detail 3D building models with panoramic street-level imagery. Their results show that low-detail building models can provide suitable planar reference surfaces for the orthorectification of panoramic images. In addition, they demonstrate that applying semantic segmentation to accurately textured low-detail façade surfaces preserves key properties required for LoD3 reconstruction, including georeferencing consistency, watertight geometry, and a compact low-polygon representation. Recently, Ma et al. (2026) presented a method for reconstructing façade openings in 3D building models by integrating street-view imagery. The approach introduces a mathematically derived method for estimating unknown intrinsic camera parameters, enabling metric 2D-to-3D projection without relying on multi-view imagery or pre-existing depth information. Furthermore, a supervised semantic segmentation model is employed to detect façades in an example study area in Amsterdam, allowing detailed façade openings to be measured and pixel coordinates to be transformed into spatial coordinates.

3. Methodology

To enable the reconstruction of façade openings from street-view imagery (SVI) and their integration into existing LoD2 building models for the generation of LoD3 models, we employ an agent-driven framework that combines the high-level reasoning capabilities of a Large Language Model (LLM) with the zero-shot image segmentation capabilities of a Large Vision Model (LVM). Within this framework, the agent coordinates the reconstruction workflow by invoking a set of expert-designed programs designed for geometric processing and model integration.

For façade segmentation from SVI, our approach utilizes an LVM because of the substantial variability in façade design, including differences in architectural style, materiality, and ornamentation. Within the agent framework, we evaluate and compare the performance of GPT-5.1[1] and Qwen3-VL 30B Instruct[2] for guiding the LoD3 reconstruction process leveraging their high-level reasoning capabilities (Xu et al., 2025). Specifically, these models orchestrate the execution of expert-designed programs that extract façade openings from the segmentation masks obtained from SAM 3 and integrate them into existing LoD2 building models, as illustrated in Figure 2.

The comparison between different LLMs as backbone for our agent framework is relevant from a practical deployment perspective. While GPT-5.1 is a commercial model associated with usage costs, Qwen3-VL is available as an open-source alternative. In our case, access to the latter is facilitated through the ChatAI system[3] (Doosthosseini et al., 2024), which provides access to large-scale models hosted on a high-performance computing infrastructure, allowing experimentation with open models in a controlled environment while avoiding additional usage costs.

To improve the LLM's reliability when performing geometrical and mathematical operations during the reconstruction process, our agent is designed to invoke expert-designed programs tailored to specific reconstruction tasks, as highlighted in Figure 2. These include, for example, façade surface-to-camera image matching based on camera location, as well as the extraction of façade openings from segmented masks and their integration into an LoD2 building model.

To evaluate our approach, we consider three of the four principal error dimensions. According to the OGC standard definition[4] (Biljecki et al., 2016b; Gröger and Plümer, 2012), these dimensions are geometry, semantics, topology, and completeness. In this work, we only focus on the semantic, topological, and completeness dimensions, which we assess using a set of state-of-the-art evaluation metrics and a newly defined metric that jointly captures semantic and topological quality within a single score.

3.1 Zero-Shot Façade Composition Segmentation

To address the challenge of identifying façade openings in SVI across diverse urban environments characterized by substantial variation in façade design, material composition, and architectural ornamentation, we employ SAM 3. SAM 3 is a LVM that achieves strong predictive performance in object detection and segmentation tasks, including on previously unseen object categories, without requiring task-specific fine-tuning or labeled training data (Carion et al., 2026). This zero-shot capability removes the dependence on annotated datasets, whose production is both resource-intensive and time-consuming, and facilitates straightforward transfer of our approach to other urban contexts where Mapillary[5] imagery is available.

SAM 3 supports text-conditioned segmentation, whereby natural language prompts describing target objects are provided to guide mask generation. Through prompting, the model was instructed to segment windows, doors and façade surfaces within each image, as illustrated in Figure 2. For every input image, SAM 3 generated multi-class segmentation masks corresponding to regions consistent with the provided textual prompts.

3.2 LLM-Driven Reasoning for Façade Shape Completion

After semantic segmentation of the façade image using SAM 3, the approximate geometry of the façade surface is derived from the class-specific color encoding of the predicted segmentation mask. More specifically, façade elements are extracted based on their RGB values, with façades being represented in blue, windows in red, and doors in green.

A refinement step is needed because the segmentation mask obtained from the SVI contains the façade openings required for LoD3 reconstruction, yet it is often degraded by noise, occlusions, and small spurious artifacts due to the inherently imperfect nature of SVI data (Hou and Biljecki, 2022; Biljecki and Ito, 2021), as illustrated in Figure 2. To improve the quality and topology of the extracted façade-opening geometries, we prompt the agent to leverage the high-level reasoning capabilities of the underlying LLM for shape completion and reprojection on the semantically segmented mask predicted by SAM 3. Furthermore, in order to preserve genuine openings that are only partially visible because of occlusions caused by vegetation or parked vehicles, the agent retains candidate openings that either have their centroid within the façade geometry or touch the façade boundary.

[1] https://openai.com/index/gpt-5-1/
[2] https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list
[3] https://docs.hpc.gwdg.de/services/saia/index.html
[4] https://www.ogc.org/standards/citygml/
[5] https://www.mapillary.com/

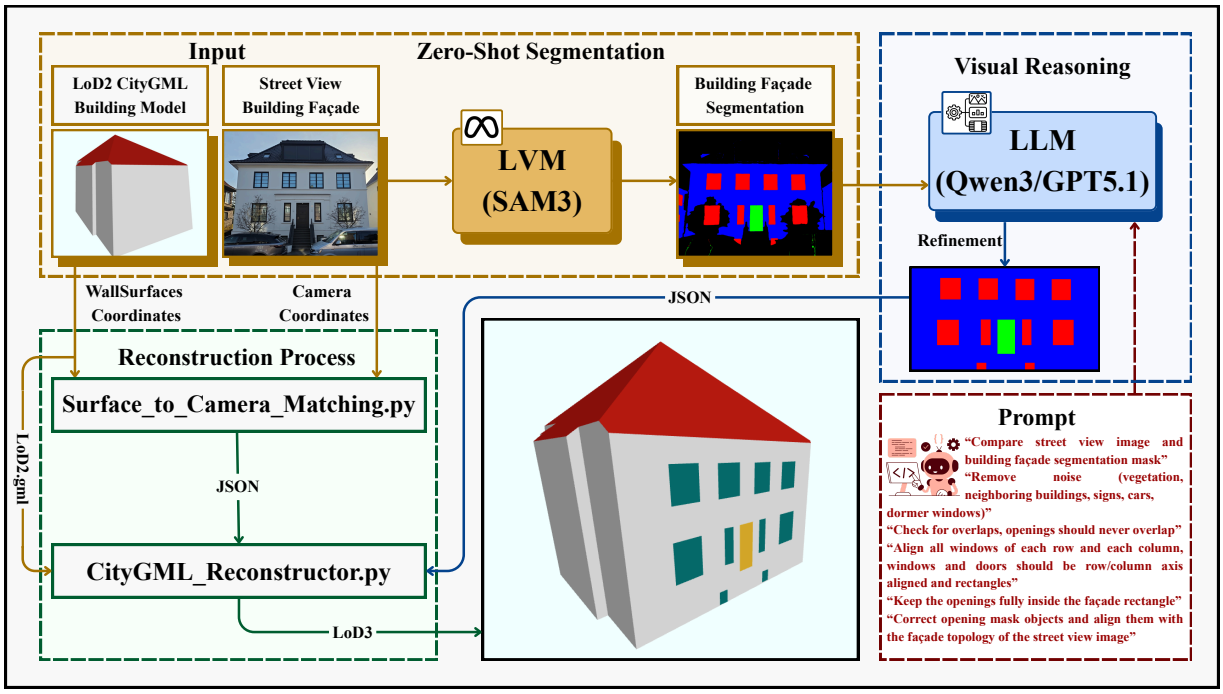


Figure 2. Overview of our agent-driven approach for reconstructing façade openings in valid LoD3 CityGML-conform models.

Moreover, the agent is instructed to remove sources of noise, including vegetation, neighboring buildings, traffic signs, vehicles and dormer windows, as shown in Figure 2. An additional prompt enforces a consistency check for overlapping façade openings, as such overlaps are not geometrically plausible in the reconstructed model. Finally, to regularize and complete the geometry of façade openings, the agent is prompted to impose structural constraints such that windows and doors are represented as axis-aligned rectangles, while openings within the same horizontal and vertical arrangements are aligned along common rows and columns. These assumptions reflect the geometric regularity typically exhibited by real-world building façades tested in this study.

3.3 Façade Surface-To-Camera Image Matching

In order to associate the façade visible in the SVI with the corresponding façade of the LoD2 model to be reconstructed, we developed an algorithm that enables the agent to identify the correct LoD2 façade surface from the camera perspective. For the sake of clarity and reproducibility, Algorithm 1 is provided as a reference to guide the reader through the methodological steps described in the following paragraphs.

First, the original camera parameters and positions of the SVI are transformed into the same Coordinate Reference System (CRS) as the LoD2 CityGML model (cf. Algo. 1 line 2). To determine which building façade faces the camera, we adopt the assumption that the nearest observable façade is also the façade facing the camera. In CityGML, a façade may consist of multiple `WallSurface` elements, which can be represented as a set W of `WallSurfaces` $ws_1, \ldots, ws_n$. Additionally, we define a set C of `WallSurface` centroids $c_1, \ldots, c_n$. If a `WallSurface` has an area of $4\,\mathrm{m}^2$ or less, we treat it as a distinct building component, such as a garage, and disregard it as a matchable façade.

To identify all façades observable in the camera image, we compute the camera viewing normal vector n_{cam} corresponding to the image vanishing direction (cf. Algo. 1 line 3). For each `WallSurface` ws_i, we compute its outward normal vector n_i in order to exclude `WallSurfaces` that are not visible in the camera image. This filtering is performed by comparing each n_i with the previously computed camera normal vector n_{cam} using a photogrammetric visibility criterion based on plane-ray intersection and normal alignment, as illustrated in Figure 3 (cf. Algo. 1 line 5-10). If two or more `WallSurface` normals yield the same matching quality with respect to n_{cam}, we identify the `WallSurfaces` that jointly constitute the full camera-facing façade by additionally evaluating their Euclidean distance to the camera position and ranking them accordingly (cf. Algo. 1 line 12-20).

The Euclidean distance d_i between the camera position, represented by the coordinates (x_{cam}, y_{cam}), and the centroid c_i, represented by the coordinates (x_i, y_i), is computed as follows:

$$\forall c_i \in C: \quad d_i = \sqrt{(x_i - x_{cam})^2 + (y_i - y_{cam})^2}. \tag{1}$$

Because the LoD2 models in our dataset frequently contain small geometric extrusions, e.g., bay windows, we additionally verify whether the closest matching `WallSurface` and the remaining matching `WallSurfaces` exhibit a perpendicular separation of at most $1\,\mathrm{m}$. This threshold was chosen, as most observed extrusions fall within this range (Jörg Schmittwilken, 2012). `WallSurfaces` whose perpendicular distance is less than $1\,\mathrm{m}$ are retained. The remaining `WallSurfaces` are then considered to jointly form the true camera-facing façade of the LoD2 model (cf. Algo. 1 line 20-26).

Figure 3 illustrates an example in which the `WallSurfaces` walls ws_1, ws_2, and ws_3 are forming the true camera-facing façade of the LoD2 model while `WallSurfaces` ws_6, ws_5, and ws_4 are excluded from the image-to-façade matching process because their normal vectors n_6, n_5, n_4 are not aligned with n_{cam}.

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-XII-4-W1-2026-203-2026 | © Author(s) 2026. CC BY 4.0 License.

Algorithm 1: Selection of the camera-facing façade from an LoD2 building model

Input : Camera parameters $\mathcal{C}$, camera position p_{cam}, `WallSurfaces` set $W = \{ws_1, \ldots, ws_n\}$
Output: Matched façade F_{match}

1 **Coordinate transformation and view direction estimation**
2 Transform $\mathcal{C}$ and p_{cam} to the LoD2 coordinate reference system
3 Compute camera normal vector n_{cam}
4 **Visibility and geometric filtering**
5 $W_{\text{obs}} \leftarrow \emptyset$
6 **foreach** $ws_i \in W$ **do**
7 Compute `WallSurface` normal n_i
8 Compute `WallSurface` area ws_i^{area}
9 **if** ws_i *satisfies the visibility criterion with respect to* n_{cam} ***and*** $ws_i^{area} > 4\,\text{m}^2$ **then**
10 $W_{\text{obs}} \leftarrow W_{\text{obs}} \cup \{ws_i\}$
11 **Candidate ranking**
12 $W_{\text{cand}} \leftarrow \emptyset$
13 Determine the highest normal-consistency score in W_{obs}
14 **foreach** $ws_i \in W_{\text{obs}}$ **do**
15 **if** ws_i *attains the highest normal-consistency score* **then**
16 Compute centroid c_i
17 Compute $d_i \leftarrow \|p_{\text{cam}} - c_i\|$
18 $W_{\text{cand}} \leftarrow W_{\text{cand}} \cup \{(ws_i, d_i)\}$
19 Sort W_{cand} by ascending distance
20 Select the nearest candidate $ws_{\min}$
21 **Façade aggregation**
22 $F_{\text{match}} \leftarrow \{ws_{\min}\}$
23 **foreach** $ws_j \in W_{\text{cand}} \setminus \{ws_{\min}\}$ **do**
24 Compute perpendicular distance $\delta_\perp(ws_{\min}, ws_j)$
25 **if** $\delta_\perp(ws_{\min}, ws_j) < 1\,\text{m}$ **then**
26 $F_{\text{match}} \leftarrow F_{\text{match}} \cup \{ws_j\}$
27 **Export**
28 Export F_{match} with CityGML `wall_id` and semantic dimensions
29 **return** F_{match}

As illustrated in Figure 2, the matched `WallSurfaces` are exported as JSON elements containing their CityGML `wall_id` and geometric information describing the physical dimensions of each segment (cf. Algo. 1 line 28). This is needed to distribute the façade openings derived from the shape-completed segmentation mask in a geometrically and topologically consistent manner when a façade consists of multiple `WallSurfaces`.

3.4 LoD3 CityGML Reconstruction and Validation

As an outcome of the reasoning process, the agent produces a JSON file containing a list of façade-opening geometries, specifically windows and doors, and the façade surface geometry together with their corresponding coordinates extracted from the predicted façade segmentation mask after shape completion. We match the façade surface geometry with the previously matched wall surface geomery, ensuring correct partonomy and scaling of openings.

In cases where a façade is represented by multiple `WallSurface` elements in the LoD2 model, the segmentation mask is subdivided proportionally to the widths of the respective `WallSurface` elements in the LoD2 geometry, thereby preserving the geometric arrangement and topological consistency of the detected openings.

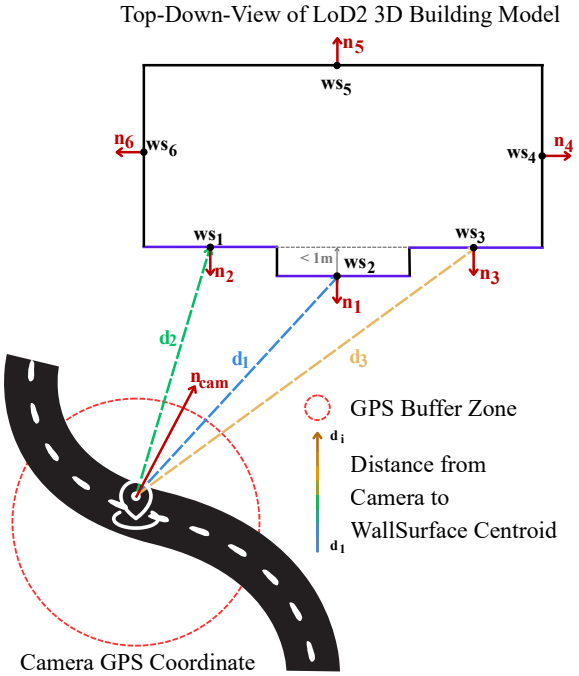


Figure 3. Illustration of our approach for façade surface-to-camera image matching, as described in Algorithm 1.

Subsequently, the LoD3 model is reconstructed by incorporating the detected and matched façade openings into the building model, as highlighted in Figure 2. The identified windows and doors are encoded as explicit CityGML elements and inserted directly into the XML structure of the CityGML document. The resulting XML representation is then validated using the CityGML schema validation framework[6]. Finally, the validated output is written as a compliant LoD3 CityGML 2.0 document.

3.5 Benchmarking Dataset & Evaluation Metrics

To benchmark our approach, we use the eTRIMS dataset introduced by Korč and Förstner (2009). The dataset contains street-view images together with corresponding semantic annotations covering eight semantic classes, namely: sky, façade, window, door, vegetation, car, road and pavement. This dataset was selected because it captures a wide range of façade images representing different architectural styles and cities across European urban environments, where we conducted our test.

Prior to evaluation, classes not relevant to the façade reconstruction task, such as trees, cars, and road and pavement surfaces, were removed from the annotations, as illustrated in Figure 4. In addition, the color encoding of the remaining classes, specifically windows, doors, and building façades, was adapted to match the class color scheme produced by SAM 3 in order to enable consistent evaluation. To evaluate our reconstructed LoD3 models, we selected two state-of-the-art metrics for the evaluation of reconstructed façade openings in 3D building models: mean Intersection over Union (mIoU) (Wang et al., 2024) and the Facade Reprojection Dice Score (FRDS) (Ma et al., 2026; Pantoja-Rosero et al., 2022). Both metrics evaluate the reconstructed LoD3 models purely based on 2D planar geometry extracted from images. However, they neglect the

[6] `https://github.com/tudelft3d/CityGML-schema-validation`

semantic and topological relationships that are critical for successful reconstruction and the generation of valid, CityGML-conform models.

For each class i, IoU measures the overlap between the predicted segmentation and the ground-truth annotation. It is defined as the ratio between correctly predicted pixels of a class (true positives) and the union of predicted and ground-truth pixels for that class, including false positives and false negatives. The mIoU is then computed by averaging the IoU values across all K classes:

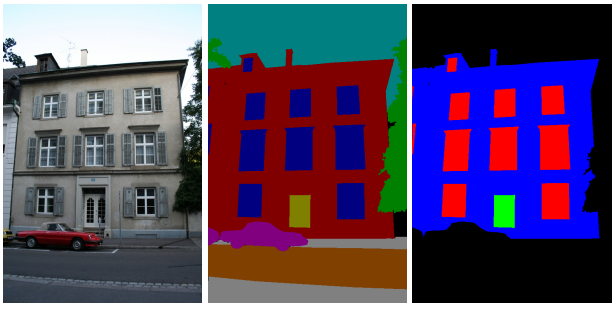
$$\mathrm{IoU_i} = \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i} + \mathrm{FN_i}}, \quad \mathrm{mIoU} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{IoU_i}. \tag{2}$$

In the context of façade image segmentation, mIoU quantifies how accurately the model segments façade elements such as windows, doors, and façade surfaces. A higher mIoU indicates better agreement between predicted and annotated regions, reflecting more precise extraction of façade openings.

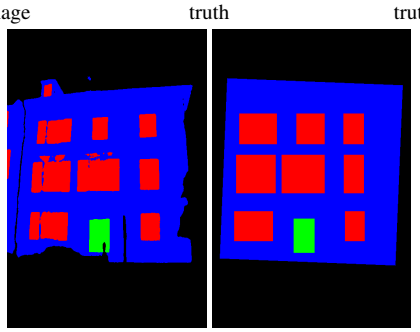
Building upon the Dice Score, the FRDS compares ground truth masks with the re-projection of modeled façades derived from terrestrial LiDAR point clouds (Pantoja-Rosero et al., 2022). This metric assesses both the precision of the surface reconstruction and the layout of the façade openings. Similar to the mIoU, a perfect reconstruction where the re-projected façade overlaps exactly with the ground truth yields an FRDS of 1.0. Following Pantoja-Rosero et al. (2022), the FRDS is defined as:

$$\mathrm{FRDS} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{3}$$

In our approach, the LLM backbone of our agent refines and re-projects the segmentation masks derived from SAM 3, followed by a shape completion step. We consider this process equivalent to the re-projection of modeled façades derived from LiDAR reconstructions. This equivalency holds because our agent extracts the façade openings using expert-designed algorithms and integrates them into the LoD2 building model, which remains static during this reconstruction phase. Consequently, our evaluation compares the ground truth with the LLM-driven refinement of the SAM 3 segmentation masks.



(a) street-view building image (b) eTRIMS ground truth (c) adapted ground truth



(d) segmentation mask (e) shape completion

Figure 4. Original ground truth mask from eTRIMS dataset (b), our ground truth adaptation (c), semantic segmentation using SAM 3 (d) and LLM-driven refinement results (e). Façades represented in blue, windows in red, and doors in green.

3.6 Facade Feature Distance

For evaluating façade reconstruction, metrics such as FRDS and mIoU are inherently limited because they operate strictly at the pixel level, measuring overlap between predicted and ground-truth masks. While effective for assessing segmentation accuracy, they are highly sensitive to small spatial misalignments and fail to capture higher-level properties such as semantic consistency or structural coherence.

In contrast, a high-level feature distance evaluates similarity in a learned latent feature space, where representations encode rich semantic and contextual information. In transformer-based vision models such as DINOv2, image representations can be compared either through the CLS token (classification token), which serves as a global representation of the entire image, or through patch-level features (Oquab et al., 2023). Using the CLS token, the metric reflects global semantic similarity, allowing it to assess whether the reconstructed façade preserves the correct architectural elements and overall scene meaning, even under geometric variations. Alternatively, using patch-level features enables comparison of local structure and spatial layout, capturing how well architectural elements, e.g., windows and doors, are arranged. This makes high-level feature distance metrics significantly better suited for façade reconstruction, where object consistency and structural fidelity are more important than exact pixel-wise correspondence. Let $f_{\mathrm{CLS}}(x) \in \mathbb{R}^d$ denote the CLS embedding of an image x, where d is the feature dimension. The global semantic distance between two images x_1 and x_2 can then be defined as the cosine distance between their CLS embeddings:

$$d_{\mathrm{global}}(x_1, x_2) = 1 - \frac{f_{\mathrm{CLS}}(x_1)^\top f_{\mathrm{CLS}}(x_2)}{\|f_{\mathrm{CLS}}(x_1)\|_2 \, \|f_{\mathrm{CLS}}(x_2)\|_2}. \tag{4}$$

Subsequently, let $f_i(x) \in \mathbb{R}^d$ denote the feature embedding of patch i for image x, with $i = 1, \ldots, N$, where N is the total number of image patches and d is the feature dimension. The patch-wise distance between two images x_1 and x_2 can then be defined as the mean cosine distance over all corresponding patch embeddings:

$$d_{\mathrm{patch}}(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{f_i(x_1)^\top f_i(x_2)}{\|f_i(x_1)\|_2 \, \|f_i(x_2)\|_2} \right). \tag{5}$$

As a result, we propose a new distance metric for evaluating reconstructed façades, termed *Facade Feature Distance (FFD)*, which is based on high-level feature similarity measured via cosine distance. FFD is defined as:

$$\mathrm{FFD}(x_1, x_2) = \frac{d_{\mathrm{global}}(x_1, x_2) + d_{\mathrm{patch}}(x_1, x_2)}{2}, \tag{6}$$

thereby combining both the preservation of the overall architectural semantics and the accurate reconstruction of the local structural layout. In other words, the metric jointly captures

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-XII-4-W1-2026-203-2026

whether the correct architectural elements are present and how well these elements are spatially arranged. Note that the resulting distance lies in the range $[0, 1]$, where values lower indicate higher similarity.

4. Experimental Results

To evaluate the effectiveness of our approach, we conducted a series of experiments using the previously introduced metrics, complemented by a qualitative analysis of a set of architecturally heterogeneous buildings in Hamburg, Germany.

First, we computed the mIoU between the segmentation masks produced by SAM 3 and the color- and class-adapted ground-truth masks from the eTRIMS dataset. This evaluation was intended to isolate and assess the pure semantic segmentation capability of our zero-shot approach. The resulting mIoU of 0.7221 indicates strong segmentation performance on heterogeneous building façades across Europe, as represented in the eTRIMS dataset, particularly considering that the model was neither fine-tuned nor specifically trained for this task. A substantial portion of the observed error in our experimental results can be attributed to the fact that dormer windows, which are not reconstructed by our approach, are present in the eTRIMS ground truth, as illustrated in Figure 4.

To evaluate the reconstruction performance with respect to the semantic and topological dimensions, we computed both the FRDS and the FFD between the LLM-refined façade masks and the corresponding color- and class-adapted ground-truth masks from the eTRIMS dataset. The FRDS is 0.7654, indicating only limited additional expressive power compared with the previously reported mIoU, as it effectively evaluates only the semantic dimension. In contrast, the FFD is 0.4945, with a d_{global} of 0.4523 and a d_{patch} of 0.5367, indicating that our approach performs substantially better in reconstructing façade semantics, that is, the individual façade elements, than in preserving layout or topological consistency. This provides deeper insight into the performance of reconstruction approaches than previous metrics allow. A noticeable share of the topological error can be attributed to the uncalibrated nature of the SVI data and the resulting propagation of uncertainties, which led to errors in the scale and geometric distortion of façade openings in the reconstructed LoD3 model.

Although the agent is guided through the denoising, shape-completion, as well as reprojection steps, and is explicitly instructed to ignore dormer windows, the approach occasionally still predicts such windows in the refined masks. A qualitative comparison between GPT 5.1 and Qwen 3 indicates that GPT 5.1 produces more accurate shape completion, as illustrated in the first and second last row of Figure 5. Furthermore, GPT 5.1 exhibits fewer hallucinations than Qwen 3. This is evident in the third row, where Qwen 3 introduces several non-existent windows in the roof area, and in the fourth row, where it hallucinates an additional basement door, likely influenced by the car window visible in the foreground of the street-view image. Overall, these observations suggest that Qwen 3 is more susceptible to visual noise present in the input images or from inaccuracies in the object segmentation produced by SAM 3 in the preceding step. However, because the expert-designed reconstruction programs provided to the agent do not model dormer windows, these erroneous predictions do not further degrade the final façade reconstruction results. Interestingly, the agent is nevertheless able to detect and successfully reconstruct

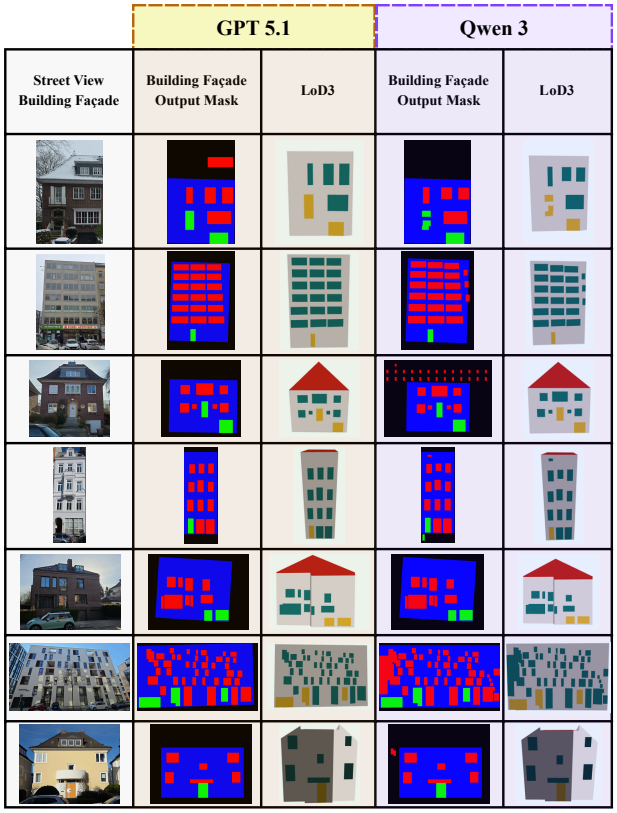


Figure 5. Exemplary results of our approach using either GPT-5.1 or Qwen 3. In the shape completed mask, façades represented in blue, windows in red, and doors in green. In the reconstructed LoD3 model, façades represented in gray, windows in green, and doors in yellow.

the shape of garage doors and small basement windows, even when these elements are partially occluded, as illustrated in the third top row of Figure 5. Furthermore, both models are capable of identifying small windows partially hidden by façade ornamentation and balconies, as shown in the last row of Figure 5.

To analyze the completeness of the reconstructed façades using the eTRIMS dataset, we counted the number of correctly identified windows and doors relative to the ground truth annotations. For windows, our approach achieved a detection rate of 76.66%, indicating that 23.34% of the annotated windows were missed. For doors, the detection rate was 106%, meaning that our approach identified more doors than were annotated in the ground truth. This discrepancy can likely be attributed to the fact that the eTRIMS ground truth contains incomplete and inaccurate annotations, e.g., for garage doors.

5. Outlook & Conclusion

In this paper, we present an end-to-end pipeline for the LoD3 reconstruction of façade openings in 3D city models in a CityGML-conform format using an agent-driven system. Our approach advances existing methods by introducing a zero-shot segmentation strategy that achieves sufficiently strong results in our benchmark on the eTRIMS dataset, while eliminating the need for manual, labor-intensive annotation of training data required by supervised semantic segmentation approaches. A further distinguishing feature of our method is the agent-driven automatic validation step, which produces ready-to-use, CityGML-conform LoD3 building models.

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-XII-4-W1-2026-203-2026

In addition, this paper introduces a novel evaluation approach for façade reconstruction based on the reprojected segmentation masks that are also available in other methods addressing this task. The proposed metric, termed *Facade Feature Distance (FFD)*, leverages high-level feature distances extracted from a vision transformer model to assess similarity in feature space rather than through pixel-wise overlap, as done by existing metrics such as FRDS or mIoU. The rationale behind FFD is that it provides a balanced assessment of both semantic correctness and architectural layout fidelity, which are equally important for façade reconstruction, instead of focusing primarily on semantic overlap as in conventional pixel-wise metrics.

Future research will focus on the semantic enrichment of additional façade characteristics, such as the reconstruction of balconies and protrusions and dormer windows, which could further strengthen the use of semantic 3D city models for urban simulation and analysis. Another promising direction is the integration of additional data sources to support more advanced reasoning strategies and to improve GPS-based map matching, thereby simplifying the correspondence between façade surfaces and camera images.

6. Acknowledgment

The authors gratefully acknowledge the computing time granted by the KISSKI project. This research was supported by the project 'Next Generation City Networking' (Grant No. 19DZ24004) at the Hanseatic Wireless Innovation Competence Center (HAWICC). The project is funded by the Federal Ministry of Transport via the German Center for Future Mobility.

References

Biljecki, F., Ito, K., 2021. Street View Imagery in Urban Analytics and GIS: A Review. *Landscape and Urban Planning*, 215, 104217.

Biljecki, F., Ledoux, H., Stoter, J., 2016a. Generation of Multi-LoD 3D City Models in CityGML with the Procedural Modelling Engine Random3DCity. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-4/W1, 51–59.

Biljecki, F., Ledoux, H., Stoter, J., 2016b. The Most Common Geometric and Semantic Errors in CityGML Datasets. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W1, 13–22.

Carion, N. et al., 2026. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719*.

Dehbi, Y., Hadiji, F., Gröger, G., Kersting, K., Plümer, L., 2016. Statistical Relational Learning of Grammar Rules for 3D Building Reconstruction: Statistical Relational Learning of Grammar Rules for 3D Building Reconstruction. *Transactions in GIS*, 21.

Doosthosseini, A., Decker, J., Nolte, H., Kunkel, J. M., 2024. Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services.

Gröger, G., Plümer, L., 2012. CityGML – Interoperable Semantic 3D City Models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 71, 12-33.

Hou, Y., Biljecki, F., 2022. A Comprehensive Framework for Evaluating the Quality of Street View Imagery. *International Journal of Applied Earth Observation and Geoinformation*, 115, 103094.

Jörg Schmittwilken, 2012. Attributierte Grammatiken zur Rekonstruktion und Interpretation von Fassaden. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Kolbe, T. H., Donaubauer, A., 2021. *Semantic 3D City Modeling and BIM*. Springer Singapore, Singapore, 609–636.

Korč, F., Förstner, W., 2009. eTRIMS Image Database for Interpreting Images of Man-Made Scenes. Technical Report TR-IGG-P-2009-01.

Ma, R., Yang, C., Chen, J., Biljecki, F., Li, X., 2026. Street View Imagery-based Method for Reconstructing 3D Building Façade Openings. *Automation in Construction*, 184, 106842.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*.

Pang, H. E., Biljecki, F., 2022. 3D Building Reconstruction from Single Street View Images using Deep Learning. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102859.

Pantoja-Rosero, B., Achanta, R., Kozinski, M., Fua, P., Perez-Cruz, F., Beyer, K., 2022. Generating LOD3 Building Models from Structure-from-Motion and Semantic Segmentation. *Automation in Construction*, 141, 104430.

Tang, W., Li, W., Liang, X., Wysocki, O., Biljecki, F., Holst, C., Jutzi, B., 2025. Texture2LoD3: Enabling LoD3 Building Reconstruction with Panoramic Images. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2007–2017.

Wang, Q., Yan, L., Zhang, L., Ai, H., Lin, X., 2016. A Semantic Modelling Framework-Based Method for Building Reconstruction from Point Clouds. *Remote Sensing*, 8(9).

Wang, Y., Jiao, W., Fan, H., Zhou, G., 2024. A Framework for Fully Automated Reconstruction of Semantic Building Model at Urban-Scale using Textured LoD2 Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216, 90-108.

Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T. H., 2024. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4-2024, 493–500.

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 . *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE Computer Society, Los Alamitos, CA, USA, 6548–6558.

Xu, F., Hao, Q., Shao, C., Zong, Z., Li, Y., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., Yan, Y., Yang, Q., Song, Y., Ren, S., Hu, X., Feng, J., Gao, C., Li, Y., 2025. Toward Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *Patterns*, 6(10), 101370.