Multi-Source Fusion of Roof Skeletons, LiDAR and Street-View Imagery for Semi-Automated LoD-2 Building Modelling

Vaibhav Rajan[1], Sander Münster[1], Jonas Bruschke[1], Ferdinand Maiwald[2]

[1] Digital Humanities, Friedrich-Schiller-Universität Jena, Germany - (vaibhav.rajan, sander.muenster, jonas.bruschke)@uni-jena.de
[2] Chair of Optical 3D-Metrology, TUD Dresden University of Technology - ferdinand.maiwald@tu-dresden.de

Keywords: LOD-2 Models, Textured 3D Models, Roof Shape Detection, Building Segmentation, LiDAR.

Abstract

LoD-2 building models are more informative and practically more useful than LoD-1 representations because they capture the roof structure that defines the essential three-dimensional form of a building. They are important for applications such as urban planning, environmental simulation, and digital heritage. Although recent roof shape extraction methods can derive vectorised 2D roof structures from very-high-resolution imagery, transforming these image-based representations into fully textured 3D buildings remains challenging. In this paper, we present a semi-automated LoD-2 reconstruction pipeline that integrates HEAT-derived roof geometry with airborne LiDAR, satellite and Google Street View imagery. The 2D outputs are reprojected into map coordinates, fused with LiDAR through a two-stage roof reconstruction strategy to derive roof shapes and combined with an adaptive, LiDAR-based ground base initialisation to create a complete 3D wireframe. Roofs are textured using VHR orthophotos while the walls are textured via a process of Street View panorama selection, geometric filtering, Mask2Former segmentation, and homography rectification. Across a large-scale evaluation on 1000 buildings, the proposed two-stage reconstruction strategy improves geometric agreement with the LiDAR reference data achieving a roof-surface RMSE of 0.445 m. The wall texturing process produces convincing facades when suitable panoramas are available. While minor challenges such as sensitivities to LiDAR outliers, incomplete roof geometry, and facade occlusions persist, this pipeline effectively bridges 2D roof parsing and textured LoD-2 model generation, providing a robust and scalable foundation for advancing toward fully automated workflows.

1. Introduction

1.1 Context

Accurate three-dimensional (3D) building models underpin cultural-heritage documentation and smart city decision-making, enabling virtual preservation, planning, and analysis on an urban scale (Münster et al., 2024; Wysocki et al., 2024). Detailed city models at Level of Detail 2 (LoD-2) with articulated roof forms and thematically distinct surfaces support analyses such as solar potential, shadow impact, and runoff, and are increasingly produced from widely available geospatial data (Rajan et al., 2025). Deep learning has further accelerated this trend by extracting structured roof geometry from overhead imagery as two-dimensional (2D) “skeletons”, providing a compact and vectorised description of roof corners and edges that is well-suited for downstream 3D reconstruction (Chen et al., 2022).

1.2 Motivation

Motivated by these developments, our goal is to turn such 2D roof skeletons into complete, textured LoD-2 models using accessible data sources rather than labour-intensive multi-view photogrammetry. Airborne Light Detection and Ranging (LiDAR) contributes reliable building heights and roof shape cues, while street-level panoramic imagery provides the facade appearance that overhead data cannot capture (Župan et al., 2023; Hoffmann et al., 2019). Combining these multiple modalities allows to retain geometric realism (from LiDAR and roof vectors) and visual realism (from panoramic facade imagery), while reducing acquisition complexity and cost (Ogawa et al., 2024). In this setting, Very High Resolution (VHR) aerial orthophotos along with airborne LiDAR, remain valuable for roof appearance, and street-view panoramas supply view-consistent facade textures across buildings.

1.3 Contribution

This paper investigates whether vectorised roof skeletons can act as reliable structural priors for semi-automated LoD-2 building reconstruction when fused with LiDAR. Rather than presenting the workflow only as an end-to-end system, we study its effectiveness through large-scale quantitative evaluation using LiDAR-based geometric consistency measures. Our contributions are as follows: (1) a geospatial pipeline that lifts 2D roof skeletons into LoD-2 roof and builds the base geometry using LiDAR-guided height fusion; (2) a facade texturing branch based on street-view panorama selection, segmentation, and wall rectification; (3) a quantitative evaluation on 1000 buildings, including comparison against simpler baselines; and (4) a focused ablation and failure analysis identifying which components of the reconstruction pipeline contribute most to performance and where the approach remains limited.

To assess the scientific value of the proposed workflow, this study addresses the following research questions: (1) Does LiDAR-guided fusion of 2D roof skeletons improve LoD-2 roof reconstruction compared with simpler height-assignment baselines? (2) Which components of the geometric pipeline contribute most to reconstruction accuracy? (3) To what extent can the street-view branch provide usable facade textures at scale?

2. Related Work

As the contribution is split into three parts, the related work gives an overview focusing on the three consecutive topics of

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-XI-2-2026-127-2026

2D roof extraction, 3D building modelling, and building texturization.

2D roof extraction has long been studied, typically focusing on detecting outer roof boundaries or polygons (Van Etten et al., 2019). Most geospatial methods remain limited to these outlines (Büyükdemircioğlu et al., 2022). Only a handful of modern models, such as LCNN, HAWP, Conv-MPN, and HEAT, capture internal roof structures (Zhou et al., 2019; Xue et al., 2020; Zhang et al., 2020; Chen et al., 2022). We chose HEAT for its transformer-based design, which recovers full roof skeletons.

Accurate and automatic 3D modelling of building geometry is a long studied problem (Dorninger and Pfeifer, 2008; Haala and Kada, 2010; Maiwald et al., 2023). Approaches using LiDAR data are capable of modelling complete polyhedral buildings at LoD-2 (Jarząbek-Rychard and Borkowski, 2016), comprehensively compiling building primitives to 3D models (Li and Shan, 2022), and generating roofs and building models through end-to-end deep-learning pipelines (Li et al., 2022; Liu et al., 2024). While these works mainly focus on one single data source such as LiDAR, our approach is to combine the information from VHR imagery and Google Street View (GSV) data for accurate reconstruction.

Early facade texturing relied on ground photographs or oblique imagery, requiring precise pose estimation and hardware-heavy mobile-mapping systems to manage occlusions (Früh and Zakhor, 2003). Recent urban-scale research increasingly exploits panoramic street-level imagery as a scalable source of facade appearance (Cinnamon and Jahiu, 2021). Advanced systems now perform automatic panorama-to-building alignment and texture extraction using semantic and geometric cues (Park et al., 2021). Modern segmentation pipelines such as Mask2Former have the potential to further refine facade isolation through transformer-based segmentation, enabling cleaner and more realistic facade mapping (Cheng et al., 2022).

However, most prior research tends to treat 3D building modelling, roof extraction, and facade texturing as independent tasks, rather than integrating them into a unified pipeline for comprehensive urban reconstruction.

3. Methodology

3.1 Overview

The proposed pipeline integrates three complementary data sources – VHR orthophotos, airborne LiDAR, and street level images – to reconstruct textured LoD-2 building models in several consecutive steps (Figure 1).

3.2 From 2D Skeletons to Geospatial 2D Vectors

The roof skeletons used in this study are predicted from VHR imagery by HEAT (Chen et al., 2022), finetuned with 600 training samples to fit our study area. These are subsequently reprojected from image space into map coordinates using preserved georeferencing metadata (Rajan et al., 2025). The resulting vector representation provides a compact and topologically consistent set of roof nodes and edges that serves as the geometric prior for subsequent LiDAR fusion and surface generation (Figure 2).

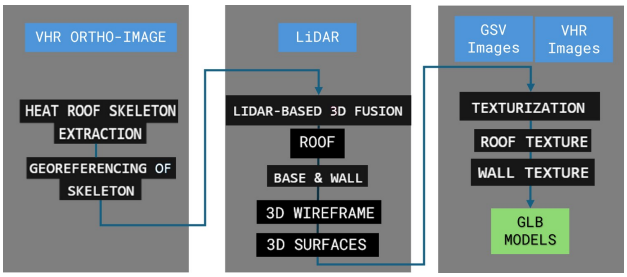


Figure 1. Workflow summarizing roof skeleton extraction, LiDAR-based 3D reconstruction, and facade/roof texturing for LoD-2 model generation.

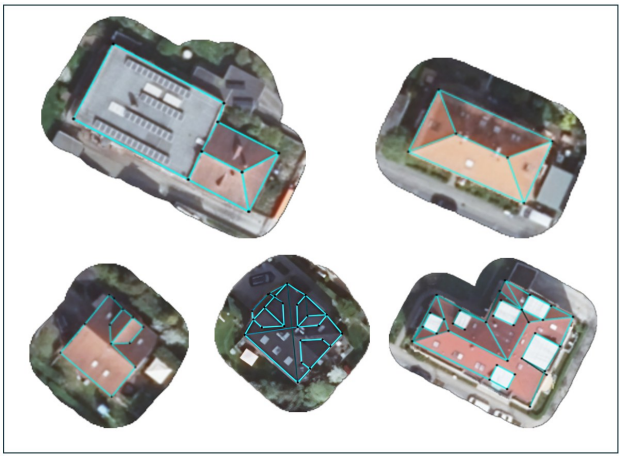


Figure 2. 2D skeletons are generated from HEAT (Chen et al., 2022) model.

3.3 LiDAR-based 3D Fusion

The study area is the city of Jena, Germany, for which LiDAR data acquired in 2011 are available. The point clouds (horizontal accuracy about 0.15 m, vertical about 0.10 m, ETRS89 / Universal Transverse Mercator (UTM) zone 32N) were tiled at 1×1 km and merged to provide continuous urban coverage. They serve as the primary 3D source for injecting elevation into the HEAT derived roof skeletons.

3.3.1 Data Preparation: The LiDAR dataset was provided as a single merged point cloud covering the entire study area. Building footprints were retrieved from OpenStreetMap (OSM) via the Overpass API, returned in geographic coordinates (EPSG:4326), and reprojected to the same projected coordinate reference system as the LiDAR data to ensure spatial correspondence. This procedure follows the same footprint-based cropping logic previously applied to the VHR imagery for HEAT input (Rajan et al., 2025). Each footprint, padded by 5 m (to guarantee full roof coverage), was used to extract individual building point clouds from the main LiDAR dataset.

3.3.2 Alignment: For each building, the georeferenced 2D roof skeletons ($z = 0$) were directly overlaid with their corresponding LiDAR point subsets (Figure 3a). Because both datasets share a common spatial reference, the roof edges coincided precisely with the LiDAR structures. In cases of slight misalignment caused by metadata inconsistencies, an optional rigid 2D Iterative Closest Point (ICP) adjustment was applied to the LiDAR points to refine translation and rotation, ensuring

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-XI-2-2026-127-2026 | © Author(s) 2026. CC BY 4.0 License.

an accurate edge-to-edge correspondence before height attribution.

3.3.3 Height Fusion for Roofs: Figure 3 illustrates the process of deriving roof heights by fusing LiDAR elevation data with the 2D roof skeletons.

We implement a two-stage height reconstruction strategy: Stage I performs edge-wise height lifting using local LiDAR support around structurally relevant roof edges, and Stage II regularises the resulting roof geometry by fitting dominant planes over closed roof regions. In Section 4, Stage I alone is used as a baseline variant, and Stage II forms the full method.

Stage I — Edge-wise height lifting: The georeferenced 2D roof skeleton is first overlaid with the corresponding cropped LiDAR points. For each structurally relevant horizontal roof edge, a narrow 2D corridor is constructed and all LiDAR points falling inside this corridor are collected (Figure 3a). Because these local point sets may contain roof, wall, ground, and other outlier points, the selected points are divided into small vertical bins (e.g., 0.2 m). The most plausible roof level is then identified based on maximum points and spatial spread (Figure 3b and 3c). Finally, a local RANSAC fit is applied so that only the dominant roof-consistent inliers are retained before assigning an initial height estimate to the edge.

$$\left|z_i - (ax_i + by_i + c)\right| \; < \; \tau_z \; = \; 0.3 \; m \qquad (1)$$

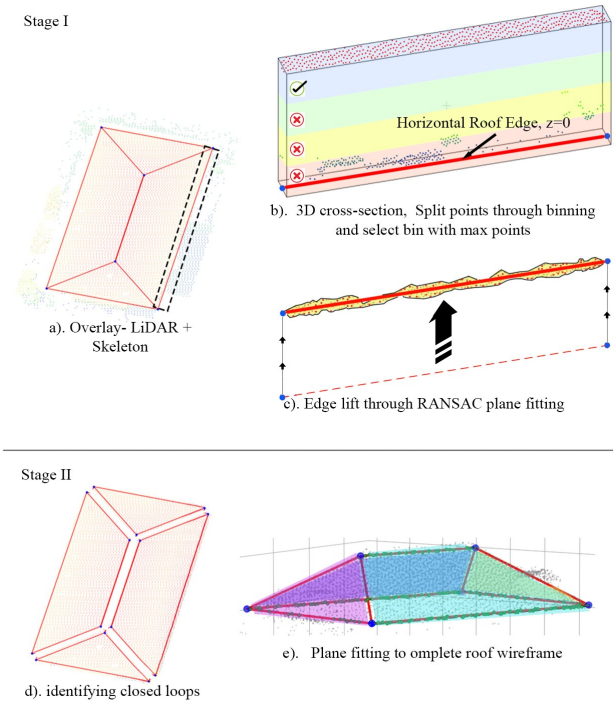


Figure 3. Height fusion process: LiDAR points near selected roof edges are locally filtered to obtain initial edge-wise height estimates, after which closed roof regions are fitted with dominant planes.

Stage II — Plane fitting on closed roof regions: The initial edge-wise height estimates are then regularised at the level of complete roof faces. For this, closed face loops are extracted separately for each connected roof component. For every polygonized roof region, LiDAR points are selected from a buffered face area and used to fit one dominant plane by robust regression. This step replaces purely local edge-wise behaviour by a region-level geometric constraint, so that each roof face is reconstructed as a coherent planar surface rather than as a set of independently lifted edges.

3.3.4 Base, Wall and Surface Generation: Following the roof reconstruction, the model is extended downward to generate the vertical walls and the building base, completing the 3D wireframe structure. For this, the outer footprint of the roof is first extracted using a face-walk algorithm to define the planimetric outline of the building. This becomes the base of the building.

1. *Adaptive radius search:* For each base vertex, the ground elevation (z) is now estimated from the LiDAR data using a simple data-dependent search. Specifically, we query LiDAR points within an adaptive radius centered at the vertex, where the radius is set to half of the shorter of the two incident base edges at that vertex (Figure 4a).
2. *Ground height extraction:* If classified ground points are available, we take the minimum z among ground-class points inside the radius; otherwise, we take the minimum z among all points after removing the roof points within the radius (Figure 4b, 4c, 4d).
3. *Base-wall wireframe construction:* After estimating all vertex bases, each roof vertex is linked vertically to its corresponding base vertex to create the wall edges. Together, these components form the complete 3D wireframe of the building (Figure 4e).

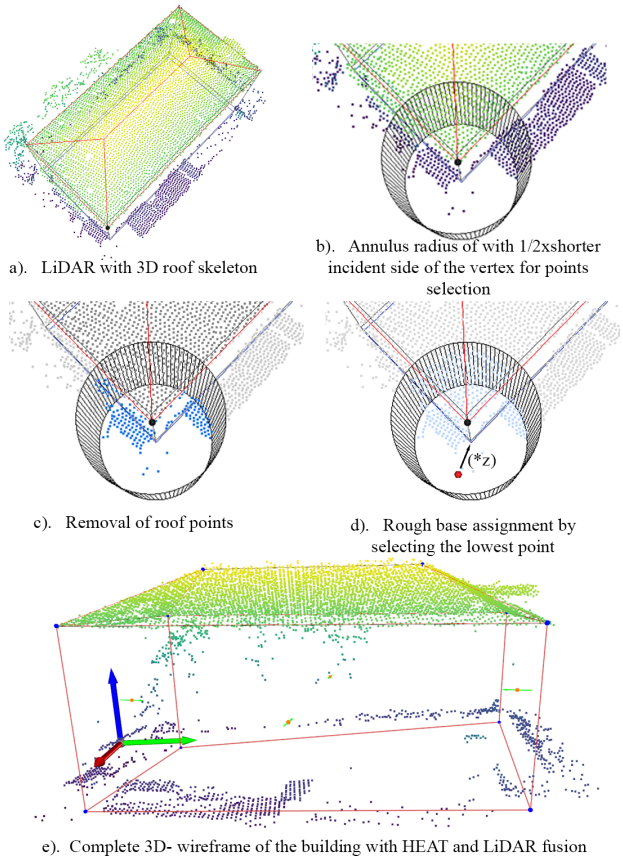
Conventional ground-surface detection methods such as slope-based filtering (Vosselman, 2000), progressive TIN densification (Axelsson, 2000), or segment-based classification were not adopted, as the cropped LiDAR tiles around buildings often contain too few ground points for these techniques to perform reliably. Instead, we use an adaptive up-shoot as a lightweight, vertex-local initialiser for base heights that scales with local footprint geometry and can operate with sparse ground returns. Alternative strategies are discussed in Section 5.

3.4 Texturization

To convert the 3D wireframes into LoD-2 models, we attach appearance from two geospatially aligned sources: VHR orthoimagery for roofs and street-level images for facades. All mappings operate in the building’s projected Coordinate Reference System (CRS) so that texture coordinates are reproducible across buildings. Surfaces are formed via Delaunay triangulation, with roof and base loops oriented consistently and wall quads split so that outward normals point away from the footprint; each triangle is exported with a persistent $face\_id$ and semantic type (roof, wall, or base) for texturing.

3.4.1 Roof Texturization: Roof surfaces are textured first using the OSM-cropped, orthorectified GeoTIFFs that were also used as HEAT input. Because the source is nadir and orthographic, no view-dependent warping is needed. Pixels outside the roof footprint are discarded, and the (Red Green and Blue) RGB bands are used directly without additional radiometric normalisation.

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-XI-2-2026-127-2026 | © Author(s) 2026. CC BY 4.0 License.

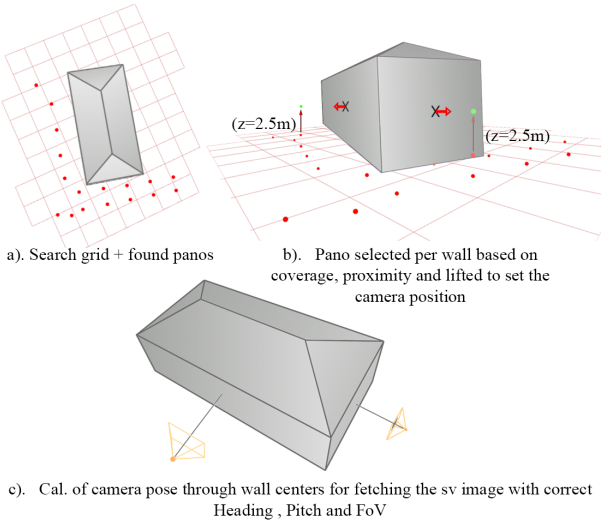


a). LiDAR with 3D roof skeleton
b). Annulus radius of with 1/2xshorter incident side of the vertex for points selection
c). Removal of roof points
d). Rough base assignment by selecting the lowest point
e). Complete 3D- wireframe of the building with HEAT and LiDAR fusion

Figure 4. Base initialisation and 3D wireframe generation: LiDAR points and the 3D roof skeleton are used to estimate a preliminary building base via adaptive radius sampling and lowest-point selection, followed by wall extrusion.

3.4.2 Wall Texturization: Wall texturing is based on Google Street View images. The following steps describe how suitable panoramas are selected, the target facade is isolated, and a rectified wall texture is obtained for mapping onto the model.

1. *Data and search grid:* Using the surfaced 3D model (roof, walls, base) saved in GeoJSON, we build a square search grid around the footprint (buffer ≈ 20 m; grid $\approx 10 \times 10$). Grid nodes are queried against the GSV metadata API. Returned pano locations are retained only if they fall inside the buffered search zone (Figure 5a).
2. *Candidate filtering and best-camera selection:* For each wall we compute its center and outward normal from the corresponding wall quad. We form a half-strip prism by offsetting the base edge slightly behind the wall and extruding outward along the normal (length $>$ footprint size). Candidate panos are kept if they lie in front of the wall (positive projection on the outward normal) and inside this half-strip. We then pick the panorama that is most centered and close to the wall (Figure 5b).
3. *Set Parameters (heading, pitch, FoV):* We assume a camera height of about 2.5 m (car-mounted). The heading is the compass angle from the panorama point to the centre of the wall (0° = north, 90° = east), and the pitch depends on whether the centre of the wall lies above or below the camera. The field of view is chosen to cover the wall span with a small margin, typically between 15° and 120° (Figure 5c).



a). Search grid + found panos
b). Pano selected per wall based on coverage, proximity and lifted to set the camera position
c). Cal. of camera pose through wall centers for fetching the sv image with correct Heading , Pitch and FoV

Figure 5. Grid-Based search and strategic selection of pano locations for calculating the parameters needed to retrieve the SV images with correct camera pose.

4. *Image fetching and wall-aligned cropping:* With the chosen pano and parameters (heading, pitch, FoV, image size 640×640), we fetch a rectilinear view from the GSV API. Using the camera intrinsics implied by FoV and the extrinsics from the pano pose, we project the 3D wall quad of the 3D building into the image to obtain a four-point polygon (Figure 6a). A left/right outward band is then constructed by offsetting the two side edges of this polygon by a fixed pixel margin (default 20 px), keeping full image height to produce a cropped wall image (Figure 6b). This removes neighbouring facades, preserving the target wall with tolerance for small pose errors. If no single view contains all four projected wall corners, additional Street View images are retrieved and mosaicked via feature-based image matching into a single composite facade view before applying the cropping step.
5. *Facade segmentation – Mask2Former:* After isolating the building from other neighbouring structures, we use a segmentation model – Mask2Former – to segment the targeted building facade from the cropped wall image. Mask2Former is a transformer-based, universal image segmentation model that predicts a set of mask–label pairs via masked attention, allowing a single architecture to handle semantic, instance, and panoptic segmentation within a unified framework (Cheng et al., 2022). In this work, we use the ADE20K-pretrained semantic segmentation variant, which is trained on diverse indoor and outdoor scenes (Zhou et al., 2017). All facade-related semantic classes, such as building, house, wall, and window are merged into a single combined mask so that only the facade belonging to the target building remains. Everything else (vegetation, cars, sky, people, neighbouring structures, etc.) is removed (Figure 6c).
6. *Rectification:* The four image points of the wall polygon (earlier projected on the image) are matched to the four corners of the segmented wall in real-world units (width $\times$ height). From these correspondences, we compute a homography that maps image pixels onto the wall plane.

https://doi.org/10.5194/isprs-annals-XI-2-2026-127-2026

The facade is then warped to a fronto-parallel canvas at a fixed pixel density (pixels per meter), producing a clean and metric image ready for texturing (Figure 6d).

Once the rectified wall textures and roof ortho-textures are available, they are projected onto the corresponding 3D surfaces via UV coordinates, completing the transition from LiDAR-based wireframes to textured LoD-2 models. The full pipeline, from roof skeleton extraction and height reconstruction to surface generation and facade rectification, produces compact GLB files suitable for web-based visualisation.

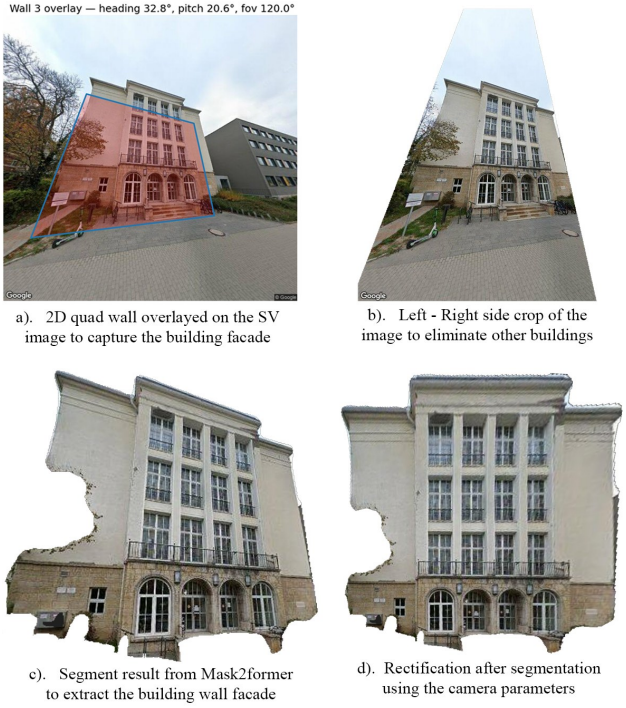


Figure 6. Wall facade generation from a) GSV image using b) cropping, c) segmentation and d) rectification.

4. Results and Evaluation

We evaluate the proposed approach at two levels. First, we assess geometric reconstruction quality on 1000 buildings using LiDAR-based geometric consistency measures. This large-scale evaluation allows us to compare the proposed method with a simpler baseline and to quantify the effect of the second stage of the roof reconstruction pipeline. Secondly, we use a manually inspected subset of 25 buildings for qualitative analysis of characteristic success and failure cases, including facade texturing behaviour.

4.1 Evaluation Setup and Metric

Because the roof height fusion method is explicitly two-stage, we compare two variants. The baseline corresponds to Stage I only, i.e. edge-wise height lifting based on local LiDAR support along roof edges. The full method corresponds to Stage II, i.e. Stage I initialisation followed by plane fitting on closed roof regions.

For quantitative evaluation, we restrict the analysis to roof surfaces and use the LiDAR point clouds as geometric reference data. We report one primary metric: the LiDAR point-to-roof-surface distance RMSE, expressed in meters. For each building, LiDAR points belonging to the roof are compared to the reconstructed roof surface by computing the shortest 3D distance from each LiDAR point to the predicted roof mesh. The RMSE is then calculated over all roof points.

Since LiDAR is also used during reconstruction, this evaluation should be interpreted as a measure of geometric consistency with the LiDAR reference data rather than as a fully independent benchmark. Nevertheless, it provides a direct and practically meaningful way to compare the baseline and full variants of the proposed roof reconstruction pipeline.

4.2 Large-Scale Roof Geometry Evaluation

Table 1 summarises the large-scale evaluation on 1000 buildings using LiDAR-based RMSE. The baseline variant (Stage I), which relies on local edge-wise height lifting, achieves a mean RMSE of 0.634 m. Incorporating Stage II plane fitting reduces the mean RMSE to 0.445 m, corresponding to an improvement of approximately 29.83%. A similar trend is observed for the median RMSE, which decreases from 0.622 m to 0.447 m (28.08% improvement).

Table 1. Large-scale comparison on 1000 buildings using LiDAR point-to-roof-surface RMSE (in meters). Lower is better.

Method	Mean RMSE	Median RMSE
Stage I only (baseline)	0.634	0.622
Stage II (full method)	0.445	0.447
Rel. Improvement (%)	**29.83**	**28.08**

These results demonstrate that the plane-fitting stage substantially improves geometric consistency. While Stage I provides reasonable local height estimates, it remains sensitive to noise, outliers, and inconsistencies across adjacent edges. Stage II regularises these estimates by enforcing dominant planar structures over closed roof regions, resulting in smoother and more coherent roof surfaces that better align with the LiDAR data.

4.3 Qualitative Roof and Base Analysis

To complement the large-scale quantitative evaluation over 1000 buildings, representative cases were manually inspected for qualitative error analysis. The examples shown in this section are intended to illustrate the typical behaviour, strengths, and remaining limitations of the proposed method.

Figure 7 compares the baseline Stage I reconstruction with the full Stage II method on representative examples. Case (a) shows that for very simple roof geometries both variants can already produce a correct reconstruction, since the local edge-wise lifting is sufficiently stable and the subsequent plane regularisation does not need to introduce substantial changes. In more articulated cases, however, clear differences emerge. Example (b) illustrates that Stage I is more vulnerable to local outliers because height assignment is performed only within narrow edge-based support regions; this can lead to deformed or inconsistent roof faces. By contrast, the full method uses closed roof regions to fit dominant planes, which regularises neighbouring edges jointly and yields a roof shape that aligns much more closely with the LiDAR reference. A similar behaviour is visible in case (c), where the Stage I result exhibits a strong geometric distortion that disappears after Stage II plane fitting. Example (d) further shows that the full method is better able to

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-XI-2-2026-127-2026 | © Author(s) 2026. CC BY 4.0 License.

preserve roof structures with distinct height-level differences across adjacent regions, which are not recovered reliably by the edge-wise baseline alone.

At the same time, the qualitative analysis also reveals a limitation of the full method. In case (e), nearby vegetation appears to contaminate the LiDAR support of a closed roof region, so that the fitted plane becomes biased and introduces a disoriented roof face. Overall, the examples indicate that the reconstruction benefits from progressing from a local initialisation to a regional regularisation: Stage I provides useful edge-wise height cues, while Stage II consolidates these cues into coherent roof faces through plane fitting on closed regions. This generally improves roof-height correctness and yields more geometrically plausible roof shapes.

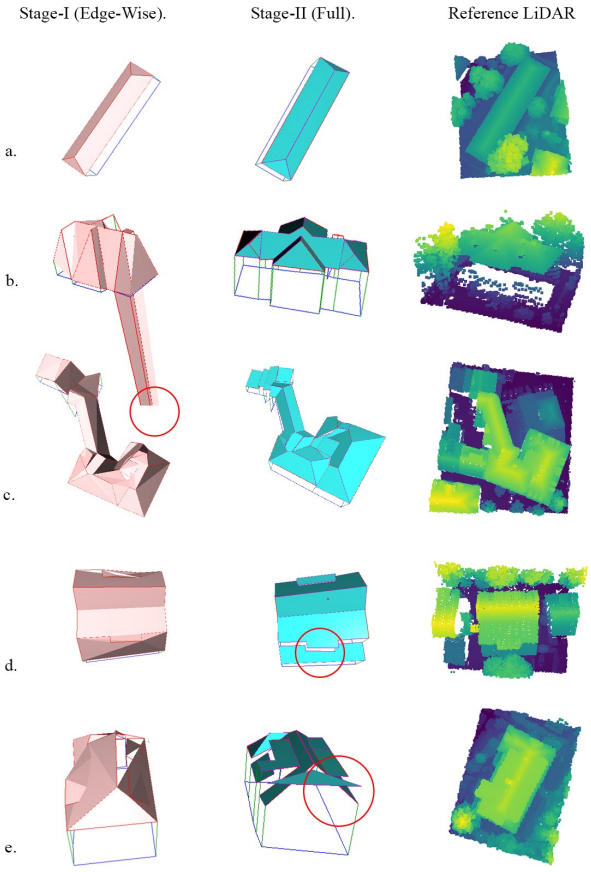


Figure 7. Representative qualitative comparison of roof reconstruction behaviour for Stage I only (left), the full Stage II method (middle), and the LiDAR reference (right).

The behaviour of the adaptive base estimation is also assessed qualitatively. Figure 8 shows representative examples of inferred base rings overlaid with the point cloud in 3D. For many buildings on relatively flat terrain, the estimated base elevations remain stable and consistent across neighbouring vertices. The method also performs reasonably well on continuous slopes, where the adaptive search radius can still capture the dominant local ground level.

However, the base estimation is more sensitive than the roof reconstruction stage. In stepped terrain, sloped driveways, or vegetated surroundings, the adaptive radius may include points from lower terraces or other locally low returns. In such cases, the “lowest point” rule can pull parts of the base ring below the visually expected ground level, producing uneven or floating wall segments. Extreme failures may occur when isolated low outliers dominate the local search neighbourhood. These examples indicate that the current up-shoot strategy provides a useful first-order base estimate, but that it remains sensitive to local outliers and terrain complexity.

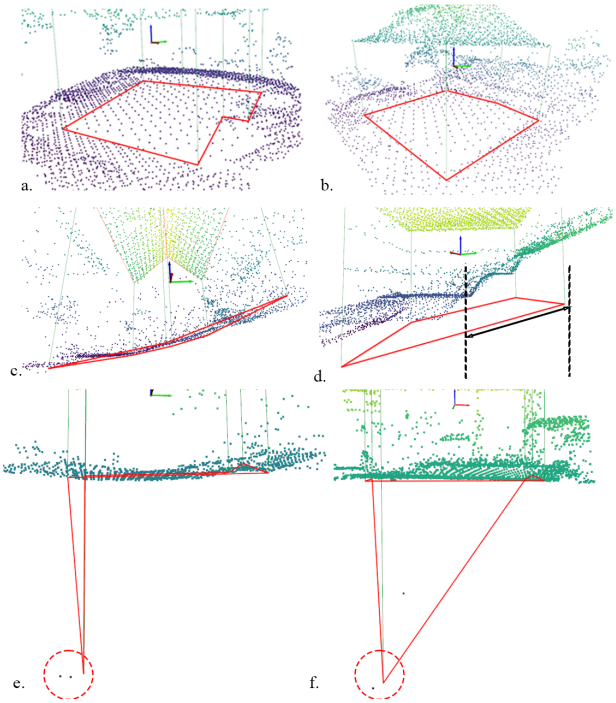


Figure 8. Representative qualitative examples of base height behaviour: (a,b) good fits on relatively flat terrain; (c) reasonable behaviour on continuous slope; and (d) a failure case where the adaptive lowest-point rule pulls the base below the expected ground level.

4.4 Wall Texturization Performance

The second evaluation axis concerns the practical usability of the facade-texturing branch. Because this stage depends not only on reconstruction quality but also on panorama availability, visibility, and occlusion, it is not evaluated with the same geometric metric as the roof reconstruction. Instead, this branch is analysed through qualitative examples and operational outcomes.

Figure 9 illustrates representative texturing results. When a well-positioned panorama is available, the combination of outward-band cropping, facade segmentation, and homography-based rectification produces wall textures that are both geometrically plausible and visually coherent. In these cases, Mask2Former preserves the main facade structure while removing a large portion of vegetation, vehicles, and street clutter. At the same time, the resulting masks may still contain small holes and missing wall fragments, which reduce the visual completeness of the final textures (Figure 9a, 9c).

The examples also reveal a scale mismatch between the projected 3D wall and the visible facade in the panorama: in some cases the wall polygon is too large and mostly empty inside, whereas in others it is too small and the facade extends beyond

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-XI-2-2026-127-2026 | © Author(s) 2026. CC BY 4.0 License.

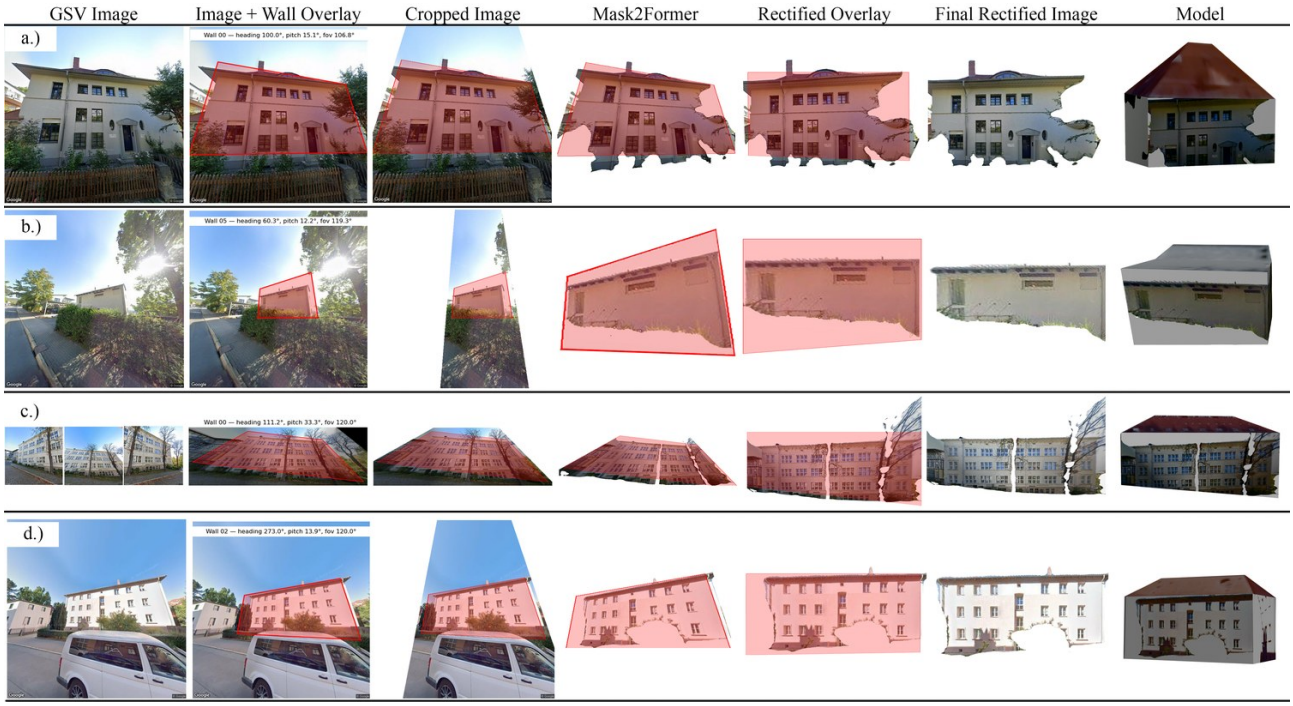


Figure 9. Examples of Street View–based wall texturing: GSV image, projected wall, cropping, Mask2Former segmentation, rectification, and the resulting textured 3D model.

the projected boundaries (Figure 9b, 9d). Despite these limitations, the selected panoramas and projection geometry consistently target the correct building, so that facades are not projected onto neighbouring structures. In many unsuccessful cases, however, the projection falls on heavily occluded facades, leaving only fragmented patches after segmentation that are visually unsuitable for texturing.

5. Discussion

The degree of automation in the proposed pipeline depends mainly on the completeness of the input roof skeleton. When Stage I is considered on its own, height assignment relies on structurally relevant horizontal edges, whose identification may require manual selection. In the full Stage II formulation, however, roof faces are reconstructed automatically by detecting closed loops in the skeleton and fitting dominant planes to the corresponding regions. As a result, the roof-height fusion itself becomes largely automatic once a complete and topologically valid skeleton is available. The remaining semi-automated aspect of the workflow therefore lies primarily in verifying that the extracted roof skeleton is sufficiently complete and suitable for closed-loop analysis. Further improvements in skeleton extraction and completion would therefore be an important step toward a fully automated pipeline.

The underlying HEAT model for skeleton extraction was trained on 600 samples covering a range of roof geometries from simple to more complex cases. In many instances, this already yielded roof skeletons that were sufficiently accurate for subsequent height fusion. However, the qualitative analysis showed that more complex roof forms remain more challenging, as some skeletons were still incomplete and required manual completion before closed-loop detection and Stage II plane fitting could be applied reliably. This indicates that the current limitation lies less in the two-stage height fusion itself than in the completeness of the initial skeleton extraction. A promising direction toward greater automation is therefore to further improve the HEAT model through training on a larger and more diverse set of roof geometries, and/or to integrate a parametric strategy for completing incomplete skeletons.

The base heights are deliberately initialised with a simple “first-hit from below” rule, taking the lowest LiDAR elevation within an adaptive radius around each footprint vertex. This keeps the method lightweight and scalable, but also means that LiDAR is not an ideal reference for rigorous base evaluation. Ground returns near facades are often sparse, noisy, or mixed with vegetation and ramps. Therefore, the minimum LiDAR point does not necessarily match the visually perceived footing line. A more reliable assessment of base height would require image-based evidence, in particular cues derived from the lower termination of rectified facade textures, and integrating such facade cues is planned as part of future work beyond the current LiDAR-only method.

The facade pipeline, in its current form, only textures walls when at least one panorama passes the geometric filters and Mask2Former yields a sufficiently coherent facade mask. However, walls with missing, heavily occluded or unusable panoramas are simply left untextured. This leads to visually incomplete buildings even when the underlying geometry is correct. A natural next step is to treat these gaps explicitly as a completion problem: generative models such as diffusion-based texture synthesis or learned inpainting could be used to fill missing or fragmented wall regions with plausible facade appearance that respects the building’s geometry and surrounding context.

Overall, the proposed pipeline demonstrates that vectorised roof skeletons, LiDAR, and street-level imagery can be integrated in a modular workflow for the reconstruction of LoD-2 building models from widely available data sources. Within this

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-XI-2-2026-127-2026

workflow, particular emphasis is placed on the roof reconstruction stage, since the correct roof geometry is the key element that determines the LoD-2 character of the final model. For this reason, the paper gives strongest weight to the extraction of geometrically consistent roof shapes and their height reconstruction through the proposed two-stage fusion strategy. The subsequent base estimation complements this by completing the overall building volume, so that the resulting model captures the essential three-dimensional form. In addition, the facade texturing component shows how street-level imagery can be linked to the reconstructed geometry to enrich the model visually, even though this part of the pipeline is still at a more preliminary stage and requires further development for greater completeness and robustness.

6. Conclusion

This paper presented a semi-automated pipeline that transforms 2D roof skeletons into textured LoD-2 building models by combining VHR imagery, LiDAR point clouds, and Google Street View images. Across a large-scale evaluation on 1000 buildings, the proposed two-stage reconstruction strategy showed improved geometric agreement with the LiDAR reference compared with the Stage I baseline, while the facade workflow demonstrated that street-level imagery can be projected onto reconstructed 3D wall surfaces to create visually plausible textures whenever suitable panoramas were available. The study also highlights areas where the approach can be strengthened, particularly in improving base-height reliability, handling missing or occluded facades, and reducing the remaining manual steps that limit full automation. These aspects outline a clear direction for advancing the pipeline toward a fully automated reconstruction workflow.

Acknowledgements

We greatly acknowledge the authors and developers of the HEAT (Chen et al., 2022) and Mask2Former (Cheng et al., 2022) model for openly sharing their source code via GitHub. Their contributions have been instrumental in enabling this research. The authors thank Samuel Glowka, student assistant at Friedrich Schiller University Jena, Germany, for his support in preparing the training samples for the HEAT model.

The research upon which this paper is based was carried out in the EU projects INDUX-R (Grant No. 101135556) and 3DBigDataSpace (Grant No. 101173385).

References

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. *International Archives of Photogrammetry and Remote Sensing*, 33(B4), 110–117.

Büyükdemircioğlu, M., Can, R., Kocaman, S., Kada, M., 2022. Deep Learning Based Building Footprint Extraction from Very High Resolution True Orthophotos and NDSM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-2-2022, 211–218. https://doi.org/10.5194/isprs-annals-V-2-2022-211-2022.

Chen, J., Qian, Y., Furukawa, Y., 2022. HEAT: Holistic edge attention transformer for structured reconstruction. *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3866–3875. https://doi.org/10.1109/CVPR52688.2022.00384.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135.

Cinnamon, J., Jahiu, L., 2021. Panoramic Street-Level Imagery in Data-Driven Urban Research: A Comprehensive Global Review of Applications, Techniques, and Practical Considerations. *ISPRS International Journal of Geo-Information*, 10(7), 471. https://doi.org/10.3390/ijgi10070471.

Dorninger, P., Pfeifer, N., 2008. A Comprehensive Automated 3D Approach for Building Extraction, Reconstruction, and Regularization from Airborne Laser Scanning Point Clouds. *Sensors*, 8(11), 7323–7343. https://doi.org/10.3390/s8117323.

Früh, C., Zakhor, A., 2003. Constructing 3D city models by merging aerial and ground views. *IEEE Computer Graphics and Applications*, 23(6), 52–61. https://doi.org/10.1109/MCG.2003.1242382.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 570–580. https://doi.org/10.1016/j.isprsjprs.2010.09.006.

Hoffmann, E. J., Wang, Y., Werner, M., Kang, J., Zhu, X. X., 2019. Model Fusion for Building Type Classification from Aerial and Street View Images. *Remote Sensing*, 11(11), 1259. https://doi.org/10.3390/rs11111259.

Jarząbek-Rychard, M., Borkowski, A., 2016. 3D building reconstruction from ALS data using unambiguous decomposition into elementary structures. *ISPRS Journal of Photogrammetry and Remote Sensing*, 118, 1–12. https://doi.org/10.1016/j.isprsjprs.2016.04.005.

Li, L., Song, N., Sun, F., Liu, X., Wang, R., Yao, J., Cao, S., 2022. Point2Roof: End-to-end 3D building roof modeling from airborne LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 193, 17–28. https://doi.org/10.1016/j.isprsjprs.2022.08.027.

Li, Z., Shan, J., 2022. RANSAC-based multi primitive building reconstruction from 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 247–260. https://doi.org/10.1016/j.isprsjprs.2021.12.012.

Liu, Y., Obukhov, A., Wegner, J. D., Schindler, K., 2024. Point2Building: Reconstructing buildings from airborne LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215, 351–368. https://doi.org/10.1016/j.isprsjprs.2024.07.012.

Maiwald, F., Komorowicz, D., Munir, I., Beck, C., Münster, S., 2023. Semi-automatic generation of historical urban 3D models at a larger scale using structure-from-motion, neural rendering and historical maps. S. Münster, A. Pattee, C. Kröber, F. Niebling (eds), *Research and Education in Urban History in the Age of Digital Libraries. UHDL 2023*, Springer, Cham, 107–127. https://doi.org/10.1007/978-3-031-38871-2_7.

Münster, S., Maiwald, F., di Lenardo, I., Henriksson, J., Isaac, A., Graf, M. M., Beck, C., Oomen, J., 2024. Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe. *Heritage*, 7(2), 794–816. https://doi.org/10.3390/heritage7020038.

Ogawa, Y., Nakamura, R., Sato, G., Maeda, H., Sekimoto, Y., 2024. End-to-End Framework for the Automatic Matching of Omnidirectional Street Images and Building Data and the Creation of 3D Building Models. *Remote Sensing*, 16(11), 1858. https://doi.org/10.3390/rs16111858.

Park, J., Jeon, I.-B., Yoon, S.-E., Woo, W., 2021. Instant Panoramic Texture Mapping with Semantic Object Matching for Large-Scale Urban Scene Reproduction. *IEEE Transactions on Visualization and Computer Graphics*, 27(5), 2746–2756. https://doi.org/10.1109/TVCG.2021.3067768.

Rajan, V., Münster, S., Bruschke, J., Maiwald, F., 2025. Towards LOD-2 Building Reconstruction: Leveraging Segmentation and Roof Shape Extraction Methods from VHR Imagery. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-M-9-2025, 1251–1256. https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-1251-2025.

Van Etten, A., Lindenbaum, D., Bacastow, T. M., 2019. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv prepint arXiv:1807.01232*. https://arxiv.org/abs/1807.01232.

Vosselman, G., 2000. Slope based filtering of laser altimetry data. *International Archives of Photogrammetry and Remote Sensing*, 33(B3/2), 935–942.

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. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-4-2024, 493–500. https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-493-2024.

Xue, N., Wu, T., Bai, S., Wang, F., Xia, G.-S., Zhang, L., Torr, P. H. S., 2020. Holistically-attracted wireframe parsing. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2785–2794. https://doi.org/10.1109/CVPR42600.2020.00286.

Zhang, F., Nauata, N., Furukawa, Y., 2020. Conv-MPN: Convolutional message passing neural network for structured outdoor architecture reconstruction. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2795–2804. https://doi.org/10.1109/CVPR42600.2020.00287.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ADE20K dataset. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5122–5130. https://doi.org/10.1109/CVPR.2017.544.

Zhou, Y., Qi, H., Ma, Y., 2019. End-to-end wireframe parsing. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 962–971. https://doi.org/10.1109/ICCV.2019.00105.

Župan, R., Vinković, A., Nikçi, R., Pinjatela, B., 2023. Automatic 3D Building Model Generation from Airborne LiDAR Data and OpenStreetMap Using Procedural Modeling. *Information*, 14(7), 394. https://doi.org/10.3390/info14070394.