ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

LoD2-Former: Multi-Modal Transformer-Based 3D Building Wireframe Reconstruction

Youssef Abdelhedi[1,3], Daniel Panangian[1], Chaikal Amrullah[1], Houda Chaabouni-Chouayakh[2], Ksenia Bittner[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany – (youssef.abdelhedi, daniel.panangian, chaikal.amrullah, ksenia.bittner)@dlr.de
[2] Sm@rts Laboratory, Digital Research Center of Sfax, Tunisia – houda.chaabouni@crns.rnrt.tn
[3] Higher School of Communication of Tunis, Tunisia – youssef.abdelhedi@supcom.tn

Keywords: LoD2 Building Reconstruction, 3D Building Modeling, Transformer, Deep Learning, Multi-Modal Remote Sensing

Abstract

This paper presents LoD2-Former, a multi-modal Transformer architecture for end-to-end 3D roof wireframe reconstruction from both light detection and ranging (LiDAR) point clouds and aerial imagery. Unlike existing methods that rely solely on point clouds, LoD2-Former leverages complementary geometric and visual information to address challenges posed by sparse and incomplete airborne LiDAR data. State-of-the-art methods for 3D roof wireframe reconstruction typically explore the search space from 3D to 2D by first generating 2D heatmaps of roof corner probabilities from point cloud features, lifting the predicted corners back to 3D, and then inferring edge connections. While effective, these purely point-cloud-driven approaches leave substantial information unexploited, particularly from complementary 2D data sources. In this work, we investigate how integrating aerial optical imagery can improve reconstruction accuracy and provide insights into optimal multi-modal fusion strategies, highlighting the advantages and limitations of combining geometric and visual cues. We also introduce a robust pipeline for collecting, cleaning and matching aerial images with LiDAR point cloud, enabling the reconstruction of complete 3D roof wireframes. Experiments on two datasets demonstrate that LoD2-Former surpasses state-of-the-art baselines and mitigates the challenges posed by sparse or incomplete point clouds. To allow further comparisons with our methodology the dataset has been made available at https://github.com/KseniaBittner/LoD2-Former

1. Introduction

Accurate reconstruction of building level of detail (LoD)2 wireframes is essential for applications such as urban mapping, digital twins, and autonomous navigation. Recent advances in deep learning have substantially improved the automation and reliability of wireframe extraction from various input modalities, including light detection and ranging (LiDAR) point clouds, multi-view images, and remote sensing imagery, enabling the detection of intricate features such as roof structures and external building details beyond basic volumetric shapes. Among these modalities, LiDAR point clouds have become a standard choice for large-scale urban modeling due to their large coverage and high accuracy. Nevertheless, when relying solely on LiDAR, significant challenges persist due to inevitable issues such as noise, occlusions, and scanning artifacts in the data, making the generation of complete and geometrically faithful LoD2 models considerably more difficult.

Recent learning-based methods for 3D roof wireframe reconstruction operate either directly on 3D point clouds or on 2.5D height maps derived from airborne LiDAR. Despite architectural differences, they share a common limitation. They ultimately rely on LiDAR alone, making them sensitive to sparsity, uneven point distributions, and occlusions. At the same time, aerial imagery provides rich complementary cues, such as edge discontinuities, texture variations, and shadow patterns that can reveal structural boundaries even when LiDAR coverage is sparse or corrupted. Integrating these two modalities therefore presents a promising direction for improving LoD2 wireframe reconstruction. To the best of our knowledge, however, existing 3D roof wireframe methods do not exploit aerial images in an end-to-end fashion.

Figure 1. Example of LoD2-Former reconstruction on Talinn area. The background shows an aerial orthoimage of a residential block, while the overlaid white polygons depict LoD2 building wireframes. The figure illustrates that the model recovers complex multi-part roofs and ridges for detached houses and building blocks.

Apart from the challenges discussed before, the lack of high-quality real-world datasets also hinders progress in this field. High-fidelity supervision is mostly found in synthetic data-

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-2-2026-187-2026 | © Author(s) 2026. CC BY 4.0 License.

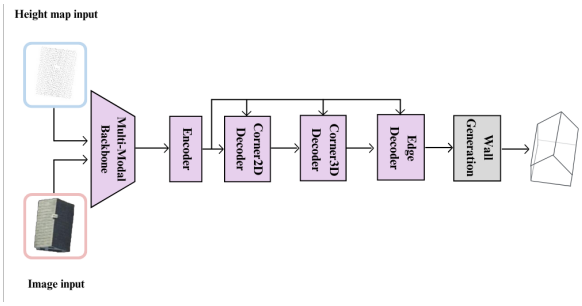

Figure 2. General architecture of the proposed LOD2-FORMER framework. Dual-modality inputs (height map and aerial image) are processed via the LOD2-FORMER dual backbone, followed by shared feature encoding, sequential 2D and 3D corner detection models, and final edge-based wireframe reconstruction.

sets such as Roof-Intuitive (Ren et al., 2021), which provide dense and more complete meshes paired with images that are largely free of topological or geometric errors. In contrast, real airborne LiDAR datasets such as the recently widely used Building3D (Wang et al., 2023) present significant limitations for multi-modal research. In particular, accurate alignment between LiDAR point clouds and available aerial imagery is often inadequate, posing a major obstacle for approaches like ours that depend on precise cross-modal correspondence.

Focusing on the challenges listed above, we propose in this paper an end-to-end multi-modal fusion framework for LoD2 roof wireframe reconstruction that jointly leverages LiDAR-derived height maps and aerial imagery. Our approach addresses the shortcomings of purely geometric methods by integrating complementary visual and structural information, enabling more robust, complete, and semantically coherent building wireframe reconstruction. Below, we list our contributions:

1. A curated subset of the Building3D dataset consisting of LiDAR point clouds paired with well-aligned aerial images, facilitating future multi-modal benchmarking.
2. An end-to-end multi-modal fusion architecture, LOD2-FORMER, that fuses LiDAR-derived height maps and aerial imagery in a multimodal-backbone Transformer and outputs 3D roof wireframes.
3. A simple yet effective LoD2 completion module that procedurally extrudes walls from the predicted roof wireframes to obtain watertight LoD2 building models.

2. Related Work

2.1 Building Reconstruction from Point Clouds

Early work on automatic building reconstruction from airborne LiDAR focused on recovering planar roof patches and assembling them into polyhedral building models. Forlani et al. (2006) perform a complete classification of raw laser points and reconstruct buildings by clustering planar regions and enforcing geometric consistency. Huang et al. (2013) formulate roof reconstruction as a generative statistical inference problem, jointly optimizing roof topology and plane parameters from airborne laser scanning point clouds. Other methods emphasize robust plane extraction using variants of the Hough transform or random sample consensus (RANSAC); for example, Maltezos and Ioannidis (2016) apply an extended 3D randomized Hough transform to detect roof planes and then regularize their intersections to form structured LoD2 models. These approaches established the pipeline of (i) detecting roof primitives, (ii) enforcing regularity priors such as orthogonality and parallelism, and (iii) assembling watertight polygonal roofs, but they rely heavily on hand-crafted rules and often struggle with incomplete data and irregular roof shapes.

2.2 Wireframe Parsing in Images

In parallel, the computer vision community has developed wireframe parsing as a structured representation of man-made scenes in images. Huang et al. (2018) introduced the wireframe parsing task and a large-scale dataset, proposing a convolutional neural network (CNN)-based baseline that detects junctions and line segments and links them into a 2D wireframe. Later, Zhou et al. (2019) presented L-CNN, which replaces heuristic post-processing with an end-to-end network that directly outputs vectorized wireframes. Xue et al. (2020) further improved robustness with holistically-attracted wireframe parsing (HAWP), using attraction-field representations and line-of-interest pooling to better couple junctions and line segments. More recently, Chen et al. (2022) proposed holistic edge attention transformer (HEAT) for structured reconstruction. HEAT takes a 2D raster image (e.g., satellite roof crops or indoor point-density maps) as input and reconstructs a planar graph by detecting corners and classifying edge candidates in an end-to-end transformer architecture. Overall, these methods show that line–junction graphs are an effective representation for man-made geometry, but they operate purely in the image domain and do not exploit 3D structure.

2.3 3D Building Wireframes Reconstruction

Closer to our setting, several recent works address 3D building wireframe reconstruction directly from airborne LiDAR. Building3D (Wang et al., 2023) provides an urban-scale dataset with point clouds, meshes, and roof wireframes, enabling learning-based building modeling at city scale. WireframeNet (Cao et al., 2023) takes unordered point clouds as input and predicts complete 3D wireframes by jointly detecting feature points and curve segments. On top of Building3D, Points2Model (Akwensi et al., 2025) introduces a neural-guided pipeline that iteratively refines edge candidates and explicitly reasons about edge topology, while PBWR (Huang et al., 2024) uses a transformer to regress parametric edge entities directly from LiDAR and performs edge non-maximum suppression in the edge space. Robust Building Wireframe Reconstruction (Gong et al., 2025) augments this line of work with a hypergraph formulation and transformer-enhanced message passing to better handle large-scale, real-world urban scenes, and BWFormer (Liu et al., 2025) extends HEAT to airborne LiDAR by exploiting its 2.5D nature: point clouds are projected into 2D height maps, roof corners and edges are predicted in the image plane, and corners are then lifted back to 3D to form roof wireframes. Our work follows this general 2D–3D lifting paradigm but departs from BWFormer in a key way: instead of relying solely on LiDAR, LOD2-FORMER introduces a dual-stream backbone with cross-modal attention that fuses aerial imagery and height maps, allowing appearance cues (edges, textures, shadows) and geometric cues to jointly drive explicit 3D junction–edge roof graph prediction.

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-2-2026-187-2026 | © Author(s) 2026. CC BY 4.0 License.

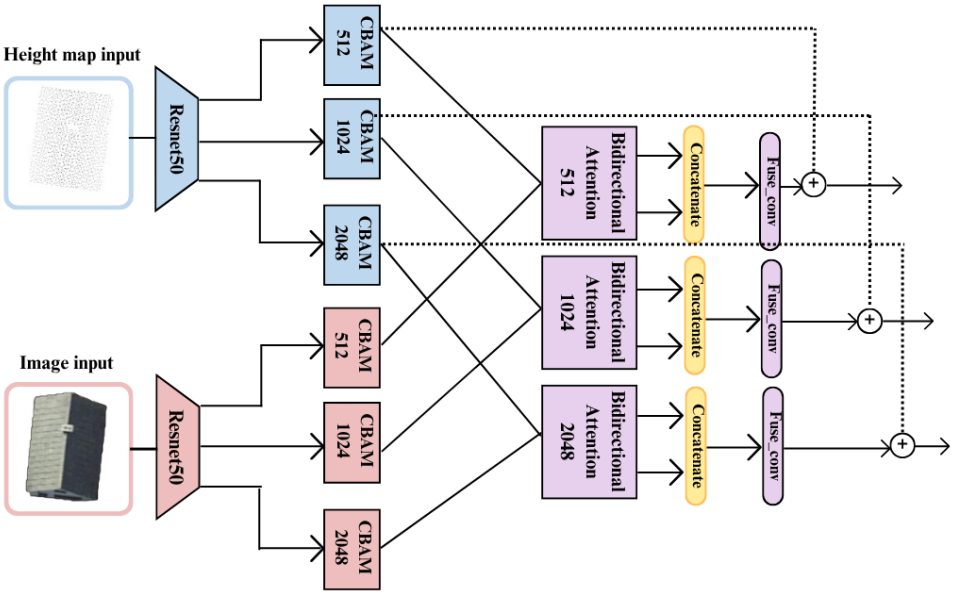

Figure 3. LoD2-Former Dual Backbone illustration. the height map and the aerial image inputs are independently processed into multiple feature scales via parallel ResNet50 blocks combined with CBAM attention. The solid lines show data flow, the dotted lines depict residual connections, and the summation points are marked by the symbole $\oplus$. This module passes fused features to the subsequent Bidirectional Attention block shown in Figure 4.

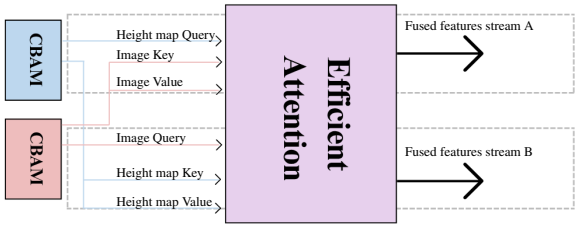

Figure 4. Bidirectional attention fusion mechanism. Features from the dual backbone are further enhanced using bidirectional attention and fusion convolutions at each scale. The CBAM blocks refine information before entering the attention modules. Outputs are merged to yield fused features streams (A and B), which are passed to the downstream encoder and corner models as shown in Figure 2.

3. Methodology

3.1 Overview

LoD2-Former takes as input an aerial image and an airborne LiDAR-derived height map of the same building and outputs a 3D roof wireframe, which is then completed to an LoD2 building model (Figure 2). The method consists of a multi-modal backbone that fuses the aerial image and height information at multiple scales, followed by three prediction heads for 2D corner detection, 3D corner lifting, and edge classification. The overall design follows the HEAT (Chen et al., 2022) and BWFormer (Liu et al., 2025) pipeline for structured wireframe prediction, but replaces the single-modality height map backbone with a dual-stream architecture that explicitly fuses appearance and geometry.

3.2 Multi-Modal Backbone

Given an input aerial image $I_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and a height map $I_{height} \in \mathbb{R}^{H\times W\times 3}$, we first extract multi-scale features using separate ResNet-50 encoders ϕ_{rgb} and ϕ_{height}:

$$F_s^m = \phi_m^{(s)}(I_m), \quad m \in \{\text{rgb}, \text{height}\},\ s \in \{8, 16, 32\}, \quad (1)$$

where $F_s^m \in \mathbb{R}^{B\times C_s\times \frac{H}{s}\times \frac{W}{s}}$ and $C_s \in \{512, 1024, 2048\}$ denote the channel dimensions at each stride.

For each scale s and modality m, we refine the features using the convolutional block attention module (CBAM) (Woo et al., 2018), which applies channel-wise and spatial attention in sequence:

$$\hat{F}_s^m = \text{CBAM}(F_s^m). \quad (2)$$

It serves to reweight informative channels and locations in each modality before fusion.

To combine aerial image and height information, we employ bidirectional cross-modal attention inspired by efficient attention mechanisms (Shen et al., 2021). At each scale s, we compute two attention streams:

$$A_s^{(A)} = \text{CrossAttn}(\hat{F}_s^{height}, \hat{F}_s^{rgb}), \quad (3)$$
$$A_s^{(B)} = \text{CrossAttn}(\hat{F}_s^{rgb}, \hat{F}_s^{height}), \quad (4)$$

where $\text{CrossAttn}(Q_{feat}, K_{feat})$ projects features to query, key, and value tensors, applies a linear-complexity attention as in Shen et al. (2021), and adds a residual connection to Q_{feat}. This design allows the aerial image to attend to geometric cues from the height map and vice versa, while keeping computational cost manageable for large feature maps.

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-2-2026-187-2026 | © Author(s) 2026. CC BY 4.0 License.

The two attention streams are then fused via concatenation and a 1×1 projection with Group Normalization and ReLU, augmented with a residual connection from the height map branch:

$$\mathcal{F}_s = \mathrm{ReLU}\big(\mathrm{GN}(\mathrm{Conv}_{1\times1}([A_s^{(A)}; A_s^{(B)}]))\big) + \mathrm{Proj}(\hat{F}_s^{height}), \tag{5}$$

where $\mathcal{F}_s \in \mathbb{R}^{B\times C_s\times \frac{H}{s}\times\frac{W}{s}}$ denotes the fused multi-modal features at scale s. The overall backbone thus defines a mapping

$$\mathcal{B} : (I_{rgb}, I_{height}) \mapsto \{\mathcal{F}_s\}_{s\in\{8,16,32\}} \tag{6}$$

that produces multi-scale fused feature maps used by the subsequent prediction heads.

3.3 Wireframe Prediction Heads

On top of the fused backbone features, LOD2-FORMER adopts the standard HEAT/BWFormer wireframe prediction stack (Chen et al., 2022; Liu et al., 2025), comprising heads for 2D corner detection, 3D corner lifting, and edge classification. A deformable transformer encoder–decoder (Zhu et al., 2020) processes the multi-scale fused features $\{\mathcal{F}_s\}$, and the 2D corner head predicts a dense corner probability map from which discrete corner candidates are obtained via non-maximum suppression. The 3D corner head then lifts these 2D corners by predicting height offsets, while the edge head classifies all candidate corner pairs as valid or invalid roof edges using features sampled along each segment. We keep the overall architecture and training strategy of these heads identical to BWFormer, and refer readers to Chen et al. (2022); Liu et al. (2025) for full design details; in our work, the main difference lies in replacing the single-modality height map backbone with the multi-modal fusion backbone described in Section 3.2.

3.4 LoD2 Completion via Wall Generation

To obtain a complete LoD2 building model from the predicted rooftop wireframe, we generate wall edges by projecting roof boundary vertices to the ground plane. We first compute the 2D convex hull of the roof vertices in the (x, y) plane to identify the building footprint and exclude interior peaks for non-flat roofs. For each boundary vertex $\mathbf{v}_i^{\text{roof}} = (x_i, y_i, z_i)$ on the hull, we create a corresponding ground vertex $\mathbf{v}_i^{\text{ground}} = (x_i, y_i, 0)$ and connect them vertically:

$$E_{\text{wall}} = \{(\mathbf{v}_i^{\text{roof}}, \mathbf{v}_i^{\text{ground}}) \mid \mathbf{v}_i^{\text{roof}} \in \mathrm{ConvexHull}(V_{\text{roof}})\}, \tag{7}$$

where V_{roof} denotes the set of predicted roof vertices. The final 3D building model consists of the roof wireframe, the generated wall edges, and the ground footprint polygon, forming a topologically consistent LoD2 structure.

3.5 Training Objective

Following HEAT (Chen et al., 2022) and BWFormer (Liu et al., 2025), we supervise 2D corners, 3D corners, and edges jointly. The overall loss is

$$L = \lambda_1 L_{c_{2D}} + \lambda_2 L_{c_{3D}} + \lambda_3 L_e, \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$ are balancing weights. The 2D corner and edge terms are binary cross-entropy losses between predicted probability maps and ground truth:

$$L_{c_{2D}} = L_{\text{BCE}}(l_{\text{pred}}, l_{\text{gt}}), \quad L_e = L_{\text{BCE}}(e_{\text{pred}}, e_{\text{gt}}). \tag{9}$$

For 3D corners, we adopt the Hungarian matching strategy of HEAT/BWFormer to associate predictions and ground-truth corners. Let N_{gt} denote the number of ground-truth corners and σ the optimal assignment that minimizes a combination of L1 distance and classification cost. The 3D corner loss is then

$$L_{c_{3D}} = \frac{1}{N_{\text{gt}}} \sum_{i=1}^{N_{\text{gt}}} \left[d_{L1}(c_i, c_{\sigma(i)}) + L_{\text{BCE}}(c_i, c_{\sigma(i)}) \right], \tag{10}$$

where c_i and $c_{\sigma(i)}$ are ground-truth and matched predicted corners, respectively. This formulation encourages both accurate 3D localization and correct existence classification, while keeping training consistent with prior wireframe reconstruction work.

4. Datasets

4.1 Roof-Intuitive

We first evaluate on the dataset introduced by Ren et al. (2021), which we refer to as Roof-Intuitive. This dataset contains images of single buildings together with reference 3D roof wireframes. In order to obtain also the height map input, we follow the method applied in AIM2PC (Turki et al., 2025) where the 3D wireframes where sampled into a consistent density of 10,000 points. In contrast to settings with severe sparsity, most buildings exhibit near complete point sampling. This makes Roof-Intuitive a suitable benchmark for assessing the performance of LOD2-FORMER under favorable geometric conditions, where incomplete LiDAR is not the dominant source of error.

For our experiments, we use 2788 training samples, 340 validation samples, and 341 test samples. We adopt the same spatial splits for both LOD2-FORMER and the BWFormer baseline to enable a fair comparison. While the original dataset was designed for point-cloud-only reconstruction, we additionally extract and align aerial images where available, allowing LOD2-FORMER to exploit both visual and geometric cues on this benchmark.

4.2 Tallinn Multi-Modal Subset

To investigate more realistic and challenging conditions, and to enable future research on multi-modal building reconstruction, we curate a dedicated subset of the Building3D Tallinn area (Wang et al., 2023) with paired aerial image and LiDAR data. Building3D provides city-scale airborne LiDAR point clouds, meshes, and roof wireframes, but does not include ready-to-use cropped aerial images aligned with individual buildings. In Tallinn, the point cloud is also considerably sparser and contains larger gaps than in Roof-Intuitive, making it a natural stress test for robust LoD2 reconstruction.

Starting from the original Building3D Tallinn tiles, we select buildings for which both point clouds and mesh-derived roof wireframes are available and then attach corresponding aerial images. Individual buildings are localized in the imagery and cropped around their footprints, and background clutter (e.g., neighbouring structures, vegetation) is suppressed using a dedicated depth- and mask-based pipeline described in Section 5.1. The resulting subset consists of aligned aerial image–height map pairs where the dominant object is a single target building.

In total, the Building3D subset comprises 2557 training samples, 141 validation samples, and 142 test samples. The

splits are defined based on the spatial distribution of buildings to avoid geographic bias between train and test areas. Compared to Roof-Intuitive, this subset combines noisier and sparser LiDAR with aerial images that can suffer from occlusions, shadows, and misalignment, providing a complementary benchmark where the benefits and limitations of multi-modal fusion become particularly visible.

5. Experimental Setup

5.1 Input Representation

For both datasets, we represent the point cloud modality as height maps and the optical modality as aerial images on a shared grid. Airborne LiDAR point clouds with the z-axis aligned to gravity are first normalized to the range $[-1.0, 1.0]$ and projected onto the xy-plane to obtain 256×256 height maps, where each pixel stores the average z-value of points projected into that pixel. Pixels with no projected points are set to zero. Aerial images are resized to the same spatial resolution to match the height map grid, and both modalities are subsequently normalized with standard ImageNet statistics (mean $[0.485, 0.456, 0.406]$, standard deviation $[0.229, 0.224, 0.225]$) after augmentation.

For Roof-Intuitive, buildings are provided as local point cloud crops and corresponding wireframes. We apply the above projection directly to the building-centric point clouds and use the available orthoimagery to obtain matching aerial images patches. For the Building3D subset, additional steps are required to obtain clean, building-centric aerial image–height map pairs. We first extract aerial images tiles in TIF format and localize individual buildings using LLMDet-Tiny (Fu et al., 2025) with building-related prompts. The resulting 2D bounding boxes are mapped to the nearest Building3D building footprints, and both the aerial image and LiDAR-derived height map are cropped accordingly. To reduce background clutter, we employ a three-stage background removal pipeline: (i) depth estimation with Depth Anything v2 (Yang et al., 2024), (ii) instance segmentation of the depth map with Segment Anything v2 (Ravi et al., 2024), and (iii) selection of the main building mask using the centrality score

$$S_c = \alpha \cdot \frac{A_{central}}{A_{mask}} + \beta \cdot \left(1 - \frac{d_{center}}{d_{max}}\right), \qquad (11)$$

where $A_{central}$ is the mask area within the central 60% region, A_{mask} is the total mask area, d_{center} is the distance from the mask centroid to the image center, d_{max} is the maximum possible distance, and α, β are weighting coefficients. The selected mask is applied to the aerial image to retain only the main building (see Figure 5), and the resulting foreground crop is paired with the corresponding height map for training.

5.2 Implementation Details

Data Augmentation We apply geometric augmentations consistently to both the aerial image and height map modalities: random horizontal flipping with 50% probability and random rotation uniformly sampled from $[0, 360)$ degrees using affine transformation. To improve robustness to the aerial images' condition variations and partial occlusions commonly encountered in aerial imagery, we introduce aerial-image-specific augmentations applied with 30% probability each: color jittering with brightness, contrast, and saturation uniformly sampled from $[0.7, 1.3]$ and hue from $[-0.05, 0.05]$ to handle varying lighting conditions and sensor characteristics, and random black box occlusions placing 1-3 boxes each covering 10-20% of the image area to simulate shadows, partial occlusions, and missing data.

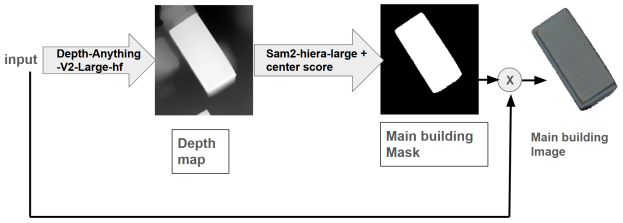

Figure 5. Background removal pipeline for building extraction. The input aerial image is processed through Depth-Anything-V2-Large-hf to generate a depth map. Sam2-hiera-large segments the depth map into multiple masks, and a centrality score (Equation 11) identifies the main building mask.The operation $\otimes$ applies the selected mask to the original image

Model Settings The maximum number of corners N, attention heads H, and edge sample points M are set to 150, 8, and 5 respectively. Loss weights λ_1, λ_2, and λ_3 are set to 1, 2000, and 100000 respectively, with corner loss weighted by 0.05. We train the model for 650 epochs with a batch size of 8, and with an initial learning rate of 2×10^{-4}, which decays by 10% in the last 50 epochs. The dual-modal backbone introduces 115.97M parameters compared to 62.09M for the single-modality baseline. Training is performed on 3 NVIDIA A100-SXM4-80GB GPUs, requiring approximately 11.3 hours on Tallinn and 13.2 hours on Roof intuitive compared to 9.5 hours and 12.0 hours for the baseline, respectively.

5.3 Evaluation Metrics

We report precision, recall, and F1-score computed separately for corners and edges. For corners, predictions are thresholded at 0.5 and matched to ground-truth corners using a distance-based assignment; Corner Precision (CP), Corner Recall (CR), and Corner F1 (CF1) denote the standard precision, recall, and harmonic-mean F1-score. To better reflect the actual performance of the reconstructed wireframes, we relaxed the corner metric computation in two respects: (i) a small spatial tolerance (radius of 3 pixels, expanded via binary dilation) is applied when matching predicted and ground-truth corners, so minor localisation offsets are not penalised as complete misses, and (ii) only those predicted corners that participate in positively classified edges are retained for the final precision, recall and F1 calculation. For edges, Edge Precision (EP), Edge Recall (ER), and Edge F1 (EF1) are defined analogously on edge-level predictions, based on the underlying candidate edge graph. During iterative refinement, edges with confidence above 0.9 are kept and those below 0.01 are discarded in subsequent iterations.

6. Experiments

6.1 Quantitative Results

Table 1 summarizes the quantitative results on the Tallinn and Roof-Intuitive datasets. On Tallinn, our approach consistently improves edge-related metrics over BWFormer, which is particularly relevant for downstream 3D reconstruction. Edge F1-score rises from 0.874 to 0.899, an absolute gain of 0.025

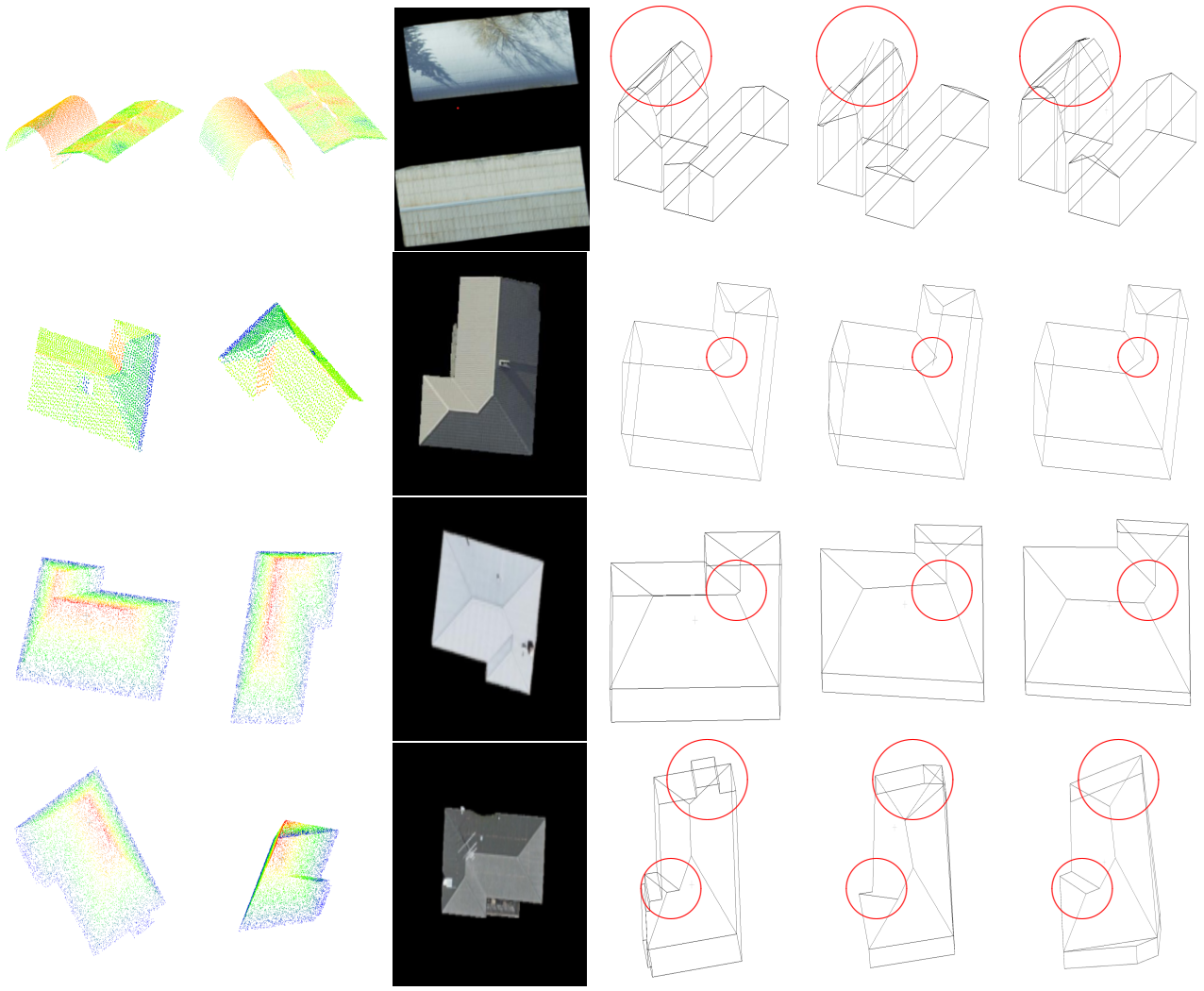

Figure 6. Quantitative evaluation results on the Tallinn dataset. The first and second columns depict point clouds from two viewpoints. The third column presents the aerial image input used in our approach. The fourth column is the ground-truth wireframe model. The fifth column displays results from the BWFormer baseline, and the sixth column shows outputs from the proposed method.

(roughly a 3% relative improvement), indicating a better balance between completeness and correctness of the recovered wireframes. Edge recall increases from 0.902 to 0.931, i.e. about three additional valid roof edges are recovered out of every hundred, despite sparse and incomplete LiDAR. Edge precision also improves, from 0.858 to 0.880, meaning that a larger fraction of the predicted edges correspond to true building structure rather than noise.

Corner metrics on Tallinn show a small but noticeable trade-off. Corner recall decreases slightly, from 0.817 to 0.799, indicating that multi-modal fusion leads to additional noise that potentially covers important features and thus missing true corners. At the same time, corner precision increases from 0.789 to 0.835, and similarly the corresponding corner F1-score from 0.793 to 0.813. In practical terms, the model detects true corners more reliably, since the detection of a corner is achieved accurately once it is supported by both modalities.However, the dual modalities introduce more noise, making the model more prune to missing a corner, especially when the latter is only emphasized in one of the modalities. Although edges are formed only from the corners actually output by the corner model, the modest drop in corner recall does not propagate to edge recall for the reason explained in the discussion section.

On the Roof-Intuitive dataset, LoD2-Former improves both corner and edge metrics across the board. The most pronounced change is observed in corner recall, which jumps from 0.460 to 0.513, i.e. nearly a 5 percentage point boost when LiDAR coverage is dense and imagery is cleaner. Corner precision increases from 0.883 to 0.984, and the corner F1-score moves from 0.599 to 0.665, so the additional corners are not obtained by sacrificing precision. Edge metrics also benefit: edge recall increases from 0.977 to 0.983, edge precision from 0.962 to 0.967, and edge F1-score from 0.968 to 0.974. Although the absolute gains on edges are smaller in magnitude than on corners, they indicate that even in this favourable setting the multi-modal model recovers a few extra correct edges per hundred predictions. In contrast to Tallinn, all metrics improve simultaneously, suggesting that in a less degraded setting LoD2-Former can enhance both completeness and precision without the corner-precision trade-off observed on sparse data.

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-2-2026-187-2026 | © Author(s) 2026. CC BY 4.0 License.

Dataset	Method	Corner			Edge		
		Recall	Precision	F1-score	Recall	Precision	F1-score
Tallinn	BWFormer	**0.817**	0.789	0.793	0.902	0.858	0.874
	Lod2Former (ours)	0.799	**0.835**	**0.813**	**0.931**	**0.880**	**0.899**
Roof-Intuitive	BWFormer	0.460	0.883	0.599	0.977	0.962	0.968
	Lod2Former (ours)	**0.513**	**0.984**	**0.665**	**0.983**	**0.967**	**0.974**

Table 1. Quantitative evaluation on Tallinn and Roof-Intuitive (2021) for corner and edge metrics, comparing the BWFormer baseline and the proposed architecture. Best values per metric are shown in bold.

Components		Corner			Edge		
Dual Backbone	Aerial image Augmentation	Precision	Recall	F1-score	Precision	Recall	F1-score
✓	×	**0.850**	0.772	0.807	**0.894**	0.908	0.896
×	✓	0.832	0.797	0.810	0.865	0.906	0.879
✓	✓	0.835	**0.799**	**0.813**	0.880	**0.931**	**0.899**

Table 2. Ablation results on the Tallinn dataset covering corner and edge detection precision, recall, and F1-score. ✓ indicates that a component is used, × indicates it is not used. Best values per metric are shown in bold.

6.2 Qualitative Results

Qualitative examples on both datasets (see Figure 6) illustrate how the improvements in edge metrics translate into more faithful roof wireframes. The first line of the figure shows an industrial area in the Tallinn dataset with a large footprint and multiple intersecting roof planes, LoD2-Former recovers a larger fraction of the true edges than BWFormer and produces more coherent polygonal structures. Visually, this is reflected in fewer missing roof sections and cleaner junctions at ridges and valleys, in line with the gains in edge recall and F1-score.

The second line, also in Tallinn dataset, is a residential area building, where LiDAR is sparser and roof contrast is lower, the fused model is still able to delineate many roof boundaries that are only weakly expressed in the height map. The resulting wireframes tend to follow facade lines and roof outlines more closely than those of the baseline, reducing broken or fragmented roof parts. In very challenging regions with extremely sparse LiDAR coverage, both methods show failure cases, particularly around isolated corners, but LoD2-Former often reconstructs more complete building outlines, consistent with the quantitative improvements observed on edges. The third and last lines, are two samples from Roof-Intuitive dataset, where we recall that the height map only baseline is not disadvantaged yet our approach still shows measurable improvement across all metrics.

6.3 Role of Cross-Modal Fusion and Image Augmentation

We next analyze how augmenting aerial imagery and the choice of fusion architecture affect performance on the Tallinn dataset. To this end, we compare three configurations that share the same training schedule, loss weights, and network depth. The first variant uses the LoD2-Former backbone with bidirectional cross-modal attention but is trained without aerial-image-specific augmentation. The second variant replaces the backbone with a simple feature-concatenation scheme while retaining aerial image augmentation. The full model combines both the LoD2-Former backbone and aerial-image-specific augmentation (color jitter and random occlusions).

Table 2 summarizes the results on the Tallinn validation set. Adding aerial-image-specific augmentation to the LoD2-Former backbone leads to a clear gain in robustness: corner recall increases from 0.772 to 0.799, indicating that simulating realistic degradations in the aerial image channel (such as shadows, occlusions, and color shifts) enables the network to exploit visual cues more reliably under sparse LiDAR conditions. The fusion architecture itself also has a pronounced effect. When the aerial image and height map features are merged by simple concatenation, edge F1-score reaches 0.8790, which remains noticeably below the 0.8988 obtained with the LoD2-Former backbone and its bidirectional cross-modal attention. This gap suggests that naive feature combination is insufficient to resolve cross-modal geometric ambiguities, whereas the proposed attention-based fusion can selectively emphasize consistent cues and suppress conflicting ones.

Overall, the full configuration, which uses the dedicated fusion backbone together with the aerial image-specific augmentation, has the best overall performance with an edge recall of 0.931 and an edge F1-score of 0.899. The above results show that the architectural design and the robustness in each modality have a synergistic effect, where each component on its own is not able to reproduce the quality achieved by the full model. On the corner side, the full model has the highest corner F1-score at 0.813. This is in accordance with the trade-off observed on the Tallinn dataset, where the multi-modal fusion has a negative impact on corner recall ($0.817 \rightarrow 0.799$) while having a positive impact on corner precision ($0.789 \rightarrow 0.835$). The reason why this negatively affects corner recall but not edge metrics is discussed in detail in the discussion section.

7. Discussion and Limitations

Across both datasets, the results highlight that edge-level predictions are the most stable and informative aspect of the proposed model. On Tallinn, all edge metrics improve over BWFormer, and on the Roof-Intuitive dataset we observe consistent gains in edge recall, precision, and F1-score. Since LoD2 reconstruction quality is primarily driven by the correctness and completeness of roof edges, these improvements indicate that multi-modal fusion delivers tangible benefits for the down-

stream 3D task, even when corner-level behaviour is less uniform.

The comparison between Tallinn and Roof-Intuitive also illustrates the data-dependent nature of multi-modal fusion. Tallinn combines sparse and incomplete LiDAR with challenging aerial imagery, and in this setting we observe a small trade-off in corner recall while edges still improve. In contrast, Roof-Intuitive provides dense LiDAR coverage and cleaner aerial images, and in this regime the model improves all corner and edge metrics simultaneously, with particularly strong gains in corner recall. This asymmetry suggests that the underlying fusion strategy is sound, and that the corner recall drop on Tallinn is largely driven by dataset-specific conditions rather than a fundamental limitation of the architecture.

Logically speaking, a small decrease in corner recall is not obliged to result in a decrease in edge recall; although edges are purely a function of connection between predicted corners. With more informative multi-modal inputs to the corner detector, it becomes more selective in that it will reject some of the more dubious corner proposals that the single modality model would have accepted. The remaining corners will be more accurately placed and more consistent with one another. And since these are the only accurate corners that are provided to the edge classifier, the line segments between these will have much stronger and more uncluttered combined geometric-visual evidence to work with. The edge model will be able to more confidently classify a larger percentage of the actual roof connections that exist between these accurate locations. In other words, the model is making a trade-off between some of the more dubious peripheral corners for a cleaner and more reliable set of junction candidates; this cleaner set of candidates will enable the subsequent edge stage to achieve higher overall roof edges (edge recall increases from 0.902 to 0.931), rather than being distracted and confused by noisy corner locations.

There are, nevertheless, several limitations. First, performance remains sensitive to extreme sparsity in the point cloud and to severe aerial imagery degradation; in scenes with heavy occlusion or very few LiDAR returns, both the baseline and our method can fail to reconstruct small or heavily occluded structures. Second, the multimodal-backbone design increases the parameter count and training time compared to a single-modality model, which may be relevant for very large-scale or resource-constrained deployments, even though the added cost is moderate in our experiments. Finally, our evaluation is restricted to two datasets with LiDAR and aerial imagery. Future work could extend the analysis to additional cities and sensing setups, and explore more adaptive fusion strategies that explicitly down-weight unreliable aerial image regions or model uncertainty in corner proposals, with the aim of retaining the benefits of multi-modal fusion while further reducing corner-level artifacts on challenging datasets.

8. Conclusion

We have presented LoD2-Former, a multi-modal Transformer architecture for 3D building wireframe reconstruction that fuses LiDAR and aerial imagery within a multimodal-backbone, bidirectional cross-attention framework. By extending the 2D-to-3D corner detection to explicitly exploit complementary geometric and visual cues, LoD2-Former achieves consistently stronger edge reconstruction on both sparse and dense light detection and ranging (LiDAR) point clouds, with edge F1 increasing from 0.874 to 0.899 on Tallinn and from 0.968 to 0.974 on the Roof-Intuitive dataset. In the complete-data setting, it also substantially boosts corner recall (from 0.460 to 0.513) while delivering a favourable precision-recall trade-off on the challenging Tallinn data and strongly improving edge metrics on both, demonstrating that multi-modal fusion can resolve geometric ambiguities that are difficult to handle from point clouds alone.

Beyond the architecture itself, we contribute a curated Building3D subset with paired, foreground-cleaned aerial image and height map inputs, designed as a challenging benchmark for multi-modal level of detail (LoD) 2 reconstruction under sparse and noisy LiDAR. Our analysis of multi-modal fusion, including ablations on image-specific augmentation and fusion strategy, shows that both a dedicated cross-modal backbone and robustness to the aerial image artifacts are necessary to fully realize the benefits of imagery. Finally, we demonstrate how the predicted roof wireframes can be lifted to complete LoD2 models by deriving wall geometry from the 3D roof graph, providing a direct link from learned wireframes to usable 3D building representations suitable for downstream urban modeling applications.

References

Akwensi, P. H., Bharadwaj, A., Wang, R., 2025. Points2Model: A Neural-Guided 3D Building Wireframe Reconstruction from Airborne LiDAR Point Clouds. *International Journal of Digital Earth*, 18(1), 1–27.

Cao, L., Xu, Y., Guo, J., Liu, X., 2023. WireframeNet: A Novel Method for Wireframe Generation from Point Cloud. *Computers & Graphics*, 115, 226–235.

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

Forlani, G., Nardinocchi, C., Scaioni, M., Zingaretti, P., 2006. Complete classification of raw LIDAR data and 3D reconstruction of buildings. *Pattern Analysis and Applications*, 8(4), 357–374.

Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.-S., 2025. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14987–14997.

Gong, H. et al., 2025. Robust Building Wireframe Reconstruction: A Hypergraph and Transformer-Enhanced Framework for Large-Scale and Real-World Urban Point Clouds. *International Journal of Remote Sensing*, 46(21).

Huang, H., Brenner, C., Sester, M., 2013. A generative statistical approach to automatic 3D building roof reconstruction from airborne laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 79, 29–43.

Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y., 2018. Learning to parse wireframes in images of man-made environments. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 626–635.

Huang, S., Wang, R., Guo, B., Yang, H., 2024. Pbwr: Parametric building wireframe reconstruction from aerial LiDAR point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27778–27787.

Liu, Y., Zhu, L., Ye, H., Huang, S., Gao, X., Zheng, X., Shen, S., 2025. Bwformer: Building wireframe reconstruction from airborne LiDAR point cloud with transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22215–22224.

Maltezos, E., Ioannidis, C., 2016. Automatic extraction of building roof planes from airborne LiDAR data applying an extended 3d randomized Hough transform. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3, 209–216.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, J., Zhang, B., Wu, B., Huang, J., Fan, L., Ovsjanikov, M., Wonka, P., 2021. Intuitive and efficient roof modeling for reconstruction and synthesis. *arXiv preprint arXiv:2109.07683*.

Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H., 2021. Efficient attention: Attention with linear complexities. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3531–3539.

Turki, S., Panangian, D., Chaabouni-Chouayakh, H., Bittner, K., 2025. AIM2PC: Aerial Image to 3D Building Point Cloud Reconstruction. *arXiv preprint arXiv:2503.18527*.

Wang, R., Huang, S., Yang, H., 2023. Building3D: An urban-scale dataset and benchmarks for learning roof structures from point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Xue, N., Wu, T., Bai, S., Wang, F., Xia, G., Zhang, L., Torr, P. H. S., 2020. Holistically-attracted wireframe parsing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2788–2797.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

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

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.