The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 "From Imagery to Understanding", Commission II, 4–11 July 2026, Toronto, Canada

CityZen: LOD2 building reconstruction with a point cloud-free model-driven approach

Mehmet Büyükdemircioğlu[1], Ibrahim Sall[1,2], Simone Rigon[1], Fabio Remondino[1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (mbuyukdemircioglu, srigon, remondino)@fbk.eu

[2] Ecole Nationale des Sciences Geographiques (ENSG), France - Email: Ibrahim.Sall@ign.fr

KEYWORDS: 3D reconstruction, Orthophotos, Roof classification, Height estimation, Deep learning

Abstract

Accurate building footprints and 3D models are nowadays essential for a wide range of urban applications, yet the generation of Level of Detail 2 (LOD2) models remains constrained by the availability of dense 3D data such as LiDAR or image matching products. While these sources provide high geometric accuracy, they are costly to acquire and update, creating a gap between data availability and the increasing demand for city-scale 3D modelling. Recent advances in deep learning enable monocular height estimation from aerial imagery, offering a potential alternative to traditional 3D data sources. However, integrated workflows that combine image-based inference with structured 3D reconstruction are still limited. This paper presents CityZen, a point cloud-free workflow for LOD2 building reconstruction from only RGB orthophotos. The proposed approach integrates monocular height estimation (evaluating DSMNet, HTC-DC-Net and TSE-Net), roof type classification and model-driven reconstruction within a unified pipeline. Building footprints are used as geometric constraints, while inferred height and semantic cues guide the generation of consistent 3D structures. The proposed framework, available at https://github.com/3DOM-FBK/citizen, enables scalable and practical LOD2 city modelling using widely available aerial orthophotos, reducing dependency on costly 3D data acquisition.

1. Introduction

Accurate and up-to-date building footprints and 3D building models are fundamental components of contemporary urban geospatial infrastructures. They support a broad range of applications, including urban planning, digital twins, solar potential analysis, environmental assessment, risk evaluation, and infrastructure management (Haala and Kada, 2010; Biljecki et al., 2015). Among the different levels of geometric representations, LOD2 building models are particularly attractive because they preserve the main roof structure and volumetric form of individual buildings while remaining computationally manageable for city-scale applications (Kolbe et al., 2005).

Despite their practical value, LOD2 models remain expensive to produce and maintain when the workflow depends on dense 3D source data such as LiDAR point clouds or dense image matching products. These data sources can provide high geometric accuracy, but they also require dedicated acquisition campaigns, substantial storage and processing resources and specialized expertise. As a result, many municipalities possess recent high-resolution orthophotos but lack equally recent elevation data suitable for large-scale 3D analyses and management. This mismatch creates an important operational gap between data availability and the growing demand for detailed 3D city models.

Recent advances in deep learning have shown that aerial imagery can support monocular height estimation and other scene understanding tasks without the need for explicit 3D input at inference time. At the same time, model-driven building reconstruction approaches remain attractive due to their geometric structure and semantic consistency on the final building representation. These two directions are complementary: image-based learning can provide dense height and semantic cues, whereas rule-based reconstruction can transform those cues into stable and interpretable 3D geometry. However, workflows that combine orthophoto-based height prediction, roof type attribution and structured LOD2 reconstruction in a single operational pipeline are still limited.

To address this gap, this paper presents CityZen, a point cloud-free, footprint-guided, model-driven workflow for LOD2 3D city modelling from RGB orthophotos. The method combines monocular height estimation, roof type classification, and model-driven building reconstruction to generate semantic LOD2 3D city models. In contrast to purely data-driven 3D reconstruction approaches (Peters et al., 2022; Jarzabek-Rychard et al., 2025), CityZen uses building footprints as geometric constraints and considers height prediction and roof attribution as complementary inputs for later reconstruction. This approach aims to provide a practical compromise between the flexibility of learning-based inference and the structural regularity of model-driven city modelling.

The proposed workflow is evaluated around two datasets and results show how LOD2 outputs can be derived from widely available orthophotos, making it relevant for areas where high-quality 3D source data are unavailable or too costly to acquire.

The remainder of the paper is structured as follows. Section 2 reviews related work on LOD2 building reconstruction, monocular height estimation, and roof type classification. Section 3 describes the proposed CityZen methodology, including height estimation, roof type classification and LOD2 building reconstruction. Section 4 introduces Çeşme and Bologna datasets and the benchmark design. Section 5 presents the experimental results for roof type classification, height estimation and final 3D reconstruction. Finally, Section 6 summarizes the main findings and outlines directions for future work.

This work addresses the need for a practical and scalable alternative to LiDAR-dependent LoD2 reconstruction by proposing CityZen, a footprint-guided pipeline that integrates orthophoto-based height estimation, roof type classification and model-driven 3D building reconstruction. Rather than attempting direct end-to-end mesh generation from imagery alone (Gao et al., 2022), the method combines learned raster inference with geometric constraints derived from building footprints to produce structured and LOD2 3D city models in CityJSON (Ledoux et al., 2019) format. The objectives of the study are multifold:

- to estimate building heights from RGB orthophotos using monocular depth estimation learning-based methods;
- to assign roof type classes to building footprints using supervised roof classification;

- to reconstruct geometrically consistent LOD2 3D city models from the combination of predicted height, roof class and footprint geometry;
- to evaluate how different height estimation backbones affect reconstruction quality.

Therefore, the study separates height estimation and roof classification into two experimental settings: the roof classifier is trained on the Çeşme dataset for seven different roof types, whereas the height estimation models are trained and benchmarked on the Bologna dataset under a common evaluation protocol. The final aim is to assess whether CityZen can provide reliable and operational 3D building reconstruction in urban environments where up-to-date LiDAR or DSM data are not available. We consider LoD-2 as it is defined by the CityGML standard (Kolbe et al., 2005).

2. Related works

Many approaches for the reconstruction of 3D city models from measurement data have been reported in the past decades, either data- or model-driven (Brenner, 2005; Lafarge et al., 2010; Huang et al., 2011; Jarzabek-Rychard et al., 2025; Schuegraf et al., 2025; Zhu et al., 2025). In particular, LOD2 building reconstruction has long been a central topic in photogrammetry and remote sensing. Compared with simpler block representations, LOD2 building models preserve the main roof structure of individual buildings and therefore provide a useful compromise between geometric detail and computational efficiency. This makes them especially suitable for urban-scale applications such as environmental simulation, solar potential analysis, urban climate studies, digital-twin development, and interactive 3D campus and city visualization (Biljecki et al., 2015; Biljecki et al., 2016; Buyukdemircioglu and Kocaman, 2018). Some works have also highlighted the growing role of deep learning in 3D building reconstruction (Wichmann et al., 2018; Stucker et al., 2022), while emphasizing the continuing importance of geometric constraints and data quality in practical workflows (Buyukdemircioglu et al., 2022).
Traditional high-quality building modeling methods are typically based on LiDAR point clouds or dense image-matching products. Nevertheless, Biljecki et al. (2017) demonstrated that 3D city models can also be generated without explicit elevation data, showing the potential of footprint- and rule-based strategies in cases where dedicated height information is unavailable. On the other hand, Farella et al. (2021), starting from historical maps, inferred missing height information using machine learning techniques.
A representative example of data-driven method is City3D, which performs automated large-scale building reconstruction from airborne LiDAR point clouds and demonstrates the scalability of LiDAR-based modelling workflows (Huang et al., 2022). These approaches benefit from direct geometric observations of roof surfaces and can achieve high reconstruction accuracy when high-density 3D measurements are available (Jarzabek-Rychard et al., 2025). Similarly, Roofer (Peters et al., 2022) can reconstruct LOD2 building models from DSM / LiDAR data through roof plane extraction, candidate face generation and optimization-based model selection. Such methods are well suited to large-scale national mapping contexts, but their dependence on dense elevation data limits their practical use in areas where LiDAR coverage is unavailable, outdated, or too expensive to acquire.
In parallel, recent learning-based approaches have shown that aerial imagery alone can support monocular height estimation. Liu et al. (2020) introduced IM2ELEVATION, a deep learning approach for deriving building height and digital surface model information from single-view aerial imagery. Cao and Huang (2021) further demonstrated the feasibility of deep learning for large-scale building height estimation from high-resolution imagery in urban environments. DSMNet (Elhousni et al., 2021) combines dense height prediction with auxiliary semantic and geometric supervision, while HTC-DC-Net (Chen et al., 2023) introduces a strong monocular framework specifically designed for remote-sensing height estimation. DSMNet combines dense height prediction with auxiliary semantic and geometric supervision, while HTC-DC-Net introduces a strong monocular framework specifically designed for remote-sensing height estimation. TSE-Net (Chen and Zhu, 2025) further extends this direction through semi-supervised learning. These methods demonstrate that orthophotos contain enough radiometric and structural information to infer useful height patterns, even though the resulting predictions remain less precise than LiDAR derived elevation products in many challenging urban settings. Recent monocular depth estimation methods applied to aerial images (Madhuanand et al., 2021; Hermann et al., 2024) could also be a valuable approach for building height estimation.
Another relevant line of research concerns roof type classification from aerial imagery. Roof classification is important because it provides semantic information that can constrain or guide geometric reconstruction. Buyukdemircioglu et al. (2021) showed that transfer learning with convolutional neural networks can achieve reliable roof type recognition from very high-resolution aerial imagery. This type of semantic roof information is especially valuable in model-driven reconstruction pipelines, where roof labels can be translated into explicit geometric rules. More recent studies have also confirmed the effectiveness of deep learning for roof type recognition from high-resolution aerial and UAV imagery (Wang et al., 2022). Single image reconstruction pipelines for urban scenes have also been investigated in recent years, showing growing interest in practical 3D reconstruction directly from optical imagery (Tripodi et al., 2022).
Although these research directions have individually matured, relatively few studies combine them into a single operational pipeline. In particular, there remains a gap between dense raster prediction from imagery and the production of structured semantic LOD2 building models. The proposed CityZen pipeline is designed to bridge this gap by combining orthophoto-based height estimation, supervised roof type classification and footprint-guided rule-based building reconstruction. The proposed method is not intended to replace high precision LiDAR based modelling where such data are available, but rather to offer a practical alternative for locations where just orthophotos (and, at most, building footprints) are available, while up-to-date 3D elevation data are not.

3. Methodology

The proposed point cloud-free building modeling pipeline, named CityZen, integrates three main components (Figure 1): height estimation, roof classification and 3D geometry generation.

3.1 Building height estimation

The first stage of the CityZen workflow estimates dense building height information from aerial orthophotos. Rather than relying on externally acquired LiDAR or Digital Surface Models, the method predicts a normalized digital surface model (nDSM) directly from RGB imagery, which is later used for footprint level height attribution and 3D reconstruction. (Section 3.2-3.3). Since this stage conditions all subsequent geometric processing, it is evaluated separately as a benchmark task on the Bologna dataset (Section 5.1).

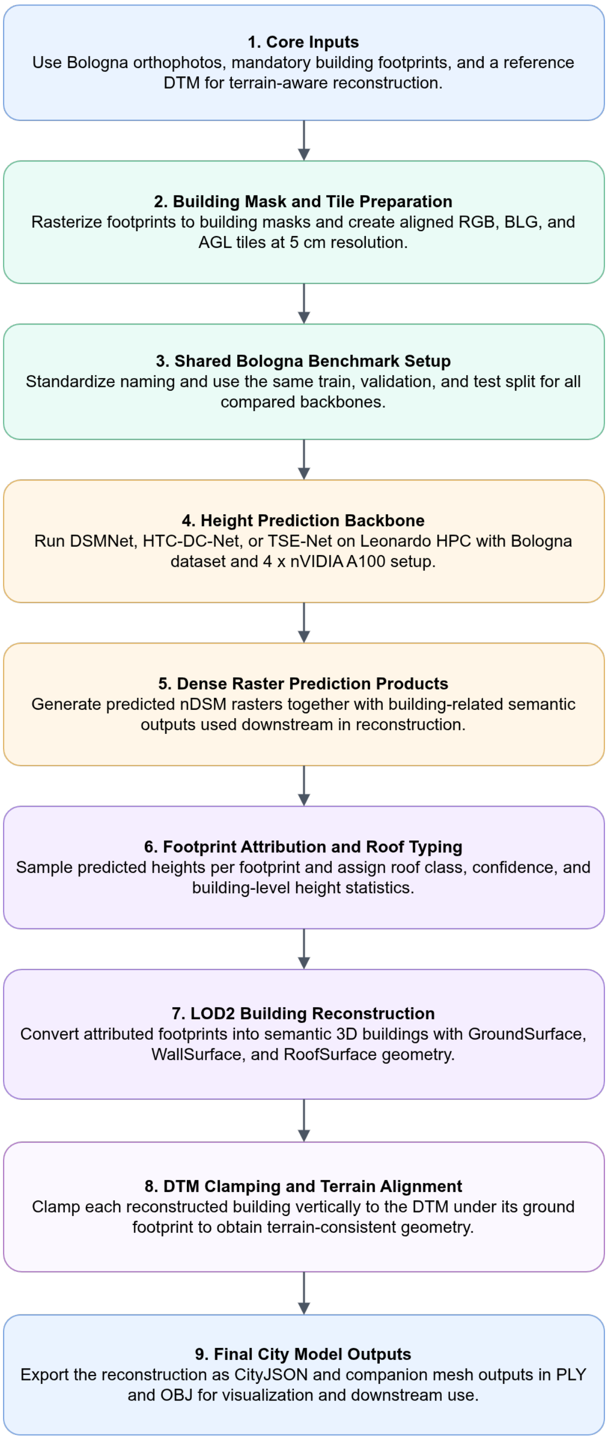

Figure 1. Overview of the proposed CityZen workflow for LOD2 building reconstruction from only orthophotos.

To ensure a fair comparison, three monocular height estimation networks are trained and evaluated on the same Bologna split: DSMNet, HTC-DC-Net and TSE-Net. The benchmark consists of aligned RGB orthophotos, normalized DSM targets and binary building masks at 5 cm GSD.
All three models are trained and tested on identical train, validation and test partitions, allowing a direct comparison across dense raster accuracy, building-only accuracy, non-building accuracy and per-building error statistics.
Within the operational CityZen pipeline, DSMNet is used as the downstream height backbone. In the current Bologna dataset, DSMNet processes 512 × 512 RGB image tiles and jointly predicts height information. The model is trained in a multi-task setting, followed by a refinement stage that improves the height output. Since unrealistic negative values may propagate into later footprint level queries, predicted nDSM values below zero are clamped to zero before raster export and before the reconstruction step. This correction improves numerical stability and prevents underground artifacts in the final CityZen outputs.
The output of this stage is a georeferenced height raster that describes the spatial distribution of building and terrain elevation patterns across the scene. Although the raster quality is important in itself, its role within CityZen is primarily functional: it provides the height evidence that is aggregated per footprint and later transformed into LOD2 building geometry. For this reason, both pixel-based and building-based error metrics are considered in the experiments.

3.2 Roof type classification

The second stage of the CityZen workflow assigns semantic roof labels to individual building footprints. In contrast to generic roof recognition as an isolated image-classification task, the purpose here is explicitly geometric: roof type predictions in the orthophotos are used as structural priors for the subsequent reconstruction stage. This stage takes as input the orthophoto, the mandatory building footprints and the predicted height raster from the previous step (Section 3.1).
Roof type prediction is performed using a VGG16based classifier trained on the Çeşme dataset. The network predicts seven roof categories, namely complex, flat, gable, hip, L-shaped, pyramid and unknown. During inference, each footprint is associated with a roof label, class probabilities, and a confidence score. To avoid propagating unstable semantic information into the reconstruction process, uncertain predictions are not forced into a valid class but are instead retained as unknown, allowing the following stage to apply fallback modelling rules. This is particularly important in the present study because the classifier is trained on the Çeşme orthophoto dataset and then transferred to the Bologna orthophoto under different spatial resolution and acquisition conditions.
In addition to the predicted roof type, each footprint is enriched with height related attributes derived from the predicted nDSM. For every building polygon, the pipeline samples raster values within the footprint and computes descriptive statistics such as mean height, minimum height, maximum height, standard deviation, and the number of supporting pixels. The resulting attributed footprint layer therefore combines planimetric geometry, semantic roof information, confidence estimates, and footprint level height descriptors. This layer represents the key intermediate product linking raster inference to geometric reconstruction.

3.3 LOD2 Building reconstruction

The final stage transforms the attributed footprint layer into semantic LOD2 building geometry. Reconstruction is performed through roof type specific geometric rules that combine footprint shape, predicted building height, and roof label. Flat roofs are generated by vertical extrusion, whereas gable, hip, pyramid, and L-shaped roofs are reconstructed using dedicated parametric formulations constrained by footprint geometry and dominant orientation. More complex or uncertain cases are handled through simplified fallback strategies to preserve robustness and avoid unstable roof generation.
Special attention is given to geometric consistency during model construction. Footprint topology is preserved, including interior rings and courtyard structures where present, and the reconstructed surfaces are organized into semantic roof, wall, and ground components. The final outputs are stored as LOD2

building models in CityJSON format, with additional export to PLY and OBJ for visualization and downstream applications.
After geometry generation, the reconstructed buildings can be vertically aligned to a terrain model by sampling DTM heights beneath their ground footprints and applying per-building vertical offsets. This post-processing step improves terrain consistency without altering the reconstructed roof topology. In this way, the final CityZen output is not merely a set of extruded blocks, but a structured semantic city model representation combining predicted height, roof type and footprint geometry in a consistent LOD2 framework.

4. Study area and datasets

The study uses two datasets with distinct roles in the CityZen workflow. The Çeşme dataset is employed for supervised roof type classification, whereas the Bologna dataset is used for training and benchmarking the monocular height estimation networks, as well as for the final LOD2 reconstruction experiments. This separation allows the two learning problems to be analysed independently and avoids conflating roof classification performance with height estimation accuracy.
The Çeşme dataset consists of roof image samples extracted from aerial orthophotos acquired over Çeşme (Türkiye), with a ground sampling distance of 10 cm. It is used exclusively to train and validate the roof type classifier. The dataset contains labelled examples of seven roof categories, namely complex, flat, gable, hip, L-shaped, pyramid and unknown. In the present study, Çeşme provides the semantic supervision required for roof attribution, but it is not used for height estimation benchmarking or for end-to-end 3D reconstruction assessment. Table 1 summarizes the two datasets used in this study and highlights their different roles within the CityZen pipeline.

Attribute	Çeşme	Bologna
Role	Roof type classification training	Height estimation benchmark and LoD2 building reconstruction
Resolution / GSD	10 cm	5 cm
Data Volume	10,007 roof image samples	1,681 aligned RGB/nDSM/buildingmask triplets
Spatial Unit	Variable size roof patches derived from building footprints	512 × 512 pixel patches
Split	7207 train / 1800 val / 1000 test	1,177 train / 168 val / 336 test
Labels / Reference	7 roof classes	Metric nDSM and binary building masks

Table 1. Summary of the datasets used in this study and their roles in the CityZen pipeline.

On the other hand, the Bologna dataset is derived from a high-resolution true orthophoto (Figure 2) acquired in 2024 over a dense historical urban area in Bologna (Italy) together with a normalized DSM reference and a binary building mask generated from building footprints[1]. The data are organized as RGB, nDSM, and building mask tiles at 5 cm ground sampling distance. Each tile has a size of 512 × 512 pixels, corresponding to approximately 25.6 × 25.6 m on the ground.

The Bologna dataset contains a total of 1,681 RGB, nDSM and building mask patches. These are divided into 1,177 training patches, 168 validation patches, and 336 test patches, corresponding exactly to a 70/10/20 split. The same partition is used for DSMNet, HTC-DC-Net and TSE-Net, thereby ensuring a direct comparison among the three height estimation models. A connected component analysis of the building masks yields 2,320 building components in the training split, 306 in the validation split, and 636 in the test split, for a total of 3,262 patch level building instances. Since buildings crossing tile boundaries may appear in more than one patch, this value should be interpreted as a patch level measure of building content rather than as a unique building count.
Beyond the patch-based benchmark, Bologna is also used as the full scene reconstruction case study for the CityZen workflow. At this scale, the current pipeline produces a final semantic LOD2 model containing 1,040 reconstructed buildings. Bologna is therefore the central dataset for evaluating both the comparative behaviour of the height estimation backbones and the downstream footprint-guided 3D reconstruction process. Table 2 summarizes the Bologna benchmark split and the distribution of patch level building content across the training, validation, and test subsets.

Split	Patches	Patches with buildings	Patch level building components
Train	1177	1080	2320
Validation	168	149	306
Test	336	299	636
Total	1681	1528	3262

Table 2. Summary of the Bologna benchmark split and patch level building content.

Figure 2. Example of the Bologna true orthophoto used for benchmark generation and reconstruction.

5. Experiments and results

5.1 Height estimation from RGB orthophotos

The revised Bologna experiments use a DSMNet model retrained on Bologna-specific data and a corrected post-processing pipeline in which negative nDSM outliers are clamped to zero before raster export and before 3D reconstruction. This

[1] https://opendata.comune.bologna.it/pages/home

correction is important because it removes unrealistic underground artifacts and makes the DSMNet raster substantially more stable.

Figure 3 presents an example nDSM tile used for qualitative comparison among the evaluated height estimation methods. Visually, DSMNet and HTC-DC-Net recover the main building mass and general scene structure, but they differ in edge sharpness and local detail. TSE-Net produces a noticeably different response, with weaker agreement to the ground truth in both object shape and height pattern distribution. Overall, this example illustrates the relative ability of the methods to reconstruct building geometry and fine scale spatial variation in nDSM prediction.

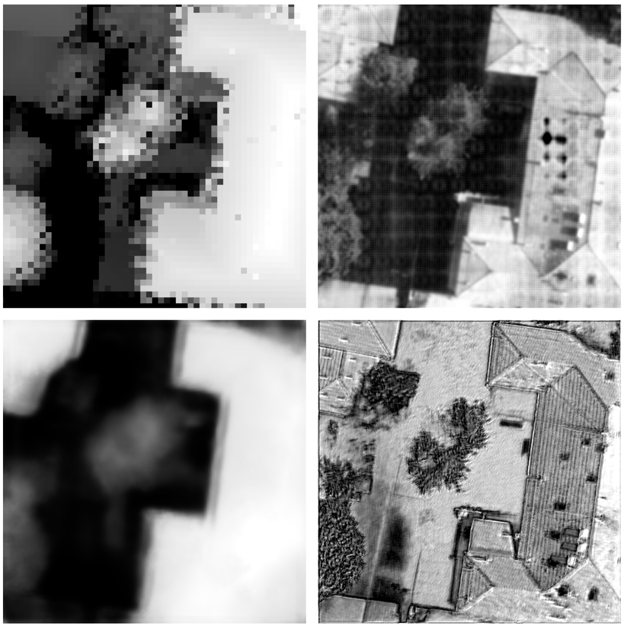

Figure 3. Qualitative comparison of nDSM prediction results for one sample tile: ground truth (top left), DSMNet (top right), HTC-DC-Net (bottom left) and TSE-Net (bottom right). DSMNet and HTC-DC-Net more closely match the ground truth, while TSE-Net shows larger deviations.

Table 3 reports the quantitative height estimation results on the Bologna benchmark and highlights the different behaviour of DSMNet, HTC-DC-Net and TSE-Net across pixel-wise and per-building metrics. The reported metrics characterize complementary aspects of height quality. MAE measures the average absolute difference between prediction and reference, whereas RMSE penalizes large errors more strongly. The metrics are computed over three spatial domains: all pixels in the overlapping test patches, building mask pixels only, and the remaining nonbuilding background pixels. The per building measures are computed after aggregating height at the building level and are therefore particularly relevant for LOD2 reconstruction, where the overall height of each building is more important than isolated local noise. The benchmark indicates that HTC-DC-Net remains the strongest model for dense pixel wise height prediction, with the lowest overall RMSE (4.52 m) and building RMSE (5.49 m). However, DSMNet becomes highly competitive at the building level after the Bologna-specific retraining: its per building RMSE is 5.05 m, which is slightly lower than the 5.18 m obtained by HTC-DC-Net. TSE-Net performs clearly worse on building pixels and on per building metrics, indicating a stronger tendency to underestimate building height on this benchmark. These results suggest that HTC-DC-Net remains preferable when the objective is the best dense DSM raster, whereas DSMNet is particularly promising when the downstream goal is building wise height estimation within the CityZen reconstruction chain.

Model	DSMNet	HTC-DC-Net	TSE-Net
MAE all	6.09 m	3.20 m	6.27 m
RMSE all	7.12 m	4.52 m	9.36 m
MAE Building	4.74 m	3.91 m	13.48 m
RMSE Building	5.89 m	5.49 m	14.56 m
MAE non-building	6.64 m	2.91 m	3.35 m
RMSE non-building	7.57 m	4.06 m	6.09 m
MAE per building	4.04 m	3.74 m	13.04 m
RMSE per building	5.05 m	5.18 m	14.01 m

Table 3. Quantitative comparison of DSMNet, HTC-DC-Net, and TSE-Net on the Bologna dataset.

5.2 Roof type classification

On the Bologna reconstruction tile, the roof type classification stage assigned explicit labels to 980 out of 1,040 building footprints, while the remaining 60 footprints are retained as unknown and subsequently handled through fallback reconstruction rules. Therefore, 94.2% of the buildings received a valid roof type prediction, with a mean confidence score of 0.890 (minimum 0.501, maximum 1.000). It is important to note that the classifier is trained on the Çeşme orthophoto dataset acquired in 2016 at 10 cm ground sampling distance, whereas inference is performed on the Bologna true orthophoto acquired in 2024 at 5 cm ground sampling distance. The predicted class distribution is dominated by L-shaped roofs with 301 cases (30.7%), followed by complex roofs with 273 cases (27.9%), gable roofs with 128 cases (13.1%), pyramid roofs with 110 cases (11.2%), flat roofs with 96 cases (9.8%), and hip roofs with 72 cases (7.3%). Since no roof type reference labels are available for Bologna, the classification performance could not be quantitatively assessed on the target area using standard accuracy measures. Nevertheless, the results indicate that the classifier transfers reasonably well across differences in acquisition year, spatial resolution, and urban context, while the explicit preservation of uncertain cases as unknown contributes to a more robust downstream LOD2 reconstruction process. The distribution of predicted roof types for the Bologna reconstruction tile is summarized in Table 4. Figure 4 illustrates the spatial distribution of the predicted roof classes.

Roof type	Count	Percentage (%)
L-shaped	301	30.7
Complex	273	27.9
Gable	128	13.1
Pyramid	110	11.2
Flat	96	9.8
Hip	72	7.3
Unknown	60	5.8
Total footprints	*1040*	*100.0*
Classified footprints	*980*	*94.2*

Table 4. Roof type classification summary

Figure 4. Roof type classification results for the Bologna reconstruction tile. Building footprints are coloured according to the predicted roof class, while uncertain cases are preserved as unknown for downstream fallback reconstruction.

5.3 LOD2 building reconstruction

In the final pipeline, DSMNet is trained using the Bologna dataset with a batch size of 16 using nDSM and true orthophotos. During the subsequent roof classification stage, 980 building footprints received roof labels, with a mean confidence score of 0.890. The most frequently predicted roof classes are L-shaped (301) and complex (273), followed by gable (128), pyramid (110), flat (96), and hip (72). Because no reference roof type labels are available for Bologna, a detailed accuracy analysis for each roof class cannot yet be reported in this benchmark. Even so, the reconstruction logs still provide useful information, showing that rare concave roof forms, especially gabled-L roofs, are the least stable cases and still require fallback handling in the current pipeline.

LOD2 results of CityZen are compared with open-source data-driven method Roofer (Figure 5). Roofer results are generated using classified (building, ground and vegetation) LiDAR point cloud as elevation input, whereas CityZen relies on nDSM prediction from orthophotos. A direct output level comparison with Roofer on the same tile provides an additional way to examine the reconstruction performance. This comparison is especially meaningful because the building counts are exactly matched: both reconstructions contain 1,040 buildings, while the Roofer model also represents these through 1,040 BuildingPart objects. The total reconstructed footprint area is also very similar, measuring 309,738 m² for CityZen and 310,978 m² for Roofer. One-to-one spatial matching produced correspondences for all 1,040 buildings. The resulting mean and median footprint IoU values are 0.967 and 0.970, respectively, and 1,028 of the 1,040 matched buildings have an IoU greater than 0.9. In addition, the mean absolute footprint area difference is only 4.14 m² while the mean relative area difference is about 1.0%. These results show that the horizontal agreement between the two reconstructions is very strong and that both methods produce almost the same building extent, while differences are more apparent in roof structure and vertical modelling.

The mean absolute building height difference is 4.59 m, while the median absolute difference is 3.99 m. The mean height bias is 1.72 m, which suggests that the remaining difference between the two outputs is mainly related to height rather than footprint geometry. In other words, the main uncertainty is no longer horizontal building extent, but the estimation of building height and roof form. This behaviour is expected, since CityZen estimates building height from orthophotos and does not rely on high precision elevation input such as LiDAR. The final Bologna model contains 1,040 buildings, 19,531 vertices, and 34,723 faces, which confirms that the pipeline can produce a complete structured output for the full tile.

At the footprint level, both methods reproduce the urban block with very similar building shapes, extents, and orientations. This agrees well with the high IoU values and the very small footprint area differences reported in the quantitative evaluation. However, the roof geometry shows clearer visual differences between the two methods. In the CityZen result, the roof surfaces appear cleaner, more regular, and more stable at the block level. The roof planes are simpler and more consistent, which gives the reconstruction a more coherent overall appearance. In the Roofer result, more internal roof detail and ridge structure are visible, but this comes together with stronger geometric fragmentation. Small breaks, thin roof fragments, and local irregularities can be seen more often, especially on larger and more complex buildings. This suggests that Roofer preserves more local roof variation, but it may also be more sensitive to noise or instability in difficult cases.

These differences can be observed particularly well in buildings with complex or concave roof shapes. In such cases, Roofer often creates additional small roof elements and more articulated internal roof patterns, while CityZen tends to preserve a more continuous and unified roof surface. A clear example is the courtyard type building near the centre right part of the scene. In the Roofer rendering, this building shows a more detailed but also less stable roof reconstruction, with visible fragmentation inside the roof structure. In the CityZen rendering, the same building appears smoother and more consolidated, with fewer local irregularities. A similar pattern can also be seen in several medium-sized blocks across the scene, where Roofer reconstructs extra roof surfaces, while CityZen produces simpler and more regular roof surfaces. This visual behaviour is consistent with the reconstruction logs, which indicate that rare and complex roof types remain the most difficult cases in the CityZen pipeline, especially for concave forms such as gabled-L roofs.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-351-2026 | © Author(s) 2026. CC BY 4.0 License.

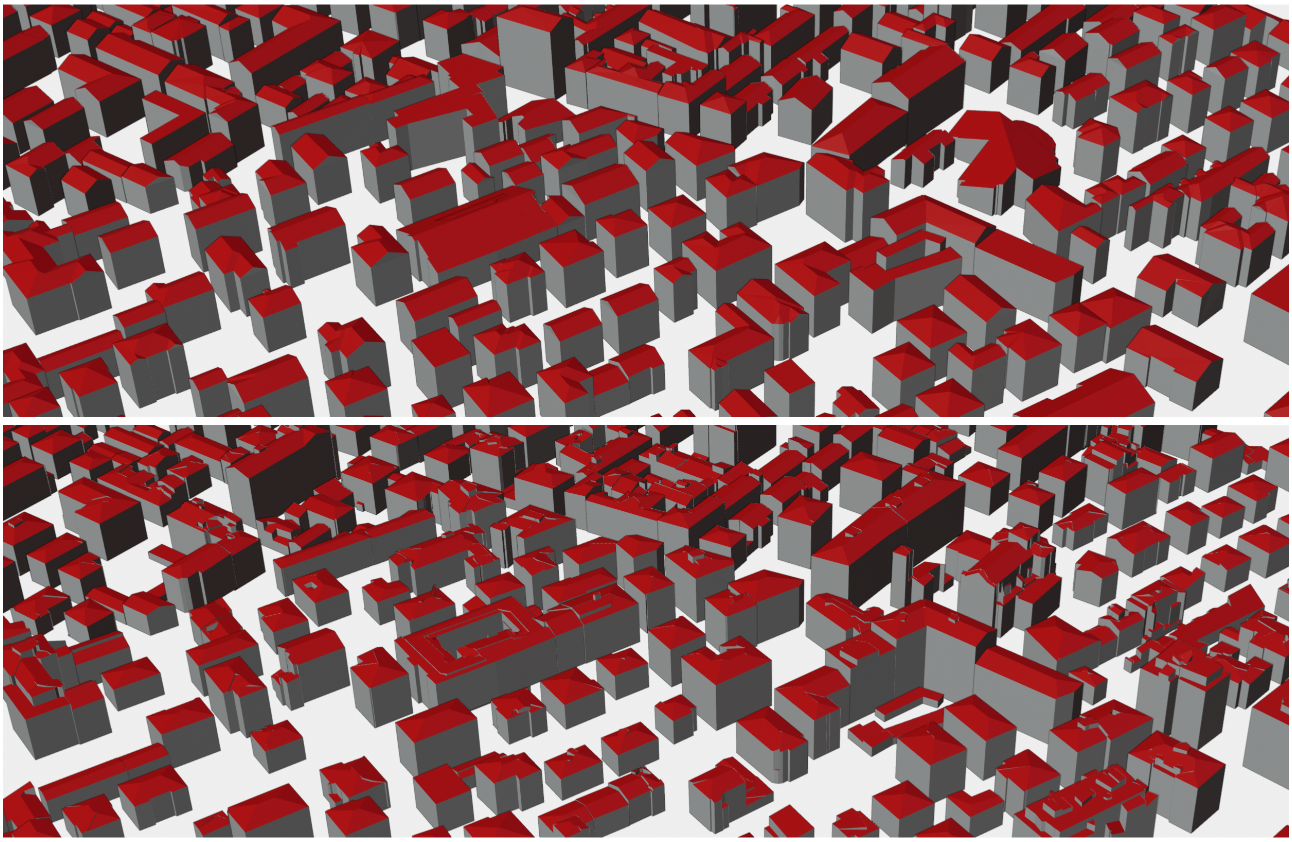

Figure 5. Visual comparison of reconstructed CityJSON 3D city models for the same Bologna tile and viewing direction, generated by CityZen using the predicted nDSM (top) and by Roofer using a LiDAR derived DSM (bottom). A mesh-to-mesh comparison between the two city models using bidirectional nearest neighbour distances revealed discrepancies of ca 3 m.

The visual comparison also suggests that the main differences between the two methods are not related to building detection itself, since both reconstructions recover the same number of buildings and show very strong footprint agreement. Instead, the most important differences are related to roof interpretation and vertical modelling. In simple roof cases, the outputs are generally very similar. In more difficult cases, especially where the roof form is complex, concave, or internally articulated, the differences become more visible. This supports the conclusion that the main challenge in the Bologna benchmark is no longer the horizontal reconstruction of building footprints, but the reliable modelling of roof shape and building height.

6. Conclusions

This paper presented CityZen, a point cloud-free and model-driven pipeline for LOD2 building reconstruction from just aerial orthophotos. The workflow combines deep learning-based height estimation, roof type classification and rule based geometric reconstruction to generate georeferenced 3D building models without requiring LiDAR input. The experiments show that this strategy can produce operational LOD2 outputs from widely available orthophotos, making it relevant for areas where high quality 3D source data are unavailable or too costly to acquire.
Overall results point to three main limitations. First, the reported errors depend strongly on the evaluation target, because the agreement is much stronger for footprints than for heights. Second, rare and complex roof types are still less robust than more common roof forms and continue to be the main source of reconstruction difficulty. Third, strong generalization across different datasets and urban scenes still requires domain specific retraining and adaptation. Taken together, the quantitative metrics and the visual comparison show that CityZen and Roofer achieve almost the same footprint reconstruction on the Bologna area, whereas the remaining differences are concentrated mainly in roof shape reconstruction, geometric stability on complex buildings and vertical height estimation. For municipalities lacking recent LiDAR, DSM, or point cloud data, these remaining differences may be acceptable in view of the practical availability of orthophotos and the resulting scalability of the workflow.
Future work will focus on improving height estimation robustness, expanding roof type coverage, testing multi-city training strategies for better domain generalization, and defining quality assessment procedures that do not depend on point-cloud reference data. Further investigation will also address large-scale deployment and the generation of more consistent roof geometries for rare and complex building forms.

References

Biljecki, F., Ledoux, H., Stoter, J., 2017. Generating 3D city models without elevation data. *Computers, Environment and Urban Systems*, 64, 1-18.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. *Comput. Environ. Urban Syst.*, 59, 25-37.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Coltekin, A., 2015. Applications of 3D city models: State of the art review. *ISPRS Int. J. Geo-Inf.*, 4, 2842-2889.

Brenner, C., 2005. Building reconstruction from images andlaser scanning. *Int. Journal of Applied Earth Observation and Geoinformation*, Vol.6(3-4):187–198.

Buyukdemircioglu, M., Kocaman, S., Kada, M., 2021. Deep learning based roof type classification using very high resolution aerial imagery. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-2-2021, 161-166.

Buyukdemircioglu, M., Kocaman, S., 2018. A 3D campus application based on city models and WebGL. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-4, 161-165.

Buyukdemircioglu, M., Kocaman, S., Kada, M., 2022. Deep learning for 3D building reconstruction: A review. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII, 359-366.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. *Remote Sensing of the Environment*, 264, 112590.

Chen, S., Shi, Y., Xiong, Z., Zhu, X.X., 2023. HTC-DC Net: Monocular height estimation from single remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 61, 1-18.

Chen, S., Zhu, X.X., 2025. TSE-Net: Semi-supervised monocular height estimation from single remote sensing images. *arXiv preprint* arXiv:2511.13552.

Elhousni, M., Zhang, Z., Huang, X., 2021. Height prediction and refinement from aerial images with semantic and geometric guidance. IEEE Access, 9, 145638-145647.

Farella, E.M., Özdemir, E., Remondino, F., 2021: 4D Building Reconstruction with Machine Learning and Historical Maps. *Applied Science*, 11(4), 1445.

Gao, X., Wu, K., Pan, Z., 2022. Low-poly mesh generation for building models. Proc. *SIGGRAPH,* pp. 1-9.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS J. Photogramm. Remote Sens.*, 65, 570–580.

Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E.K., Remondino, F., Jutzi, B., Ruf, B., 2024. Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 13, 100065.

Huang, H., Brenner, C., Sester, M., 2011. 3D building roof reconstruction from point clouds via generative models. Proc. 19th *SIGSPATIAL*, pp. 16-24.

Huang, J., Stoter, J., Peters, R., & Nan, L., 2022. City3D: Large-scale building reconstruction from airborne LiDAR point clouds. *Remote Sensing*, 14(9), 2254.

Jarzabek-Rychard, M., Rigon, S., Boguslawski, P., Remondino, F., 2025. Automating 3D building modeling: A comparative study of data- and model-driven methods. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-G-2025, 701–707.

Kolbe, T.H., Gröger, G., Plümer, L., 2005. CityGML: Interoperable access to 3D city models. *In: Geo-information for Disaster Management*, 883–899. Springer, Berlin, Heidelberg.

Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2010. Structural approach for building reconstruction from a single DSM. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol.32, pp. 135-147.

Ledoux, H., Arroyo Ohori, K., Kumar, K., Dukai, B., Labetski, A., Vitalis, S., 2019. CityJSON: A compact and easy-to-use encoding of the CityGML data model. *Open Geospatial Data, Software and Standards*, 4(1), 1-12.

Liu, C.-J., Krylov, V.A., Kane, P., Kavanagh, G., Dahyot, R., 2020. IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sensing*, 12(17), 2719.

Madhuanand, L., Nex, F., Yang, M.Y., 2021. Self-supervised monocular depth estimation from oblique UAV videos. *ISPRS J. Photogramm. Remote Sens.*, 176, 1-14.

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

Schuegraf, P., Gui, S., Qin, R., Fraundorfer, F., Bittner, K., 2025. SAT2BUILDING: LoD-2 Building Reconstruction from Satellite Imagery using Spatial Embeddings. *PE&RS*, Vol. 91(4), pp. 203-212.

Stucker, C., Ke, B., Yue, Y., Huang, S., Armeni, I., and Schindler, K., 2022. Implicity: city modeling from satellite images with deep implicit occupancy fields. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.,* V-2-2022, 193-201.

Tripodi, S., Girard, N., Fonteix, G., Duan, L., Mapurisa, W., Leras, M., Trastour, F., Tarabalka, Y., Laurore, L., 2022. BRIGHTEARTH: Pipeline for on-the-fly 3D reconstruction of urban and rural scenes from one satellite image. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-3/W1-2022, 263-270.

Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., Cai, H., 2022. Improved Mask R-CNN for rural building roof type recognition from UAV high-resolution images: A case study in Hunan Province, China. *Remote Sensing*, 14(2), 265.

Wichmann, A., Agoub, A., Kada, M., 2018. RoofN3D: deep learning training data for 3d building reconstruction. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2, 1191-1198.

Zhu, X. X. and Chen, S. and Zhang, F. and Shi, Y. and Wang, Y., 2025. GlobalBuildingAtlas: an open global and complete dataset of building polygons, heights and LoD1 3D models. *Earth System Science Data*, Vol. 17(12), pp. 6647--6668.