Unit-level LoD2 Building Reconstruction from Satellite-derived Digital Surface Model and Orthophoto

Shengxi Gui[1], Philipp Schuegraf[2], Ksenia Bittner[2]*, Rongjun Qin[1]†

[1]The Ohio State University, Columbus, Ohio, United States
{gui.55, qin.324}@osu.edu
[2]Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany
{philipp.schuegraf, ksenia.bittner}@dlr.de

Keywords: LoD2 building modeling; Satellite photogrammetry; unit-level building semantic segmentation;

Abstract

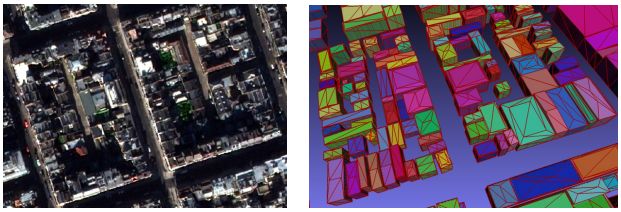
Recent advancements in deep learning have enabled the possibility to identify unit-level building sections from very high resolution satellite images. By learning from the examples, deep models can capture patterns from the low-resolution roof textures to separate building units from duplex buildings. This paper demonstrates that such unit-level segmentation can further advance level of details (LoD)2 modeling. We extend a building boundary regularization method by adapting noisy unit-level segmentation results. Specifically, we propose a novel polygon composition approach to ensure the individually segmented units within a duplex building or dense adjacent buildings are consistent in their shared boundaries. Results of the experiments show that, our unit-level LoD2 modeling has favorably outperformed the state-of-the-art LoD2 modeling results from satellite images.

1. Introduction

1.1 Background

Level of Detail (LoD)2 building models describe architectural features and topological structures of building roofs (Gröger et al., 2008; Biljecki et al., 2016), therefore, are of high interest in various applications such as mapping, urban planning, architectural design, virtual reality environments, and risk management. Typically, creating high-quality LoD2 models involves a manual and very expensive process, while recent research efforts aim to automate this process. Out of many sources, very-high-resolution (VHR) satellite stereo imagery (with ground sampling distance (GSD) < 1 m) is beneficial due to its global coverage and low cost per unit area (Facciolo et al., 2017; Li et al., 2023b). Previous works have shown that it is possible to reconstruct LoD2 (Gui and Qin, 2021; Gui et al., 2022; Partovi et al., 2019) from such data, which typically follow a standard process takes pre-processed digital surface model (DSM) and orthophotos from stereo satellite imagery as input data: first, perform building detection to obtain building masks; second, vectorize individual building masks with topologically consistent line primitives, third, determine the types of roofs and then join individual small buildings into more complex building models. Although these methods produce reasonable results for individual and single structured buildings, due to the lack of resolution of satellite images, reconstructing models in densely built areas and duplex buildings remains a significant challenge (Chen et al., 2018).

Challenges of reconstructing duplex buildings, or buildings in densely built regions, arise from the difficulties of building segmentation algorithms to identify distinct boundaries for duplex and adjacent buildings (Huang et al., 2023) based on the mere low-resolution orthophoto and DSM. Duplex building consists of two or more separate units, typically side-by-side or stacked on top of each other, sharing a common wall but operating independently. Complex buildings usually consist of several rectangular units with similar roof materials and textures. Complex/duplex buildings are often constructed from smaller, and contextually separate building units, while it is traditionally extremely challenging to infer information at the unit-level. The recent work in (Schuegraf et al., 2023) has shown that, by learning from examples, it is possible to infer unit-level segmentation from the low-resolution image textures, which, if successfully applied, can be used to extract unit-level models for LoD2 model reconstruction.



(a) Orthophoto (b) 3D building models

Figure 1. Sample figure for unit-level LoD2 building reconstruction. (a) Orthophoto for weak texture buildings; (b) 3D building models by using unit-level reconstruction

In this study, we integrate unit-level building segmentation with building model reconstruction, introducing an effective composition method for level of details (LoD)2 building model generation. This method preserves boundary consistency among segmented units in duplex buildings, and then apply unit-level building shapes into LoD2 3D building model. Upon evaluation in seven distinct regions, the experiments show that our unit-level LoD2 modeling approach significantly outperforms existing LoD2 models derived from satellite imagery in terms of accuracy and detail.

* Corresponding author
† Corresponding author

1.2 Related works

As mentioned earlier, reconstructing building models from photogrammetric data typically entails a multi-stage process that starts with the detection of building masks. This is followed by the extraction of 2D parameters (regularized footprint), and concludes with the derivation of 3D parameters (roof primitive) using specialized algorithms (Alidoost et al., 2019; Partovi et al., 2019).

Building Segmentation: Building segmentation has moved towards unit-level segmentation. Hence, recent studies have not only tackled the segmentation of each pixel as building or non-building, but also into instances. PolyMapper (Li et al., 2019) directly predicts building and road vectors on the instance level (or unit level for buildings). Approximating shapes in images with polygons (ASIP) (Li et al., 2020) surpasses the performance of PolyMapper on the CrowdAI (Mohanty et al., 2020) benchmark dataset for building instance segmentation. Then, Frame Field Learning set new standards for building instance segmentation on the CrowdAI benchmark by first predicting a pixel-level segmentation of buildings and building borders together with a map of two tangent directions per pixel. The tangent directions, called Frame Field, are used in an iterative optimization procedure to produce building polygons with regular appearance. Exceeding the performance of Frame Field Learning on CrowdAI, PolyWorld (Zorzi et al., 2022) is an end-to-end trainable building instance segmentation approach. It includes multiple steps of extracting vertices and learning the adjacency matrix that is used to connect the vertices. This procedure is error-prone, since a false negative vertex can strongly alter the appearance of the predicted polygon. Missing links in the adjacency matrix can cause missing polygons. Furthermore, PolyWorld does not separate directly adjacent buildings. Tackling this issue, Schuegraf et al. (2023) predict separation lines between buildings together with the building segment and use the watershed transform to robustly predict building instances. The results of Schuegraf et al. (2023) surpass those of Frame Field Learning for complex urban scenarios.

Footprint regularization: The process of extracting regularized building footprints begins with the vectorization of images into regularized polylines, designated as building boundaries, subsequently generating rectangular-shaped building footprints. The preliminary processing of building segments employs shape reconstruction methods, such as alpha-shape (Kada and Wichmann, 2012) and Hough Transform (**?**), to establish initial building boundary formation. These are further refined through polyline simplification techniques, including the Random Sampling Consensus (RANSAC) (Fischler and Bolles, 1981; Schnabel et al., 2007) and the Douglas–Peucker algorithm (Douglas and Peucker, 1973). Next, post-processing of line segments from polyline can be facilitated based on orthophoto and algorithms like line segment detector (LSD) (Von Gioi et al., 2008), KInetic Polygonal Partitioning of Images (KIPPI) (Bauchet and Lafarge, 2018), and PolyCity (Li et al., 2023c). Subsequent steps involve further decomposition to delineate individual buildings, aligning them with preliminary rectangular or circular 2D models. An illustrative method is the orthogonal line-based 2D rectangle extraction technique by Partovi et al. (2019), which decomposes building footprints into rectangle shapes starting from the longest boundary lines.

Model Reconstruction: The methods for building 3D reconstruction from images are generally classified into two distinct strategies: bottom-up and top-down, and both depend on 3D elevation generated by photogrammetric methods (Xu et al., n.d.; Han et al., 2020) or LiDAR (Jayaraj and Ramiya, 2018). The bottom-up, or data-driven approach, treats buildings as collections of roof planes and other elements, assembling them based on geometric relationships observed in DSMs and point clouds. This strategy may employ techniques such as feature filling (Zhou et al., 2016) and region growing (Sun and Salvaggio, 2013) to merge the structural components. Conversely, the top-down, or model-driven approach, relies on a predefined library of 3D building models (Lafarge et al., 2008; Huang et al., 2013). It selects the most suitable model for a given set of data (like DSMs or point clouds), but this method often requires complex processes or adaptable parameters to match the diverse nature of building architectures. Many advanced techniques adapt deep learning methods for object recognition and meshing, and are increasingly being incorporated into primitives estimation (Wang et al., 2021; Li et al., 2023a; Mao et al., 2023).

2. Method

The method described in this paper employs pre-processed satellite-derived DSM and Orthophoto data, along with image processing techniques and deep models, to create 3D geometric models of buildings (LoD). The input data, DSM and Orthophoto, can be generated through standard photogrammetric workflow using the provided Rational Polynomial Coefficients (RPC), as for example, our input data are generated by using the RSP (RPC stereo processor) software (Qin, 2016). As shown in Figure 2, with this input data, the proposed workflow initiates with unit-level building segmentation and then reconstructing building models in 2D and 3D. Specifically, the unit-level semantic segmentation process aims to detect and segment discernable building units from Orthophoto and DSM, which stands for standard single-unit buildings, or multiple units of duplex buildings. The LoD2 building footprint extraction process extracts regularized rectangular building footprints from these individual building segments, further dividing bigger segments into basic building units. Finally, it utilizes the most appropriate building model with 3D primitives to represent the building units at 3D level.

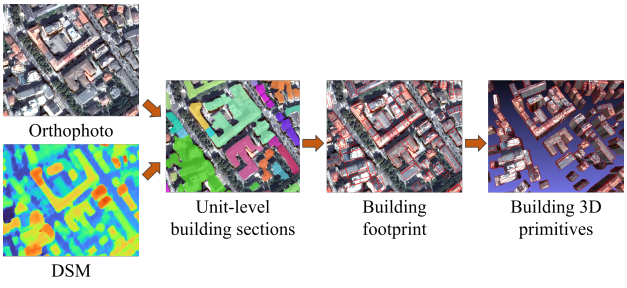


Figure 2. Workflow of unit-level building LoD2 model reconstruction

2.1 Unit-level semantic segmentation

Buildings in large cities exhibit complex structures comprising of various interconnected units and components. To perform as detailed as possible 3D reconstruction, the modeling of each building component as a separate unit is the correct way to proceed. Supporting the LoD2 modeling methodology, this paper employs a unit-level semantic segmentation strategy previously developed by Schuegraf et al. (2023).

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-X-2-2024-81-2024 | © Author(s) 2024. CC BY 4.0 License.

Our method uses a deep convolutional neural network at its core for a 3-class problem: building component, separation line and background. The inputs to the network are the DSM and orthorectified RGB satellite images tiled to patches of size 512×512 px. We employ the well-proven U-Net shape architecture (Ronneberger et al., 2015), consisting in our case of two ResNet34 encoders (He et al., 2016) for each input modality and one decoder. In order to maintain detailed spatial information, we aggregate feature maps acquired at four distinct scales from the two encoders through summation. These aggregated feature maps then serve as input for the full-scale skip-connections.

We follow the same training procedure as in (Schuegraf et al., 2023) and employ a combination of segmentation and regularization losses. To minimize the dissimilarity between predicted and true probability distributions, we incorporate the *weighted cross-entropy* loss function

$$\mathcal{L}_{CE}(x, y, p, w) = -\sum_{i} y_i w_i \cdot log(p(x_i)), \quad (1)$$

to achieve accurate and meaningful outcomes in the context of multi-class semantic segmentation. Here, y denotes the ground truth, x is the input tensor, $p(.)$ is the *softmax* output of the neural network model, i is the respective class and w is an array of manually selected loss weighting coefficients, which we set to $[1, 1, 4]$. Often, the utilization of cross-entropy loss leads to smoothed or indistinct boundaries for objects. To suppress this issue and push the model towards more precise delineation of boundaries, we used the *generalized dice loss* (Sudre et al., 2017)

$$\mathcal{L}_{DICE}(x, y, p) = 1 - 2 \cdot \frac{\sum_i v_i \sum_n y_{in} \cdot p(x_i)_n}{\sum_i v_i \sum_n y_{in} + p(x_i)_n}, \quad (2)$$

where v_i is the inverse frequency of the class i. $\mathcal{L}_{DICE}$ is developed for precise boundary detection.

To penalizes structural irregularities like curved corners or uneven edges we employed the topological loss (Mosinska et al., 2018)

$$\mathcal{L}_{TOP}(x, y, p)_C = \sum_{n=1}^{N} \sum_{m=1}^{M_n} ||l_n^m(y_C) - l_n^m(p(x)_C)||_2^2, \quad (3)$$

which minimizes the differences between the VGG19 (Simonyan and Zisserman, 2014) descriptors of the ground-truth images and the corresponding predicted delineations, in our case for both the building and separation line classes separately. In eq. (3), we denote the class on which to apply the term as C, l_n^m describes feature map m of layer n of a pre-trained VGG19.

The final objective function combines three above-described losses

$$\begin{aligned}\mathcal{L}_{TOTAL} = \mathcal{L}_{CE} + \mathcal{L}_{DICE} + \lambda_{BM} \cdot (\mathcal{L}_{TOP})_{BM} \\ + \lambda_{TB} \cdot (\mathcal{L}_{TOP})_{TB},\end{aligned} \quad (4)$$

where λ controls the influence of the topological term on the overall training procedure, abbreviations BM and TB relate to building mask and touching border classes.

Following this, a map representing instances of building sections is created through the application of the watershed transform (Beucher and Meyer, 2018) in a post-processing stage. Essentially, the watershed transform interprets the obtained three-class map, consisting of background, building, and separation line, along with a seed image and a mask, as a topographical surface. The seed map and mask are derived from the predicted information related to buildings and separation lines. Subsequently, the watershed transform simulates a flooding scenario, wherein water begins flooding from the seeds and settles into basins. These basins are delineated by watershed lines, aligning with high image intensities. The mask confines the virtual water flow to specific regions, and the enclosed regions marked by watershed lines are then identified as objects.

2.2 LoD2 building reconstruction

Upon obtaining unit-level building segments, we apply 2D footprint extraction process and 3D primitives computation process for each building unit to generate rectangular-based 3D building models. The Orthophoto and DSM, derived from very high-resolution satellite imagery, typically have spatial resolutions ranging from 0.3 m to 1 m. Due to this resolution constraint, accurately detecting small buildings and detailed roof structures from complex buildings remains a challenge. To address this, we employ a model-driven approach for 3D building reconstruction from studies (Gui and Qin, 2021; Partovi et al., 2019). This approach assumes that a complex building footprint can represented by 2D rectangles, thus it turns the LoD2 modeling problems into a topology fusion problems from 3D primtives buildings (extended from the 2D rectangle footprints).

In order to represent building footprints as regularized 2D shapes, unit-level building segments are vectorized into polygons and subsequently refined into rectangular footprints. This process contains three steps from satellite-derived data: initially, coarse boundary delineation is achieved using the Douglas-Peucker algorithm (Douglas and Peucker, 1973), effectively primary vectorizing building segments into initial polylines of building boundaries. This is followed by a polyline adjustment step, where the main orientations of each building unit are calculated, and shorter line segments with similar orientations are merged into more extended line segments. The final step involves polyline regularization with the LSD algorithm (Von Gioi et al., 2008), aligning the orientations of line segments with detected line segments with texture information from Orthophotos. The culmination of this process not only identifies the main orientation of each building segment but also accurately vectorizes the building polygons from rasters, facilitating the extraction of rectangular building footprints without DSM data and refining building shapes from satellite images with enhanced precision.

The vectorized building footprints may still be in arbitrary polygon with a number of vertex. To facilitate the process of generating 3D primitives, it is necessary to decompose these polygons into multiple simple rectangles. Our approach employs a grid-based decomposition approach (Gui and Qin, 2021), predicated on the concept that complex building polygons can be fundamentally broken down into multiple, simpler rectangular entities, which then serve as the regularized building 2D model for the subsequent 3D reconstruction stage. The procedural workflow of this decomposition is divided into four distinct parts: First, for each unit, the 2D building polygon is rotated to align its primary local orientation orthogonally. Second, initial separation of the building mask is performed, using DSM and Orthophoto gradients. Third, a three-tier image pyramid approach is applied to iteratively identify and refine the largest possible inner rectangles, progressing from the coarsest to the finer layers.

Table 1. Study areas basic information

Region	Image size (pixel)	GSD (pixel size)	Location	Building area proportion	Building instances
Columbus 1	1003×890	0.5 m	USA, North America	0.2305	224
Columbus 2	1646×1118	0.5 m	USA, North America	0.2866	151
Buenos Aires 1	2000×2000	0.3 m	Argentina, South America	0.2704	352
Buenos Aires 2	2000×2000	0.3 m	Argentina, South America	0.0587	111
London 1	3000×3000	0.5 m	UK, Europe	0.5985	676
London 2	3000×3000	0.5 m	UK, Europe	0.2968	910
Trento	3680×3309	0.5 m	Italy, Europe	0.3114	1556

Finally, excessively segmented adjacent rectangles are consolidated, leveraging both Orthophoto and DSM data to ensure accurate and efficient footprint reconstruction. To determine if two neighboring rectangles should be merged, the following criteria are proposed:

$$\begin{cases} merge, & |\overline{C_1} - \overline{C_2}| < T_d \\ & \bigcap |\overline{H_1} - \overline{H_2}| < T_{h1} \\ & \bigcap \max|\Delta H_{edge}| < T_{h2} \\ not\ merge, & otherwise \end{cases} \tag{5}$$

As in equation 5, there are multiple thresholds based on color and height information of buildings used to control the merging process, including 1) color threshold T_d set for color difference between two rectangles projected onto the Orthophoto, where the absolute difference in mean color values $|\overline{C_1} - \overline{C_2}|$ of the two rectangles (projected onto the Orthophoto); 2) height threshold T_{h1} set for mean height difference $|\overline{H_1} - \overline{H_2}|$ between the rectangles; and 3) gap threshold T_{h2} set as the threshold for significant height variations in a buffered area encompassing the shared edge of two adjacent rectangles.

After determining the rectangular footprint of each building unit, 3D roof structure can be fitted based on rectangular models derived from satellite-based DSM. These primilary model shapes include five types of rectangular building roof models: flat, gable, hip, pyramid, and mansard, and each roof model represents a specific architectural style and primitives. A set of 3D parameters, including ridge height, eave height, and hip structure, is utilized to characterize the detailed roof primitives across all five model types (Gui and Qin, 2021). These parameters are computed and optimized through an exhaustive search strategy designed to identify the parameters set that minimizes the root mean square error (RMSE) between the fitted roof height and DSM data. The optimization includes iterative parameter updates informed by DSM, starting with the determination of terrain height as the local minimum of the building height. Despite the DSM data noise since resolution or stereo matching limitation (Ling and Qin, 2022; Huang and Qin, 2020), our exhaustive search approach efficiently selects the most accurate roof type and parameter set, maintaining computational accuracy even for buildings only with a few hundred pixels. The final output of this process includes detailed 3D parameters for the building model, which reconstructs buildings into LoD2 levels.

3. Experiments

3.1 Study areas

Our experiments include four cities, each exemplifying unique geographical locations and distinctive urban landscapes, including 1) Columbus, Ohio, a typical U.S. city characterized by low-density residential and industrial areas; 2) Buenos Aires, Argentina, a South American megacity, that contains a mix of sparsely populated residential areas and densely inhabited slums; 3) London, UK, a European megacity with a compact urban structure and high-density development; 4) Trento, Italy, a medium-sized European city with numerous adjacent buildings.

The accuracy of model-driven 3D building reconstruction, which deduces a set of 3D primitives based on texture and height data from individual building sections, depends entirely on the accuracy and comprehensiveness of the regularized building footprint in producing the final LoD2 model. Furthermore, the density or urban complexity represents the difficulty of 3D building reconstruction. In areas characterized by a high concentration of buildings, accurately delineating individual building perimeters becomes particularly challenging. This challenge is compounded in scenarios where adjacent buildings feature roofs with low texture contrast, often leading to the aggregation of multiple structures into a single reconstruction section, thereby adversely affecting the accuracy of the LoD2 model. Hence, the success of building reconstruction in densely populated regions is heavily dependent on the segmentation effectiveness for Orthophotos with weak textures.

The 3-band (RGB) Orthophotos and DSMs for all study areas are generated using a multi-view stereo matching approach (RSP, Qin 2016, 2019) from multiple World-view-2 stereo pairs for the Columbus, London, and Trento dataset, and Worldview-3 for Buenos Aires dataset.

Table 1 shows the information of each study area. In total, there are three low building density regions with small numbers of buildings, and most buildings are isolated, and four high building density regions have dense building distribution with large numbers of buildings, and many buildings are densely located with a sharing wall to their neighborhoods.

3.2 Evaluation in 2D and 3D level

The evaluation of 2D segmentation and 3D geometry are computed separately using both a 2D intersection over union (IOU_{2D}) and 3D intersection over union (IOU_{3D}) based on manually created reference data for building footprint and light detection and ranging (LiDAR)-based DSM for 3D geometry (Kunwar et al., 2020). IOU_{2D} assesses the accuracy of 2D building footprint extraction, while IOU3D evaluates the accuracy of 3D model fitting. The IOU_{2D} and IOU_{3D} are defined following as follows:

$$IOU_{2D} = \frac{TP}{TP + FP + FN} \tag{6}$$

$$IOU_{3D} = \frac{TP_{3D}}{TP_{3D} + FP + FN} \tag{7}$$

where TP is the number of true positive pixels that are determined as extracted and manually labeled building footprint simultaneously, FP is the number of false positives and FN is the number of false negatives. TP_{3D} is TP pixels whose 3D vertical difference from the ground-truth LiDAR is within 2 m.

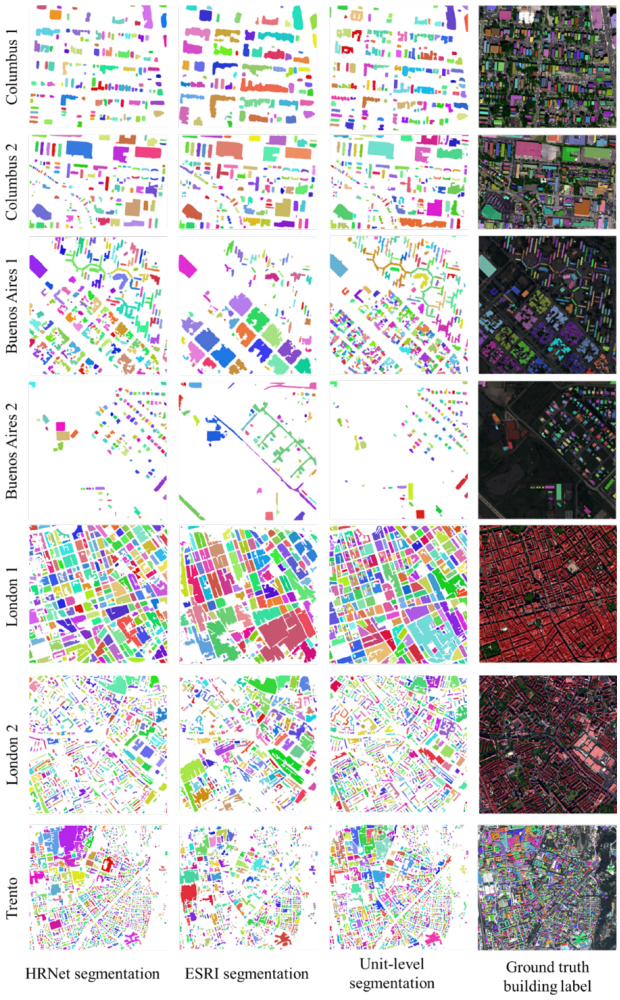


Figure 3. Building semantic segmentation results for each study region and the ground truth

Other than unit-level building segmentation, two publicly available semantic segmentation methods for building footprint detection are compared to evaluate the performance at 2D and 3D levels. The first one is based on HRNetV2 (Wang et al., 2020) to get building segments by using Orthophoto with RGB bands. The training and validation datasets were combined with satellite and aerial imagery (Gui et al., 2022). The second one is High Resolution Land Cover Classification – USA (Ronneberger et al., 2015; Robinson et al., 2019), developed by ESRI for ArcGIS multi-classes semantic segmentation. This approach uses the UNet model architecture and is trained based on aerial imagery with 0.8m-1m resolution, which can also be used to segment buildings.

Figure 3 displays the segmentation results from three semantic-based segmentation methods and ground truth building mask. The visual comparison indicates that compared to normal semantic segmentation methods, our unit-level segmentation method can extract building sections from very dense urban and complex structure buildings. Besides, since the training dataset for ESRI’s segmentation method is aerial imagery, the segmentation result in Buenos Aires regions (from Worldview-3) is not as good as other study regions (from Worldview-2).

The numerical results comparing 2D and 3D levels are presented in Table 2. These findings reveal that unit-level building segmentation performs best in three regions for 2D building masks and in five regions for 3D building models. The overall accuracy for IOU metrics indicate that the performance of building model reconstruction is largely contingent upon the initial building segmentation accuracy. Nonetheless, unit-level segmentation significantly enhances the accuracy of 3D primitives for each building section.

Figure 4 and Figure 5 display building 3D models in two high building density regions, London area 1 and Trento area. From instance-level comparison for each building unit, it indicates that unit-level segmentation of building masks effectively divides complex or densely packed buildings in areas with weak textures into individually segmented units, maintaining consistency along their shared boundaries, and then computing fine 3D roof parameters. In contrast, building models generated by the other two methods often treat complex structures or buildings in close proximity as a single and large building section.

4. Discussion

The experimental results indicate that unit-level building reconstruction method obviously improves the granularity and accuracy of the building LoD2 model, particularly in densely populated urban environments. This approach significantly improves the delineation of building boundaries and the calculation of 3D roof primitives by segmenting complex and adjacent buildings into distinct sections. Such precise urban modeling is crucial for creating more accurate and reliable representations of building structures.

The enhanced detail is particularly advantageous for planning and analysis in urban development, notably in areas characterized by dense, irregularly shaped buildings with non-distinct textures, such as slums and poorly maintained neighborhoods—areas that previous reconstruction methods struggled to accurately model. The unit-level method has considerable capacity for simulating highly intricate and business-oriented structures. Proficiently analyzing and precisely depicting the complex formations of these edifices can significantly assist in diverse urban planning and architectural implementations. This approach can offer a more intricate comprehension of the urban environment, particularly in the case of commercial structures that frequently showcase distinctive and intricate architectural styles.

5. Conclusion

This paper introduces an effective level of details (LoD)2 building reconstruction approach at the unit-level, leveraging unit-level building segmentation results from satellite-derived Orthophoto and digital surface model (DSM) and a model-derived approach for 3D modeling. This method initiates with the segmentation to get unit-level building segments, followed by a polygon composition strategy designed to distinguish duplex or dense buildings as separate entities equipped with 3D primitives. Our technique effectively segments complex buildings and immediately adjacent buildings in densely populated urban areas with low-texture quality, subsequently reconstructing 3D building models utilizing a comprehensive library of predefined

Table 2. Accuracy comparison in 2D label (semantic segmentation), 2D footprint, and 3D model (reconstruction) for all regions. The difference in the method is the input of building mask.

IoU	HRNet 2D label	ESRI 2D label	Unit-level 2D label	HRNet 2D footprint	ESRI 2D footprint	Unit-level 2D footprint	HRNet 3D model	ESRI 3D model	Unit-level 3D model
Columbus 1	0.7140	0.6647	**0.7360**	0.5389	0.6151	**0.6248**	0.4881	0.5342	**0.5801**
Columbus 2	**0.8492**	0.8306	0.7980	0.7526	**0.7815**	0.7309	0.7403	**0.7649**	0.7196
Buenos Aires 1	**0.6706**	0.4434	0.6324	**0.6149**	0.4043	0.5662	**0.5610**	0.3296	0.5314
Buenos Aires 2	0.5635	0.1534	**0.6010**	0.4965	0.1233	**0.5255**	0.4480	0.0932	**0.4970**
London 1	0.6826	0.5993	**0.7471**	0.5668	0.5222	**0.6265**	0.3857	0.3067	**0.4382**
London 2	0.5974	0.4846	**0.6115**	0.4882	0.4404	**0.5377**	0.4154	0.3348	**0.4728**
Trento	**0.6578**	0.4071	0.6400	0.5808	0.3573	**0.6021**	0.3010	0.1418	**0.3311**

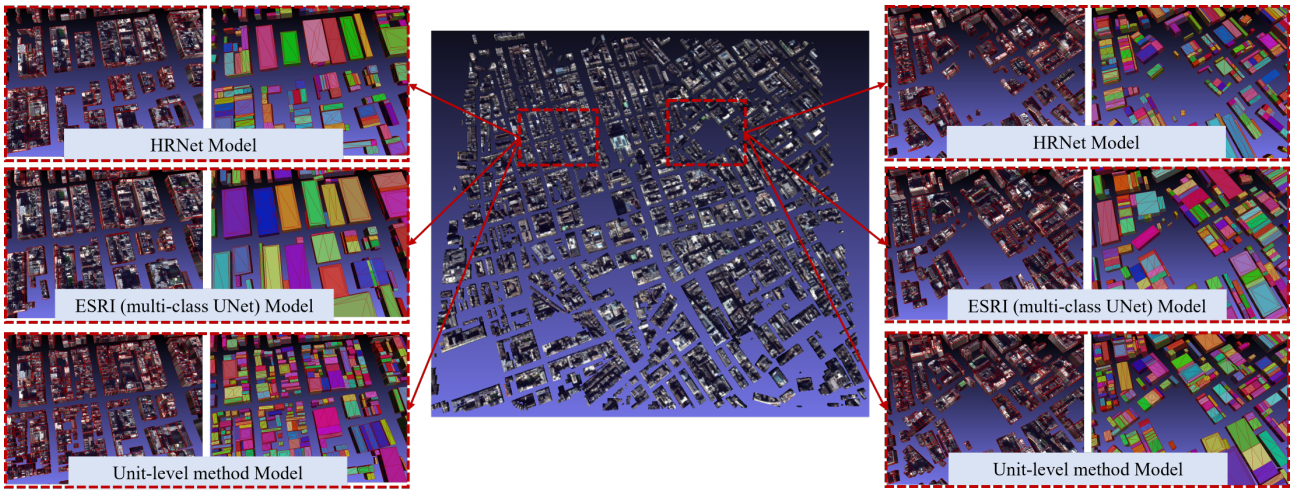


Figure 4. Building 3D models in London 1 region with dense urban structures

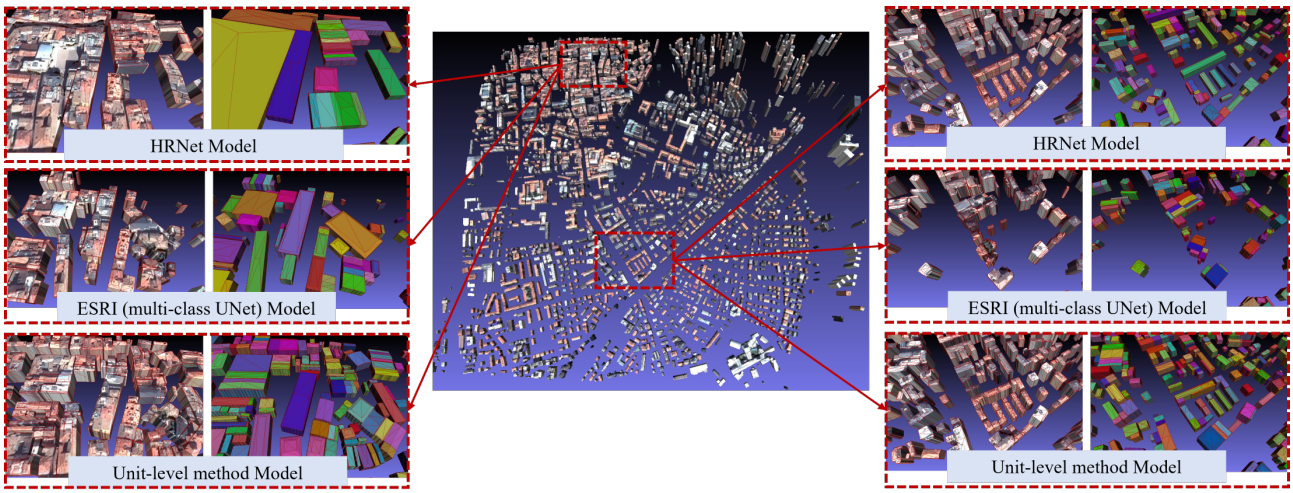


Figure 5. Building 3D models in Trento region with both dense and sparse buildings

models. The empirical evaluation of experiments shows that our unit-level LoD2 modeling surpasses the construction result from publicly available building segmentation methods, especially in dense and complex urban environments.

References

Alidoost, F., Arefi, H., Tombari, F., 2019. 2D image-to-3D model: Knowledge-based 3D building reconstruction (3DBR) using single aerial images and convolutional neural networks (CNNs). *Remote Sensing*, 11(19), 2219.

Bauchet, J.-P., Lafarge, F., 2018. Kippi: Kinetic polygonal partitioning of images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Beucher, S., Meyer, F., 2018. The morphological approach to segmentation: the watershed transformation. *Mathematical morphology in image processing*, CRC Press, 433–481.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59, 25–37.

Chen, K., Lu, W., Xue, F., Tang, P., Li, L. H., 2018. Automatic building information model reconstruction in high-density urban areas: Augmenting multi-source data with architectural knowledge. *Automation in Construction*, 93, 22–34.

Douglas, D. H., Peucker, T. K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2), 112–122.

Facciolo, G., De Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 57–66.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Gröger, G., Kolbe, T. H., Czerwinski, A., Nagel, C., 2008. OpenGIS city geography markup language (CityGML) encoding standard, version 1.0. 0.

Gui, S., Qin, R., 2021. Automated LoD-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 1–19.

Gui, S., Qin, R., Tang, Y., 2022. SAT2LOD2: a Software for Automated LOD-2 Building Reconstruction from Satellite-Derived Orthophoto and Digital Surface Model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 379–386.

Han, Y., Liu, W., Huang, X., Wang, S., Qin, R., 2020. Stereo dense image matching by adaptive fusion of multiple-window matching results. *Remote Sensing*, 12(19), 3138.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, H., Brenner, C., Sester, M., 2013. A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data. *ISPRS Journal of photogrammetry and remote sensing*, 79, 29–43.

Huang, X., Chen, K., Tang, D., Liu, C., Ren, L., Sun, Z., Hänsch, R., Schmitt, M., Sun, X., Huang, H. et al., 2023. Urban Building Classification (UBC) V2-A Benchmark for Global Building Detection and Fine-grained Classification from Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*.

Huang, X., Qin, R., 2020. Post-filtering with surface orientation constraints for stereo dense image matching. *The Photogrammetric Record*, 35(171), 375–401.

Jayaraj, P., Ramiya, A. M., 2018. 3D CityGML building modelling from lidar point cloud data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 175–180.

Kada, M., Wichmann, A., 2012. Sub-surface growing and boundary generalization for 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1, 233–238.

Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2008. Structural approach for building reconstruction from a single DSM. *IEEE Transactions on pattern analysis and machine intelligence*, 32(1), 135–147.

Li, M., Lafarge, F., Marlet, R., 2020. Approximating shapes in images with low-complexity polygons. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8630 ff.

Li, Q., Mou, L., Hua, Y., Shi, Y., Chen, S., Sun, Y., Zhu, X. X., 2023a. 3DCentripetalNet: Building height retrieval from monocular remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 120, 103311.

Li, S., He, S., Jiang, S., Jiang, W., Zhang, L., 2023b. WHU-Stereo: A Challenging Benchmark for Stereo Matching of High-Resolution Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–14.

Li, W., Zhao, W., Yu, J., Zheng, J., He, C., Fu, H., Lin, D., 2023c. Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 201, 26–37.

Li, Z., Wegner, J. D., Lucchi, A., 2019. Topological Map Extraction From Overhead Images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1715 ff.

Ling, X., Qin, R., 2022. A graph-matching approach for cross-view registration of over-view and street-view based point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 2–15.

Mao, Y., Chen, K., Zhao, L., Chen, W., Tang, D., Liu, W., Wang, Z., Diao, W., Sun, X., Fu, K., 2023. Elevation Estimation-Driven Building 3D Reconstruction from Single-View Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*.

Mohanty, S. P., Czakon, J., Kaczmarek, K. A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S. et al., 2020. Deep Learning for Understanding Satellite Imagery: An Experimental Survey. *Frontiers in Artificial Intelligence*, 3.

Mosinska, A., Marquez-Neila, P., Koziński, M., Fua, P., 2018. Beyond the pixel-wise loss for topology-aware delineation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3136–3145.

Partovi, T., Fraundorfer, F., Bahmanyar, R., Huang, H., Reinartz, P., 2019. Automatic 3-D building model reconstruction from very high resolution stereo satellite imagery. *Remote Sensing*, 11(14), 1660.

Qin, R., 2016. Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 77–82.

Qin, R., 2019. Automated 3D recovery from very high resolution multi-view satellite images. *arXiv preprint arXiv:1905.07475*.

Robinson, C., Hou, L., Malkin, K., Soobitsky, R., Czawlytko, J., Dilkina, B., Jojic, N., 2019. Large scale high-resolution land cover mapping with multi-resolution data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12726–12735.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 234–241.

Schnabel, R., Wahl, R., Klein, R., 2007. Efficient RANSAC for point-cloud shape detection. *Computer graphics forum*, 26(2), 214–226.

Schuegraf, P., Zorzi, S., Fraundorfer, F., Bittner, K., 2023. Deep Learning for the Automatic Division of Building Constructions Into Sections on Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, Springer, 240–248.

Sun, S., Salvaggio, C., 2013. Aerial 3D building detection and modeling from airborne LiDAR point clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3), 1440–1449.

Von Gioi, R. G., Jakubowicz, J., Morel, J.-M., Randall, G., 2008. LSD: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4), 722–732.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. et al., 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 3349–3364.

Wang, Y., Zorzi, S., Bittner, K., 2021. Machine-learned 3d building vectorization from satellite imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1072–1081.

Xu, N., Qin, R., Huang, D., Remondino, F., n.d. Multi-tiling neural radiance field (NeRF)—geometric assessment on large-scale aerial datasets. *The Photogrammetric Record*.

Zhou, G., Cao, S., Zhou, J., 2016. Planar segmentation using range images from terrestrial laser scanning. *IEEE Geoscience and Remote Sensing Letters*, 13(2), 257–261.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal Building Extraction With Graph Neural Networks in Satellite Images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1848 ff.