From Pixels to Polylines: Extracting City-scale Vectorized Roof Structures with Line Segment Detection Networks

Mehmet Büyükdemircioğlu[1], Fabio Remondino[1], Martin Kada[2], Sultan Kocaman[3]

[1] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy – mbuyukdemircioglu, remondino@fbk.eu
[2] Technische Universität Berlin, Institute of Geodesy and Geoinformation Science, Berlin, Germany – martin.kada@tu-berlin.de
[3] GeoPlato Engineering Inc., Bilkent Cyberpark, Ankara 06450, Türkiye – sultan@geoplato.com

Keywords: Deep Learning, Line Segment Detection, LOD2.2, True Orthophoto, Building Reconstruction

ABSTRACT
Automatic extraction of vectorized roof structures above LOD2.0 remains challenging due to their geometric complexity and the presence of small and occluded elements over the roofs. Detecting fine-scale roof objects such as chimneys and dormer windows in very high-resolution aerial imagery is still an active research topic. This study presents a workflow for automated detection and vectorization roof structures at city scale using Line Segment Detection (LSD) networks. Compared to model-based building reconstruction approaches, LSD networks do not rely on pre-defined roof typologies and are able to extract complex roof structures and small objects over the building roofs. For this purpose, a dataset comprising approximately 139,000 buildings with LOD2.2 roof structures and more than 2.2 million roof segments is generated using 8 cm GSD aerial imagery. An automated end-to-end workflow is developed, trained and tested from the available data. Experimental results indicate that roof structures suitable for LOD2.2 3D roofs can be extracted and vectorized with high accuracy, achieving 58.4% msAP and 73.1% mAP^J with ULSD network. Robustness is further assessed by visual inspection in areas affected by roof-blocking objects such as trees and cast shadows. Source code of the proposed workflow is publicly available at: https://github.com/3DOM-FBK/pix2poly.

1. Introduction

Roofs are key components of buildings and play a central role in 3D building reconstruction. In recent years, applications of 3D city models have expanded across a wide range of domains and are expected to grow further in the future (Biljecki et al., 2015). Most roof types worldwide follow standard shapes such as flat, hip, and gable. Three-dimensional building models with such roof types can be generated manually or semi-automatically at Level of Detail (LOD) 2.0 (Haala and Kada, 2010). However, automatic extraction of roof structures beyond LOD2.0, especially for complex roof structures and occluded geometries remains challenging and is still actively investigated.

Roof geometries can be derived from laser scanning (Kada, 2022), satellite imagery (Wang et al., 2025), and stereo aerial imagery (Buyukdemircioglu, 2023), using geometric- (Nex and Remondino, 2012) or learning-based (Gui et al., 2022) methods. Each data source exhibiting specific advantages and limitations. During laser scanning, critical points such as roof corners or nodes may be missed. In satellite imagery, roof details may be insufficiently resolved depending on spatial resolution. Very high-resolution stereo aerial imagery is often preferred for 3D city model production because roof details are clearly visible and dense digital surface models (DSMs) can be generated. Moreover, such imagery is typically more cost-effective than light detection and ranging (LiDAR) data (Buyukdemircioglu et al., 2018).

Building reconstruction methods for city modelling can broadly be categorized as model-driven and data-driven (Jarząbek-Rychard et al., 2025). Model-driven approaches reconstruct roof models by matching predicted roof line structures to roof types stored in a building library. Data-driven methods avoid such libraries and aim to reconstruct complex buildings directly from the data by estimating their parameters. Existing automatic and semi-automatic approaches for extracting roof details and geometries often lack robustness in the presence of roof-blocking objects. Consequently, in many operational projects, roof structures are still digitized manually by photogrammetric operators from stereo aerial imagery, which is both labour intensive, time-consuming and costly.

Building roofs can be reconstructed at various LODs. The Open Geospatial Consortium (OGC) standard City Geography Markup Language (CityGML) (Gröger and Plümer, 2012) specifies five building LODs between 0 and 4. Biljecki et al. (2016) further refined the LOD2 definition into four sublevels: 2.0, 2.1, 2.2, and 2.3. An LOD2.0 model represents only the main roof structures and omits small objects such as windows and chimneys. Compared to LOD2.0, LOD2.1 includes smaller roof parts and extensions such as alcoves, large wall indentations, and external flues. LOD2.2 imposes additional requirements because roof superstructures larger than 2 m and 2 m^2 must be represented. At LOD2.3, overhangs are explicitly modelled when they exceed 0.2 m in length, ensuring that roof edges and building footprints are represented in their true positions.

State-of-the-art performance in various photogrammetry and remote sensing tasks, including segmentation, classification, and object detection, can be achieved with deep learning methods (Heipke and Rottensteiner, 2020). Deep learning has also been applied to extract roof line structures at LOD2.0 (Xu et al., 2024; Hensel et al., 2021), but there remains considerable room for improvement, particularly for complex and occluded roof structures. This study introduces a framework based on line segment detection (LSD) networks for extracting vectorized LOD2.2 roof structures at city-scale. The framework investigates four state-of-the-art LSD networks and achieves state-of-the-art performance compared to previous roof structure extraction studies. ULSD (Li et al., 2021), L-CNN (Zhou et al., 2019), HAWP v2 (Xue et al., 2022), and F-Clip (Dai et al., 2022) are implemented and benchmarked. The proposed approach differs from existing work in which building footprint vectors are frequently combined with RGB images and DSMs, and image tiles are clipped strictly along building footprints so that each tile contains a single building. By contrast, the proposed method does

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-2-2026-439-2026

not require footprint-based clipping and can predict roof line structures directly from RGB image tiles.

The main contributions of this study can be listed as following:

(1) An end-to-end, fully automated pipeline that converts existing 3D city models and very high resolution true orthophotos into training and test data for LSD networks, including vectorization, georeferencing and topological post-processing at city scale.
(2) A framework, based on LSD networks, to extract roof segments and produce vector results suitable for LOD2.2 roof models.
(3) A comparative evaluation of four state-of-the-art LSD networks - L-CNN, HAWP v2, F-Clip, and ULSD - on LOD2.2 roof structure extraction, using both quantitative (msAP, mAP^J) and qualitative analyses, including robustness under occlusion.
(4) An ablation study on Bezier curve order and data augmentation, demonstrating that ULSD with fourth-order curves and 4× augmentation yields state-of-the-art performance on the employed dataset.
(5) A city-scale LOD2.2 roof-structure dataset comprising more than 139,000 buildings and 2.2 million roof line segments tailored to line segment detection networks.

2. Related Work

Line segment detection is a widely studied problem in computer vision, aiming to generate vectorized line representations from RGB images. Line segments can be extracted using conventional methods such as Line Segment Detector (LSD) (Von Gioi et al., 2008). However, deep learning methods are now predominantly employed due to their superior accuracy and robustness compared to conventional methods. Most line segment detection networks are evaluated on the Wireframe (Huang et al., 2018) and YorkUrban (Denis et al., 2008) benchmark datasets. However, only a limited number of large-scale datasets are available for extracting LOD2.2 roof structures from very high resolution (< 10cm) aerial images.

Zhou et al. (2019) proposed End-to-End Wireframe Parsing (L-CNN) for vectorizing wireframes in RGB images. The network consists of four modules. First, a backbone network extracts features from the input image and provides intermediate feature maps for subsequent modules. A junction proposal module then produces candidate junctions, followed by a line sampling module that generates line proposals based on pairs of candidate junctions. Finally, a line verification module classifies the proposed line segments. Holistically-Attracted Wireframe Parsing (HAWP), proposed by Xue et al. (2022), comprises three main components in version 2: generating line segments and proposing junctions, binding line segments to junctions, and verifying line segments. The method first lifts line segments in the image wireframe into attraction regions and parameterizes each pixel within these regions, resulting in a representation known as Holistic Attraction (HAT) fields. Compared to L-CNN, HAWP v1 requires approximately four times fewer line segment proposals, and HAWP v2 further reduces this to a factor of eight, substantially accelerating wireframe parsing and improving inference speed.

The Fully Convolutional Line Parsing network (F-Clip) proposed by Dai et al. (2022) is a one-stage line segment detector that bypasses explicit junction detection and predicts lines directly. A convolutional neural network is used to extract image features, and two additional convolutional layers regress center, length, and angle score maps. For each high-confidence line center, a line segment is generated by associating the corresponding length and angle estimates. Unlike many other line segment detection networks that rely on a single backbone, F-Clip offers six different backbone variants. PPGNet (Zhang et al., 2019) is another convolutional neural network designed to extract line segment graphs from RGB images. PPGNet comprises four steps: a convolutional backbone to extract features, a Junction Detection Module, a Line Segment Alignment Module that identifies line segment candidates based on junction pairs by extracting a feature tensor, and an Adjacency Matrix Inference Module to determine whether each junction pair is connected. An adjacency matrix representation is used to jointly predict junction locations and their connectivity.

Li et al. (2021) introduced Unified Line Segment Detection (ULSD) to detect line segments in both distorted and undistorted images acquired from pinhole, fisheye, and spherical cameras. The network directly generates vectorized line segments from input images and consists of three components: a feature extraction backbone, a Line Proposal Network (LPN), and a Line of Interest (LoI) head. ULSD can extract not only straight line segments but also curved segments parameterized as Bezier curves. Conventional (non learning-based) methods can also be used to match line segments from aerial imagery in urban areas (Ok et al., 2012). Neural networks, however, are not only capable of detecting edges in images but also of assembling them into graph structures. Several studies have applied both conventional and deep learning approaches to roof-line detection and 3D building reconstruction. Buyukdemircioglu et al. (2022) provided a detailed review of deep learning methods for 3D building reconstruction.

Using very high resolution orthophotos of the city of Detmold, Hensel et al. (2021) vectorized building roof structures at LOD2.0 with PPGNet. The reported F1 scores for junction detection and edge detection are 0.93 and 0.87, respectively. Conv-MPN, a message-passing neural network architecture proposed by Zhang et al. (2020), reconstructs outdoor buildings as planar graphs from single RGB images. This method is strongly dependent on pre-processing steps, such as corner detection, and cannot be trained in a fully end-to-end manner. Multi-stage strategies of this type are computationally expensive and can be inefficient for both training and inference.

Deep Roof Refiner proposed by Qian et al. (2022) is specifically designed to compute roof structure lines. Using level-18 Google Earth satellite (GES) imagery, the method achieved optimal F1 scores of 60.89% and 63.48% in quantitative and qualitative experiments, respectively. Gui and Qin (2021) proposed a framework for automatically reconstructing LOD2.0 models from high resolution multi-view satellite stereo images using a “decomposition-optimization-fitting” paradigm based on a model-driven approach. Due to the limited number of roof model types in their library, the applicability of this strategy to complex roof structures is restricted.

Existing line segment detection networks have primarily been developed and evaluated on generic wireframe benchmarks, while applications targeting detailed roof structures in 3D city modelling remain limited. Previous roof-line extraction studies largely focus on LOD2.0 representations, often rely on predefined roof libraries or extensive pre-processing, and are not designed for end-to-end, city-scale LOD2.2 reconstruction. Dehbi et al. (2021) proposed a hierarchical method based on probability density functions and SVM classification to separate dormer and roof points from LiDAR data for reconstruction, but their approach is not tailored to image-based, end-to-end LOD2.2 extraction. In contrast, the present study investigates state-of-the-art LSD networks for extracting vectorized LOD2.2 roof structures from very high resolution true orthophotos and introduces an end-to-end pipeline for city-scale extraction of vectorized and georeferenced LOD2.2 roof structures.

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-2-2026-439-2026 | © Author(s) 2026. CC BY 4.0 License.

3. Study Area

Experiments are conducted on a dataset covering approximately 1010 km² in Ankara, Türkiye. The dataset contains more than 139,000 buildings with a total of 2.2 million line segments describing roofs at LOD2.2. Although the roof types in the study area are diverse, they are representative of roof structures commonly observed across Türkiye. The site is selected due to the availability of high-quality data provided by the General Directorate of Land Registry and Cadastre (GDLRC) of Türkiye for the purposes of this study. Photogrammetric flight missions are carried out by the GDLRC as part of a nationwide 3D city modelling project (Dursun et al., 2022). Roof structures are manually delineated in LOD2.2 by photogrammetry operators using stereo viewing. In addition to the LOD2.2 roof models, high resolution true orthophotos with 8 cm ground sampling distance and digital surface models (DSMs) are produced within the same project. Provided roof vectors are preserved in their original form without any elimination or simplification. An overview of the randomly selected training (red) and test (green) tiles over the study area is shown in Figure 1.

Figure 1. A close view of the randomly selected training (red) and test (green) tiles over the study area.

4. Methodology

4.1 Data preparation

The data preparation phase aimed at generating training and test samples to be used as input to the line segment detection networks. The entire pipeline for image tiling, coordinate transformation, ground truth generation, and merging of output predictions is automated in Python. As used networks require image tiles of 512 × 512 pixels, non-overlapping tiles are produced from the true orthophotos. Data are generated as non-overlapping tiles to better reconnect junctions at tile boundaries and avoid spurious merges within roofs, which would deform roof geometries. The same tiling scheme is applied to the vector data to generate corresponding masks. The final dataset consists of 30,446 image tiles with approximately 2.2 million roof line segments at LOD2.2. Both the number of tiles and the number of line segments are split into 90% for training and 10% for testing, ensuring a balanced distribution between them. The total counts of tiles and line segments in the training and test sets are reported in Table 1. An overview of the image tiling approach for data is given in Figure 2. In addition to a large unseen area, further test tiles are randomly sampled from different parts of the study area to cover a wide range of roof typologies within the city. Roofs partially occluded by trees and cast shadows are explicitly included in the test set to enable a robust evaluation of the methods.

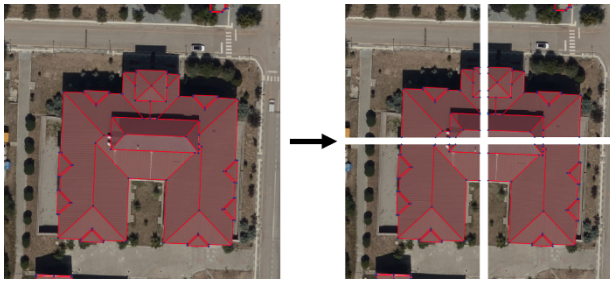

Figure 2. A building example before (left) and after (right) tiling with roof lines and junction points.

Data type	Percentage	Nr. of lines	Nr. of tiles
Training	90%	1,982,313	27,402
Test	10%	220,257	3,044
Total	100%	2,202,570	30,446

Table 1. Number of lines and image tiles used in the study.

4.2 Line Segment Detection networks

Four state-of-the-art line segment detection networks - HAWP, L-CNN, F-Clip, and ULSD - are employed to predict roof line segments in this study. The selection of these networks is motivated by their shared use of a stacked hourglass backbone (Newell et al., 2016) and their comparable performance on the Wireframe and YorkUrban benchmark datasets. The stacked hourglass network has a U-shaped architecture and is predominantly used for human pose estimation. Training of all networks is performed from scratch using the generated LOD2.2 roof structure dataset. For a fair comparison, all models adopt an identical training configuration and input data, and roof segments and junctions are predicted using the held-out test set. An overview of the overall workflow is given in Figure 3.

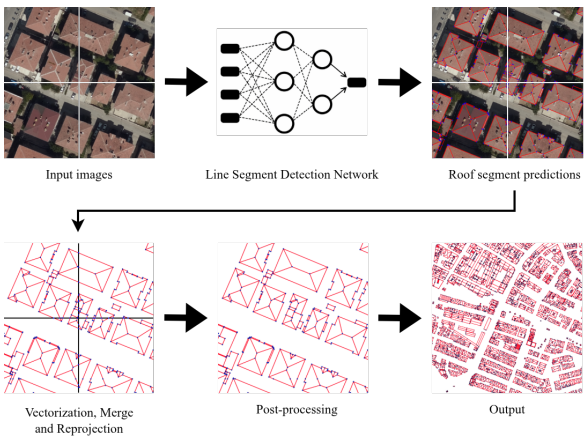

Figure 3. Overall workflow of the study.

Non-overlapping RGB tiles of size 512 × 512 pixels serve as input in both training and testing. Generated tiles pass through the stacked hourglass backbone to produce feature maps with an output size of 128 × 128 × 256. The hyperparameter configuration of the stacked hourglass backbone followed that of L-CNN and HAWP. The initial learning rate and weight decay are set to 4×10^{-4} and 1×10^{-4}, respectively, with the learning rate decaying at the 25th epoch. A batch size of 32 is used during training phase. A separate data augmentation experiment, reported in the ablation study section, examines its impact; however, the main training runs presented here do not employ data augmentation. Preliminary experiments indicated that 30 training epochs with the stacked hourglass backbone sufficed to achieve stable performance, with additional epochs providing no

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-2-2026-439-2026 | © Author(s) 2026. CC BY 4.0 License.

substantial improvement. Consequently, all networks ran for 30 epochs using the Adam optimizer. All experiments ran on a single NVIDIA Quadro RTX 8000 GPU with 48 GB of memory.

4.3 Evaluation Metrics

Heat map-based metrics originate from boundary detection but fail to penalize overlapping or fragmented lines and do not explicitly assess wireframe connectivity, which makes them unsuitable for line segment detection. To address this issue, Zhou et al. (2019) introduced structured Average Precision (sAP), computed as the area under the precision-recall curve from a scored list of detected line segments over all test images. Line segments are matched to ground truth under different endpoint distance thresholds (5, 10, and 15 pixels), yielding sAP^5, sAP^{10}, and sAP^{15}; their average is reported as mean structural Average Precision (msAP), which summarizes overall line segment quality under both strict and relaxed matching criteria.
Junction quality is assessed using the mean Average Precision for junctions (mAP^J). Junction predictions are matched to ground-truth junctions based on Euclidean distance at several thresholds (typically 0.5, 1.0, and 2.0 pixels), and the corresponding AP values are averaged. While msAP captures the geometric and structural quality of the detected line segments, mAP^J reflects the accuracy of junction localization and roof topology, providing a complementary measure of performance for extracting detailed LOD2.2 roof structures.

4.4 Vectorization and post-processing

The vectorization process transforms the pixel coordinates of the predicted line segments into projected map coordinates with sub-pixel precision. In the final step, all roof line segments are merged into a single vector dataset and stored in standard geospatial formats such as ESRI Shapefile and GeoJSON. Post-processing reduces artifacts introduced by tiling and model predictions. The tiled outputs contain small gaps along tile boundaries as well as redundant junctions, duplicate lines, and unnecessary vertices. A distance-based snapping procedure closes gaps between tiles, followed by topological cleaning that removes redundant entities. Subsequently, the Douglas-Peucker simplification algorithm eliminates additional superfluous vertices while preserving the main roof geometry. After post-processing, the total number of junctions and line segments decreases by 24.8% and 35.4%, respectively. A before-and-after view of the post-processing results is given in Figure 4.

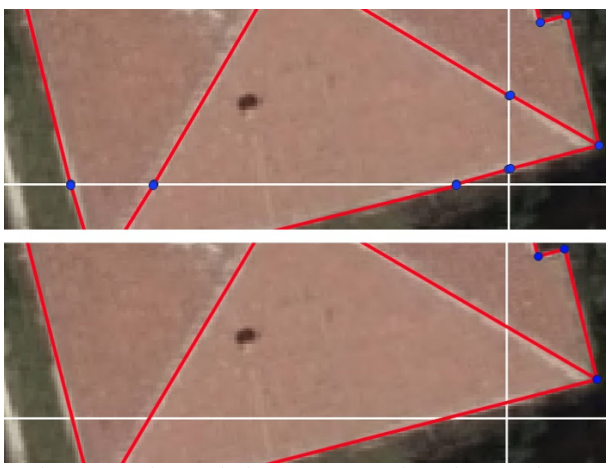

Figure 4. Before and after post-processing of vector data.

5. Results

5.1 Quantitative Results

The quantitative results for all line segment detection networks are reported in Table 2. ULSD achieved the highest performance across all evaluated metrics. Compared to HAWP v2, which ranked second overall, ULSD improved msAP by 4.8%. Since F-Clip does not explicitly model junctions, the junction metric mAP^J cannot be computed for this method. F-Clip (HG2-LB backbone) and L-CNN reached similar performance in terms of msAP, while HAWP v2 outperformed both L-CNN and F-Clip—despite sharing the same stacked hourglass backbone—by 3.7% in msAP.

Method	sAP^5	sAP^{10}	sAP^{15}	msAP	mAP^J
F-Clip	43.5	48.0	50.0	47.2	-
L-CNN	44.7	47.7	49.2	47.2	53.7
HAWP	48.2	51.5	53.1	50.9	54.7
ULSD	**52.5**	**56.3**	**58.3**	**55.7**	**69.0**

Table 2. Quantitative results for roof segment extraction.

ULSD represents line segments with Bézier curves, with the curve order treated as a tunable parameter. In this study, a fourth-order Bézier formulation is adopted, as it yielded the best results among the tested configurations. Section 5.4.1 provides a detailed comparison of alternative curve orders, highlighting the sensitivity of performance to this design choice. Since all evaluated networks originate from development and tuning on the Wireframe benchmark, their relative ranking shifts when transferred to the roof line segment extraction task. Quantitative analysis identifies ULSD as the most suitable LSD network for extracting LOD2.2 roof line segments, achieving the highest overall performance among the evaluated methods. The ablation study in Section 5.4 highlights that applying data augmentation further enhances the performance of the ULSD network, yielding additional gains in both msAP and mAP^J.

5.2 Qualitative Results

Figure 5 illustrates predicted and merged roof segments for the LSD networks considered in this study. For all networks except F-Clip, roof segments and junctions with scores above 80% are visualized. For F-Clip, a lower threshold of 40% is applied, as no segments exceed 80% confidence. Due to the large number of missing roof segments and incomplete predictions, F-Clip shows clearly lower performance than the other methods in the visual assessment. Since F-Clip bypasses explicit junction detection and predicts line segments directly, line endpoints are treated as junctions for visualization purposes.
L-CNN and HAWP v2 produced visually similar results, consistent with their comparable quantitative scores. L-CNN performed slightly better in detecting buildings without explicit roof structures (LOD1) and buildings whose roof colours are similar to surrounding structures. However, relative to ULSD and HAWP v2, L-CNN generated more redundant line segments over non-building objects. HAWP v2 showed marginally better visual quality than L-CNN but also produced redundant lines and incomplete roof polygons in some cases. HAWP v2 is generally more reliable in capturing complex LOD2.2 roofs with small superstructures such as chimneys and small windows, yet still failed to recover some simpler roof types, including hip and gable roofs.

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-2-2026-439-2026 | © Author(s) 2026. CC BY 4.0 License.

Figure 5. Visual results achieved with the LSD networks on 56 merged image tiles.

Both quantitative metrics and visual inspection indicate that ULSD is the most effective method for detecting roof segments and creating vectors suitable for LOD2.2 roof 3D models. ULSD successfully delineated roof structures smaller than 1 m² and captured line segments in complex roof configurations. Furthermore, the qualitative results show that ULSD produced fewer redundant lines outside building footprints compared to the other networks. Its ability to model line segments as Bezier curves also contributed to improved delineation of curved roof structures.

5.3 Robustness against occlusion

This section evaluates the robustness of the networks for roofs that are completely or partially occluded by trees and cast shadows. Many computer vision methods struggle to correctly reconstruct or detect building roofs when they are heavily occluded. To assess robustness of the LSD networks under such conditions, a subset of test tiles containing roofs with full or partial tree occlusions is manually selected for visual inspection. As in the qualitative analysis, roof segments with probability scores above 40% for F-Clip and 80% for the other methods are considered.

Among the evaluated methods, F-Clip exhibited the lowest accuracy, even under the more permissive score threshold, with numerous missing roof segments and frequent false detections. L-CNN detected a larger proportion of roof structures but introduced redundant line segments outside roof areas. HAWP v2 successfully delineated several occluded roofs, although its performance remained generally inferior to that of L-CNN in these cases. ULSD achieved slightly higher completeness and correctness in roof delineation than L-CNN, yet still produced several mis-predicted line segments under strong occlusion. Figure 7 presents examples of test images affected by tree occlusions alongside the corresponding predictions and ground-truth annotations.

5.4 Ablation studies

5.4.1 Order of the Bezier Curve

ULSD differs from the other evaluated networks by representing line segments with Bezier curves, which enables the modelling of both straight and curved roof edges. This subsection analyses the impact of Bezier curve order on performance, followed by a separate examination of data augmentation effects. A key factor influencing the ability to accurately fit roof lines is the order of the Bezier curve in ULSD. With first-order curves, the model can represent straight segments, whereas higher-order curves allow the approximation of curved structures. The influence of Bezier curve order on the LOD2.2 roof line dataset is evaluated for orders 1 through 6. Although the differences between configurations are relatively small, Bezier curve order 4 achieved the best overall performance. Quantitative results for different Bezier curve orders are given in Table 3.

Order	sAP^5	sAP^{10}	sAP^{15}	msAP	mAP^J
1	51.9	55.8	57.8	55.2	68.3
2	52.1	56.0	58.0	55.4	68.6
3	51.9	55.9	57.9	55.2	68.8
4	**52.5**	**56.3**	58.3	**55.7**	**69.0**
5	52.1	56.2	58.3	55.5	68.9
6	52.2	**56.3**	**58.4**	55.6	68.9

Table 3. Comparison of Bezier curve order results for ULSD.

In the original wireframe study by Li et al. (2021), fifth-order curves yielded the highest accuracy on indoor wireframe data. The discrepancy with the present findings can be attributed to differences in scene characteristics: the LOD2.2 roof line dataset predominantly contains building roof structures, whereas the Wireframe dataset mainly comprises indoor layouts. These results indicate that the optimal Bezier curve order is data-dependent and varies with the geometric properties of the target objects.

5.4.2 Data Augmentation

Data augmentation refers to the process of synthetically enlarging a training dataset through label-preserving transformations. This strategy is commonly adopted to improve network generalization by increasing the diversity and effective size of the training data. In this study, data augmentation is applied by horizontally and vertically flipping the tiles, effectively quadrupling the number of training samples. The impact of augmentation is assessed by comparing performance against a baseline model trained on the original (non-augmented) dataset.

As a result of augmentation, the size of the training set increased from 27,402 to 109,608 images, along with their corresponding masks. Both ULSD models (with and without augmentation) employ identical hyperparameters to ensure a fair comparison. Data augmentation yields a consistent performance gain: compared to the baseline, msAP increases by 2.7% and mAP^J by 4.1%. Visual inspection also indicates that data augmentation improves the delineation of small roof details. The quantitative results for both ULSD configurations are summarized in Table 4. An example of augmented image tile is given in Figure 6.

Augmentation	sAP^5	sAP^{10}	sAP^{15}	msAP	mAP^J
None	52.5	56.3	58.3	55.7	69.0
4x	**55.3**	**59.0**	**60.8**	**58.4**	**73.1**

Table 4. Quantitative results with data augmentation

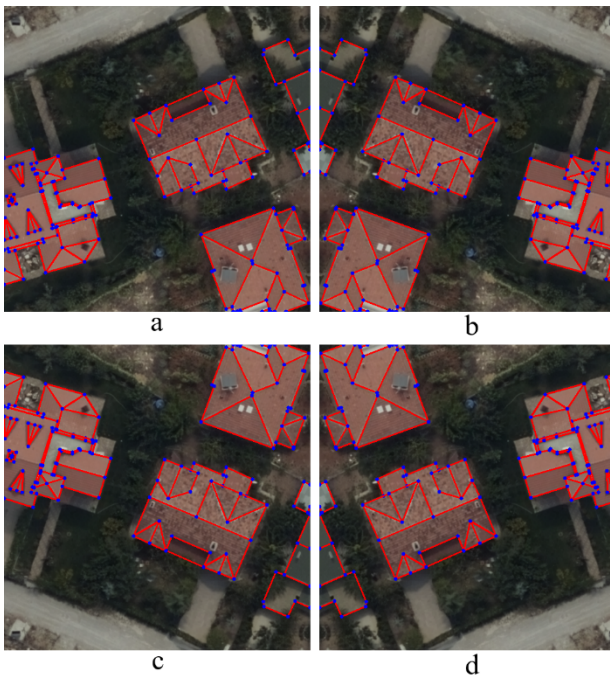

Figure 6. Data augmentation: (a) original, (b) horizontal flip, (c) vertical flip and (d) mirroring.

Figure 7. Robustness comparison of LSD networks against occluded roofs.

6. Conclusions

This study introduces a deep learning-based framework for fully automatic, city-scale vectorization of roof structures from very high-resolution aerial imagery suitable for LOD2.2 3D roof models. A dedicated LOD2.2 training dataset for line segment detection networks is presented, encompassing more than 139,000 buildings and 2.2 million roof line segments. The pipeline automates image tiling, ground-truth generation, prediction merging, and topological cleaning via custom Python scripts, converting existing 3D roof models and true orthophotos into inputs suitable for line segment detection. Among the evaluated methods, ULSD attains the strongest performance (58.4% msAP; 73.1% mAP^J) and maintains robustness under challenging conditions such as partial or full occlusion by trees and cast shadows. The vectorized outputs are directly usable for subsequent 3D building reconstruction when combined with height information from LiDAR or multi-view stereo-derived DSMs.

All networks are trained from scratch on the proposed dataset to ensure a fair, task-specific comparison. Results indicate that the ULSD-based pipeline produces fewer redundant off-roof segments and demonstrates greater resilience to roof-blocking objects than competing approaches.

Future work will lead us to: extend the framework to direct 3D roof reconstruction by fusing true-orthophoto planimetry with elevations from point clouds or raster DSM; enlarge and diversify the dataset to strengthen cross-city generalization; test stronger topology-aware rules during prediction and post-processing; and use DSM as a fourth input band with tailored backbone models.

Acknowledgements

The authors would like to thank General Directorate of Land Registry and Cadastre of Türkiye for providing the data used in this study.

References

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.

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

Buyukdemircioglu, M., 2023: Automatic Reconstruction and Efficient Visualization of 3D City Models. *Hacettepe University Graduate School of Science and Engineering*

Buyukdemircioglu, M., Kocaman, S., Isikdag, U., 2018: Semi-automatic 3D city model generation from large-format aerial images. *ISPRS Int. J. Geo-Inf.*, 7, 339.

Buyukdemircioglu, M., Kocaman, S., Kada, M., 2022: Deep learning for 3D building reconstruction: a review. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 43, 359–366.

Dai, X., Gong, H., Wu, S., Yuan, X., Ma, Y., 2022: Fully convolutional line parsing. *Neurocomputing*, 506, 1-11.

Dehbi, Y., Koppers, S., & Plümer, L., 2021: Looking for a needle in a haystack: Probability density based classification and reconstruction of dormers from 3D point clouds. *Transactions in GIS*, 25(1), 44-70.

Denis, P., Elder, J.H., Estrada, F.J., 2008: Efficient edge-based methods for estimating Manhattan frames in urban imagery. Proc. *ECCV*, 197-210.

Dursun, İ., Aslan, M., Cankurt, İ., Yıldırım, C., Ayyıldız, E., 2022: 3D city models as a 3D cadastral layer: the case of TKGM model. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B4-2022, 507-512.

Gröger, G., Plümer, L., 2012: CityGML – interoperable semantic 3D city models. *ISPRS J. Photogramm. Remote Sens.*, 71, 12-33.

Gui, S., Qin, R., 2021: Automated LOD-2 model reconstruction from very high-resolution satellite-derived digital surface model and orthophoto. *ISPRS J. Photogramm. Remote Sens.*, 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. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 43, 379-386.

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

Heipke, C., Rottensteiner, F., 2020: Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spat. Inf. Sci.*, 23, 10-19.

Hensel, S., Goebbels, S., Kada, M., 2021: Building roof vectorization with PPGNet. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 46, 85-90.

Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y., 2018: Learning to parse wireframes in images of man-made environments. *Proc. IEEE CVPR,* 626-635.

Jarząbek-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.

Kada, M., 2022: 3D reconstruction of simple buildings from point clouds using neural networks with continuous convolutions (ConvPoint). *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 48, 61-66.

Li, H., Yu, H., Wang, J., Yang, W., Yu, L., Scherer, S., 2021: ULSD: unified line segment detection across pinhole, fisheye, and spherical cameras. *ISPRS J. Photogramm. Remote Sens.*, 178, 187-202.

Newell, A., Yang, K., Deng, J., 2016: Stacked hourglass networks for human pose estimation. Proc. *ECCV 2016,* 483-499.

Nex, F., Remondino, F., 2012: Automatic roof outlines reconstruction from photogrammetric DSM. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, I(3), pp. 257-262.

Ok, A.O., Wegner, J.D., Heipke, C., Rottensteiner, F., Soergel, U., Toprak, V., 2012: Matching of straight line segments from aerial stereo images of urban areas. *ISPRS J. Photogramm. Remote Sens.*, 74, 133-152.

Qian, Z., Chen, M., Zhong, T., Zhang, F., Zhu, R., Zhang, Z., Zhang, K., Sun, Z., Lü, G., 2022: Deep roof refiner: a detail-oriented deep learning network for refined delineation of roof structure lines using satellite imagery. *Int. J. Appl. Earth Obs. Geoinf.*, 107, 102680.

Von Gioi, R.G., Jakubowicz, J., Morel, J.M., Randall, G., 2008: LSD: a fast line segment detector with false detection control. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32, 722-732.

Wang, J., Chen, G., Zhang, X., Wang, T., Tan, X., Yang, Q., Zhou, W., Zhu, K., 2025: RoofMapNet: utilizing geometric primitives for depicting planar building roof structure from high-resolution remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.*, 141, 104630.

Xu, Y., Jubanski, J., Bittner, K., Siegert, F., 2024: Roof plane parsing towards LOD-2.2 building reconstruction based on joint learning using remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.*, 133, 104096.

Xue, N., Wu, T., Bai, S., Wang, F.D., Xia, G.S., Zhang, L., Torr, P.H., 2022: Holistically-attracted wireframe parsing: from supervised to self-supervised learning. *arXiv preprint* arXiv:2210.12971.

Zhang, F., Nauata, N., Furukawa, Y., 2020: Conv-MPN: convolutional message passing neural network for structured outdoor architecture reconstruction. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2798-2807.

Zhang, Z., Li, Z., Bi, N., Zheng, J., Wang, J., Huang, K., Luo, W., Xu, Y., Gao, S., 2019: PPGNet: learning point-pair graph for line segment detection. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 7105–7114.

Zhou, Y., Qi, H., Ma, Y., 2019: End-to-end wireframe parsing. *Proc. IEEE/CVF ICCV,* 962-971.