The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W6-2025
3rd International Workshop on Evaluation and BENCHmarking of Sensors, Systems and
GEOspatial Data in Photogrammetry and Remote Sensing (GEOBENCH), 20–21 November 2025, Wroclaw, Poland

Benchmarking Vectorized Building Footprint Extraction from Very High Resolution Aerial Imagery

Mehmet Büyükdemircioğlu[1], Salim Malek[1], Elisa Mariarosaria Farella[1], Sultan Kocaman[2], Martin Kada[3], Fabio Remondino[1]

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

Keywords: Building Footprint, Deep Learning, Line Segment Detection, Photogrammetry, GeoAI

Abstract
Accurate, topologically consistent building footprints are essential for building reconstruction and GIS applications. But high-resolution orthophotos often contain occlusions (trees, cast shadows, etc.) or dense roof structures that challenge pixel-based segmentation and polygonization. In recent years, Line Segment Detection (LSD) networks have gained popularity as they can directly extract vectorized building footprints. This study benchmarks three line-segment detection (LSD) networks - L-CNN, ULSD, and F-Clip - against a strong semantic segmentation network - DeepLabV3+ - for building footprint extraction from very high resolution orthophotos across multiple regions with varied built-up morphology. Our evaluation on the considered urban areas revealed that LSD approaches generally deliver cleaner boundaries and more reliable roof topology than segmentation methods, whose high pixel scores mask boundary breaks. These findings indicate that when polygonal fidelity and downstream GIS usability are priorities, LSD pipelines could be superior for vectorized building footprint extraction compared to segmentation methods.

1. Introduction

Object detection and segmentation from satellite and airborne imagery have reached remarkable efficiency in the last years (Gui et al., 2024; Kocaman et al., 2022; Heipke and Rottensteiner, 2020). In particular, building footprints are fundamental for urban management, 3D city modelling, urban digital twins and GIS applications. Their production has traditionally relied on manual digitization of drone, airborne or satellite orthophotos, which is time-consuming, subjective, and labour-intensive. Various automated methods have been developed to address these challenges, ranging from traditional approaches (Ok, 2013; Nex et al., 2013), machine learning approaches (Thottolil and Kumar, 2022), semantic segmentation of imagery (Du et al., 2025; Yi et al., 2019) and more recently, deep learning-based frameworks that can directly extract vectorized footprints (Wei et al., 2023; Tejeswari et al., 2022; Chen et al., 2020). Recent studies demonstrate that deep learning methods provide state-of-the-art results, consistently outperforming traditional approaches in accuracy and scalability. However, most of the workflows for building footprint extraction with deep learning methods follow a multi-stage process, typically involving building segmentation followed by vectorization of raster outputs (Buyukdemircioglu et al., 2022). While effective in principle, this indirect strategy introduces additional processing steps, potential geometric inaccuracies and higher computational costs, making it less suitable and often unscalable for streamlined workflows. Moreover, converting segmented pixels into clean vector polygons is not an easy task: roof outlines often deviate from straight lines, and neighbouring buildings that share the same roof structure or are closely adjacent are frequently merged into a single region. These shortcomings boosted the development of methods for the direct extraction of polygonal building footprints as a more efficient and practical alternative. In this context, Line Segment Detection (LSD) networks have emerged as a promising direction. Unlike segmentation-based approaches, these networks are capable of directly extracting line segments and junction points of building footprints and roof structures from aerial imagery, offering the potential for fully automated polygon-level footprint delineation (Buyukdemircioglu, 2023).
This study benchmarks various LSD networks - namely F-Clip (Dai et al., 2022), ULSD (Li et al., 2021), L-CNN (Zhou et al., 2019) which are mainly developed for detecting line segments in indoor environments - and a segmentation method - DeepLabV3+ (Chen et al., 2018). We use both quantitative and qualitative evaluations, to extract building footprint vectors from very high-resolution aerial true orthophotos using three dedicated datasets.

2. Related Works

Common benchmark datasets for building footprint extraction include Inria Aerial Image Labeling dataset (Maggiori et al., 2017), CrowdAI Mapping Challenge dataset (Mohanty et al., 2020), WHU building dataset (Ji et al., 2018), BONAI dataset (Wang et al., 2022) or the P^3 dataset (Sulzer et al., 2025). These datasets do not match the aim of this study, i.e. work with sub-decimeter resolutions and diverse roof structure complexity, hence we collected aerial orthophotos (5 cm, 8 cm and 10 cm) with diverse building roof types, including manually digitized footprints as ground truth.
Recent approaches for vectorized building outlined follow two main lines. The first approach augments segmentation with shape priors before polygonization. As an example, SANET integrates a transformer-based boundary module with a Fourier shape–descriptor loss to regularize footprints, improving completeness and geometric fidelity (Hu et al., 2024). A complementary direction directly predicts polygons, avoiding raster-to-vector error accumulation. PolyR-CNN casts outline extraction as end-to-end vertex prediction from RoI features and reports competitive accuracy with higher efficiency than multi-stage designs (Jiao et al., 2024). RoIPoly extends this idea with RoI-query attention and a learnable logit embedding to curb redundant vertices, yielding strong results - especially on small buildings - without post-processing (Jiao et al., 2025). Graph-based models have also emerged; the spatial-cognitive shaping framework (SCShaping) regresses vector coordinates directly

with graph convolutions and reports gains on both mask-wise and edge-wise metrics (Du et al., 2024).

3. Study Areas

The study considers four different regions to capture variability in building typologies and image acquisition characteristics, providing a diverse basis for benchmarking the four considered methods. For the training/validation split, the total number of lines and 512 x 512 pixel sized image tiles for each study area are split by ~90% for training and ~10% for validation. The test dataset was held out entirely and remained unseen during the training phase. The details for each study area and test area are given in Table 1.

	Area A	Area B	Area C	Test Area
GSD	8 cm	10 cm	5 cm	5 cm
Area (km^2)	24.33	5.65	1.08	0.055
Training Lines	168,419	72,375	10,784	-
Test Lines	18,713	8,041	1,198	1,289
Training Image Tiles	4,145	883	1,358	-
Test Image Tiles	467	88	163	90
Built-up complexity	Medium	Hard	Medium	Very Hard

Table 1. An overview of the employed datasets.

Study area A is characterized by predominantly regular and standardized roof forms, such as hip and gable roof shapes. Study area B contains more irregular and complex roof geometries. Study Area C also contains spaced buildings with common roof types and some occluded roofs. Test area features medieval built-up areas with very cluttered buildings and irregular roof shapes. By combining these datasets, this study aims to create a heterogeneous dataset that enables training and evaluation of the models under diverse conditions, ensuring generalization across varying building geometries, density and geometric/radiometric image resolutions. This combination and diversity are crucial for developing models that are transferable beyond a single urban context and capable of handling both standardized and complex roof morphologies. For study areas A and B, building footprint vectors (ground truth), covering all structures larger than 10 m², were digitized through manual stereo digitization by professional photogrammetry operators, ensuring high geometric accuracy. This combination of high-resolution imagery and reliable vector data establishes a robust ground truth reference for training and validation of the results.

For Study Area C and the test area, orthophotos were created by combining nadir and oblique images. Building footprints were then digitized by experienced operators directly on these images. Cadastral data were not used during digitization; the operator traced the footprint borders based on what is visible in the orthophotos. The test area was intentionally chosen from an old district with dense, irregular buildings and frequent occlusions, where deciding the exact footprint is difficult even for a human. This setting provides a strict test of model performance in complex urban scenes.

4. Methodology

The focus of the work is assessing the potential of the LSD networks - F-Clip, ULSD, and L-CNN - which has not yet fully exploited for building footprint extraction from high resolution

Figure 1. Study area A, B, C and test area (from top to bottom).

aerial orthophotos. L-CNN is an end-to-end wireframe parser that detects junctions and line segments through a four-stage pipeline. F-Clip is a one-stage method that bypasses junction detection by directly regressing line parameters from feature maps, offering greater efficiency and backbone flexibility. ULSD is an unified framework capable of detecting both straight and curved segments across pinhole, fisheye, and spherical imagery using Bézier curve representations. For the training tasks of both approaches (LSD and semantic segmentation), we first tiled the true orthophotos into fixed-size 512x512-pixel patches, together with their corresponding building footprints. Importantly, instead of extracting isolated building patches, we divided the full orthophotos into tiles to preserve spatial context and ensure that roofs in dense urban areas remain represented with their neighbors. This strategy allows the network to learn boundaries in realistic scenarios where adjacent buildings are tightly packed or partially occluded.

We compared LSD with DeepLabV3+ segmentation network to see whether vectorization should start from lines or masks - two dominant, complementary methods for building footprint extraction. This comparison on the same test area clarifies what are pros and cons of each approach in an operational pipeline. While segmentation is widely used and robust, it produces raster masks that must be thresholded and polygonized, where LSD targets the primitives that needed—straight edges and junctions—producing vector-ready outlines with less post-processing and rectilinear building geometry.

All models are trained with all available datasets to expose the learning phase to a broader spectrum of roof types and imaging conditions. By combining regular roof forms (such as hip and gable) from study area A with the more complex geometries present in study areas B and C, we ensured diversity in both building typologies and Ground Sampling Distances (5 cm, 8 cm, and 10 cm). To convert the reference building footprints into training labels compatible with the evaluated neural networks, a set of in-house Python scripts has been developed to automatically transform GeoJSON geometries into pixel-based coordinates within each image tile and store them in JSON format. This pipeline handles polygon-to-line conversion and edge subdivision into segments according to each networks' input format.

To enable a fair comparison across methods, all networks were trained under an identical configuration. Non-overlapping RGB tiles of 512×512 pixels were used for both training and testing. Each tile was processed by the same stacked-hourglass backbone configuration (as in L-CNN, ULSD, and F-Clip with HG2-LB backbone), producing feature maps of size 128×128×256. Optimization settings were held constant: learning rate 4×10^{-4}, weight decay 1×10−4 at epoch 25, and a total of 50 training epochs with a batch size of 32. For visualization, detected line segments carry confidence scores in [0,100], consistent with common practice in line-segment detection. All figures depicting LSD outputs therefore display only segments with scores ≥80, ensuring comparable visual interpretations of the predictions.

5. Evaluation

5.1 Metrics

Traditional heat-map–based boundary measures are not well suited to line segment detection tasks since they neither penalize overlapping/duplicate segments nor evaluate whether segments connect to form a valid graph. To address this, Zhou et al. (2019) proposed Structured Average Precision (sAP), which treats each method's output as a ranked list of detected segments pooled over all test images. Detections are matched to ground truth within a pixel tolerance, a precision–recall curve is computed, and the area under this curve yields sAP at each tolerance (e.g., 5, 10, 15 px). The mean structural AP (msAP) averages these sAP values across tolerances to summarize segment quality.

Wireframe detection differs from plain line detection by explicitly representing junctions, which carry 3D meaning (corners/occlusion points) and encode line connectivity. Junction quality is evaluated with junction (mAP^J) on vectorized junctions, rather than heat maps. Given a ranked list of predicted junctions, each prediction is matched to the nearest unused ground truth junction within a distance threshold; unmatched or duplicate predictions count as false positives. Precision–recall is then computed, the junction AP is the area under this curve, and the mean junction AP is the average over multiple thresholds (0.5, 1.0, and 2.0 pixels at 128×128 resolution). 512x512 pixel images are downscaled by ×1/4 before matching (so τ at 128 equals 4τ at 512).

A detected segment $\hat{L}_j = \left(\hat{p}_j^1, \hat{p}_j^2\right)$ is a true positive if (Eq.1):

$$\min_{(u,v)\in E} \left(\left\|\hat{p}_j^1 - p_u\right\|_2^2 + \left\|\hat{p}_j^2 - p_v\right\|_2^2 \right) \le \tau, \qquad \text{(Eq.1)}$$

with one-to-one matching so that lower-ranked duplicates are counted as false positives. The precision-recall curve over the ranked detections gives $AP_{seg}(\tau)$. Tolerances are $\tau \in \{5,10,15\}$ (at 128 x 128 resolution).

Mean structural AP (msAP) (Eq.2):

$$\text{msAP} = \frac{1}{3}\left[\text{AP}_{\text{seg}}(5) + \text{AP}_{\text{seg}}(10) + \text{AP}_{\text{seg}}(15)\right] \qquad \text{(Eq.2)}$$

Junction AP and mAP^J :

A predicted junction $\hat{p}$ is a true positive if its Euclidean distance to the nearest unmatched ground truth junction is within a threshold δ; precision-recall over the ranked junctions yields $AP_{junc}(\delta)$. Thresholds are $\delta \in \{0.5,1.0,2.0\}$ (at 128 x 128). The mean junction AP is (Eq.3):

$$\text{mAPJ} = \frac{1}{3}\left[\text{AP}_{\text{junc}}(0.5) + \text{AP}_{\text{junc}}(1.0) + \text{AP}_{\text{junc}}(2.0)\right] \qquad \text{(Eq.3)}$$

Segmentation was assessed using standard measures—IoU, F1, precision, recall, and accuracy. Precision is the share of predicted building pixels that are correct; recall is the share of true building pixels recovered; F1 is their harmonic mean. IoU quantifies region overlap (intersection/union). Accuracy is the fraction of all pixels classified correctly, but because background dominates, it should be interpreted alongside IoU and F1.

5.2 Results on validation dataset

Figure 2 summarizes validation on three combined datasets dominated by regular, well-spaced roofs - hip and gable types with long, sunlit eaves, and limited mutual occlusion - so the scene places a premium on straight-edge continuity rather than gap-bridging under clutter. In this regime, F-Clip achieves the strongest segment accuracy, and the visuals corroborate why: edges along planar, texture-rich tiles are traced as long, clean strokes with few discontinuities, and repetitive roof patterns are followed reliably across adjacent buildings. L-CNN ranks third on segment measures and achieves the lowest junction quality; corners at ridge–eave and eave–gable intersections are localized tightly, producing polygonal outlines that are geometrically tidy even when a few wall edges are slightly offset. ULSD resulted as follows: output remains largely rectilinear but displays occasional over-linking across low-contrast areas (e.g., near homogeneous roofs) and short, zig-zag breaks around small roof

Figure 2. Performance comparison on the validation dataset of Area A.

details or mild shadows, which reduce segment matches without severely disturbing the overall shape.

The validation set combines three sources but is dominated by Study Area A, where buildings are well spaced and roofs are mostly regular roof types. This scene includes cast shadows along façades—edges that line-segment detectors can easily mistake for roof lines.

In our tests, however, all three models handled shadows well and did not trace shadow borders as structure. They also captured most fine details of the building footprints, with only a few small elements missed. Overall, the data favor methods that track long straight edges and clean corners, and all networks behaved robustly under these conditions. Across methods, performance improves monotonically with increasing tolerance, indicating that most residual errors are small localization shifts rather than missed detections.

	sAP^5	sAP^{10}	sAP^{15}	msAP	mAP^J
L-CNN	44.9	47.6	49.0	47.2	48.0
ULSD	46.3	51.2	53.7	50.4	57.6
F-Clip	**51.7**	**57.8**	**60.5**	**56.6**	-

Table 2. Quantitative results for validation dataset.

Overall, the ranking aligns with the dataset's characteristics: methods that lean on local appearance cues and straight-line continuity (F-Clip) benefit most from clean, regular rooftops, while explicit junction modeling (L-CNN) preserves topological fidelity; ULSD is comparatively less favored in this easier, less cluttered setting. Junction mean average precision metric (mAP^J) is not calculated for F-Clip because junction predictions are not available in the released configuration Exact scores are reported in Table 2.

5.3 Results on test dataset

The test area has dense, irregular roofscapes where manual delineation of building footprints is challenging. Figure 3 summarizes qualitative observations from the model outputs, while Table 3 reports the quantitative metrics computed against these manually curated references. L-CNN achieves the strongest average segment precision across tolerances (highest msAP in Figure 3) and produces the most spatially coherent roof outlines.

	sAP^5	sAP^{10}	sAP^{15}	msAP	mAP^J
L-CNN	**25.9**	**29.0**	**29.8**	**28.2**	35.1
ULSD	24.5	27.4	29.1	27.0	**36.1**
F-Clip	18.8	24.8	27.6	23.7	-

Table 3. Quantitative results for test dataset.

Error patterns are dominated by small localization offsets rather than gross false positives, which suggests that the detector generally places segments correctly but with minor endpoint or alignment deviations relative to the manually delineated edges. ULSD achieves the best junction-aware performance (highest mAPJ in Table 3), indicating accurate corner localization on rectilinear roofs. At the same time, the qualitative assessment in Figure 3 shows characteristic failure modes: occasional mis-reconstructed, redundant long spans (over-linking across low-contrast regions) and gaps where roof boundaries fall within shadowed or textureless areas, leading to incomplete and partly

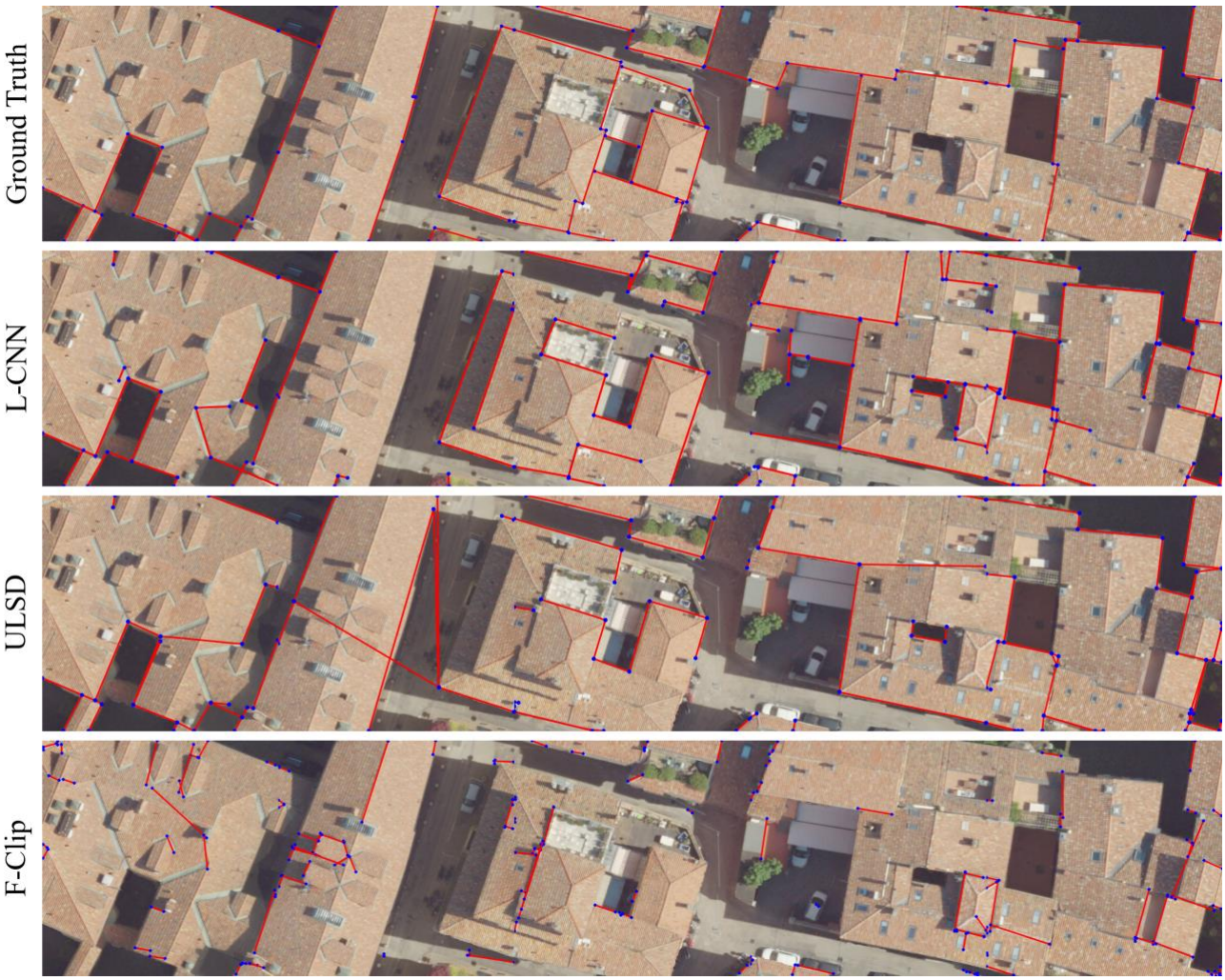

Figure 3. Performance comparison on test dataset.

fragmented contours relative to the manual ground truth. F-Clip demonstrates a pronounced generalization gap: despite yielding the strongest outcomes on the validation set with more isolated and regularly structured buildings, it performs weakest on the test area. Figure 3 shows outputs that are frequently fragmented into short segments, failing to trace long eaves and continuous façade edges. Again, mAP^J is not calculated for F-Clip because junction predictions are not available in the released configuration, limiting direct comparison on topology fidelity. F-Clip doesn't predict junctions, so clarifying that segment metrics are still comparable, and that topology comparison is limited.

Overall, the results indicate that architectures with explicit junction modeling (L-CNN, ULSD) better preserve the rectilinear topology needed for downstream GIS uses (e.g., polygonization, snapping, topology checks) under heavy clutter and shadowing. The contrast between validation and test outcomes underscores the sensitivity of line-segment detectors to urban morphology: methods that excel on separated, regular buildings may degrade on dense, historic buildings where occlusions and complex roof adjacency drive both quantitative errors and the qualitative artifacts visible in Figure 3.

5.4 Robustness against occluded roofs

Occlusion is a major problem for extracting building footprints and roof structures from high resolution imagery. Trees, nearby buildings, rooftop objects, and deep shadows hide parts of the roof and make edges hard to see, so outlines break or go missing even on simple roofs. These effects are common in older neighbourhoods with narrow streets and heavy tree cover, where deciding the true footprint is difficult even for a human. To examine this challenge, validation tiles in which roofs are partly covered by trees or shadow are evaluated and compared to determine how each method performs under these conditions.

Figure 4 examines robustness to vegetation occlusion by including validation tiles where tree crowns partially cover rooftops.

Visual inspection indicates that ULSD is the most robust in these conditions: roof-edge candidates stay under tree cover and line up into clear outlines once gaps are bridged, giving the closest match to the manually delineated footprints. L-CNN, by contrast, often emphasizes edges along the tree canopy itself; when crowns intersect roof planes, false segments appear inside or along the vegetation boundary, leading to over-segmentation and misplaced junctions. F-Clip struggles most under occlusion, often failing to activate on the hidden roof edges and producing few, if any, usable line segments on heavily covered buildings.

Comparison of the 3 networks can be summarized as follows:

- L-CNN: it detects many segments on sunlit roofs but often traces along the tree canopy when crowns touch the roof, creating false edges inside/along vegetation. It tends to over-segment near foliage; junctions around occlusion boundaries are frequently misplaced. It maintains reasonable outlines where occlusion is light, but gaps appear across heavier cover

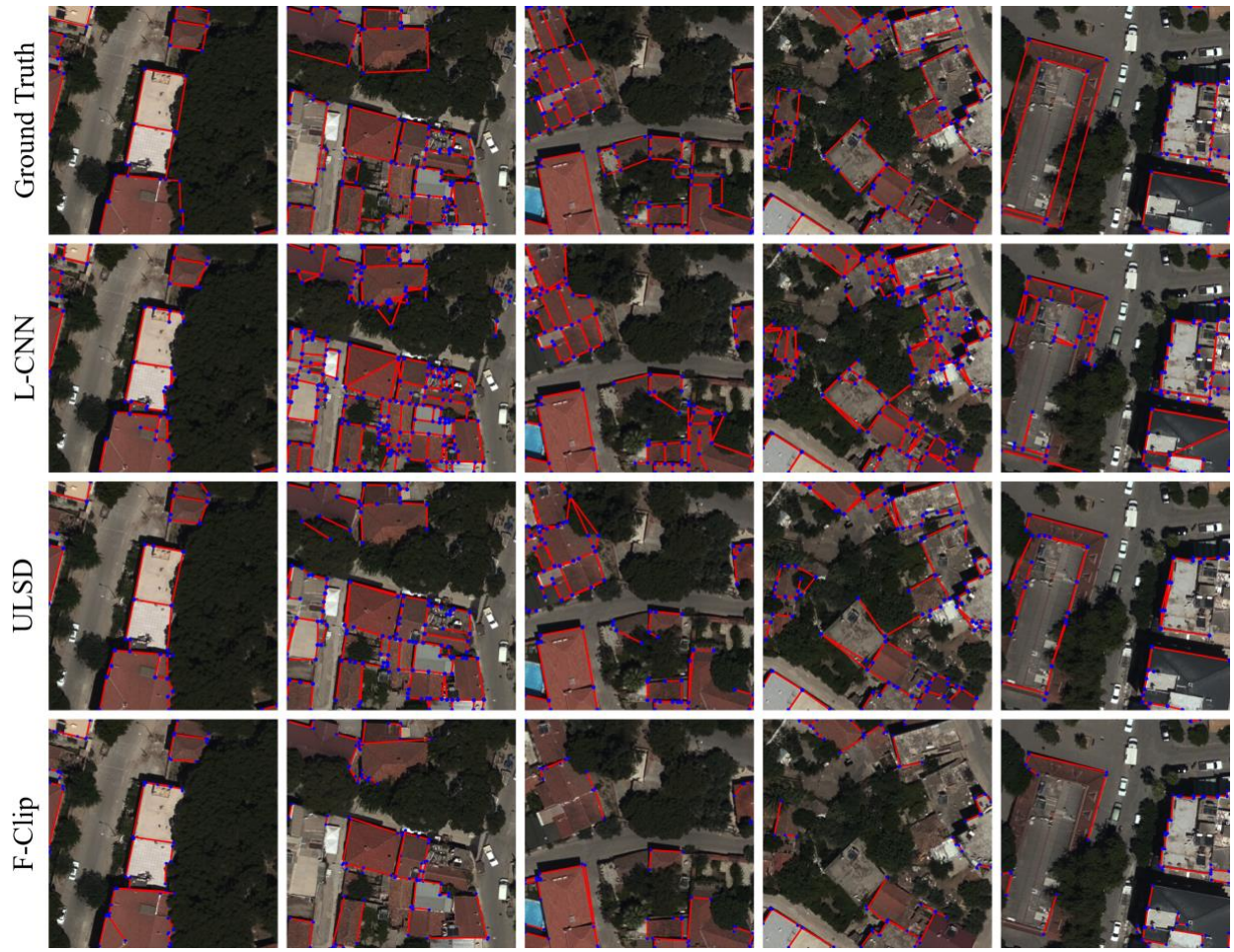

Figure 4. Robustness comparison of selected networks on validation dataset of Area B.

Figure 5. Segmentation results with DeepLabV3+ on test dataset.

and deep shadow.

- ULSD: It is the most robust under tree cover: roof outlines continue across partially hidden areas and align well once small gaps are bridged. It produces fewer vegetation-induced false positives and edges remain largely rectilinear and consistent with ground truth. Residual issues are limited to short breaks in the deepest shadows and occasional minor over-linking.
- F-Clip: it is the least robust, frequently fails to activate on occluded roof edges, yielding sparse or missing segments on heavily occluded building roofs. Detections are short and fragmented; polygons remain incomplete even when some edges are visible. It relies strongly on local appearance, performing mainly on strong, sunlit edges while ignoring weak or interrupted boundaries.

These outcomes are consistent with architectural biases: ULSD's junction-centric representation favors the assembly of longer, rectilinear structures from sparse evidence, whereas L-CNN's stronger response to textural gradients near foliage increases false positives, and F-Clip's reliance on local cues hampers recovery when roof edges are weak or discontinuous.

5.5 Segmentation results

Using DeepLabV3+ for building segmentation on the test area produces strong aggregate pixel metrics—precision 98.41%, recall 92.57%, F1 95.40%, overall accuracy 93.97%, and IoU 91.20%. Inspection of Figure 5 highlights limitations that are structural rather than purely pixelwise. Shadows and narrow roof openings lower recall because they hide roof edges, so the model misses parts of the boundary and produces broken outlines even when the interior of the building is mostly correct. Segmentation networks are tuned to maximize overlap between predicted and true regions (IoU), not to trace boundaries precisely. As a result, small boundary shifts that barely affect overlap can become major topological errors once a clean polygon is needed.

The raster-to-vector conversion process makes these problems worse. Converting masks to polygons typically involves finding connected regions, tracing their outer and inner boundaries, adding extra vertices, and then simplifying the shapes. Each step can cause a loss of details or distort geometry: thin roofs and narrow courtyards can disappear; stair-step pixels along slanted edges get over-simplified into biased straight lines; small bays or dormers are dropped; and gaps from shadows leave open outlines that must be "snapped" together, creating slivers and false holes. Using a tighter simplification threshold preserves more detail, but it also introduces more noise and many more vertices, making it harder to keep polygons simple.

In contrast, line-segment detection methods focus directly on the building boundaries themselves. They first pinpoint likely corners and straight edges, then keep only the strongest edge candidates in the right directions and finally connect these pieces into complete roof outlines using simple geometric rules about angles and lengths. This produces closed, ready-to-use footprints without relying on a separate, error-prone conversion from pixels to vectors.

On the test dataset, these methods generate cleaner edges with less noise, preserve straight, rectilinear structure even in shadows—where a few confident edge cues can bridge low-texture areas—and can be trained end-to-end to output polygons that keep correct connectivity and interior courtyards. Although pure segmentation can be improved with shadow-aware data augmentation, reweighted losses, extra boundary heads, model ensembling, and smarter vectorization (for example, careful snapping and validity checks), a basic mismatch remains: segmentation is tuned to maximize region overlap, while footprint extraction requires very accurate, topologically consistent boundaries. Therefore, when polygon accuracy and downstream GIS usability are the priority in dense, shadow-rich historic areas, line-segment detection pipelines provide a better and a more reliable solution than semantic segmentation approach.

6. Conclusions

This study compares three line-segment detection (LSD) networks—L-CNN, ULSD, and F-Clip—against a semantic segmentation baseline (DeepLabV3+) for vectorized building-footprint extraction from very-high-resolution orthophotos. Under a unified training protocol, LSD methods produced cleaner boundaries and more reliable polygon topology, particularly where segmentation suffered from shadow-induced gaps and raster-to-vector artifacts. Among LSD models, L-CNN proved most effective on challenging, complex urban fabric (dense adjacency, deep shadows), ULSD was the most robust under occlusion (e.g., roofs partially covered by trees), and F-Clip delivered the strongest results on scenes with well-separated buildings and regular roof types (hip, gable, etc.). Overall, when polygonal fidelity and downstream GIS usability are paramount, LSD pipelines constitute the more geometrically faithful approach.

Line-segment detection networks in the literature remain few, and most were developed for generic wireframe parsing in indoor scenes rather than for extracting building footprints or roof structures. As a result, these models must be retrained from scratch on roof-edge datasets to address outdoor radiometry, scale, and occlusion patterns. Key directions include strengthening generalization through targeted data augmentation and enlarging the training dataset (especially with shadowed and tree-occluded roofs), hyper-parameter tuning of each network beyond default configurations, and moving from a global, fixed visualization threshold (80) to an adaptive scoring method for each tile. In addition, incorporating height cues as a fourth band (e.g., nDSM/DSM) is expected to improve roof–tree separation and boundary continuity, thereby enhancing both segment accuracy and junction fidelity.

Another direction is to develop a model that performs well for multi-resolution building footprint extraction. Most current methods assume a single image resolution. We plan to generate both datasets and methods for multi-resolution training by automatically generalizing high-resolution building footprints to lower resolution ones. The goal is to develop a single model that performs well across different types and resolution of imagery.

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

Ok, A.O., 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.*, 86, 21–40.

Buyukdemircioglu, M., Can, R., Kocaman, S., and Kada, M., 2022. Deep Learning Based Building Footprint Extraction from Very High Resolution True Orthophotos and nDSM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.,* V-2-2022, 211–218.

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

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *In Proceedings of the European conference on computer vision (ECCV)* (pp. 801-818).

Chen, Q., Wang, L., Waslander, S.L., Liu, X., 2020. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS J. Photogramm. Remote Sens.*, 170, 114–126.

Dai, X., Gong, H., Wu, S., Yuan, X., Yi, M., 2022. Fully convolutional line parsing. *Neurocomputing* 506, 1–11.

Du, J., Li, B., Yang, J., 2025. Boundary-aware graph convolutional network for building roof detection from high-resolution remote sensed imagery. *PFG – J. Photogramm. Remote Sens. Geoinf. Sci.,* 1–13.

Du, Z., Sui, H., Zhou, Q., Zhou, M., Shi, W., Wang, J., Liu, J., 2024. Vectorized building extraction from high-resolution remote sensing images using spatial cognitive graph convolution model. *ISPRS J. Photogramm. Remote Sens.*, 213, 53–71.

Gui, S., Song, S., Qin, R., Tang, Y., 2024. Remote Sensing Object Detection in the Deep Learning Era - A Review. *Remote Sensing*, 16, 327.

Heipke, C., Rottensteiner, F., 2020. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spatial Information Science*, Vol. 23(1), pp. 10-19.

Hu, A., Wu, L., Xu, Y., Xie, Z., 2024. SANET: A shape-aware building footprints extraction method in remote sensing images by integrating Fourier shape descriptors. *IEEE Trans. Geosci. Remote Sens.*, 62, 5632215.

Kocaman, S., Akca, D., Poli, D., Remondino, F., 2022. *3D/4D City Modelling - From Sensors to Applications*. Whittles Publishing, 224 pages, ISBN: 978-184995-475-4.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.*, 57(1), 574–586.

Jiao, W., Cheng, H., Vosselman, G., Persello, C., 2025. RoIPoly: vectorized building outline extraction using vertex and logit embeddings. *ISPRS J. Photogramm. Remote Sens.*, 224, 317–328.

Jiao, W., Persello, C., Vosselman, G., 2024. PolyR-CNN: R-CNN for end-to-end polygonal building outline extraction. ISPRS J. Photogramm. Remote Sens., 218, 33–43.

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.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark. *IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 3226–3229.

Nex, F., Rupnik, E., Remondino, F., 2013. Building footprints extraction from oblique imagery. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-3/W3, 61–66.

Mohanty, S.P., Czakon, J., Kaczmarek, K.A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S., Göpfert, J.P., Tandon, A., Mollard, G., Rayaprolu, N., Salathe, M., Schilling, M., 2020. Deep learning for understanding satellite imagery: an experimental survey. *Front. Artif. Intell.*, 3, 534696.

Sulzer, R., Duan, L., Girard, N., Lafarge, F., 2025. The P^3 dataset: Pixels, Points and Polygons for Multimodal Building Vectorization. arXiv preprint arXiv:2505.15379.

Tejeswari, B., Sharma, S. K., Kumar, M., and Gupta, K., 2022. Building footprint extraction from space-borne imagery using deep neural networks. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2022, 641–647.

Thottolil, R., Kumar, U., 2022. Automatic building footprint extraction using random forest algorithm from high-resolution Google Earth images: a feature-based approach. *IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul 2022, 1–6.

Wang, J., Meng, L., Li, W., Yang, W., Yu, L., Xia, G.S., 2022. Learning to extract building footprints from off-nadir aerial images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1), 1294–1301.

Wei, S., Zhang, T., Ji, S., Luo, M., Gong, J., 2023. BuildMapper: A fully learnable framework for vectorized building contour extraction. ISPRS J. Photogramm. Remote Sens., 197, 87–104.

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., Zhao, T., 2019. Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network. *Remote Sensing*, 11(15):1774

Zhou, Y., Qi, H., Ma, Y., 2019. End-to-end wireframe parsing. *Proc. ICCV.,* pp. 962-971.