PCFootprint: A Large-Scale Dataset and Benchmark for Vectorized Building Footprint Extraction From Aerial LiDAR Point Clouds

Haoyuan Shen, Kuihao Wang, Ruisheng Wang, *Senior Member, IEEE*, and Yujun Liu, *Graduate Student Member, IEEE*

***Abstract*—Building footprint extraction is a fundamental task in photogrammetry, remote sensing, and computer vision. Recent image-based methods have achieved remarkable progress in extracting vectorized footprints from high-resolution optical imagery. However, optical imagery is inherently susceptible to occlusions, perspective distortions, and residual relief displacement, yielding incomplete or misaligned extractions. Furthermore, the lack of explicit elevation information limits its direct applicability to level of detail building modeling. To overcome these fundamental limitations, we present PCFootprint, the first large-scale public dataset dedicated to building footprint extraction from real aerial laser scanning point clouds. Significantly surpassing existing LiDAR-based datasets in scale, PCFootprint comprises 33 000 tiles featuring systematically aligned vectorized annotations across diverse urban and rural landscapes. Furthermore, in order to push the robustness of future algorithms, we construct a unique 3000-tile cross-domain benchmark to evaluate model generalization across unseen geographic regions. Comprehensive evaluations of mainstream methods reveal critical unresolved challenges in the 3-D domain, including high intraclass variance, severe data imbalance, and noise in complex geospatial environments. As a comprehensive public benchmark, we believe PCFootprint will facilitate future research in 3-D building modeling, urban scene understanding, and geospatial analysis.**

***Index Terms*—Aerial laser scanning, benchmark datasets, building footprint extraction, point clouds, remote sensing.**

I. Introduction

BUILDING footprint extraction serves as a core information creation and applied remote sensing process, standing at the intersection of photogrammetry, remote sensing, and computer vision to empower digital twin and smart city infrastructures [1], [2], [3], [4]. Building footprints are typically represented in two formats: pixelwise rasterized masks or vectorized polygons defined by ordered vertices. However, buildings exhibit diverse architectural styles, complex structures, and varied geometric configurations across different urban scenes. Therefore, automated and accurate extraction of building footprints remains a challenging and important research problem.

Recent advances in image-based datasets and methods have achieved remarkable progress in vectorized footprint extraction [5], [6], [7], [8], [9], [10], [11], [12]. Large-scale datasets, such as WHU Building [13], Inria [14], and CrowdAI [15] provide high-resolution optical imagery for extracting building geometries. These datasets have enabled methods to produce precise footprint boundaries under optimal imaging conditions. Despite these advances, image-based extraction faces inherent limitations that constrain practical applications. First, optical imagery is susceptible to spectral interference from shadows and occlusions from overhanging vegetation [16], [17], [18]. These factors often obscure building edges and result in incomplete or fragmented extractions. Second, extracted footprints lack elevation information, which limits their direct use in level of detail 1 (LoD1) modeling [19], [20], [21].

Airborne laser scanning (ALS) has emerged as a primary source for large-scale urban 3-D data acquisition [22]. By providing point clouds with precise 3-D coordinates, absolute scale, and penetration capability [23], [24], ALS enables reliable geometric representation of complex urban and rural environments and has been widely adopted for 3-D perception, understanding, and reconstruction tasks [25], [26]. However, a public benchmark dataset for extracting vectorized building footprints from ALS data is still lacking, which significantly hinders the development and systematic evaluation of point cloud-based footprint extraction methods. Existing large-scale 3-D datasets [27], [28], primarily focus on pointwise semantic or instance labeling, and thus fail to provide structured, vectorized footprint representations required by geographic information systems (GIS) applications, including spatial analysis and urban planning. Moreover, most existing footprint extraction studies in point cloud domain rely on private datasets, limiting reproducibility, fair performance comparison, and generalization across diverse urban scenes [29], [30], [31], [32], [33], [34], [35]. Therefore, a standardized and large-scale ALS dataset dedicated to building footprint extraction is imperative. Such a benchmark would facilitate accurate boundary delineation in cluttered urban landscapes and occluded dense forested areas. As demonstrated in Fig. 1(d), the inherent elevation information

Received 13 March 2026; revised 27 May 2026; accepted 17 July 2026. Date of publication 22 July 2026; date of current version 24 August 2026. This work was supported in part by the Key Technological Innovation Program of Ningbo City under Grant 2024Z297. *(Haoyuan Shen and Kuihao Wang contributed equally to this work.) (Corresponding authors: Ruisheng Wang; Yujun Liu.)*

The authors are with the School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China (e-mail: haoyuan.shen2024@mails.szu.edu.cn; wangkuihao2025@mails.szu.edu.cn; ruiswang@szu.edu.cn; yujunliu2024@mails.szu.edu.cn).

Dataset is publicly available online at https://huggingface.co/datasets/Haoyuan-Shen/PCFootprint.

Digital Object Identifier 10.1109/JSTARS.2026.3715686

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

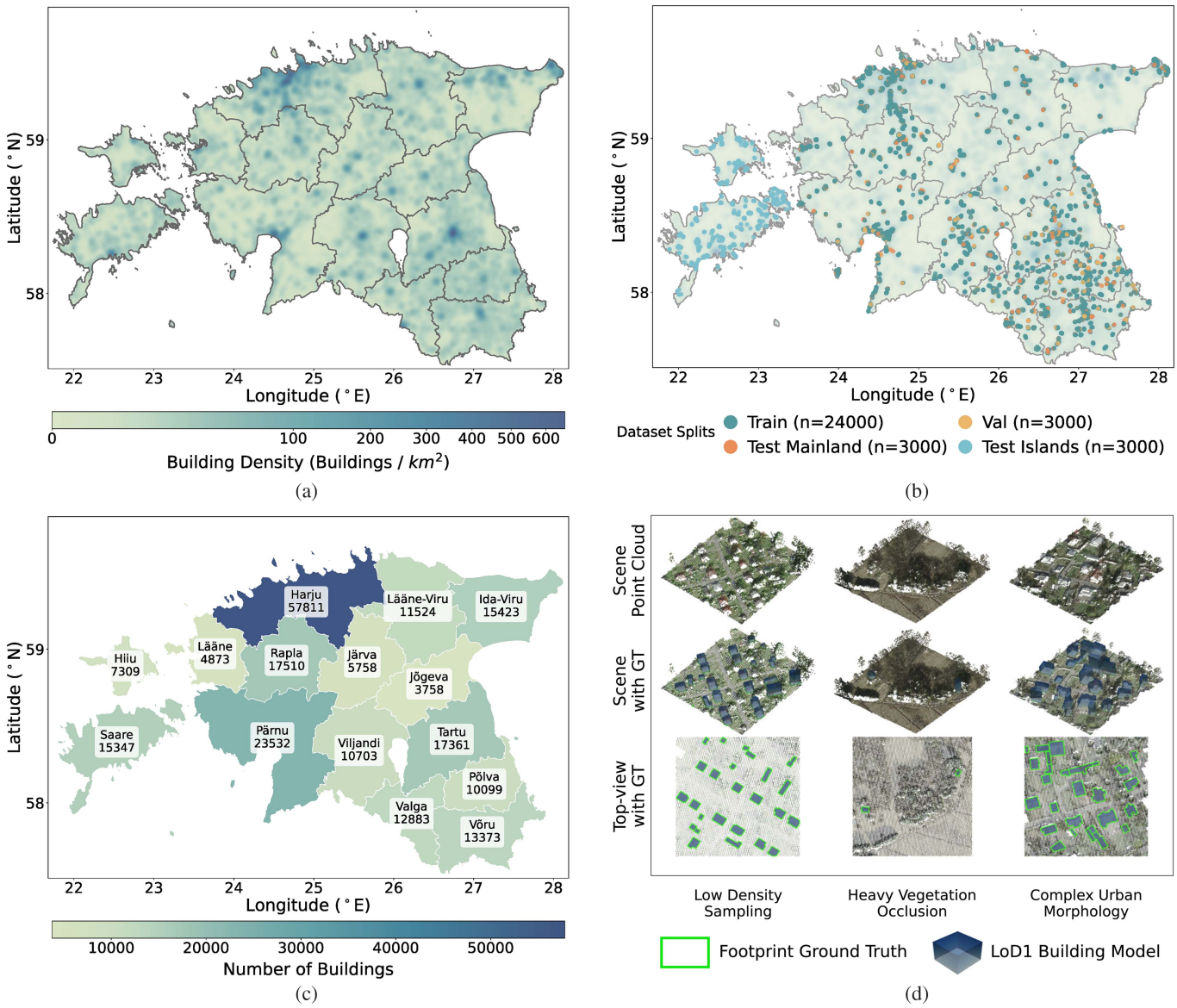


Fig. 1. Geographic sampling strategy and illustrative visualization of the PCFootprint dataset. (a) Nationwide building density heatmap of Estonia, serving as the geographic prior for data acquisition. (b) Spatial distribution of sampled point cloud tiles, which is strategically designed to be highly correlated with the actual building density shown in (a). (c) Cumulative building instances aggregated by count, demonstrating the dataset's massive scale, and validating that our sampling strategy accurately reflects real-world geographic patterns across Estonia. (d) Illustrative visualization of structured data representations, establishing the strong spatial correspondence between raw ALS point clouds (top), geographically aligned 3-D LoD1 models (middle), and structured 2-D vectorized footprints (bottom).

within the point clouds was exploited for the seamless LoD1 reconstruction, and further serve as reliable boundary constraints to improve LoD2 building modeling accuracy [23], [34], [36], [37], [38].

In this article, we present PCFootprint, the first large-scale public dataset for vectorized building footprint extraction from ALS point clouds. These data are derived from the Estonian Land and Spatial Development Board. It encompasses diverse urban and rural landscapes across multiple regions. The geographical distribution of these acquired point cloud tiles across Estonia is visualized in Fig. 1(b). We initially selected 500 large point cloud tiles characterized by significant height variations and structural complexity. These tiles exhibit varying point densities across different geographic regions. The 500 tiles were subdivided into 30 000 smaller point cloud tiles for processing. We provide an additional cross-domain generalization test set comprising 3000 point cloud tiles. Building instances fractured by tile boundaries were identified and removed from this test set. Ultimately, 33 000 point cloud tiles were obtained for the complete dataset. Each tile covers a standard spatial extent of 128 m $\times$ 128 m. We further establish a comprehensive benchmark for building footprint extraction from ALS point clouds and evaluate a wide range of mainstream methods. The results reveal substantial challenges arising from intraclass variation, data imbalance, and noise across complex geospatial environments. We believe PCFootprint will support future research in building modeling,

3-D reconstruction, urban scene understanding, and geospatial analysis.

The primary contributions of this work are summarized as follows.

1) We present PCFootprint, the first large-scale dataset dedicated to the vectorized extraction of building footprints from ALS point clouds, which facilitates the automated city-scale LoD1 building reconstruction.
2) The dataset features unprecedented scale with nationwide coverage across Estonia, comprising 33 000 tiles with systematically aligned vectorized annotations for 227 264 building instances to ensure national representativeness.
3) We establish a standardized benchmarking platform featuring robust intradomain evaluation and cross-domain generalization protocols, providing a verified baseline through the assessment of mainstream algorithms.
4) Our open-source release of the PCFootprint benchmark bridges a critical gap in high-quality public resources, fostering reproducible research and community-wide advancement within the geospatial domain.

II. Related Works

A. Building Footprint Extraction Datasets

Building footprint extraction benchmarks are categorized by input modality and structural annotation format. Early research framed building footprint extraction as a pixelwise semantic segmentation task. This paradigm is exemplified by datasets, such as Massachusetts [39], ISPRS [40], and Inria [14], which represent ground truth in the form of rasterized binary masks. However, these rasterized representations frequently suffer from aliasing artifacts and irregular boundaries [6], hindering their seamless integration with GIS workflows. Transforming these raster masks into regularized polygons necessitates computationally intensive postprocessing. To address these geometric limitations, a subsequent generation of datasets has pivoted toward native vectorized polygon annotations. This transition enables the development of end-to-end building vectorization and contour-based extraction methods that directly regress building corner vertices. For instance, the SpaceNet challenge series [41] released a corpus of high-resolution satellite imagery featuring over 685 000 annotated polygonal building footprints. Similarly, the WHU Building [13] and the WHU-Mix (Vector) [6] established robust benchmarks for footprint extraction by providing high-fidelity vector labels for large-scale aerial and satellite imagery. Despite their structural advantages, these image-centric frameworks remain fundamentally susceptible to spectral interference, relief displacement, and physical obstructions, such as overhanging vegetation. Such optical artifacts frequently lead to fragmented extractions and inaccurate boundary delineations, particularly in cluttered urban environments [42].

Geometric fidelity and occlusion handling represent the inherent advantages of 3-D point clouds over 2-D imagery. Optical sensors often suffer from vegetation obstructions and shadow artifacts that blur semantic boundaries [16], [43]. In contrast, 3-D LiDAR systems feature multireturn and side-view acquisition. These capabilities enable the direct capture of underlying structures by effectively penetrating upper-level canopies [44]. Moreover, unlike standard orthophotos, which are processed derivatives susceptible to residual relief displacement and perspective distortions, 3-D sensing preserves absolute spatial coordinates. Although modern deep learning attempts to generate True Digital Orthophoto Maps (TDOMs) without explicit digital surface model (DSM) priors [45], [46], [47], these methods remain reliant on implicit radiometric consistency to rectify displacements. Consequently, direct 3-D sensing remains indispensable for achieving the structural integrity required in GIS and digital twin frameworks.

Despite their high quality, prominent 3-D urban datasets like DALES [27], WHU-Urban3D [28], and Toronto-3D [48] are predominantly optimized for pointwise semantic labeling or instance segmentation. While these benchmarks enable the classification of discrete points into categories, such as "building" or "vegetation," they lack the structured 2-D polygon vectors necessary for direct geometric modeling. Consequently, research focusing on vectorized footprint extraction from LiDAR point clouds often resorts to localized, private, or small-scale collections, such as the Aitkenvale or Hervey Bay collections [29], which offer limited building instances. Other datasets rely on manually refined national inventories or datasets restricted to specific regions, such as Fredericton [42], Depok [30], or Hyderabad [31]. The lack of a standardized, large-scale public benchmark for building footprint extraction from ALS persists as a critical bottleneck, hindering both the objective assessment of geometric fidelity and the development of domain-invariant algorithms for complex 3-D environments.

B. Building Footprint Extraction Methodologies

Current methods for building footprint extraction, largely driven by the availability of optical benchmarks, typically follow two primary paradigms: pixelwise segmentation and polygonal regression. Pixelwise segmentation approaches leverage the strong generalizability of vision foundation models like SAM [49]. Within this paradigm, MSA-SAM [50] addresses object scale variations, while SegEarth-OV3 [51] introduces open-vocabulary capabilities. Other advancements have integrated dynamic dictionary learning [52], style mixing for domain-invariant feature learning [53], diffusion-based generative modeling [54], and frequency-guided structural encoding [55] to refine extraction quality. However, the inherently discrete nature of these pixelwise masks often yields irregular boundaries and lacks the sharp geometric topology required for high-precision mapping. To mitigate raster artifacts such as staircase effect, polygonal regression methods bypass intermediate rasterization by directly predicting vertex sequences or geometric primitives. Architectures, such as P2PFormer [8], Pix2Poly [11], and HoliTracer [12], leverage global Transformer-based sequence prediction or holistic tracing to capture topological relationships. In contrast, methods driven by target detection, such as PolyR-CNN [56] and RoIPoly [10], focus on efficient local vertex regression within regions of interest, while other works achieve GIS-compatible outputs through contour evolution [6] or bottom-up line-primitive reconstruction [57]. Furthermore,

TABLE I
DETAILED SPECIFICATIONS OF THE PROPOSED PCFOOTPRINT DATASET

Property	Details
Subsets	Mainland, Islands
Data Type	ALS Point Cloud
Annotation Format	Vectorized Polygon
Accessibility	Open-source
Point Spacing [m]	0.22–1.09 m
Tile Size	128× 128 m
Total Tiles	33,000
Building Instances	227,264
Coverage [km^2]	540.67

SAM's segmentation power is integrated into polygonal regression paradigm via SAMPolyBuild [7]. Nevertheless, these purely 2-D-driven approaches lack intrinsic 3-D geometric awareness, making them susceptible to spatial misalignments caused by varying terrain elevations. To bridge the gap between 2-D imagery and 3-D space, multimodal frameworks, such as MFNet [58], incorporate DSM elevation data into unified fine-tuning architectures to compensate for optical deficiencies. Despite these sophisticated advancements, image-based workflows remain fundamentally constrained by perspective distortions and the geometric information loss incurred during the projection process.

To overcome these limitations, methodologies operating directly within the 3-D domain hold immense promise for ensuring absolute geometric fidelity and resolving complex occlusions. By exploiting raw spatial coordinates, 3-D-centric approaches can theoretically achieve superior robustness against the shadows and perspective distortions that typically compromise 2-D orthophotos. Research in this field, however, is currently dominated by traditional geometric heuristics, such as LasBuildSeg [59]. Alongside other classical algorithms [29], [31], [33], [34], [60], these methods rely on plane fitting and morphological operations that frequently necessitate manual parameter tuning. Recent deep learning attempts have addressed these issues by introducing terrain-aware self supervised learning to learn structural features from LiDAR data [61]. While this approach enhances structural representation learning, it primarily generates pixel-level segmentation masks that lack direct vectorized structure and require extensive postprocessing. Similarly, GAN-based frameworks have been proposed for parameter free footprint extraction from gridded images to eliminate manual tuning [32]. However, these methods still rely on intermediate raster representations rather than achieving end-to-end 3-D vectorization. Such direct extraction of vectorized building footprints from point clouds is primarily stifled by a lack of large scale benchmarks with high-quality annotations. To bridge this gap, this article introduces PCFootprint, the first extensive dataset for building footprint extraction from ALS designed to support the automated generation of structured, high-fidelity building footprints directly from 3-D space.

III. PCFOOTPRINT DATASET

PCFootprint introduces the first extensive, open-source benchmark specifically engineered for the vectorized extraction of building footprints from ALS point clouds. Sourced from the Estonian Land and Spatial Development Board, the dataset spans a vast geographic extent of 540.67 km^2, comprising 33 000 standardized tiles and 227 264 systematically aligned building instances across diverse urban and rural landscapes. The detailed specifications of the dataset are summarized in Table I.

A. Dataset Construction

The sampling strategy of PCFootprint is meticulously curated to reflect the geospatial diversity of Estonian urban and rural landscapes, fundamentally driven by the actual geospatial distribution of buildings to ensure national representativeness.

To make the procedure rigorous and reproducible, we implemented the following county-based proportional quota sampling strategy.

Target and quota allocation: We have determined the target dataset size to consist of 30 000 mainland crops and 3000 island crops, each with a size of 128 m × 128 m. Since each 1 km^2 tile contains 64 crops, the minimum number of tiles required is 469. We increased this initial tile quota to ensure an adequate amount of valid data. Subsequently, we allocated the total tile quota among the counties in proportion to their total number of buildings.

Tile ranking and selection: Within each county, we ranked the 1 km^2 tiles based on the area covered by buildings, followed by the number of LiDAR points. We selected the top-ranked tiles to focus on areas with dense building distributions and high-quality point clouds. Tiles without any buildings were eliminated, and the next highest ranked tiles were used to meet the county's tile quota.

Cropping and compensation: We divided the selected 1 km^2 tiles into 128_m× 128 m crops and removed any crops that lacked valid building annotations. If the number of valid crops did not meet the required sample quota for a county, we obtained additional 1 km^2 tiles from that county's ranked list to make up the shortfall.

Island dataset processing: We applied the same process to the two island counties (Saare and Hiiu) to create the independent 3000-sample test set.

As demonstrated by the correlation between the nationwide building density heatmap [see Fig. 1(a)] and the spatial allocation of tiles [see Fig. 1(b)], the distribution of these 33 000 tiles directly aligns with building concentration. This ensures that the dataset captures a representative cross-section topography and a diverse range of architectural styles. This statistical alignment is further supported by a closed validation of the cumulative instance counts in Fig. 1(c), confirming that the dataset reflects the real-world density patterns of Estonian counties.

To illustrate the diverse data characteristics captured through this geographically informed sampling, Fig. 1(d) presents an illustrative visualization of representative scene alongside their high fidelity structural annotations. These examples showcase

a broad spectrum of data diversity within PCFootprint, ranging from fluctuating point densities and heavy vegetation occlusion to complex scene compositions with highly intricate building morphologies. By incorporating such geospatial heterogeneity, PCFootprint facilitates the rigorous evaluation of model robustness against varying land use patterns and enables researchers to better address the inherent challenges of cross-regional generalization. To ensure compatibility with mainstream deep learning architectures, all annotations adhere to the Microsoft Common Objects in Context (MS-COCO) [62] format. Each footprint is represented as a vectorized polygon defined by an ordered sequence of corner vertices, $V = \{v_1, v_2, \ldots, v_n\}$, where each vertex maps to a precise spatial coordinate within the ALS point cloud. As shown in Fig. 1(d), this establishes a rigorous structural correspondence between the raw 3-D input and the final structured 2-D vectorized footprints.

Beyond horizontal geometry, the dataset incorporates a density-connectivity climbing procedure to calculate accurate building heights. By analyzing the vertical point density distribution, this approach effectively mitigates height distortions induced by sensor noise and vegetation occlusion. These calculated elevations subsequently serve as a rigorous criterion for data refinement. Specifically, instances with implausible height attributes were systematically removed, as such discrepancies typically stem from temporal inconsistencies between the LiDAR acquisition and the official footprint updates. Furthermore, to maintain structural integrity, we eliminated building instances intersected by tile boundaries. In such cases, both the vector annotations and their associated point clusters were discarded to preserve the quality of the remaining samples.

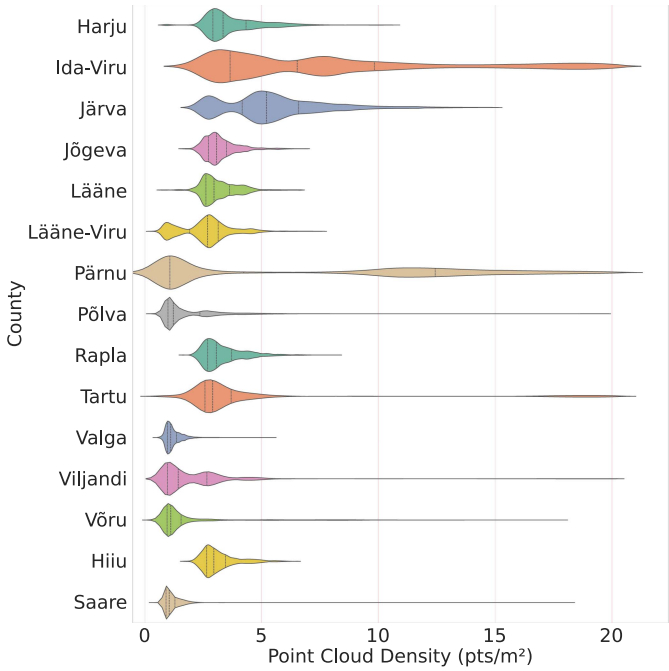


Fig. 2. Comparison of point cloud densities ($\mathrm{pts/m^2}$) across different counties in Estonia. Each violin plot shows the distribution of point cloud density, revealing the spread and concentration of points over building areas within each region. The multimodal nature of these distributions indicates significant variations in LiDAR scanning patterns.

B. Dataset Statistics

This section provides a detailed quantitative characterization of the PCFootprint dataset, emphasizing the structural, geometric, and geospatial diversity that underpins the benchmark. To demonstrate the representativeness of the sampled data, statistics are organized across 15 administrative regions, capturing a broad spectrum of urban and rural profiles.

Point cloud density variations: A critical challenge in extracting building footprints from ALS lies in the substantial heterogeneity of point cloud density and quality. As shown in the density distribution analysis in Fig. 2, the point density within PCFootprint ranges from 0.9 $\mathrm{ptsm^2}$ to 20.9 $\mathrm{ptsm^2}$. This variance is primarily attributed to different flight altitudes and sensor configurations used across various geographic regions. High-density areas provide rich geometric details for boundary localization, whereas low-density regions present substantial challenges for the detection of small structures and the preservation of topological integrity. By covering such a wide density range, this benchmark enables rigorous evaluation of algorithm robustness against nonuniform data quality.

Multiscale characteristics: The structural complexity of PCFootprint is characterized by both vertical elevation and horizontal physical area. Regarding vertical dimensions, the noise from aerial LiDAR sensors and vegetation interference can lead to the appearance of outlier data points, which can significantly distort the estimation of building heights. Consequently, we eschew the naive use of maximum elevation within a footprint boundary to estimate building heights. Instead, we leverage the inherent scanning characteristics of airborne LiDAR—where roof surfaces are typically dense, facades are sparse, and noise points are isolated—to accurately extract the true building height through detailed density analysis across the vertical axis.

As depicted in Fig. 3, we begin by filtering out vegetation (represented by green points) based on the characteristics of laser returns. To establish a robust ground reference (Z), we expand the horizontal search radius to include surrounding ground returns. Subsequently, we identify potential windows for the highest building plane within the search range $[Z_{\min}, Z_{\max}]$. By defining a fixed window height w and a step size s, we divide this range into a sequence of sliding windows $\mathcal{W} = \{W_j \mid j = 1, 2, \ldots, J\}$. Each window W_j covers the interval $[t_j, t_j + w]$, where t_j denotes the starting elevation of the jth window. The total number of windows J is determined by the formula $J = \lfloor (Z_{\max} - Z_{\min}/s) \rfloor$. Since a building may have multiple roof levels, multiple potential peak windows will be identified by calculating the point count for each window. However, for the purpose of supporting subsequent accurate height estimation, only the highest peak window $W^* = [t_{j^*}, t_{j^*} + w]$ is retained, where $j^* = \arg\max_{j \in \mathcal{M}} t_j$.

To accommodate varying point cloud densities, we introduce an adaptive connectivity threshold τ. This threshold is determined by drawing m sample points from the set W^*

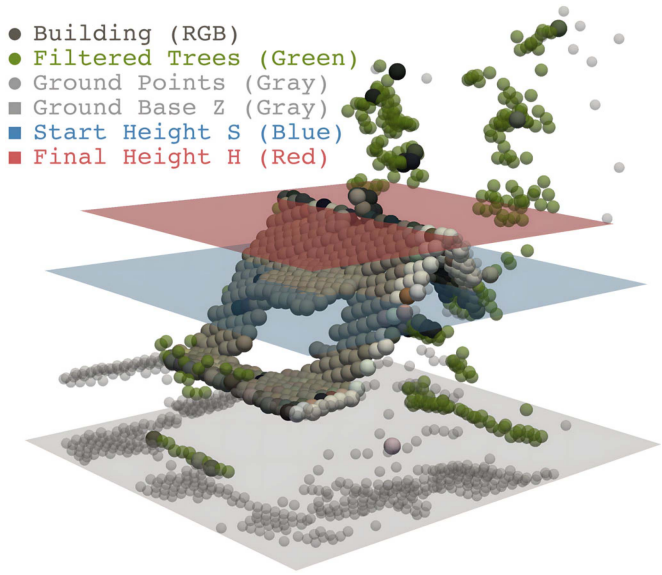


Fig. 3. Conceptual workflow of the building height estimation. Following the suppression of vegetation interference (green points), the algorithm establishes a robust ground reference (Z) from filtered terrain returns (grey points). The primary building structure is then anchored at S (blue plane) via hierarchical peak detection, with the final height (H, red plane) subsequently determined through density connectivity climbing.

Algorithm 1: Density-Connectivity Climbing.

Input: Building points $\mathcal{P}$, ground reference Z, climb starting point S, adaptive threshold τ
Output: Estimated building height H
$\mathbf{z}_{\text{upper}} \leftarrow \text{sort}(\{z \mid \mathbf{p} \in \mathcal{P}, z > S\})$
$\hat{H} \leftarrow S$
for each $z \in \mathbf{z}_{\text{upper}}$ **do**
  if $z - \hat{H} \leq \tau$ **then**
    $\hat{H} \leftarrow z$
  else
    break
  end if
end for
$H \leftarrow \hat{H} - Z$
return H

and calculating the distance $d_k^{(i)}$ to the kth nearest neighbor for each point. The adaptive threshold is then estimated as $\tau = \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} d_k^{(i)}$, where α is a tolerance coefficient that controls the sensitivity of the algorithm to point density variations. Subsequently, the points in W^* are sorted in ascending order, and the estimated roof elevation $\hat{H}$ is iteratively updated. The algorithm terminates when a vertical gap exceeding τ is encountered, indicating the presence of a structural discontinuity or the top of the roof. Finally, we assign $H = \hat{H} - Z$ as the estimated building height. The core logic of this approach is described in Algorithm 1.

The vertical profiles in Fig. 4 faithfully reflect the geographic diversity of Estonia. Harju County, the urbanized core, exhibits a prominent concentration of tall buildings, evidenced by its

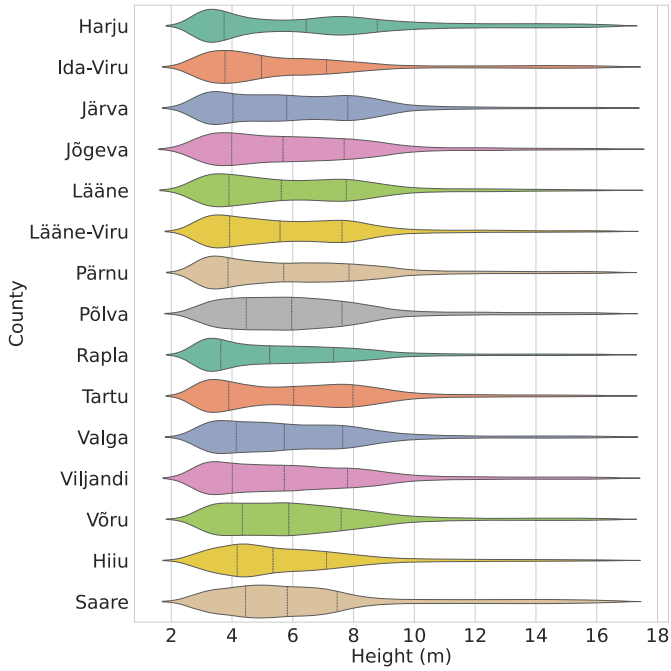


Fig. 4. Statistical distribution of building heights across Estonian counties. The violin plots characterize the predominant building scales by focusing on the height distribution up to the 95$^{\text{th}}$ percentile (P_{95}). Although the maximum height among all building instances reaches 118.82 m, the visualization is optimized to exclude these extreme outliers.

TABLE II
MAPPING OF PIXEL AREA THRESHOLDS TO PHYSICAL AREA SCALES

Scale	Pixel Area Threshold	Physical Area Threshold
Small	$A_{\text{pix}} < 32^2$ (1,024 px)	$A_{\text{phy}} < 64\,\text{m}^2$
Medium	$32^2 \leq A_{\text{pix}} \leq 96^2$	$64\,\text{m}^2 \leq A_{\text{phy}} \leq 576\,\text{m}^2$
Large	$A_{\text{pix}} > 96^2$ (9,216 px)	$A_{\text{phy}} > 576\,\text{m}^2$

elongated upper tail. Conversely, multipeak distributions in Järva, Lääne, and Tartu signify a structural mix of low-rise residential and large industrial complexes. Notably, the striking vertical consistency between the islands (Saare and Hiiu) and mainland counties (e.g., Põlva) demonstrates the feasibility for cross-domain generalization. Such inherent height heterogeneity ensures that models learn a broad spectrum of vertical features, essential for high-fidelity 3-D extraction across varied landscapes.

Horizontally, building scales are stratified into three categories including small, medium, and large, following the criteria established by the MS-COCO object detection benchmark [62]. Specifically, we mapped pixel-based area thresholds defined by MS-COCO (Area $< 32^2$ pixels for small, $32^2 \leq$ Area $< 96^2$ for medium, and Area $\geq 96^2$ pixels for large) [62] back to actual physical ground areas based on a predefined projection resolution. Consequently, building instances are categorized by their real-world footprint areas based on these derived physical thresholds, as detailed in TableII. The resulting regional distribution, visualized as a percentage stacked bar chart in

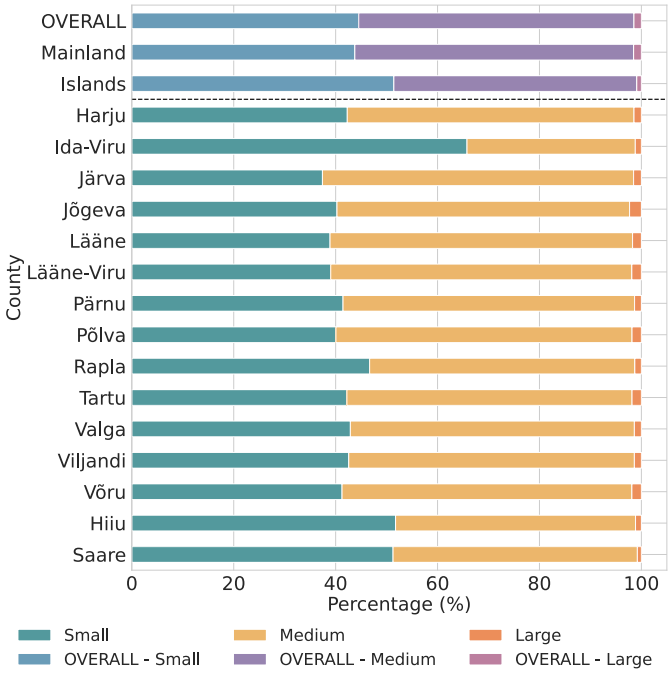


Fig. 5. Regional distribution of building scales across Estonian counties. The horizontal scale is categorized into small, medium, and large. The stacked bars illustrate the percentage composition of these categories for each county, as well as aggregated statistics for Mainland, Islands, and the entire dataset (OVERALL).

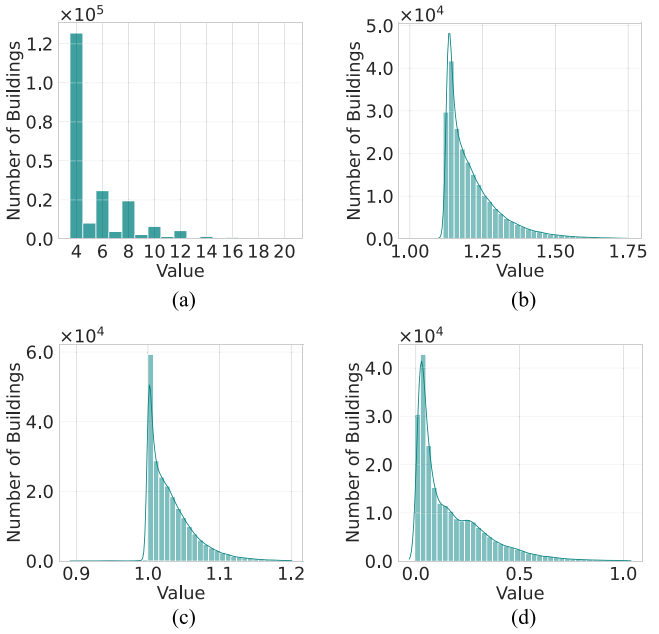


Fig. 6. Statistical distribution of building footprint complexity metrics. (a)–(c) Distributions of individual geometric dimensions, including normalized vertex count, shape index, and fractal dimension. (d) Final composite complexity index derived from the weighted combination of the three metrics. The results highlight a right-skewed pattern, indicating that while simple rectangular structures predominate, a portion of the dataset consists of irregular and complex geometries. (a) Vertex Count (V). (b) Shape Index (SI). (c) Fractal Dimension (FD). (d) Complexity Index.

Fig. 5, reveals that small and medium-scale structures constitute the vast majority of the dataset, collectively accounting for over 98 % of the total instances. While most counties exhibit a relatively balanced distribution between small and medium categories, notable regional heterogeneities emerge. For instance, Ida-Viru displays a significantly higher concentration of small buildings (approximately 65 %), whereas the Islands generally maintain a higher proportion of small-scale units compared to the Mainland. In contrast, large-scale buildings remain a marginal fraction across all counties.

Geometric complexity: To further quantify the geometric morphology of building footprints, we calculate a composite complexity index by integrating three parameters including vertex count (S_V), shape index (S_{SI}), and fractal dimension (S_{FD}). Fig. 6(a)–(c) displays the statistical distributions of these three metrics across the dataset, separately. The composite complexity index is formulated as $\text{Score} = 0.5 \cdot S_V + 0.3 \cdot S_{SI} + 0.2 \cdot S_{\text{FD}}$. The weight distribution for these components is determined based on geometric characteristics and empirical data patterns. The number of vertices (S_V) is assigned the highest weight (0.5) due to its significant impact on the difficulty of polygon extraction [63], [64], [65]. However, building footprints inherently exhibit a bias towards even numbers of vertices (e.g., 4, 6, 8, ...), which can lead to discontinuous, multipeak complexity distributions. To address this issue, the Shape Index (S_{SI}) and the fractal dimension (S_{FD}) are utilized as smoothing factors, assigned weights of 0.3 and 0.2 respectively. We confirmed that this specific assignment effectively suppresses the prominent discrete peaks introduced by S_V. This results in a continuous metric with a smooth, natural downward trend that robustly reflects the 2-D morphological complexity of the dataset. Based on the final scores, buildings are stratified into "Simple" ([0.00,0.15]), "Moderate" ((0.15,0.45]), and "Complex" ((0.45,1.00]) levels. Consequently, the regional distribution in Fig. 7 reveals a standardized architectural landscape where "Simple" structures consistently constitute 60 %–80% of the building stock. While "Complex" geometries remain a sparse minority representing less than 5 % of the total samples, the high structural consistency between Mainland and Islands aggregates demonstrates a uniform geometric character across Estonia, confirming that while irregular building instances exist, the core urban fabric is dominated by predictable geometric forms. The combination of these structural factors ensures that the PCFootprint dataset provides a rigorous testbed for evaluating the geometric fidelity and spatial robustness of vectorized footprint extraction methods.

Tile-level spatial density variance: To assess the characteristics of the dataset at the algorithmic input stage, we determined the number of building instances and the building footprint density (BFD) for each of the 128 m × 128 m tiles. As illustrated in Fig. 8, the statistical analysis revealed that both metrics exhibited a distinct right-skewed, long-tailed distribution. First, the low median values of both distributions independently reveal a severe foreground–background class imbalance. Specifically, half of the tiles in the dataset contain 5 or fewer building instances, and the median BFD is merely 3.37%. This statistically indicates that for at least half of the input patches, over

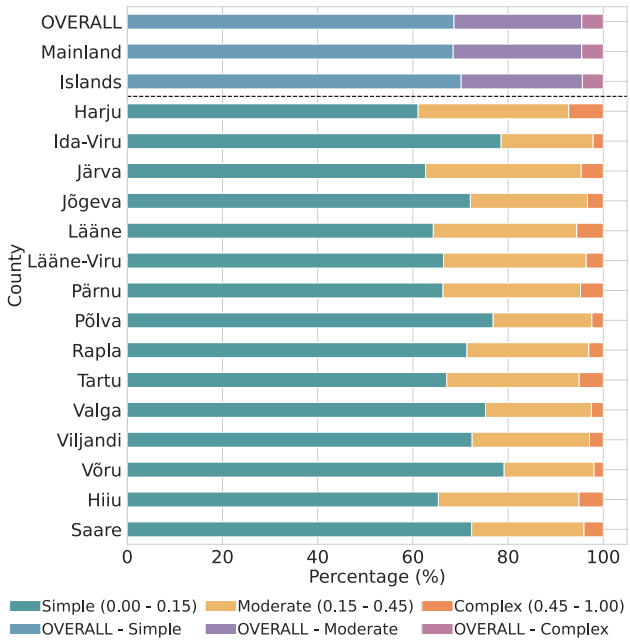


Fig. 7. Compositional analysis of building complexity levels across Estonian counties. The horizontal stacked bars illustrate the percentage distribution of simple, moderate, and complex geometries based on the composite complexity index. Aggregated metrics for Mainland, Islands, and the entire dataset (OVERALL) are provided at the top.

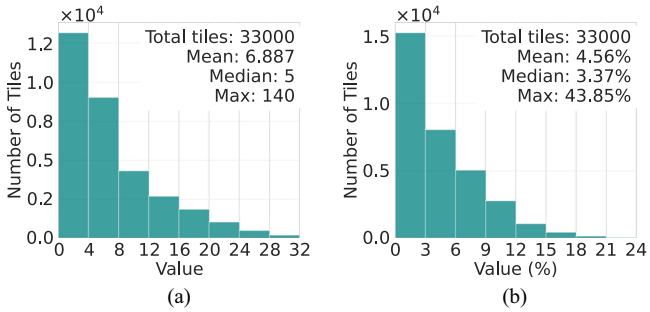


Fig. 8. Tile-level instance density illustration of the PCFootprint dataset analyzed on 128 m × 128 m tiles. (a) Histogram depicts the number of building instances per tile. (b) Histogram illustrates the BFD. For improved visualization, the x-axes in both histograms have been truncated; however, the actual maximum values can exceed 140 instances, with a BFD of 43.85%. Each bar represents the total number of tiles whose building instance count (a) or footprint density (b) falls within the respective numerical range on the *x*-axis. (a) Building Instance Counts. (b) BFD.

96% of the spatial area constitutes background clutter. Such an overwhelming background ratio substantially increases the risk of false positives, demanding that the algorithm maintain robust feature extraction and background suppression capabilities. Second, the considerable variation in local density—ranging from completely empty tiles to densely populated patches containing up to 140 instances and reaching a BFD of 43.85%—highlights the high spatial variability. The algorithm must be capable of accurately identifying isolated buildings in sparse areas to prevent false negatives, while also properly decoupling dense clusters of buildings to avoid undersegmentation. This variability in tile-level density serves as a practical and objective criterion for assessing the algorithm's adaptability.

C. Dataset Comparison

To evaluate the positioning of the proposed PCFootprint dataset within the current research landscape, we conducted a comprehensive comparison with existing building footprint extraction benchmarks as summarized in Table III. The comparative analysis focuses on data modality, accessibility, and scale to highlight the unique contributions of this work.

Most established benchmarks for building footprint extraction rely on optical sensors, including high-resolution aerial imagery, such as WHU Building [13] and Inria [14], or satellite imagery, such as SpaceNet [41] and CrowdAI [15]. While these datasets [6], [13], [14], [15], [39], [40], [41], [53] often feature massive instance counts and are generally open-source, they primarily offer 2-D spectral information that lacks the elevation cues essential for precise 3-D structural interpretation. Furthermore, as shown in Table III, a majority of existing optical datasets [14], [39], [40], [53] are provided in rasterized annotations, which necessitate additional postprocessing to satisfy GIS structural requirements.

Within the point cloud domain, a critical observation from Table III is that most existing ALS datasets specifically annotated for vectorized footprint extraction remain proprietary or restricted, including the Australian Datasets [29], the Depok Dataset [30], and the Hyderabad Study Area Data [31]. This lack of accessibility hinders the objective benchmarking of new algorithms and limits the reproducibility of research in footprint extraction from ALS. In contrast, PCFootprint is released as a fully open-source benchmark, aiming to bridge this gap in the community.

Crucially, PCFootprint establishes the first cross-domain benchmark for footprint extraction from ALS point cloud. While certain optical imagery datasets, such as Inria [14] and WHU-Mix [6], [53], have successfully incorporated cross-city or cross-regional evaluation, this capability has remained absent in the ALS research domain. Unlike existing point cloud datasets that focus on localized study areas, our benchmark is structured to support both standardized benchmarking workflow and cross-domain evaluation. This is achieved by utilizing tiles from the Estonian mainland alongside a geographically disjoint subset from major islands, enabling the rigorous evaluation of model robustness and zero-shot generalization across varying building morphologies and regional characteristics.

The scale of PCFootprint significantly exceeds that of existing point cloud benchmarks in spatial coverage and instance count. While the Australian Datasets [29] and the Hyderabad Data [31] provide high-fidelity annotations, their coverage is restricted to 0.236 and 8 km^2, respectively. PCFootprint offers a total coverage of 540.67 km^2, which represents an area over 60 times larger than the Hyderabad collection [31] and an order of magnitude higher than the individual tiles of the Depok [30] or Australian [29] series. This extensive coverage encompasses 227 264 building instances, presenting the high-volume data required for training robust deep learning architectures. By delivering 33 000 standardized tiles, PCFootprint offers a unified platform for assessing both intradomain accuracy and the inherent challenges of geographic domain shift.

TABLE III
COMPARISON OF DIFFERENT BUILDING FOOTPRINTS EXTRACTION DATASETS AND OUR PROPOSED PCFOOTPRINT DATASET

Data Type	Dataset	Subsets	Vectorized Annotation	Open Source	Elevation Information	Cross-domain Benchmark	Tiles	Building Instances	Coverage [km^2]
Imagery	Massachusetts [39]	–	✗	✓	✗	✗	151	–	340
	Inria [14]	–	✗	✓	✗	✓	180	–	405
	ISPRS Dataset [40]	Vaihingen	✗	✓	✗	✗	24	few	2
		Potsdam	✗	✓	✗	✗	16	few	11
	WHU-Mix (Raster) [53]	–	✗	✓	✗	✓	52,129	–	1,213
	WHU Building [13]	Satellite I	✗	✓	✗	✗	204	–	–
		Satellite II	✗	✓	✗	✗	17,388	29,085	550
		Aerial Imagery	✓	✓	✗	✗	8,189	220,000	450
	SpaceNet [41]	–	✓	✓	✗	✗	24,586	685,235	5,555
	CrowdAI [15]	–	✓	✓	✗	✗	341,058	–	–
	WHU-Mix (Vector) [6]	–	✓	✓	✗	✓	64,387	754,126	1,100
Point Cloud	Fredericton [42]	–	✗	✗	✓	✗	1,440	–	–
	Australian [29]	Aitkenvale	✓	✗	✓	✗	1	5	
		Hervey Bay	✓	✗	✓	✗	1	28	
		Eltham	✓	✗	✓	✗	1	75	0.236
		Hobart	✓	✗	✓	✗	1	69	
		Knox	✓	✗	✓	✗	1	52	
	Depok [30]	–	✓	✗	✓	✗	–	–	–
	Hyderabad [31]	–	✓	✗	✓	✗	–	1,300	8
	PCFootprint (Ours)	Mainland	✓	✓	✓	✓	30,000	204,608	491.52
		Islands					3,000	22,656	49.15

TABLE IV
COMPARISON OF DIFFERENT INTERPOLATION METHODS

Method	Time Cost (s/tile)	Void Pixels
w/o interpolation	0.038	67.60 %
Bilinear	0.133	39.73 %
Bicubic	5.589	0.42 %
Kriging	107.804	0.00 %
IDW (adopted)	4.541	0.07 %

The time cost metric was calculated using an amd r7-9900x computer

IV. BENCHMARK

A. Benchmarking Protocols and Data Splitting

To establish a standardized benchmark for model training and evaluation, the PCFootprint dataset is partitioned into two distinct evaluation protocols: an intradomain standard benchmark and a cross-domain generalization benchmark. The proportional distribution of tiles across Estonian counties and their allocation into specific subsets are illustrated in Fig. 9.

1) *Standard benchmark:* We designate the Mainland subset as the foundation for our standard benchmark to establish a standardized benchmarking workflow. This protocol utilizes the 30 000 valid mainland tiles. Extending the density driven sampling strategy used during dataset construction, these tiles were partitioned into training, validation, and

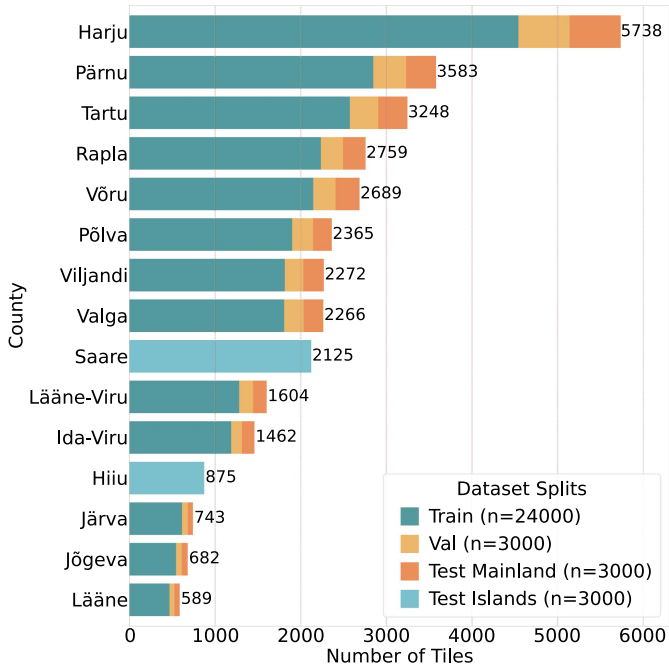


Fig. 9. Proportional distribution and subset splitting of point cloud tiles across Estonian counties. The stacked bars illustrate the allocation for training, validation, and standard benchmark sets within mainland regions, while dedicated Islands subset (Saaremaa and Hiiumaa) constitute the cross-domain generalization benchmark.

testing splits according to a strict 8 : 1 : 1 ratio within each individual county. This stratified approach ensures that the geospatial distribution of all data splits remains synchronized with the actual building density patterns across the Estonian mainland. Specifically, the mainland subset includes 24 000 samples for training, 3000 for validation, and 3000 for the standard benchmark set to evaluate in domain performance. While these tiles are strictly nonoverlapping to prevent data leakage, they share consistent domain characteristics to facilitate a rigorous assessment of prevailing architectures.

2) *Generalization benchmark:* To evaluate the robustness and extrapolation capabilities of algorithms on unseen scenarios, we leverage the Islands subset as the cornerstone of our generalization benchmark. This island-based subset comprises 3000 tiles from the major islands of Saaremaa and Hiiumaa, which are geographically isolated from the mainland domain. This cross-domain testbed simulates real-world applications where target inference scenarios often differ significantly from the source training domain in terms of building morphologies and LiDAR sampling patterns. This enables a rigorous assessment of zero-shot generalization performance.

B. Evaluation Metrics

To establish a standardized and objective evaluation of extraction performance, the final predictions from all evaluated models are unified into 2-D geometric representations. Regardless of whether the input modality is 3-D ALS point clouds or 2-D optical imagery, the predicted building footprints are formatted as vectorized polygons or pixel-level masks. This unification allows for the adoption of the MS-COCO evaluation protocol [62], which serves as a rigorous benchmark for instance-level detection and segmentation tasks. By grounding the evaluation in the 2-D geometric domain, the framework ensures consistency across distinct model architectures while bypassing the technical complexities of cross-modal alignment during the metric definition stage.

The fundamental criterion for evaluating the spatial correspondence between a predicted footprint and its corresponding ground truth is the intersection over union (IoU). A prediction is considered a true positive only if its IoU with a ground truth instance exceeds a specified threshold. This mechanism enables the quantification of model performance across varying degrees of localization stringency.

The primary evaluation metrics utilized in this study comprise mean average precision (mAP) and average recall (AR). mAP is calculated by integrating the precision–recall curve across multiple IoU thresholds ranging from 0.50 to 0.95 with an incremental step of 0.05. This composite metric provides a comprehensive assessment of both detection accuracy and boundary fidelity. We specifically report AP_{50} and AP_{75} to represent performance under relaxed and stringent spatial constraints, respectively. Furthermore, AR is evaluated to measure the model's ability to successfully recover all building instances within a given scene. Evaluating both mAP and AR ensures a balanced analysis of the model's precision in avoiding false positives and its completeness in mitigating missed detections.

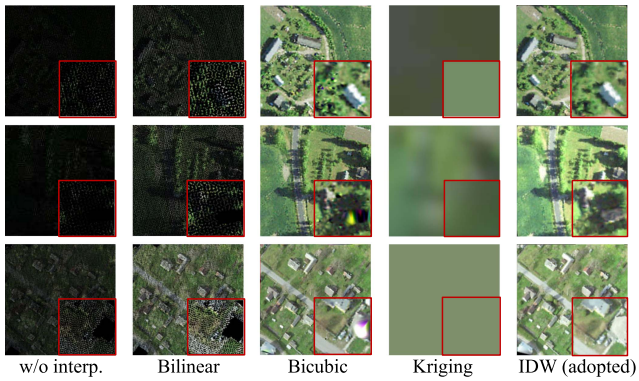


Fig. 10. Visual comparison of different interpolate methods. Compared to direct projection and bilinear interpolation (notable void artifacts), bicubic interpolation (color distortions), and Kriging (excessive smoothing), the adopted IDW method optimaly balances missing-pixel filling and local boundary preservation.

In accordance with the MS-COCO standard [62], the extraction performance is further stratified across different building scales, denoted as AP_S, AP_m, and AP_L. Instances are categorized into Small (Area $< 32^2$ pixels), Medium ($32^2 \leq$ Area $< 96^2$ pixels), and Large (Area $\geq 96^2$ pixels) classes based on their pixel area. This multiscale assessment provides granular insights into the robustness of models, which is particularly critical for the PCFootprint dataset given the significant size variation between small residential structures and massive industrial complexes. By employing this unified and comprehensive evaluation suite, the benchmark facilitates an objective comparison of mainstream algorithms in achieving high-fidelity 2-D building footprint reconstruction.

C. Representative Baselines and Results

To adapt 2-D image-based methodologies to the 3-D ALS point cloud data, we implemented a standardized projection and interpolation pipeline. Each point cloud tile covering a 128 m $\times$ 128 m^2 ground area was projected onto a 512 m $\times$ 512 m pixels grid with a ground sampling distance of 0.25 m^2 per pixel. The RGB spectral information from the original point clouds was preserved during the projection to form three-channel images. Since the average point density of most tiles is significantly lower than 16 $\mathrm{pts}/\mathrm{m}^2$ which is the theoretical threshold required to occupy every pixel at a 0.25 resolution, an effective spatial interpolation algorithm is essential. After examining different interpolation methods (as detailed in Table IV and Fig. 10), we ultimately decided to use the inverse distance weighting (IDW) algorithm for spatial interpolation. This choice was based on the algorithm^{\prime }s excellent balance between efficient computation and the preservation of high-quality local features, which resulted in the creation of smooth and continuous image representations.

For a rigorous evaluation of prevailing building footprint extraction methods on the PCFootprint dataset, we selected representative baselines encompassing two primary paradigms:

TABLE V
QUANTITATIVE COMPARISON OF BUILDING FOOTPRINT EXTRACTION ON PCFOOTPRINT

Paradigm	Method	mAP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)	AR (%)
Mainland Subset								
Pixel-wise Segmentation	SegEarth-OV3 [51]	8.2	21.2	4.5	1.3	14.4	15.3	28.2
	MSA-SAM [50]	15.7	40.4	7.3	1.7	20.5	17.5	21.2
	D2LS [52]	22.9	40.4	24.1	4.2	41.7	44.7	47.2
	MFNet [58]	**49.8**	76.1	**57.8**	**29.7**	**62.4**	56.7	54.1
Polygonal Regression	HoliTracer [12]	21.6	49.4	15.5	4.7	30.6	39.6	37.1
	Pix2Poly [11]	31.2	50.9	34.1	19.4	46.3	26.4	43.1
	P2PFormer [8]	46.3	**77.0**	50.1	17.3	52.8	60.5	**56.7**
	SAMPolyBuild [7]	47.6	74.4	54.0	20.9	53.3	**66.1**	54.2
Islands Subset (zero-shot)								
Pixel-wise Segmentation	SegEarth-OV3 [51]	1.5	4.5	0.7	0.2	4.0	0.6	18.5
	MSA-SAM [50]	11.9	33.3	4.3	1.4	15.5	16.9	16.6
	D2LS [52]	24.3	45.6	24.6	5.2	39.1	42.0	45.6
	MFNet [58]	33.2	54.8	37.4	17.1	47.2	30.3	35.7
Polygonal Regression	HoliTracer [12]	19.8	47.7	13.0	4.2	28.2	39.5	34.1
	Pix2Poly [11]	29.9	52.7	31.4	19.4	45.0	22.2	48.2
	P2PFormer [8]	40.8	71.1	43.1	15.8	47.0	60.2	51.3
	SAMPolyBuild [7]	**44.7**	**72.3**	**50.1**	**21.2**	**49.9**	**67.7**	**51.2**

Results are grouped by Mainland and Islands subsets and categorized by method paradigm. Bold values indicate the best performance for each metric within each subset.

pixelwise segmentation and polygonal regression. Since the deterministic nature of the traditional LasBuildSeg [59] precludes the generation of probabilistic confidence scores required for MS-COCO [62] metrics, this approach is restricted to qualitative assessment to facilitate a holistic comparison of structural fidelity. Quantitative performance across the Mainland and Islands subsets is summarized in Table V, while qualitative visualizations contrasting deep learning baselines with traditional geometric algorithms are presented in Fig. 11.

As illustrated in Fig. 11 and quantified in Table V, the comparative analysis across representative scenarios including low density sampling regions, areas with heavy vegetation occlusion, and cluttered urban areas reveals distinct performance gaps between traditional heuristics, pixelwise segmentation, and polygonal regression paradigms. The traditional LasBuildSeg [59] method performs poorly across all scenarios by generating numerous false positive predictions, arising from the inherent sensitivity of heuristic rules to manual parameter tuning. Similarly, SegEarth-OV3 [51] and MSA-SAM [50] exhibit significant false detections and omissions. Although SegEarth-OV3 [51] can successfully identify buildings in favorable local scenes (shown in Fig. 11), its overall quantitative precision is severely degraded by pervasive instability and false predictions across the broader, more complex full dataset. Which leads to a mAP of only 8.2% on the Mainland and drops to 1.5% on the Islands. This failure is primarily due to its training-free nature, which relies on SAM 3 [66] weights that cannot be fine-tuned for the distinct radiometric and textural characteristics of imagery generated from ALS. MSA-SAM [50] also struggles with robustness, yielding an mAP of 15.7% (Mainland) and frequently misidentifies road surfaces or vegetation as buildings, indicating limited robustness to point cloud derivatives. Despite these localized failures, the relative stability of most models in sparse regions confirms the feasibility of ALS as a reliable source for footprint extraction. Specifically, the pixelwise segmentation leader MFNet [58] achieves the highest Mainland accuracy (49.8% mAP). Yet pixelwise segmentation methods exhibit pronounced staircase effect and corner degradation. While these artifacts are less prominent in the Mainland subset, they become highly visible in the Islands generalization set. In contrast, polygonal regression models, such as SAMPolyBuild [7] and P2PFormer [8], achieve the top two results on the Islands with 44.7% and 40.8% mAP, respectively. They consistently maintain sharp and regularized boundaries across both geographic domains.

Evaluating model generalization necessitates an assessment of performance degradation across the geographically disjoint Islands subset. This dual-testing configuration is vital for identifying the limitations of current architectures under domain shift. When confronted with unseen regional layouts and sampling patterns, the pixelwise segmentation paradigm exhibits a substantial reduction in mAP ranging from 3.8 to 16.6 absolute points. Conversely, the polygonal regression paradigm demonstrates superior spatial robustness by maintaining significantly higher relative precision across distinct geographic domains. The absolute mAP reduction within the polygonal paradigm is notably more contained, with decreases restricted to between 1.3 and 5.5 points. This evidence suggests that explicitly modeling geometric primitives is more effective for accommodating

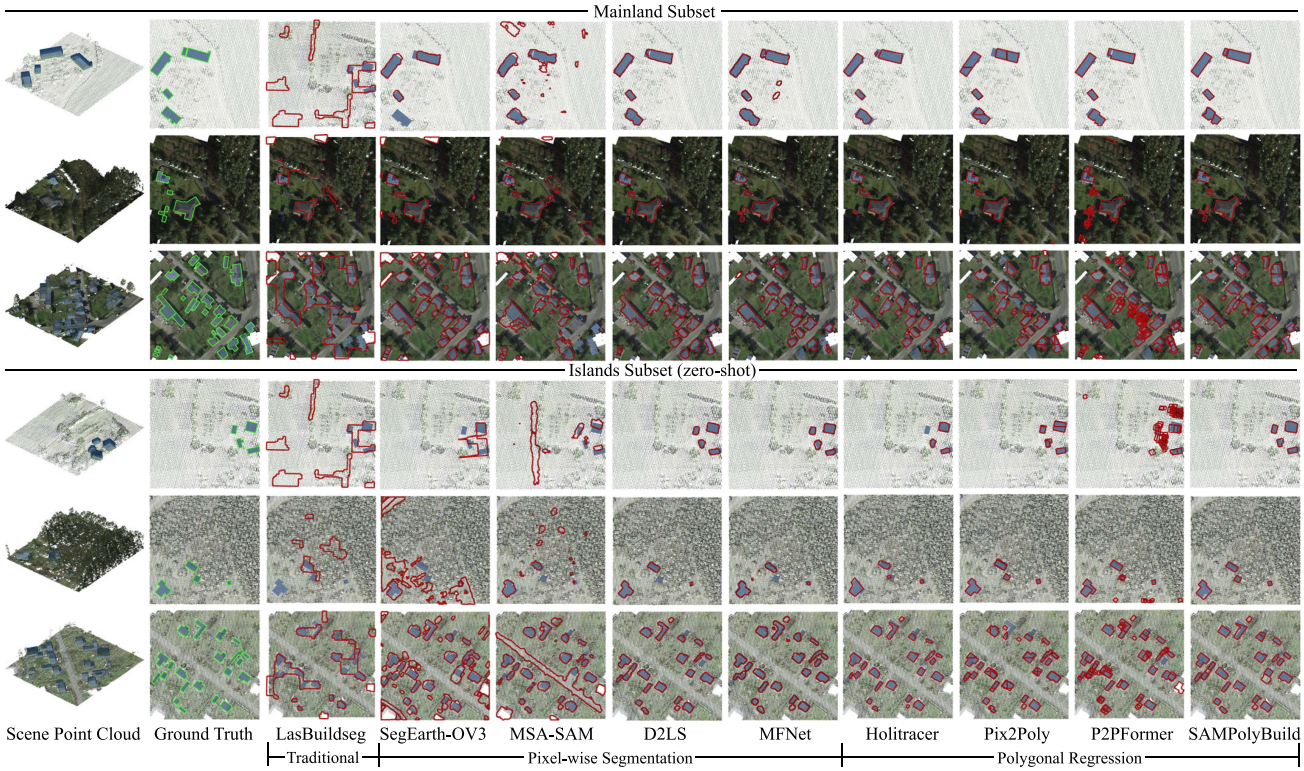


Fig. 11. Qualitative comparison of building footprint extraction results across various methodologies. Representative scenarios characterized by low density sampling, heavy vegetation occlusion, and complex urban morphology were selected to evaluate model robustness. The leftmost column provides 3-D visualizations with blue extruded buildings. Subsequent columns present top-view ground truth (green outlines) and detection results (red outlines). The first three rows represent mainland scenarios, whereas the subsequent three rows display zero-shot generalization results on the islands.

variations in building morphology and sampling density than predicting independent pixel labels, thereby providing a more stable solution for zero-shot generalization tasks.

Specific failure modes yield critical insights for future algorithm development. P2PFormer [8] suffers from severe redundant detections, where multiple overlapping polygons are predicted for a single building instance. This redundancy likely stems from insufficient nonmaximum suppression or failures in the primitive-matching logic when processing imagery generated from ALS. Similarly, Pix2Poly [11] fails to capture complete building structures and exhibits mild redundancy, suggesting that its postprocessing modules are not fully optimized for the nonuniform sampling density of ALS data. Furthermore, the low AP_S values across all models indicate a pervasive difficulty in detecting small-scale buildings because sparse point clouds provide insufficient geometric evidence to resolve these structures.

D. Challenges

The quantitative results and qualitative assessments derived from the PCFootprint benchmark highlight several persistent challenges in vectorized building footprint extraction from ALS point clouds. A primary obstacle involves geographic domain shift, as evidenced by the significant performance discrepancy between the Mainland and Islands evaluation protocols. For instance, prevailing models exhibit a substantial performance degradation, with absolute mAP reductions ranging from 1.3 to 16.6 percentage points when transitioning from the Mainland to the Islands subset. This degradation is primarily attributed to variations in architectural styles and land use patterns between the Estonian mainland and the disjoint regions of Saaremaa and Hiiumaa. Such regional heterogeneity constrains the zero-shot generalization capacity of models trained on geographically concentrated data, as they struggle to achieve spatial robustness in unseen urban and rural landscapes.

These regional differences are further compounded by substantial fluctuations in point density, which are fundamentally tied to the heterogeneity of acquisition settings across different survey campaigns. Specifically, varying flight configurations and sensor specifications result in a density range from $0.9\ \mathrm{pts/m^2}$ to $20.9\ \mathrm{pts/m^2}$ throughout the PCFootprint dataset. The inherent difficulty of extraction is significantly intensified by the transition from 3-D ALS point clouds to 2-D image-based workflows, which inevitably sacrifices critical spatial geometric attributes. Since most evaluated methods are image-based, they remain highly sensitive to the quality of 2-D projections. In low density regions, the sparsity of the ALS point cloud, a direct consequence of acquisition variability, introduces radiometric artifacts and textural blurring during interpolation. This information loss complicates precise boundary localization particularly for small structures, which is reflected in the pervasive scale sensitivity and lower APS scores observed throughout the benchmark. Furthermore, image-based methods remain vulnerable to spectral interference and the loss of vertical elevation cues caused by shadows or vegetation occlusion. In complex urban

scenes, these limitations frequently cause adjacent building instances to merge. This suggests that a paradigm shift toward the direct exploitation of 3-D geometry is essential to achieve the structural integrity and precision required for high-level modeling.

V. Conclusion

In this article, we presented PCFootprint, the first extensive public dataset dedicated to the vectorized extraction of building footprints from ALS point clouds. By providing 33 000 standardized tiles with systematically aligned annotations, this benchmark fills a critical gap in the field of structured building footprint extraction and cartographic production. Our comprehensive evaluation across diverse geographical landscapes reveals that while pixelwise segmentation and polygonal regression paradigms achieve significant advancements in geometric fidelity, persistent challenges remain regarding cross-domain generalization. Specifically, the performance degradation observed in geographically disjoint regions underscores the technical necessity for learning domain-invariant geometric features. Furthermore, the limitations inherent in image-based workflows highlight the urgent need for a paradigm shift toward the direct exploitation of 3-D spatial geometry. By bridging the gap between raw point clouds and GIS compatible vector outputs, the PCFootprint dataset not only serves as a rigorous testbed for current algorithmic constraints but also paves the way for the next generation of high precision building modeling and digital twin applications.

A. Discussion

Although PCFootprint has established a foundational standard for LiDAR-based building extraction, its current approach has certain limitations. First, the standardized 128 m $\times$ 128 m tile partition represents a well-balanced compromise among point cloud density, network input capacity, processing efficiency, and the representational completeness of geographical features. Consequently, this configuration may occasionally result in the omission of exceptionally large structures, although such instances are relatively rare within this dataset. To address this issue, future dataset construction could explore multiscale or dynamic cropping techniques to inherently preserve the complete geometry of these oversized building instances. Second, due to the limitations of the current algorithmic landscape, our benchmark study could not incorporate a broader range of open-source end-to-end point cloud algorithms for building footprint extraction. As a result, the evaluated methods primarily rely on indirect approaches that convert 3-D point clouds into 2-D raster images. Therefore, future research should focus on developing deep learning networks capable of directly extracting vectorized footprints from raw 3-D point clouds.

Acknowledgment

The authors would like to thank the Estonian Land and Spatial Development Board for providing access to the raw ALS point cloud data and administrative building records utilized in this study. The authors would like to thank the research community for sharing the open-source baseline models and code architectures which were instrumental in the benchmarking process. They would like to thank Editor-in-Chief and anonymous reviewers are acknowledged for their constructive feedback and insightful suggestions that significantly enhanced the scientific rigor of this manuscript. The authors would also like to thank the Intelligent Computing Center of Shenzhen University.

References

[1] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, pp. 6169–6181, Jul. 2021.

[2] Y. Li, D. Hong, C. Li, J. Yao, and J. Chanussot, "HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 51–65, Mar. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271624000315

[3] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote. Sens.*, vol. 11, 2019, Art. no. 403.

[4] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMS using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[5] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[6] S. Wei, T. Zhang, S. Ji, M. Luo, and J. Gong, "BuildMapper: A fully learnable framework for vectorized building contour extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 87–104, Mar. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271623000217

[7] C. Wang, J. Chen, Y. Meng, Y. Deng, K. Li, and Y. Kong, "SAMPolyBuild: Adapting the segment anything model for polygonal building extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 218, pp. 707–720, Dec. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271624003563

[8] T. Zhang, S. Wei, Y. Zhou, M. Luo, W. Yu, and S. Ji, "P2PFormer: A primitive-to-polygon method for regular building contour extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4414012.

[9] W. Yu et al., "P2PFormerV2: Improving primitive-based regular building contour extraction methods via contour feature enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4421016.

[10] W. Jiao, H. Cheng, G. Vosselman, and C. Persello, "RoIPoly: Vectorized building outline extraction using vertex and logit embeddings," *ISPRS J. Photogrammetry Remote Sens.*, vol. 224, pp. 317–328, Jun. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271625001364

[11] Y. K. Adimoolam, C. Poullis, and M. Averkiou, "Pix2Poly: A sequence prediction method for end-to-end polygonal building footprint extraction from remote sensing imagery," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.* 2025, pp. 8484–8493.

[12] Y. Wang, B. Dang, W. Li, W. Chen, and Y. Li, "Holitracer: Holistic vectorization of geographic objects from large-size remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 8482–8491.

[13] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8444434/

[14] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 3226–3229. [Online]. Available: https://ieeexplore.ieee.org/document/8127684/

[15] S. P. Mohanty et al., "Deep learning for understanding satellite imagery: An experimental survey," *Front. Artif. Intell.*, vol. 3, 2020, Art. no. 534696.

[16] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 240–252, 2022.

[17] Y. Yu et al., "Enhancing building segmentation with shadow-aware edge perception," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 1–12, 2025.

[18] D. Yang, X. Gao, Y. Yang, K. Guo, K. Han, and L. Xu, "Advances and future prospects in building extraction from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6994–7016, 2025.

[19] C. Ma, Y. Zhang, J. Guo, G. Zhou, and X. Geng, "FusionHeightNet: A multi-level cross-fusion method from multi-source remote sensing images for urban building height estimation," *Remote. Sens.*, vol. 16, 2024, Art. no. 958.

[20] W. Xu et al., "Building height extraction from high-resolution single-view remote sensing images using shadow and side information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6514–6528, 2024.

[21] D. Yu, S. Ji, S. Wei, and K. Khoshelham, "3-D building instance extraction from high-resolution remote sensing images and DSM with an end-to-end deep neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4406019.

[22] I. Tomljenovic, B. Höfle, D. Tiede, and T. Blaschke, "Building extraction from airborne laser scanning data: An analysis of the state of the art," *Remote. Sens.*, vol. 7, pp. 3826–3862, 2015.

[23] Y. Zang, W. Mi, X. Xiao, H. Guan, J. Chen, and D. Li, "Compound 3D building modeling with structure-aware partition and primitive assembly from airborne laser scanning point clouds," *Int. J. Digit. Earth*, vol. 17, 2024, Art. no. 2375112.

[24] H. You et al., "Tree extraction from airborne laser scanning data in urban areas," *Remote Sens.*, vol. 13, 2021, Art. no. 3428.

[25] E. M. Domí nguez, P. Brotzer, E. Casalini, and D. Small, "Mapping urban areas and infrastructure through fusion of airborne SAR 3-D images: A comparative study with ALS sensors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6164–6181, 2025.

[26] W. Li, F. Wang, and G. Xia, "A geometry-attentional network for ALS point cloud classification," *Isprs J. Photogrammetry Remote Sens.*, vol. 164, pp. 26–40, 2020.

[27] N. Varney, V. K. Asari, and Q. Graehling, "Dales: A large-scale aerial liDAR data set for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 186–187.

[28] X. Han, C. Liu, Y. Zhou, K. Tan, Z. Dong, and B. Yang, "WHU-Urban3D: An urban scene LiDAR point cloud dataset for semantic instance segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 500–513, Mar. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271624000522

[29] M. Awrangjeb and G. Lu, "Automatic building footprint extraction and regularisation from LIDAR point cloud data," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, Nov. 2014, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/7008096/

[30] A. Gamal, F. Zulkarnain, S. Indratmoko, L. A. Rohmah, and A. Mushfiroh, "Semi-automatic model for detecting the discrepancy between cadastral data and building footprint extraction using unmanned aerial vehicles (UAV) LiDAR in the perspective of smart city," in *Proc. Int. Conf. Technol., Eng., Comput. Appl.*, Dec. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10490944/

[31] J. Nalini, S. Saravanan, B. Narender, and S. Muralikrishnan, "Automatic building boundary extraction from point cloud data," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. X-5/W2-2025, pp. 429–438, Dec. 2025. [Online]. Available: https://isprs-annals.copernicus.org/articles/X-5-W2-2025/429/2025/

[32] G. Kong, H. Fan, and G. Lobaccaro, "Automatic building outline extraction from ALS point cloud data using generative adversarial network," *Geocarto Int.*, vol. 37, pp. 15964–15981, 2022.

[33] G. Kong and H. Fan, "Ph-shape: An adaptive persistent homology-based approach for building outline extraction from ALS point cloud data," *Geospatial Inf. Sci.*, vol. 27, pp. 1107–1117, 2023.

[34] A. Nurunnabi, N. Teferle, J. Balado, M. Chen, F. Poux, and C. Sun, "Robust techniques for building footprint extraction in aerial laser scanning 3D point clouds," in *Proc. Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. XLVIII-3/W2-2022, pp. 43–50, 2022.

[35] B. P. Hrutka, Z. Siki, and B. Takács, "Voxel-based point cloud segmentation and building detection," in *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 102, pp. 67–74, 2022.

[36] J. Huang, J. Stoter, R. Peters, and L. Nan, "City3D: Large-scale building reconstruction from airborne LiDAR point clouds," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2254.

[37] G. Kong, C. Zhang, and H. Fan, "Large-scale 3-D building reconstruction in LoD2 from ALS point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, 2025, Art. no. 6500305.

[38] B. Yang, R. Huang, J. Li, M. Tian, W. Dai, and R. Zhong, "Automated reconstruction of building LoDs from airborne liDAR point clouds using an improved morphological scale space," *Remote. Sens.*, vol. 9, 2016, Art. no. 14.

[39] V. Mnih, "Machine learning for aerial image labeling," *Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada*, 2013.

[40] International Society for Photogrammetry and Remote Sensing (ISPRS), "2D semantic labeling contest," 2026, Accessed: Feb. 10, 2026. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/semantic-labeling.aspx

[41] A. V. Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232v3*.

[42] F. S. Vostikolaei and S. Jabari, "Multimodal building footprint extraction from orthophotoa and lidar point clouds using deep learning framework," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 8214–8217 . [Online]. Available: https://ieeexplore.ieee.org/document/10641225/

[43] X. Dong, J. Cao, and W. Zhao, "A review of research on remote sensing images shadow detection and application to building extraction," *Eur. J. Remote Sens.*, vol. 57, no. 1, 2024, Art. no. 2293163.

[44] T. Raj, F. H. Hashim, A. B. Huddin, M. F. Ibrahim, and A. Hussain, "A survey on LiDAR scanning mechanisms," in *Electronics*, vol. 9, 2020, Art. no. 741.

[45] S. Chen, Q. Yan, Y. Qu, W. Gao, J. Yang, and F. Deng, "Ortho-NeRF: Generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images," *Geo-spatial Inf. Sci.*, vol. 28, no. 2, pp. 741–760, 2025.

[46] Q. Wang, Z. Zhan, J. He, Z. Tu, X. Zhu, and J. Yuan, "High-quality spatial reconstruction and orthoimage generation using efficient 2D Gaussian splatting," *Signal, Image Video Process.*, vol. 20, no. 8, 2026, Art. no. 456.

[47] X. Wang et al., "Tortho-Gaussian: Splatting true digital orthophoto maps," *Geo-spatial Inf. Sci.*, 2026, pp. 1–24.

[48] W. Tan et al., "Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit. Workshops*, 2020, pp. 202–203.

[49] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Computer Vis.*, 2023, pp. 4015–4026.

[50] S. Chen, Y. Yu, Y. Li, Z. Wang, X. Li, and J. Han, "Multiscale adapter based on SAM for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6806–6819, 2025.

[51] K. Li, S. Zhang, Y. Deng, Z. Wang, D. Meng, and X. Cao, "SegEarth-OV3: Exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images," 2025, *arXiv:2512.08730*.

[52] X. Zou et al., "Dynamic dictionary learning for remote sensing image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 22457–22466.

[53] M. Luo, S. Ji, and S. Wei, "A diverse large-scale building dataset and a novel plug-and-play domain generalization method for building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4122–4138, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10104584/

[54] T. Han et al., "CityInsight: Incorporating dual-condition based diffusion model into building footprint segmentation from remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5632417.

[55] S. Yuan, Z. Wang, J. Li, N. Xu, Z. You, and D. Huang, "FDENet: Frequency-guided dual-encoder network for building footprint extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 22403–22420, 2025.

[56] W. Jiao, C. Persello, and G. Vosselman, "PolyR-CNN: R-CNN for end-to-end polygonal building outline extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 218, pp. 33–43, 2024.

[57] S. Wei, T. Zhang, D. Yu, S. Ji, Y. Zhang, and J. Gong, "From lines to polygons: Polygonal building contour extraction from high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 213–232, Mar. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271624000388

[58] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "A unified framework with multimodal fine-tuning for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5405015.
[59] M. Erdem and B. Anbaroglu, "Reproducible extraction of building footprints from airborne LiDAR data: A demo paper," in *Proc. 31st ACM Int. Conf. Adv. Geographic Inf. Syst.*, 2023, pp. 1–4.
[60] P. Rottmann, J.-H. Haunert, and Y. Dehbi, "Automatic building footprint extraction from 3D laserscans," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 10, pp. 233–240, 2022.
[61] A. Vats et al., "Terrain-informed self-supervised learning: Enhancing building footprint extraction from LiDAR data with limited annotations," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5703810.
[62] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Eur. Conf. Computer Vis.* 2014, pp. 740–755.
[63] Y. Shen et al., "V-DETR: DETR with vertex relative position encoding for 3D object detection," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 3368–3386.
[64] Y. Zheng, O. Andrienko, Y. Zhao, M. Park, and T. Pham, "DPPD: Deformable polar polygon object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 78–87.
[65] X. Zhang, Y. Siddiqui, A. Avetisyan, C. Xie, J. Engel, and H. Howard-Jenkins, "Vertexregen: Mesh generation with continuous level of detail," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 12570–12580.
[66] N. Carion et al., "SAM 3: Segment anything with concepts," 2025, *arXiv:2511.16719*.



Haoyuan Shen received the B.Eng. degree in geomatics engineering from the Anhui University of Science and Technology, Huainan, China, in 2024. He is currently working toward the M.S. degree in geomatics engineering with the School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China.

His research interests include building footprint extraction, point cloud processing, urban scene modeling, and computer vision.



Kuihao Wang received the B.Eng. degree in software engineering from the Hubei University of Technology, Wuhan, China, in 2025. He is currently working toward the master's degree in artificial intelligence with Shenzhen University, Shenzhen, China.

His research interests include 3-D reconstruction and building footprint extraction, as well as fundamental visual perception tasks, such as image segmentation and object detection.



Ruisheng Wang (Senior Member, IEEE) received the B.Eng. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1993, the M.Sc.E. degree in geomatics engineering from the University of New Brunswick, Fredericton, NB, Canada, in 2004, and the Ph.D. degree in electrical and computer engineering from McGill University, Montreal, QC, Canada, in 2011.

He is currently a Distinguished Professor with the School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China. Before joining Shenzhen University, he was a Professor with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada, which he joined in 2012. His research interests include mobile LiDAR data processing, 3-D reconstruction, and intelligent mapping for next-generation digital mapping and navigation.



Yujun Liu (Graduate Student Member, IEEE) received the B.Eng. degree in computer science and technology from the School of Computer Engineering, Jimei University, Xiamen, China, in 2021, and the M.S. degree in information and communication engineering from the School of Computer Engineering, Jimei University, Xiamen, China, in 2024. He is currently working toward the Ph.D. degree in surveying and mapping science and technology with the School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China.

His research interests include computer vision, point cloud processing, and urban scene modeling and understanding.