ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

The P^3 Dataset: Pixels, Points and Polygons for Multimodal Building Vectorization

Raphael Sulzer[1,2], Liuyun Duan[2], Nicolas Girard[2], Florent Lafarge[1]

[1] Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France.
(raphael.sulzer, florent.lafarge)@inria.fr

[2] LuxCarta Technology, Mouans-Sartoux, France.
(lyduan, ngirard)@luxcarta.com

Keywords: 3D Point Clouds, Aerial Imagery, Building Segmentation, Building Vectorization, Data Fusion.

Abstract

We present P^3, a large-scale multimodal dataset for building vectorization, including aerial LiDAR point clouds, aerial images, and vectorized 2D building outlines, collected across three continents. P^3 contains over 10 billion LiDAR points with decimeter-level accuracy and RGB images at a ground sampling distance of 25 centimeters. While many existing datasets focus on the image modality, P^3 offers a complementary perspective by incorporating dense 3D information. We demonstrate that LiDAR point clouds serve as a robust modality for predicting building polygons, both in hybrid and end-to-end learning frameworks. Moreover, fusing LiDAR and imagery further improves accuracy and geometric quality of predicted polygons. The P^3 dataset is publicly available, along with code and pretrained weights of three state-of-the-art models for building polygon prediction at https://github.com/raphaelsulzer/PixelsPointsPolygons.

1. Introduction

Cadastral maps that register the footprints of buildings as vector representations – typically 2D polygons – constitute an important source of information in numerous application fields, going from urban planning and navigation to environmental monitoring, and disaster response through defense and intelligence. In the past decades, the construction and update of these maps was mainly done by human operators at some specific locations in the world only. Today, this tedious task is about to be replaced by automatic algorithms that learn to capture building outlines from satellite and aerial data, with the promise of vectorizing the several billion buildings in the world despite the strong variability of shape, density, and regularity of these objects.

Several scientific communities such as Computer Vision, Geoscience, Remote Sensing and Photogrammetry have actively worked on the building vectorization problem during the past years. For fully automatic building vectorization, deep learning architectures are commonly trained on datasets that (i) exploit the image modality only, (ii) have little architectural and regional variability, as well as little variability in terms of acquisition characteristics (Hensel et al., 2021, Roscher et al., 2020, Rottensteiner et al., 2012), and (iii) use pixel masks instead of direct polygon annotations (Maggiori et al., 2017, Schuegraf et al., 2023). Unfortunately, these restrictions tend to make the efficiency of recent state-of-the-art (SOTA) methods (Xu et al., 2023, Adimoolam et al., 2025, Zorzi and Fraundorfer, 2023) stagnant and, in the end, not ready to operate effectively, robustly and at large scale. (i) Using only overhead images can lead to false positive detections or missing building parts due to shadows and occlusions. Occlusions are particularly frequent in the form of trees hiding a building or part of it, or due to *leaning buildings* induced by image distortion. (ii) Small variability in the dataset can cause models to overfit to architectural styles of a country or region (Maggiori et al., 2017), to acquisition geometries, or to a ground truth annotation source, which is often a (semi-)automatic extraction pipeline itself. Finally, (iii) using pixel-level supervision only can lead to noisy, overly complex and geometrically inconsistent polygons compared to official cadastral maps (Xu et al., 2023).

To address these issues, we propose P^3, a dataset that provides two input modalities and precise vectorized building outlines collected across three continents (see Fig. 1). P^3 includes the traditional image modality with aerial ortho- or near-nadir images, and an aerial LiDAR modality in the form of 3D point clouds. Aerial LiDAR point clouds become increasingly openly available from National Mapping Agencies (NMAs) at the scale of entire countries (IGN, 2024, TGD, 2024, ahn.nl, 2024) and can thus constitute a complementary data source to imagery. In particular, aerial LiDAR point clouds provide direct height information with decimeter-level accuracy (TGD, 2024), and their acquisition is resilient to image distortion and meteorological and seasonal effects, as laser pulses can penetrate clouds and vegetation (Pearse et al., 2018). A second key specificity of our dataset is its scale and variability. The data cover a total area of 638km² across seven different regions on three different continents. Lastly, the dataset provides ground truth polygon annotations with geometric guarantees. In particular, polygons do not overlap with neighboring polygons and are either simple, or contain holes, *e.g.,* to represent courtyards.

In addition to the P^3 dataset, a second contribution of our work is a benchmark that analyzes the impact of the data modalities on the building vectorization problem. In particular, we propose a general pipeline to adapt the overpopulated image-based methods to the LiDAR modality and to the combination of image and LiDAR. We conduct experiments with three recent SOTA methods: FFL (Girard et al., 2021), HiSup (Xu et al., 2023) and Pix2Poly (Adimoolam et al., 2025). Besides the traditional accuracy metrics used in the field, we also introduce a metric that measures the geometric quality of polygons in terms of symmetry and regularity. This aspect is often ignored in the commonly-used evaluation protocol, but is crucial for practitioners who want to exploit the potential of building maps. Our dataset is designed to facilitate further research on multimodal building vectorization. It is available to download

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-349-2026

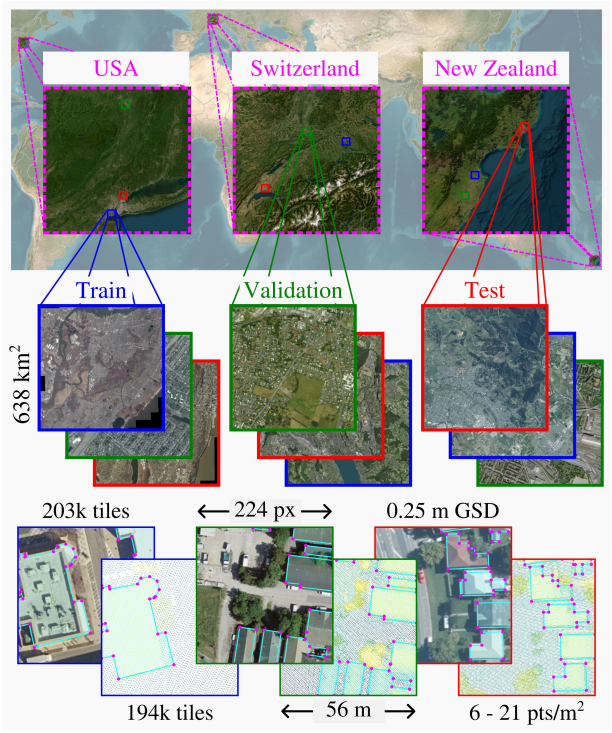


Figure 1. **Overview of P³.** We collect aerial images, LiDAR point clouds and vectorized building outlines from the USA, Switzerland and New Zealand[1]. We harmonize and tile the data to create a large benchmark dataset for building vectorization.

and can be used directly with a Python library for training, testing, and evaluating SOTA building vectorization methods.

2. Related work

Our review of previous work covers existing datasets and methods for building detection and polygonization.

2.1 Related datasets

Datasets for building detection can be grouped into three categories, with either the exploitation of 2D data modality, 3D data modality or a combination of the two. Tab. 1 lists the existing datasets for each of these three categories and provides data and annotation specifications for each of them.

2D datasets that exploit the image modality constitute the dominant category. Inria (Maggiori et al., 2017), CrowdAI (Mohanty et al., 2020), WHU (Ji et al., 2019) and Shanghai (DevGlobal, 2022) are the most popular datasets used in the literature. Data are satellite or aerial images acquired at near-nadir with a ground sampling distance (GSD) between 5 to 80 cm.

The Inria dataset (Maggiori et al., 2017) offers good urban diversity with a total area of 810km² covering different urban landscapes and countries, but requires a lossy conversion of the annotated pixel masks to polygons. Only the Microsoft (Microsoft, 2025) and Google (Google, 2025) datasets exhibit higher diversity by covering millions of square kilometers. However, these two datasets provide building polygons with a low precision and without the original images. Other datasets such as WBD (Nauata and Furukawa, 2020), RoofSat (Boyer et al., 2024) and RoofVect (Hensel et al., 2021) provide a more complete 2D roof wireframe in the form of planar graphs. These datasets are however small and poorly diversified. Note that several datasets on this list have been artificially enriched by data augmentation (Mohanty et al., 2020, Boyer et al., 2024, Hensel et al., 2021). For the CrowdAI dataset, this enrichment strategy has been shown to cause biases in the results (Adimoolam et al., 2023). Other datasets suffer from low-quality annotations typically due to either the low resolution of the original data source or a lack of harmonization (Labs, 2020, Etten et al., 2019).

3D. Building3D (Wang et al., 2023), City3D (Huang et al., 2022a) and RoofN3D (Wichmann et al., 2018), which exploit 3D data, mostly target the 3D building reconstruction problem from LiDAR point clouds (Bauchet et al., 2024). Each point cloud represents a single building and not an urban scene or a part of it. Unfortunately, annotated 3D models that are created by automatic methods or human operators present strong geometric inconsistencies that prevent these datasets from being used effectively for the building footprint extraction task.

2D and 3D. Only a few datasets present 2D and 3D modalities, *i.e.* Roof3D (Schuegraf et al., 2023), Potsdam, and Vaihingen (Rottensteiner et al., 2012). These datasets are relatively limited in size and urban diversity. They also do not offer vectorized annotations, but pixel masks only. For Roof3D and Potsdam, the complementary modality is a digital surface model (DSM) which provides 2.5D information only. In contrast, our dataset combines both images and 3D LiDAR at larger scale in various urban landscapes. In particular, it is 400 times the size of the Vaihingen dataset and provides 2D outline polygons as annotations.

2.2 Vectorization methods

Methods for detecting and polygonizing building outlines largely exploit the image modality. They generally fall into two categories: hybrid methods and direct building polygon prediction.

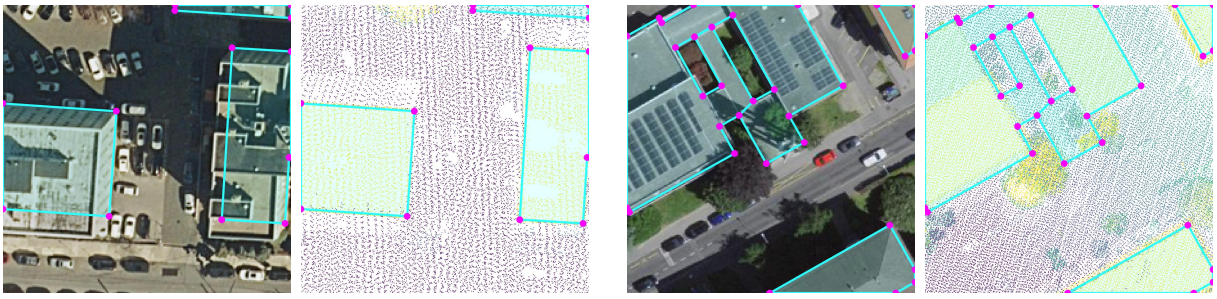
Hybrid methods typically involve multiple stages, often combining segmentation or detection with polygon extraction and refinement steps. Many recent approaches in this category incorporate structured priors, semantics, or preexisting databases. For instance, semantically encriched point clouds can be used to validate and correct existing building databases of NMAs (Roche, 2024), and elevation information provides geometric primitive clues for building rectangles extraction (Ortner et al., 2007). ASIP (Li et al., 2020) splits and merges cells of a polygonal partition according to a deep semantic segmentation map to generate compact polygons. Frame Field Learning (Girard et al., 2021) directly learns the directional fields that align with building boundaries, and obtains more regular polygons for urban scenes. HiSup (Xu et al., 2023) tackles this challenge by integrating directional fields, semantics, and geometry learning with hierarchical supervision to enhance the geometry precision. The sophisticated hierarchical design preserves a good balance between complexity and fidelity.

Direct building polygon prediction approaches typically leverage advanced neural network architectures to process aerial or satellite imagery and directly output vectorized building polygons without intermediate raster representations. A common strategy is to extract vertices from heatmaps produced with convolutional or recurrent neural networks and apply a graph neural

[1] Licensed by Toitū Te Whenua Land Information New Zealand for reuse under CC-BY-4.0.

	Dataset	Data specifications					Annotation				
		Modality	Ground sampling	Area (km²)	Tiling	Diversity	Pixel label	2D outline polygon	2D roof wireframe	3D building wireframe	Origin
2D	Inria (Maggiori et al., 2017)	AI	0.3m	810	T+	+	✓				M
2D	SemiCity (Roscher et al., 2020)	SI	0.5m	50	T+	-	✓				H
2D	WHU (Ji et al., 2019)	AI	0.075m	450	T	-	✓				H
2D	CrowdAI (Mohanty et al., 2020)	SI	0.3m	2.3k	T	-		✓			H
2D	OpenCities (Labs, 2020)	DI	0.4-0.8m	415	T	+		✓			H
2D	Shanghai (DevGlobal, 2022)	SI	0.3m	21	T	-		✓			H
2D	WHU 2(Ji et al., 2019)	SI	0.45m	550	T	+		✓			H
2D	SpaceNet2 (Etten et al., 2019)	SI	0.3m	3k	T	+		✓			H
2D	UBC (Huang et al., 2022b)	SI	0.5-0.8m	66.1	T	+		✓			M
2D	Microsoft (Microsoft, 2025)	SI*	0.3-1.0m	N/A	-	++		✓			A
2D	Google (Google, 2025)	SI*	0.5m	58M	-	++		✓			M
2D	WBD (Nauata and Furukawa, 2020)	SI	0.3m	32	SB	- -			✓		H
2D	RoofSat (Boyer et al., 2024)	SI	0.3m	14.9	T	-			✓		H
2D	RoofVect (Hensel et al., 2021)	AI	0.1m	2	SB	-			✓		H
3D	Building3D (Wang et al., 2023)	ALS	30pts/m²	998	SB	+				✓	M
3D	City3D (Huang et al., 2022a)	ALS	4-50 pts/m²	N/A	SB	+				✓	A
3D	RoofN3D (Wichmann et al., 2018)	ALS	4.7 pts/m²	1010	SB	-				✓	A
2D+3D	Roof3D (Schuegraf et al., 2023)	AI+DSM	0.3m	22.4	T	-	✓				M
2D+3D	Potsdam (Rottensteiner et al., 2012)	AI+DSM	0.05m	3.4	T	-	✓				H
2D+3D	Vaihingen (Rottensteiner et al., 2012)	AI+ALS	0.08m, 4pts/m²	1.5	T	-	✓				H
2D+3D	**P³ (ours)**	AI+ALS	0.25m, 16pts/m²	638	T	+		✓			M

Table 1. **Related datasets.** Modalities typically include satellite images (SI), aerial images (AI), drone images (DI), airborne LiDAR scans (ALS) and digital suface models (DSM). SI* specfies datasets that use satellite imagery for their annotation but the imagery itself is not included. Tiling specifies whether data is a single building (SB) only, an image tile adapted to deep learning frameworks (T) or a large georeferenced tile bigger than 1000×1000 px (T+). Diversity approximates the urban variability of the dataset with four categories: ”- -” for a similar type of buildings from a single location, ”-” for several types of buildings from a single location, ”+” for several types of buildings from multiple locations in the world, and ”++” if it covers at least 1% of the total urban scenes in the world. ”H”, ”A” and ”M” refer to the annotation origin, *i.e.* either from human operators, automatic algorithms or a mix of these two.



(a) Building outline polygon

(b) Building outline polygon with holes

Figure 2. **Image/LiDAR tiles with polygon annotations.** In (a), we show an example tile of Jersey, NY, US. The image exhibits distortion leading to the *leaning building* effect. The annotation is nonetheless accurately placed on the base of the building. The LiDAR acquisition does not suffer from the *leaning building* effect. In (b), we show an example tile of Zurich, Switzerland. The building has two interior cutouts. The annotation has thus two interior rings connected to the exterior ring at their closest vertex.

network to form polygons (Li et al., 2019, Zorzi et al., 2022, Zorzi and Fraundorfer, 2023, Yang et al., 2023, Amrullah et al., 2025). PolyBuilding (Hu et al., 2023) shows promising performance by applying transformers that reduces the vertex redundancy in this CNN-GCN paradigm. Recently, image-to-vertex sequence methods (Zhang et al., 2024, Khomiakov et al., 2024b, Adimoolam et al., 2025) represented by Pix2Poly employ a transformer to produce building vertex tokens as sequences, eliminating the non-differentiable heatmap to vertex step. To reduce computational complexity while increasing the patch resolution some methods focus on a region of interest (Khomiakov et al., 2024a, Jiao et al., 2024). However, these methods mainly handle relatively simple and small urban areas.

Most of the aforementioned methods are designed to work with images and do not leverage 3D information. We present a method to adapt existing neural networks in the field to exploit both image and LiDAR modalities, making them more accurate in complex urban environments.

3. The P³ dataset

We collect data from nine cities across seven regions in three countries. P³ covers a total area of 638 km² and includes ap-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-349-2026 | © Author(s) 2026. CC BY 4.0 License.

Region			*Image*		*LiDAR*	*Building Annotations*		
Country	**City**	**Type**	**Split**	**GSD** [cm]	**Density** [pts/m²]	**Source**	**Type**	**Count**
United States	New York City	city	train	15	21	semi-automatic image extraction	2D outline polygon	88k
	Albany	residential	val		6			1.3k
	Yonkers	industrial	test		6			31k
Switzerland	Zurich	city	train	10	17	manual image extraction	2D roof wireframe	43k
	Basel	city	val		11			0.3k
	Jouxtens-Mézery	residential	test		13			11k
New Zealand	Twyford	rural	train	12.5	16	semi-automatic image extraction	2D outline polygon	31k
	Waipukurau	rural	val		19			1.6k
	Gisborne	city	test		19			16k

Table 2. **Data details**. We collect data from nine different cities in three countries, ranging from urban to rural areas, and including a variety of building types. The dataset is split into training, validation, and test sets. The image GSD varies between 10 and 15 cm, while the average LiDAR point density varies between 6 and 21 points per m^2. We downsample the images to a GSD of 25 cm for the benchmark dataset. The annotations are either semi-automatically extracted from orthorectified images or constitute cadastral data.

proximately 10B pixels, 10B LiDAR points, and 224k building polygon annotations. Table 2 provides details about characteristics of the collected input data. We perform a broad visual inspection along with random checks to ensure overall quality of the ground truth polygons and their alignment with the input images and LiDAR point clouds. To optimize the raw remote sensing and cadastral data for deep learning frameworks, we resample, tile and harmonize images, point clouds and building polygons.

Aerial images. The raw input images are near-nadir acquisition with a GSD between 10 and 15 cm. The images are not orthorectified, resulting in some distortion in off-nadir image regions (see Fig. 2a). To standardize the images across all regions, we resample them to a GSD of 25 cm and clip them to non-overlapping tiles of 224×224 pixels, equivalent to 56 m×56 m on the ground. This process yields a total of 203k image tiles.

Aerial LiDAR point clouds. The LiDAR point clouds exhibit a density ranging from 6 to 21 points per square meter across different regions. The point clouds are georeferenced to the same coordinate system as the images. We clip them to equally sized non-overlapping tiles of 56 m×56 m. This results in 194k LiDAR tiles – slightly fewer than image tiles due to missing LiDAR data in certain areas, such as deep water. On average, the LiDAR tiles contain between 20k and 70k points. We preserve the original point cloud density without resampling. Additionally, we do not color the point clouds, as it would first require to orthorectify the images before projecting the image colors on them.

Building outline polygon annotations. Our dataset includes annotations as 2D polygons representing the orthogonal projections of the complete building perimeter onto the ground. For buildings without roof overhang, these polygons typically coincide with the building footprint – the 2D polygon where the building touches the ground. In off-nadir image regions, roof outlines may appear slightly offset from the building footprint due to image distortion (see Fig. 2a). We do not provide annotations for roof partitions or interior walls, as this information is only available for the Switzerland part of the dataset. Where two or more building outline polygons overlap in the input data we merge them into a single polygon. We clip the building polygons to match the same tile dimensions as the image and LiDAR tiles. Polygons spanning multiple tiles are split at the tile boundaries. We then convert the polygon annotations to the widely adapted MS-COCO format (Lin et al., 2014).

In the urban environment, larger buildings commonly feature interior cutouts such as courtyards, lightwells, air shafts, atriums, passageways, or alleys. The MS-COCO format does not directly support storing interior rings as polygons with holes in vector format. Additionally, most current deep learning based vectorization methods do not specifically address interior ring prediction. We thus adopt the convention established by Xu *et al* (Xu et al., 2023) to connect interior and exterior rings of a polygon at their closest vertex (see Fig. 2b). The upside of this approach is that it allows deep networks to predict interior rings without architecture modifications. However, it can create invalid geometry through self-intersection and complicate the explicit distinction between exterior and interior rings. To facilitate more robust solutions for future work, we include an additional attribute *hashole*, in the MS-COCO annotation, enabling distinction between interior and exterior rings.

4. Experiments

We now present the baseline methods used for comparison, as well as evaluation metrics used to assess their performance. We then present two experiments that introduce the characteristics of our dataset: an ablation study on input modalities on a subset of the dataset and a multimodal building outline prediction on the full dataset.

4.1 Baseline methods

We test the following three methods on our dataset. We re-implement all methods in one common Python library that we make publicly available.

FFL (Girard et al., 2021) employs a CNN-based decoder to predict per-pixel orientation (frame) fields and segmentation masks. Polygonization is performed by tracing the predicted frame fields to extract structured, direction-aligned polygon boundaries, followed by geometric regularization.

HiSup (Xu et al., 2023) predicts (i) a point feature map for detecting convex and concave polygon vertices, (ii) an edge feature map in the form of attraction field maps, (iii) and a segmentation mask to represent the area enclosed by the polygon. These outputs are jointly processed and input into an optimization pipeline to reconstruct the final polygons.

Pix2Poly (Adimoolam et al., 2025) uses a transformer-based decoder to predict a sequence of vertex coordinates along with

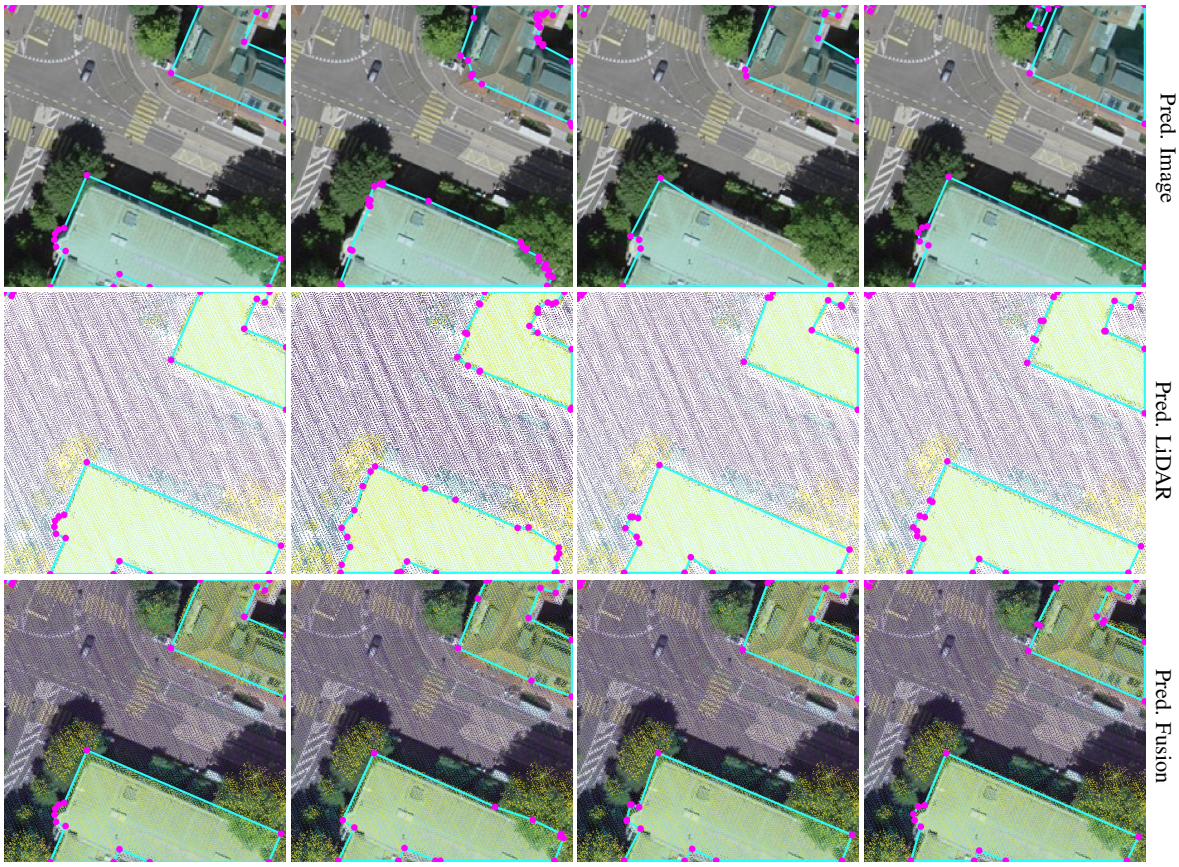


Figure 3. **Modality ablation.** We present a sample tile displaying predicted and ground truth building polygons from the Switzerland subset. The first column shows ground truth polygons, while subsequent columns show predicted polygons from baseline models trained on different input modalities, *i.e.* images only (first row), LiDAR only (second row), and the fusion of image and LiDAR data (third row). Note, in the bottom right corner, where a tree obscuring a building corner adversely affects the image-only prediction, while LiDAR-only and multimodal polygon predictions remain unaffected by this occlusion. Across all models, polygons predicted using multimodal inputs demonstrate superior simplicity and accuracy, especially with Pix2Poly.

a permutation matrix ordering and grouping them to produce polygons. This enables direct, end-to-end polygon reconstruction without relying on intermediate per-pixel segmentations.

We use the following image encoder, point cloud encoder and fusion encoder for all methods.

Image encoder. Our image encoder is a standard Vision Transformer (ViT) (Dosovitskiy et al., 2021) pretrained on ImageNet (Deng et al., 2009) with image input size of 224×224 pixels and a patch size of 8×8 pixels. The transformer extracts patch-level features without global pooling, thus producing a sequence of 784 patch embeddings. For processing these embeddings with the CNN-based decoders of Frame Field Learning (FFL) (Girard et al., 2021) and HiSup (Xu et al., 2023) we reshape and upsample the patches into a spatial map.

Point cloud encoder. We design a novel point cloud encoder to process aerial LiDAR point clouds with a variable point count and density based on a combination of PointPillars (Lang et al., 2019) and ViT. We first scale the point cloud to a fixed height of 1 m. We then voxelize the point cloud into pillars with size 2×2×1 m. This results in 784 pillars, aligned with the 784 patch embeddings of the ViT. We use a maximum number of points per voxel of 64. We then extract pillar features for each voxel (Lang et al., 2019) and process them with a small PointNet (Charles et al., 2017). To further process the features with an attention module we use the same ViT architecture as for the image encoder by replacing the image patch embedding with our point cloud embedding layer. The output of our point cloud encoder is a sequence of 784 patch embeddings, which are then directly input to the Pix2Poly (Adimoolam et al., 2025) decoder, or reshaped and upsampled to a spatial map for processing with the CNN-based decoders of FFL (Girard et al., 2021) and HiSup (Xu et al., 2023).

Fusion encoder. Setting the point pillar size to match the image patch size (8×8 pixels on 25 cm imagery translates to a 2×2 m point pillar size) allows us to design a fusion encoder that concatenates the image and point cloud features along the channel dimension. The resulting feature map is then processed with a small 2D CNN to reduce the channel dimension, and again fed into the same ViT architecture as before.

4.2 Metrics

We use three different types of metrics to evaluate the performance of the baseline methods: (i) boundary- and area-based accuracy and completeness metrics, (ii) complexity metrics, and (iii) efficiency metrics, *i.e.* average prediction time and the number of trainable parameters of the models. All metrics are computed per tile and then averaged over all tiles.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-349-2026 | © Author(s) 2026. CC BY 4.0 License.

		Boundary				*Area*			*Complexity*	*Efficiency*	
Model	**Modality**	**POLIS [m] ↓**	**CD [m] ↓**	**HD [m] ↓**	**MTA [°] ↓**	**AP ↑**	**AR ↑**	**IoU ↑**	**NR=1**	**Time [s] ↓**	**Params ↓**
ViT (Dosovitskiy et al., 2021) + **FFL** (Girard et al., 2021)	**Image**	3	2.7	12	41	0.275	0.46	0.839	0.847	0.532	23.7M
	LiDAR	2.35	1.94	9.66	44.8	0.359	0.545	0.87	0.829	0.582	23.7M
	Fusion	2.14	1.88	8.9	39.7	0.376	0.569	0.877	0.874	0.58	26.4M
ViT (Dosovitskiy et al., 2021) + **HiSup** (Xu et al., 2023)	**Image**	2.46	2.48	11.4	35.2	0.287	0.493	0.85	0.885	0.124	30.8M
	LiDAR	2.05	2.01	9.58	37	0.347	0.54	0.87	0.882	0.165	30.8M
	Fusion	1.91	1.9	8.94	35.2	0.355	0.568	0.872	0.89	0.129	33.5M
ViT (Dosovitskiy et al., 2021) + **Pix2Poly** (Adimoolam et al., 2025)	**Image**	2.46	2.5	10.8	34.3	0.317	0.492	0.845	0.906	1.44	31.9M
	LiDAR	1.88	1.9	8.5	34.1	0.379	0.552	0.869	0.913	1.15	31.9M
	Fusion	1.8	1.82	8.16	33.4	0.398	0.578	0.87	0.915	1.15	34.6M

Table 3. **Modality ablation**. We compare the baseline models trained and tested on different modalities of the Switzerland subset. For each metric, we highlight the best and second best scores.

	Boundary				*Area*			*Complexity*		
Model	**POLIS [m] ↓**	**CD [m] ↓**	**HD [m] ↓**	**MTA [°] ↓**	**AP ↑**	**AR ↑**	**IoU ↑**	**C-IoU ↑**	**NR=1**	**DoF ↓**
ViT (Dosovitskiy et al., 2021) + **FFL** (Girard et al., 2021)	2.41	2.11	9.34	40.4	0.286	0.43	0.832	0.763	0.848	0.808
ViT (Dosovitskiy et al., 2021) + **HiSup** (Xu et al., 2023)	2.07	2.03	9.1	36.5	0.309	0.474	0.844	0.789	0.881	0.758
ViT (Dosovitskiy et al., 2021) + **Pix2Poly** (Adimoolam et al., 2025)	1.91	1.92	8.17	34.4	0.347	0.49	0.842	0.801	0.901	0.741

Table 4. **Multimodal polygon prediction** of baseline models with fusion encoder on the full dataset.

Boundary- and area-based metrics. (i) The intersection over union (IoU) metric provides a combined measure of accuracy and completeness of the predicted polygons. However, it does not provide a good measure of the boundary distance between the predicted and ground truth polygons, as it is mainly influenced by the polygon area (Cheng et al., 2021). (ii) Average Precision (AP) and average recall (AR) computed over different IoU thresholds measure the accuracy and completeness of the predicted polygon sets (Lin et al., 2014).

(iii) The POLIS metric (Avbelj et al., 2015) measures the symmetric distance between each predicted polygon vertex and its closest point on the ground truth polygon boundary, and vice versa. (iv, v) The Hausdorff and Chamfer distance (HD and CD) measure the average and maximum boundary distance of predicted and ground truth polygons, irrespective of vertex locations. (vi) Finally, the maximum tangent angle (MTA) (Girard et al., 2021) measures the maximum angular error between the predicted and ground truth polygon edges. POLIS, HD, CD and MTA are computed for ground truth and predicted polygon pairs with a minimum IoU of 0.5. They are important measures for practitioners because they correlate with a high visual similarly of predicted and ground truth polygons.

Complexity metrics. (i) NR measures the relative difference of the number of predicted and ground truth vertices. (ii) C-IoU, defined as the product of NR and IoU measures the similarity between predicted and ground truth polygons in terms of both shape *and* complexity. And finally, (iii) the normalized degree of freedom (DoF) measure the regularity and symmetry of polygon edges (Boyer et al., 2024). A lower DoF score indicates that the predicted polygons are simpler and more regular, while a higher DoF score indicates that the predicted polygons are more complex and less regular.

4.3 Results

Modality ablation. We first conduct an ablation study on a subset of our dataset containing only data from Switzerland. This subset is less challenging than the full dataset because it exhibits lower variability. We use it to evaluate the impact of different input modalities on the polygon prediction task. We train each baseline method with a point, an image, and a fusion encoder. Tab. 3 demonstrates that models utilizing LiDAR data consistently predict more complete (∼0.55 AR) and accurate polygons (0.35-0.38 AP) compared to models using only images (0.46-0.49 AR, 0.28-0.32 AP). The fusion encoder efficiently integrates the strengths of both input modalities to achieve the best performance for all models. Interestingly, for angular error MTA and vertex ratio NR, FFL and HiSup perform better for image-predicted polygons. This can be attributed to the presence of more prominent corner and edge features in the images compared to LiDAR point clouds. The recently proposed end-to-end polygon prediction approach, Pix2Poly, exhibits significantly slower runtime compared to hybrid methods. Fig. 3, confirms the higher accuracy and geometry simplicity of LiDAR- and multimodal-predicted polygons compared to those predicted from images alone.

Multimodal building outline prediction. We now evaluate the performance of the baseline methods on the full dataset. Each model is trained with a fusion encoder using both images and LiDAR point clouds as input. Evaluation metrics for the full dataset are shown in Tab. 4.

The full dataset is more challenging due to its greater variability, leading to a drop in all metrics. The mean IoU decreases by about 3 percentage points compared to the Switzerland subset. Pix2Poly produces the best polygons in terms of complexity and accuracy, exhibiting the best or second best results for all metrics. Fig. 4 shows examples of Pix2Poly's more compact and accurate polygon predictions compared to other methods. However, the figure also shows that Pix2Poly still has some shortcomings. The prediction in the first row shows how an incorrect connection in the building outline leads to a significant boundary error. The second row reveals entirely missing building outlines for the Pix2Poly prediction, and the third row shows a missing interior cutout. Notably, the HiSup-predicted polygon is the only one that accurately captures the hole present in the polygon outline. A common limitation of the CNN-based methods is their tendency to produce excessive vertices for building polygons, falling to fully address the regularity and symmetry present in urban areas. In terms of relative performance, our results align with those reported on other datasets (Maggiori et al., 2017, DevGlobal, 2022, Labs, 2020). However, using identical encoders across all methods results in smaller performance gaps between the tested methods.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-349-2026 | © Author(s) 2026. CC BY 4.0 License.

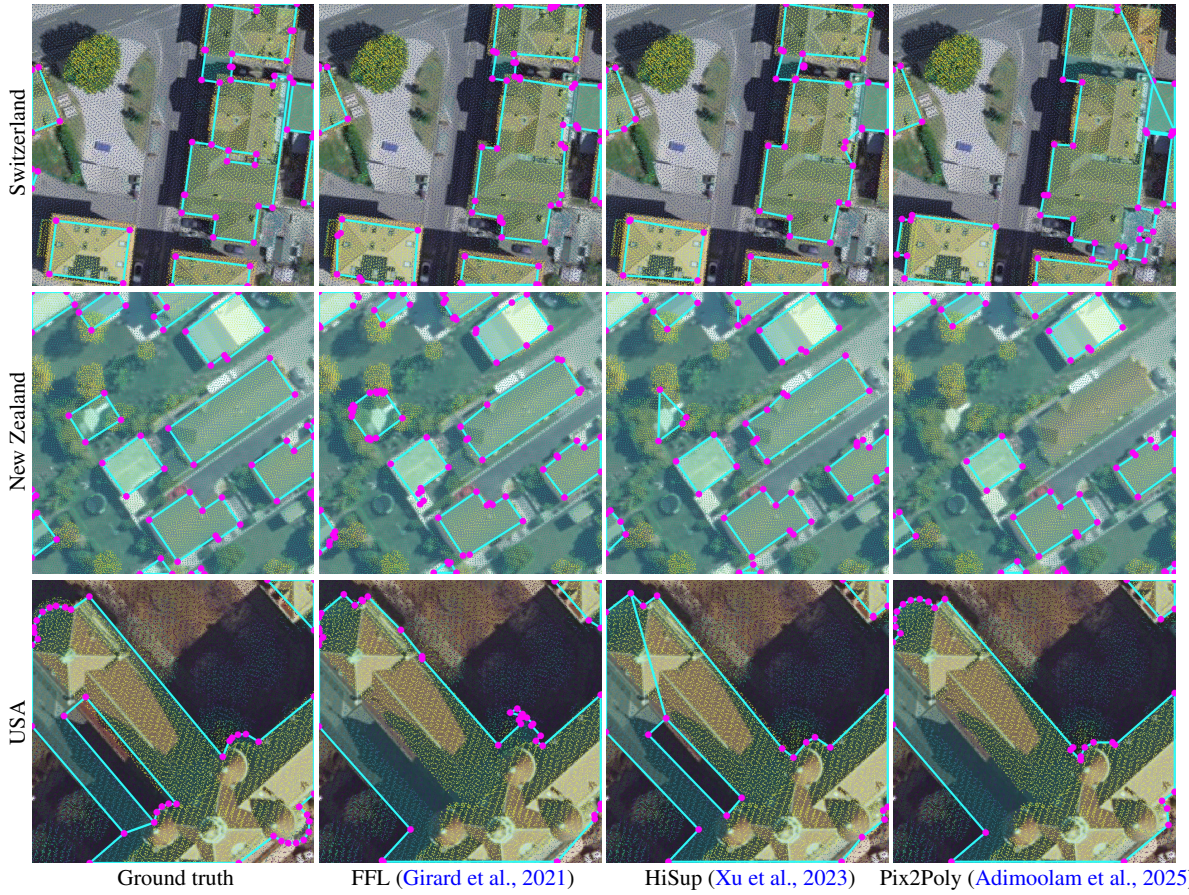


Figure 4. **Multimodal polygon prediction.** We show ground truth and predicted building outlines from our full dataset, from top to bottom for Swizerland, New Zealand [1] and the USA, with input LiDAR point clouds superimposed on aerial images. The first column shows the ground truth reference polygons, while the second to forth column present predicted polygons generated by FFL (Girard et al., 2021), HiSup (Xu et al., 2023) and Pix2Poly (Adimoolam et al., 2025) utilizing both input modalities.

5. Conclusion, Limitations and Future Work

We present a new dataset and benchmark for building vectorization from high resolution aerial images and LiDAR pointclouds. Because such high resolution data is expensive to gather, it is mainly available from developed countries. Currently, our dataset is limited to data from Switzerland, the USA and New Zealand. We plan to expand it to other countries in the future if new data becomes available. At its current state, our data set is focused on developed countries. For example, the dataset does not include informal or improvised buildings found in both rural and urban areas of underdeveloped countries. Furthermore, the dataset only includes building outlines, and does not provide annotations for roof partitions, building heights or other 3D annotations. Creating 3D annotations in a reliable manner is a challenge that we plan to tackle in the future to increase the application spectrum of our dataset.

We also present a benchmark of state-of-the-art methods for building vectorization. Equipped with recent image and LiDAR encoders all tested methods produce accurate and complete building polygons. We show that (i) LiDAR point clouds are a valuable source for building outline prediction, that (ii) the combination of image and LiDAR point clouds leads to the best predictions, and that (iii) our dataset is challenging enough to require multimodal approaches.

References

Adimoolam, Y. K., Chatterjee, B., Poullis, C., Averkiou, M., 2023. Efficient Deduplication and Leakage Detection in Large Scale Image Datasets with a focus on the CrowdAI Mapping Challenge Dataset. https://arxiv.org/abs/2304.02296.

Adimoolam, Y. K., Poullis, C., Averkiou, M., 2025. Pix2poly: A sequence prediction method for end-to-end polygonal building footprint extraction from remote sensing imagery. *WACV*.

ahn.nl, 2024. Actueel Hoogtebestand Nederland. *ahn.nl*. https://www.ahn.nl/. Accessed: 2025-05-07.

Amrullah, C., Panangian, D., Bittner, K., 2025. Polyroof: precision roof polygonization in urban residential building with graph neural networks. *JURSE*.

Avbelj, J., Müller, R., Bamler, R., 2015. A Metric for Polygon Comparison and Building Extraction Evaluation. *IEEE Geoscience and Remote Sensing Letters*.

Bauchet, J.-P., Sulzer, R., Lafarge, F., Tarabalka, Y., 2024. Simplicity: Reconstructing buildings with simple regularized 3d models. *CVPRW*.

Boyer, M., Youssefi, D., Lafarge, F., 2024. Linefit: A geometric approach for fitting line segments in images. *ECCV*.

Charles, R. Q., Su, H., Kaichun, M., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*.

Cheng, B., Girshick, R., Dollár, P., Berg, A. C., Kirillov, A., 2021. Boundary IoU: Improving object-centric image segmentation evaluation. *CVPR*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *CVPR*.

DevGlobal, 2022. Ramp Building Footprint Training Dataset - Shanghai, China, Version 1.0. http://dx.doi.org/10.34911/rdnt.grvh9e. Accessed: 2025-05-09.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929.

Etten, A. V., Lindenbaum, D., Bacastow, T. M., 2019. SpaceNet: A Remote Sensing Dataset and Challenge Series. https://spacenet.ai/spacenet-buildings-dataset-v2/.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. *CVPR*.

Google, 2025. The Open Buildings dataset. https://sites.research.google/gr/open-buildings/. Accessed: 2025-05-09.

Hensel, S., Goebbels, S., Kada, M., 2021. Building roof vectorization with ppgnet. *3D GeoInfo Conference*.

Hu, Y., Wang, Z., Huang, Z., Liu, Y., 2023. PolyBuilding: Polygon transformer for building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Huang, J., Stoter, J., Peters, R., Nan, L., 2022a. City3D: Large-Scale Building Reconstruction from Airborne LiDAR Point Clouds. *Remote Sensing*.

Huang, X., Ren, L., Liu, C., Wang, Y., Yu, H., Schmitt, M., Hänsch, R., Sun, X., Huang, H., Mayer, H., 2022b. Urban building classification (ubc) - a dataset for individual building detection and classification from satellite imagery. *CVPRW*.

IGN, 2024. Lidar HD. *IGN*. https://geoservices.ign.fr/lidarhd. Accessed: 2025-05-07.

Ji, S., Wei, S., Lu, M., 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*.

Jiao, W., Persello, C., Vosselman, G., 2024. PolyR-CNN: R-CNN for end-to-end polygonal building outline extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Khomiakov, M., Andersen, M. R., Frellsen, J., 2024a. Gast: Geometry-aware structure transformer. *WACV*.

Khomiakov, M., Andersen, M. R., Frellsen, J., 2024b. Geoformer: A multi-polygon segmentation transformer. *BMVC*.

Labs, G., 2020. Open Cities AI Challenge Dataset, Version 1.0. https://doi.org/10.34911/rdnt.f94cxb. Accessed: 2025-05-09.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. Pointpillars: Fast encoders for object detection from point clouds. *CVPR*.

Li, M., Lafarge, F., Marlet, R., 2020. Approximating shapes in images with low-complexity polygons. *CVPR*.

Li, Z., Dirk Wegner, J., Lucchi, A., 2019. Topological map extraction from overhead images. *ICCV*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *ECCV*.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *IGARSS*.

Microsoft, 2025. The Global ML Building Footprint Dataset. https://github.com/microsoft/GlobalMLBuildingFootprints. Accessed: 2025-05-07.

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

Nauata, N., Furukawa, Y., 2020. Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. *ECCV*.

Ortner, M., Descombes, X., Zerubia, J., 2007. Building Outline Extraction from Digital Elevation Models Using Marked Point Processes. *IJCV*.

Pearse, G. D., Dash, J. P., Persson, H. J., Watt, M. S., 2018. Comparison of high-density LiDAR and satellite photogrammetry for forest inventory. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Roche, F., 2024. 3D AI in the Lidar HD Production Process. *LIDAR Magazine*. https://lidarmag.com/2024/05/04/3d-ai-in-the-lidar-hd-production-process/. Accessed: 2025-04-20.

Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J., 2020. Semcity toulouse: A benchmark for building instance segmentation in satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS Benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., Bittner, K., 2023. Roof3d: A real and synthetic data collection for individual building roof plane and building sections detection. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

TGD, 2024. swissSURFACE3D. *Federal Office of Topography swisstopo*. https://lidarmag.com/2024/05/04/3d-ai-in-the-lidar-hd-production-process/. Accessed: 2025-05-07.

Wang, R., Huang, S., Yang, H., 2023. Building3d: A urban-scale dataset and benchmarks for learning roof structures from point clouds. *ICCV*.

Wichmann, A., Agoub, A., Kada, M., 2018. ROOFN3D: Deep Learning Training Data For 3D Building Reconstruction. *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Xu, B., Xu, J., Xue, N., Xia, G.-S., 2023. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Yang, B., Zhang, M., Zhang, Z., Zhang, Z., Hu, X., 2023. Topdig: Class-agnostic topological directional graph extraction from remote sensing images. *CVPR*.

Zhang, T., Wei, S., Zhou, Y., Luo, M., Yu, W., Ji, S., 2024. P2PFormer: A Primitive-to-polygon Method for Regular Building Contour Extraction from Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. Polyworld: Polygonal building extraction with graph neural networks in satellite images. *CVPR*.

Zorzi, S., Fraundorfer, F., 2023. Re:polyworld - a graph neural network for polygonal scene parsing. *ICCV*.