Building Footprint Regularization : From Vectorization to Deep Learning

This manuscript (permalink) was automatically generated from kshitijrajsharma/building-regularization-research@7d5dede on July 31, 2025.

Authors

- **Kshitij Raj Sharma**
  0000-0002-2123-3917 · kshitijrajsharma · @kshitijrajsharma@mastodon.social
  Department of Geoinformatics, Paris Lodron University, Salzburg, Austria

✉ — Correspondence possible via GitHub Issues

Abstract

It often believed that many cartographic and Geographic Information System (GIS) applications require building footprints in the form of clean vector polygons, rather than raw raster masks, to facilitate direct use in maps and spatial analysis workflows [1]. However, outputs from automated methods, including those based on satellite imagery or LiDAR, often produce noisy or overly complex polygons with an excessive number of vertices. OpenStreetMap (OSM), frequently used as a reference, contains features that are not always naturally occurring and often reflect a more human-like interpretation, which is not the case when mapped automatically. Human mappers typically apply cartographic judgment, favoring orthogonality, symmetry, and geometric simplicity while digitizing buildings.

The primary objective of this study is to review the existing efforts to generate building footprints that tries to mimic human-cartographic quality. We focus on building footprint regularization, defined as the process of converting rough or noisy outlines into clean vector shapes that adhere to expected geometric constraints (e.g., straight edges, right angles). This process enhances both the visual and analytical quality of building data. A common approach involves using deep learning to generate building masks, followed by a postprocessing step to convert these masks into vector polygons. However, many existing methods either lack generalization across different geographies or fail to enforce sufficient regularity [2].

This review traces the evolution of 2D building footprint regularization techniques from early rule-based vectorization in the 1990s to recent deep learning models in the 2020s. The focus is specifically on planimetric (2D) building outlines, excluding full 3D reconstruction and roof modeling. For each generation of methods, we highlight core ideas, algorithms, and their suitability for integration into GIS workflows. We compare classical and deep learning-based methods in terms of accuracy, flexibility, cartographic quality, and real-world applicability. Where relevant, we emphasize how these methods can be used to improve or augment OpenStreetMap-style datasets, aligning outputs more closely with the standards of human-made map features.

Introduction

Manual mapping by humans typically creates clean, regular building shapes with straight walls and right angles. This is because human mappers naturally apply cartographic principles like symmetry and geometric simplicity when drawing buildings. In contrast, automated methods often struggle to produce such clean, map-ready outputs.The challenge of converting rough building outlines into clean vector shapes is known as building footprint regularization. This process has evolved significantly over the past three decades, from simple geometric rules in the 1990s to sophisticated deep learning models today. Early methods relied on mathematical techniques like the Hough Transform to detect straight lines and enforce right angles. Modern approaches use neural networks to learn building patterns directly from training data.

Our goal is to understand the strengths and limitations of different approaches and provide guidance for practitioners working on building extraction projects.The question remains: what makes a “good” building footprint? Should it have perfect right angles, accurately represent the real building shape, or simply look aesthetically pleasing on a map? Through this review, we explore how different methods address these competing requirements and discuss the ongoing challenges in automated building footprint extraction.

Literature Review

Geometric and Heuristic Methods (1990s - 2000s)

Edge Detection and Line Fitting:

Early building extraction in the 1990s relied on low-level image processing and geometric heuristics. For example, Huertas and Nevatia (1988) [3] developed a system to detect buildings in aerial images by finding rectangular clusters of edges (lines) and using shadow cues to distinguish buildings from other structures . Building polygons often consist of jagged lines. Guercke and Sester [4] use Hough-Transformation (Mathematically formalized by Duda, R.O., & Hart, P.E.[5]) to refine such polygons.

Those approach and similar ones could identify simple rectangular building footprints, but often produced polygons with jagged (bearing in mind they don't take into account the building shape itself rather the outline), noisy outlines. To clean such outlines, researchers applied line simplification algorithms from cartography, notably the Ramer–Douglas–Peucker algorithm : to remove small zig-zags and reduce vertex count while approximating the shape (which is still used to the date) [6/].

The Douglas–Peucker algorithm (originally from 1973) [7] became a common post-processing step to "compress" or simplify building polygon geometry.

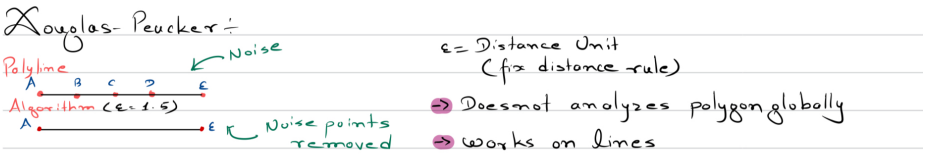

Figure 1: A simple illustration of Douglas-Peucker algorithm

Overall, early methods were largely rule-based: edges and corners were detected via image filters, and building shapes were assembled by connecting these primitives under geometric constraints defined by human experts.

Regularization via Hough Transform:

By the 2000s, more sophisticated heuristics were introduced to enforce regularity in building outlines. A prominent tool was the Hough Transform for line detection. Hough transform is a feature extraction method used in image analysis. Hough transform can be used to isolate features of any regular curve like lines, circles, ellipses, etc. Hough transform in its simplest from can be used to detect straight lines in an image.[8] For instance, Guercke and Sester [9] proposed a footprint simplification method that takes an initial digitized outline (which might be jagged) and uses a Hough Transform to identify the dominant line orientations; close-to-collinear segments are merged and adjusted by least-squares to align with those dominant directions [3].

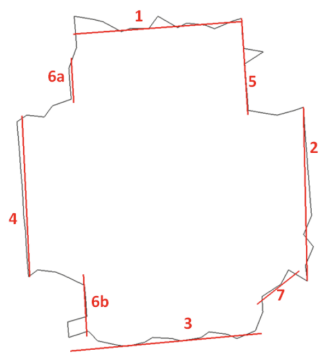

Figure 2: Initial hough transofrmation line segment explained by Guercke and Sester (2011)

The result is a cleaner, rectilinear footprint where spurious bends are straightened and most angles are ~90° or 180° [2].The Hough transform was applied to grouping line segments into two perpendicular families corresponding to a building's principal directions . An intial graph of line segments was constructed, pruned edges that lacked image contrast (assuming they were false boundaries), and then closed cycles were detected in the graph to form building polygons [2].

This yielded neatly rectangular footprints for buildings aligned to the two main axes, although the method was inherently limited to rectilinear structures. Tian and Reinartz (2013) [2] extended the idea to allow two arbitrary dominant orientations (not necessarily parallel/perpendicular to the image axes), enabling footprints with an oblique alignment (e.g. buildings rotated on the ground).

These Hough-based methods exemplify how prior knowledge of building shape (e.g. most buildings have parallel opposite walls and right-angle corners) was hard-coded into algorithms well before machine learning became common. The advantage was that the output polygons were regular by design : straight lines, right or consistent angles; making them immediately usable for mapping. However, the success of these methods depended on reliable low-level edge detection. In practice, missing or spurious line segments could cause incomplete or incorrect polygons. Methods like Cui's required a clear dominance of two perpendicular directions; complex or curved buildings, or those with more than two prevailing orientations, fell outside their scope. Hough transform is considered as a computational complex in terms of algorithm itself & often require postprocessing techniques like snapping/merging lines or form cycles to create valid polygons[8]

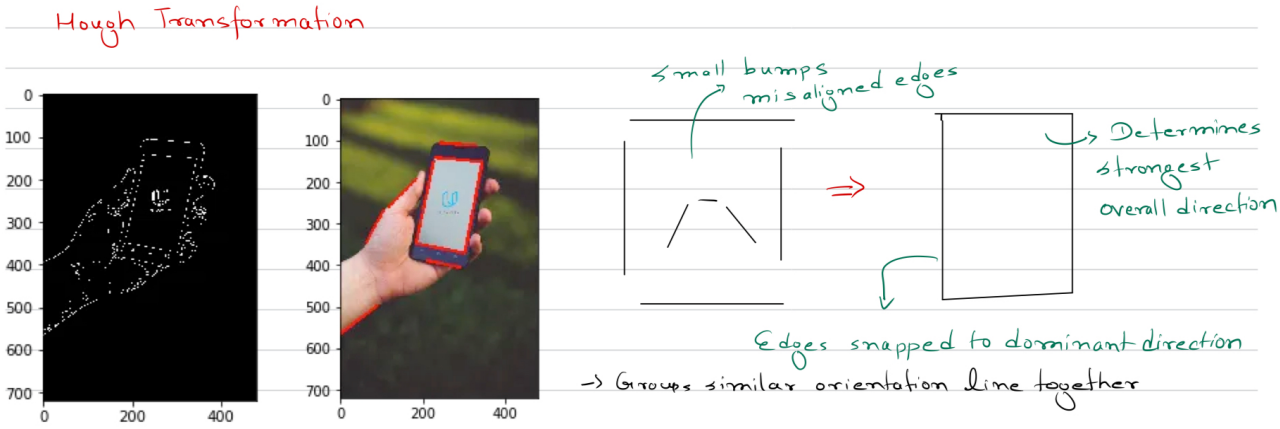

Figure 3: A simple Hough transformation explaination

Model-Based Fitting and Constraints:

Beyond Hough transforms, researchers explored explicit shape fitting. Zebedin et al. (2008)[2] introduced an approach to reconstruct building footprints by first detecting numerous line segments and then filtering and clustering these lines by orientation. Here initial lines are filtered by forming a histogram of orientation and then removing outliers. The filtered line directions are used to reconstruct the building with regular appearance. This approach is flexible, as it is not restricted to 90° angles.

This flexibility to allow non-90° angles was a strength like the footprint could, in principle, follow a building that isn't perfectly orthogonal but it still assumed buildings have a limited set of principal directions (which may not hold for very irregular architectures).

Other methods employed *snakes/active contours* and energy minimization to refine building shapes. For example, an active contour model (snakes) was applied to building roof images, optimizing an

energy that favored straight edges and right-angle corners. While this improved initial detections, A drawback of the proposed method is that the weighting functions favor right angles and therefore only work for buildings with simple rectangular shapes.

Bastian et al. (2014) combined data-driven edge detection with a global regularization step: they used an alpha shape algorithm to get an initial footprint from LiDAR point data, then a variant of Douglas–Peucker that was formulated as an energy minimization focusing on polygon complexity (number of vertices). The output was further processed in two modes one maximizing geometric accuracy, another maximizing topological simplicity to balance detail vs. regularity[10].

Energy Formulation : (Basically way to formulate errors on those lines detected) $E = \alpha Edist + \beta Eangle + \gamma Elength$

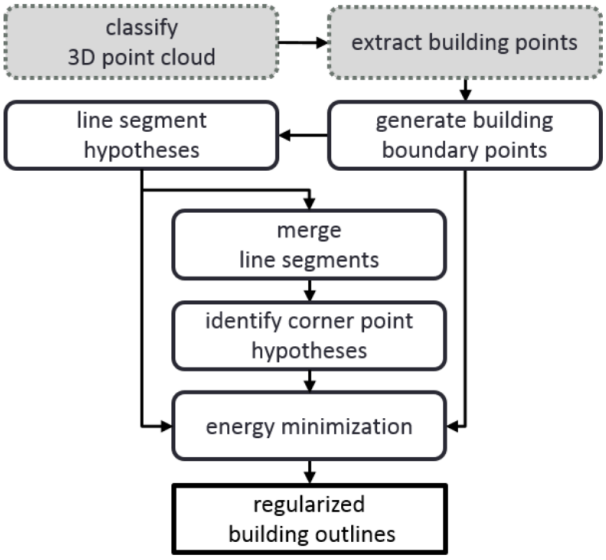

Figure 4: Workflow of building regularization using energy formulation by Albers (2016)

These model-fitting approaches introduced the idea of globally optimizing a footprint shape (e.g., via dynamic programming or least-squares) to satisfy regularity constraints.

Strengths and Limitations:

Traditional methods were mostly computationally lightweight and interpretable. They often ran in a couple of sequential steps (edge detection, line grouping, polygon formation) and could be tuned by adjusting thresholds (for line length, angle tolerance, etc.) When assumptions held e.g., a building was clearly rectangular and image data was clean ,these methods produced very clean footprints. For instance, a study by Guercke and Sester [4] showed that applying Hough-based regularization removed minor zig-zag artifacts and yielded impressively straight building edges.

However, these approaches struggled as building shapes grew more complex or data quality worsened. Irregular or curved buildings (round towers, L- or T-shaped footprints, etc.) did not fit neatly into a two-orientation assumption or a single rectangle model. Many algorithms were fragile: failing to detect a single key edge could cause entire sides of a polygon to be missed. They were also scenario-specific often tailored to isolated buildings with simple roofs and would require retuning for different environments or data sources. It is often said that while such classical methods work in some cases, they are “not applicable to many complex building structures” and they rely heavily on human-engineered features and parameters [3].

In summary, the pre-2010s state-of-the-art could produce “regular” building outlines under favorable conditions, but lacked the robustness and generality needed for broad, automated mapping tasks. These limitations set the stage for machine learning, which promised to learn building shape patterns directly from data and reduce the need for ad hoc rules.

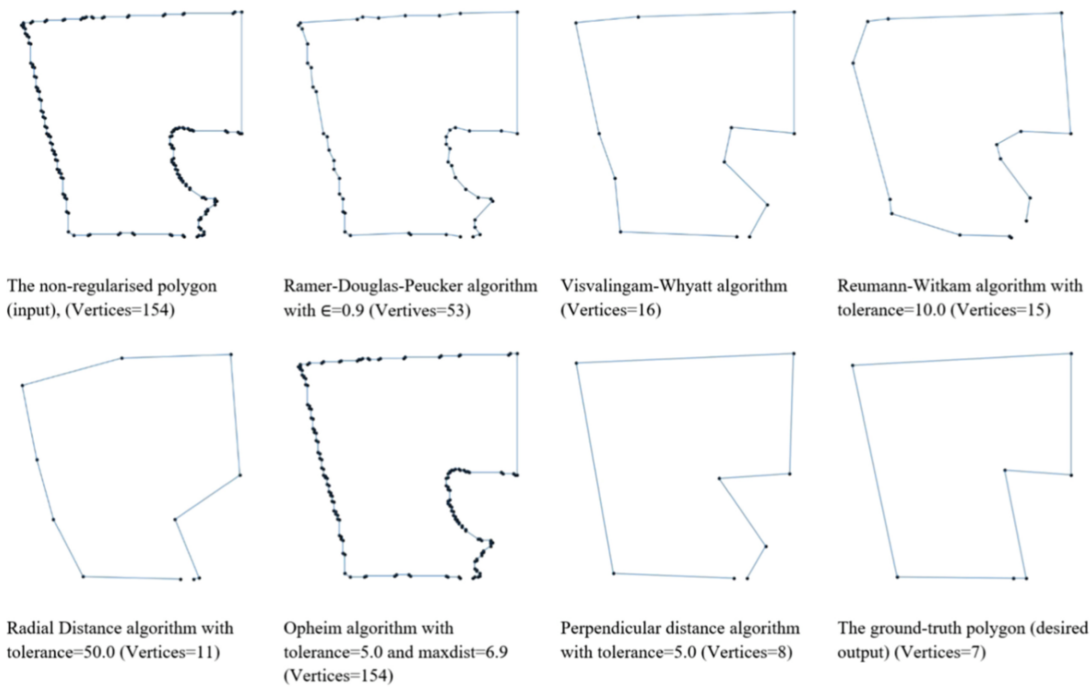

Figure 5: A comparison of traditional regularization algorithms on a noisy polygon in terms of node reduction, shape simplification, and edge smoothness [11]

Learning-Based Methods (2010s)

By the mid-2010s, the rise of deep learning fundamentally changed how building footprints were extracted. Instead of manually defining edges and shape rules, researchers began training convolutional neural networks (CNNs) to recognize buildings and output them in raster or vector form. The typical pipeline circa 2015–2017 was to use a semantic segmentation network (such as U-Net or DeepLab) to produce a binary mask of building pixels, then apply a vectorization algorithm to convert that mask into polygons [6].

This two-step approach : CNN segmentation followed by geometric post-processing was a direct evolution of earlier workflows, swapping out hand-coded image filters for learned CNN features. For example, Philipp et al. (2019) [3], mentioned that fully convolutional networks could outperform traditional techniques in detecting building regions from aerial images.

Once a clean building mask was obtained, off-the-shelf polygonization (e.g., marching squares to trace outlines) and Douglas–Peucker simplification would yield a polygon vector. A problem with this approach is that semantic segmentation models are unable to delineate the boundaries between objects of the same class. This means that a single polygon will be drawn around a group of buildings that share walls, such as a block of rowhouses. To handle this case, the semantic segmentation model can be replaced with an instance segmentation model such as Mask R-CNN. This model generates a separate raster mask for each instance of a class that is detected [6]. Beyong which additional smoothing or regularization was needed, and many practitioners continued to apply tolerance-based simplification or mild “squaring” adjustments to make the polygons map-ready.

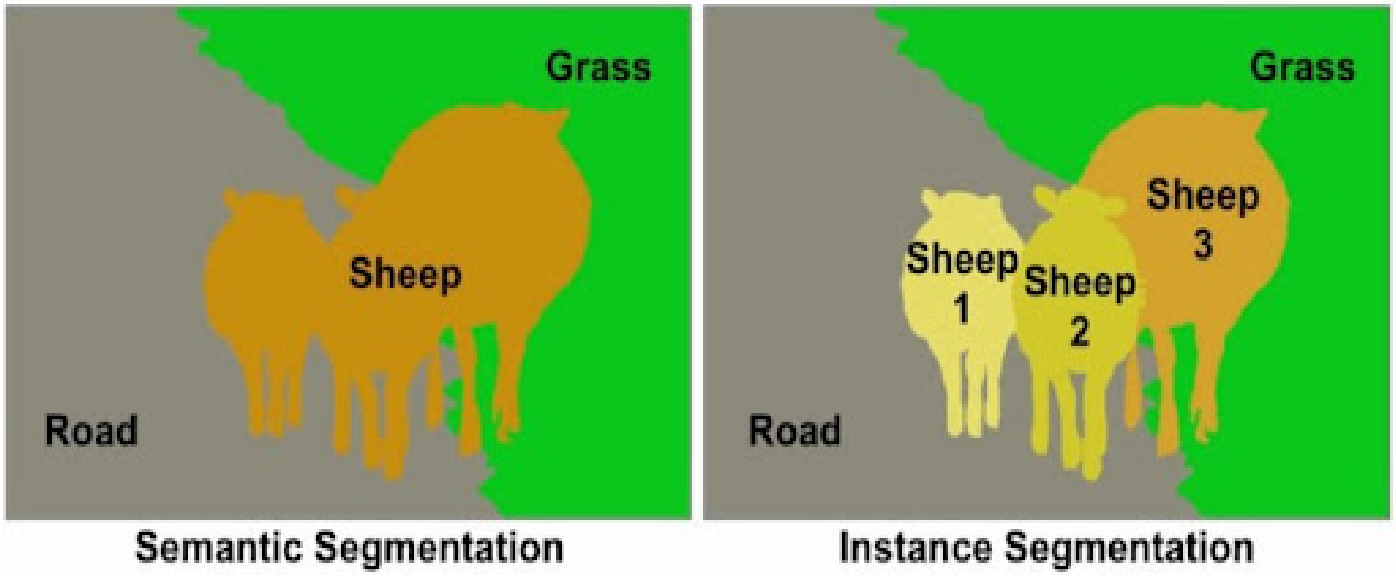

Figure 6: Semantic Segmentation to Instance Segmentation Aprooaches , source

Deep Structured Models (Active Contours)

A significant development in bridging classical regularization and deep learning was the integration of active contour models into neural networks. Zhang et al. (2018)[12] introduced Deep Structured Active Contours (DSAC), a hybrid approach where a CNN learns to predict the parameters of an active contour that locks onto building edges . In their framework, the network output is not a raster, but rather coefficients that define the shape and tension of an active contour (snake) which then deforms to fit the building boundary.

Gur et al. (2019) [13] extended this concept by iteratively updating a polygon outline in an end-to-end trainable manner. Their pipeline starts with an approximate polygon (like a coarse outline of the building) and uses a neural network to repeatedly adjust the vertices, analogous to how one would iteratively relax an active contour. While effective, the polygons produced by Gur et al. were not explicitly enforced to be rectilinear the focus was on aligning to image evidence, not necessarily making right angles.

Hatamizadeh et al. (2020) [14] proposed a multi-building active contour model: a CNN first predicts initial contours for many buildings in a scene, and then a learned energy function refines all of them simultaneously. This allowed processing dense urban scenes with many buildings at once, something earlier active-contour methods (which often assumed one building at a time) didn't handle. Hatamizadeh's model was end-to-end (it directly outputs vector polygons from an image), but like its predecessors, its regularization was implicit it preferred smooth, compact shapes but did not guarantee, say, all angles = 90°.

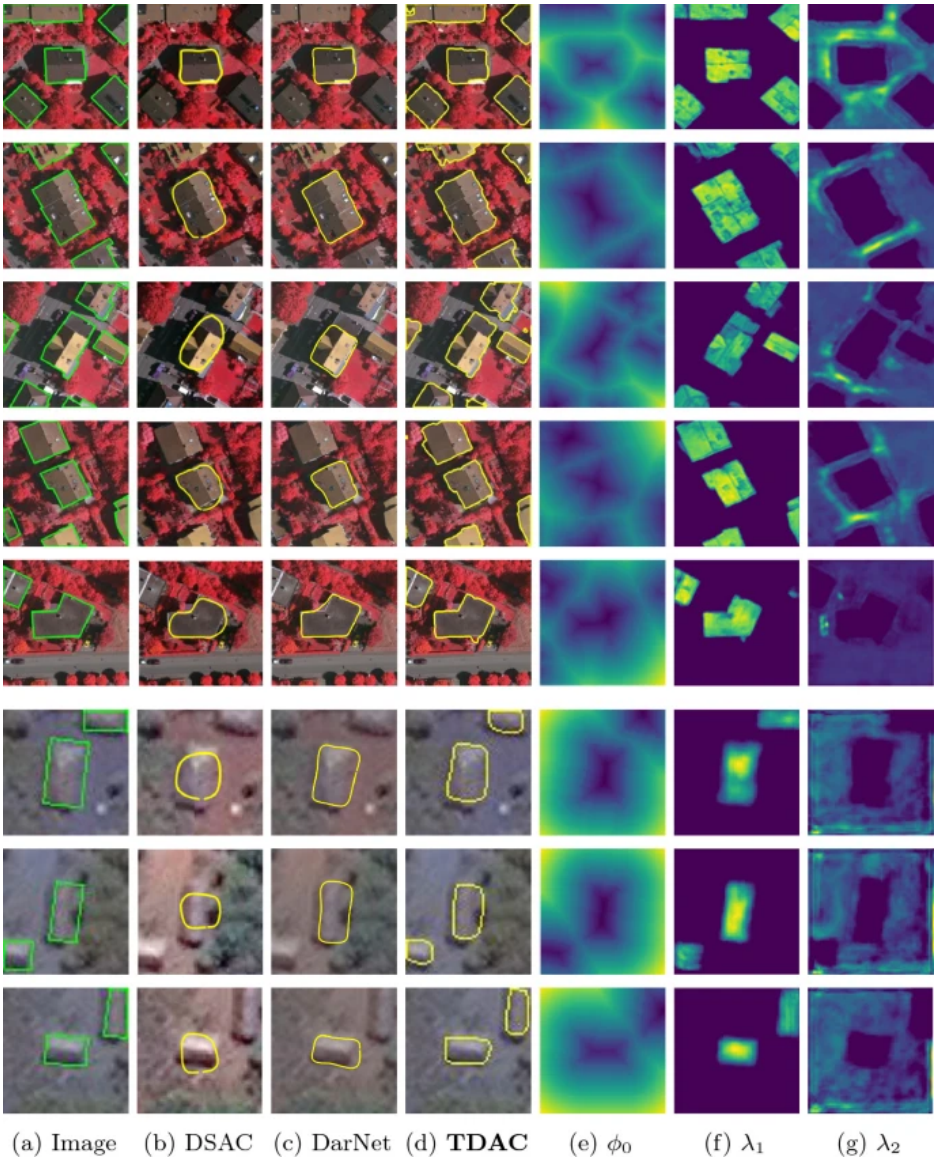

Figure 7: Comparative visualization of the labeled image and the outputs of DSAC, DarNet, and our TDAC for the Vaihingen (top) and Bing Huts (bottom) datasets. (a) Image labeled with (green) ground truth segmentation. (b) DSAC output. (c) DarNet output. (d) TDAC output. (e) TDAC's learned initialization map and parameter maps (f) and (g)

Source Code : DSAC , ACDRNet, DALS

Recurrent Vertex Prediction (Polygon RNNs) : PolyMapper

Instead of converting segmentation masks into polygons as a post-processing step, Recurrent Vertex Prediction models approach polygon extraction as a sequence prediction problem. In this framework, the model outputs a series of vertices one at a time, similar to writing out coordinate lists.

Polygon-RNN demonstrated that a recurrent neural network (RNN) could learn to draw object outlines by sequentially predicting polygon vertices, using image features to guide the process at each step.

A notable extension of this idea is PolyMapper, which integrates convolutional and recurrent modules in an end-to-end architecture. First, a CNN component (similar to Mask R-CNN) detects building instances and predicts coarse masks along with boundary and corner probability maps. Next, the most likely vertices from the corner map are selected and passed, together with image features, into an LSTM-based recurrent module. This module outputs vertices in sequence to trace the building outline, stopping when an end-of-sequence token is predicted, which signals the polygon should close.

Figure 8: A comparison of polygons generated by instance segmentation (left) and PolyMapper, which uses fewer vertices and preserves right angles (right) , source

This approach has distinct advantages: the RNN can learn to skip over minor irregularities, resulting in cleaner and simpler polygons with fewer vertices. It can also learn to favor structural regularities, such as right angles, due to its exposure to training data. PolyMapper demonstrated that such models produce more regular and human-like building footprints than traditional instance segmentation pipelines.

However, this modeling approach brings complexity. The network must learn when to terminate the sequence (when to stop adding vertexes), and the loss function must account for sequence prediction dynamics. Early Polygon-RNN models also faced issues such as generating self-intersecting polygons or incorrectly ordering vertices unless constraints were explicitly enforced.

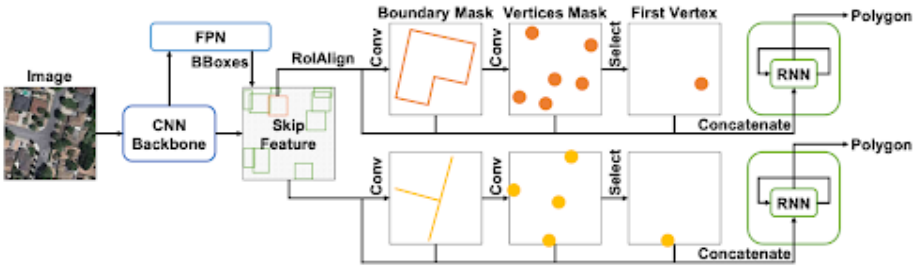

Figure 9: Overview of PolyMapper for Building and Roads : Source

By the end of the 2010s, two main deep learning approaches emerged for extracting building footprints from imagery:

1. Segmentation-based methods focused on generating accurate masks of buildings and then used advanced post-processing techniques such as learned active contours (“snakes”) to clean and regularize the shapes.

2. Direct polygon prediction methods aimed to output building outlines directly as sequences of vertices and edges, using models like recurrent neural networks (RNNs) or parameterized shape representations.

These approaches marked a significant improvement over older heuristic techniques. Convolutional neural networks (CNNs) could generalize better across diverse geographies and imaging conditions.

For instance, a model trained on buildings in one city could often perform reasonably well in another, whereas hand-tuned algorithms often failed when conditions changed.

Despite this progress, early deep learning models still had limitations. The building shapes they produced were often *almost* clean but not perfectly geometric for example, a nearly straight wall might still have a slight jitter in the predicted vertices. This lack of geometric precision posed challenges for GIS applications that require clean vector shapes.

Source Code : NA

Modern Deep Learning Approaches (2020s)

Polygonal Building Segmentation by Frame Field Learning

In addition to direct polygon prediction, researchers also explored ways to inject geometric structure into the deep learning process. One notable approach by Girard et al. (2021) [15] involved predicting not only a segmentation mask for buildings, but also a frame field : a set of orthogonal vectors at each pixel along the boundary indicating local edge directions.

A frame field acts like a directional map around a building's edges: it shows which way walls run and where corners should be. Using this directional information, the method first extracts a rough outline from the mask and then snaps and refines it by aligning it with the frame field and detected corner points. The post-processing pipeline includes multiple geometric steps such as skeletonization, corner detection, and line simplification each algorithmically defined rather than learned.

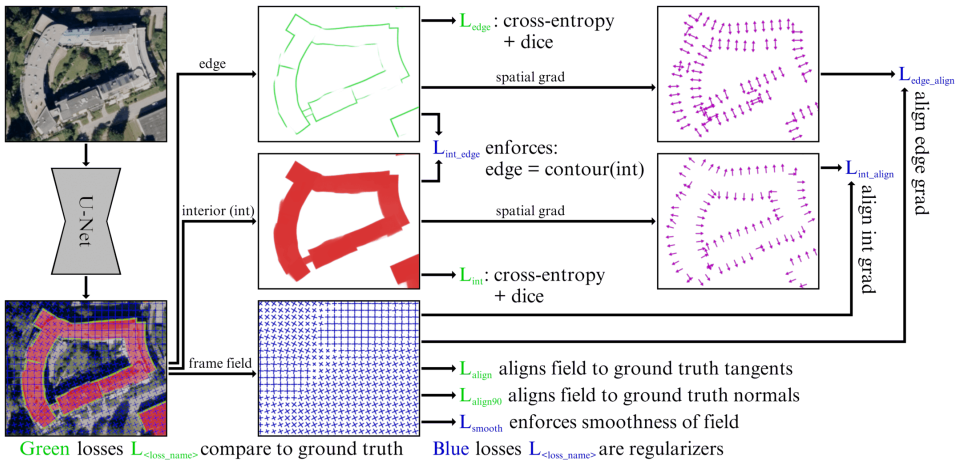

Figure 10: Explaination of framefield : Source

This method is able to handle buildings that are touching and buildings with courtyards by explicitly representing shared walls and generating polygons with holes. In addition, it runs about 10x faster than PolyMapper at inference time. The downside of this method is that the polygon extraction routine is complex and lacks the elegance of a model trained end-to-end.[6/]

In the last few years, deep learning models for building footprint regularization have reached new levels of maturity. These models are characterized by end-to-end training (the network learns to output a final polygon with minimal post-processing) and by the integration of architectural elements that explicitly handle the polygon's structure (such as graph neural networks, transformers, or differentiable geometric algorithms). Below we highlight several state-of-the-art approaches, including CNN/RNN hybrids, graph-based models, and transformer-based models, and discuss how they improve upon prior methods.

Source Code : GitHub

PolyWorld: End-to-End Polygon Extraction via CNN and GNN

PolyWorld [16] introduces a novel end-to-end deep learning architecture for extracting vector building footprints directly from satellite imagery. Unlike earlier methods such as Polygon-RNN or PolyMapper, which rely on sequential vertex prediction or post-processing of segmentation masks, PolyWorld formulates the problem as a graph-based polygon matching task.

The pipeline involves three main stages:

1. Vertex Detection: A fully convolutional neural network outputs a vertex confidence map from which likely building corners are identified. Each vertex is paired with a learned visual descriptor encoding local image features.

2. Graph-Based Learning: Detected vertices are embedded in a fully connected graph. An attentional Graph Neural Network (GNN) evaluates pairwise relationships between vertices to learn "connection strengths" i.e., the likelihood that a pair of vertices should be connected by an edge.

3. Polygon Assembly via Differentiable Matching: The final polygon structure is determined by solving a graph matching problem, formulated as an optimal cycle through the vertices. This is achieved using a differentiable relaxation of the Hungarian algorithm (Sinkhorn algorithm), enabling gradient-based learning.

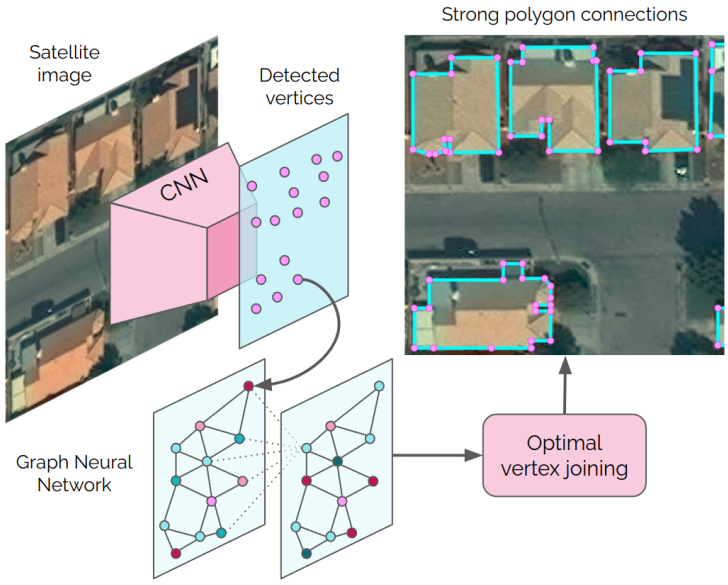

Figure 11: Explanation of how PolyWorld works: source

Despite having better performance than the frame fields models on the CrowdAI dataset, PolyWold does not have the ability to generate polygons with holes, or handle buildings with shared walls. However, the authors offer some ideas for how the model could be modified to handle these cases. The model and inference (but not the training) source code is open source, but has a restrictive license that only permits its use for research.[6/]

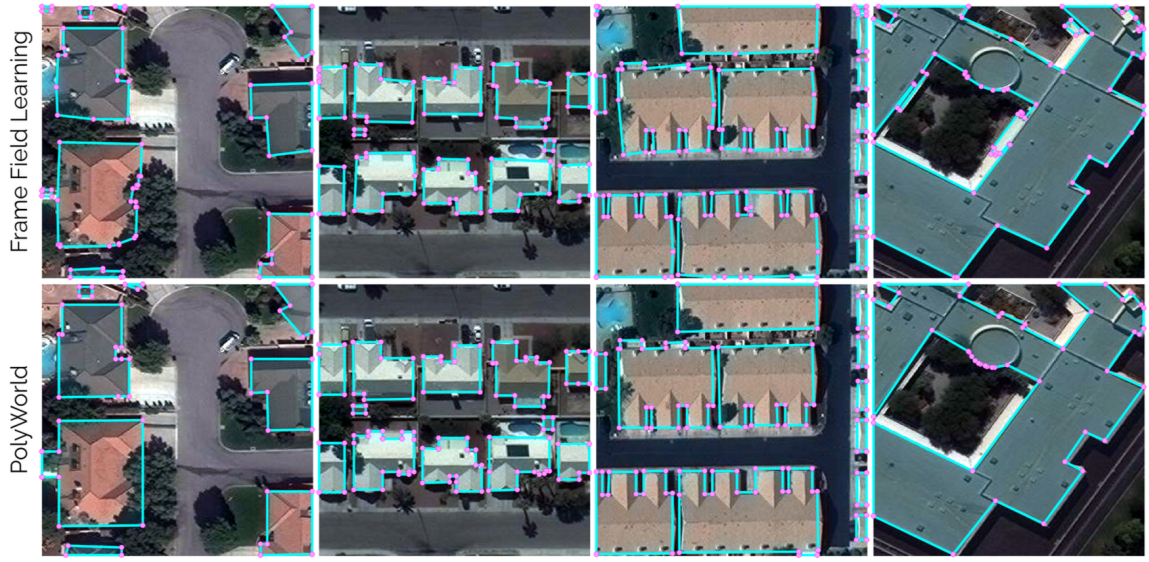

Figure 12: PolyWorld vs Frame Field Learning on CrowdAI test dataset : source

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_S	AR_M	AR_L
Mask R-CNN [13]	41.9	67.5	48.8	12.4	58.1	51.9	47.6	70.8	55.5	18.1	65.2	63.3
PANet [21]	50.7	73.9	62.6	19.8	68.5	65.8	54.4	74.5	65.2	21.8	73.5	75.0
PolyMapper [16]	55.7	86.0	65.1	30.7	68.5	58.4	62.1	88.6	71.4	39.4	75.6	75.4
FFL (no field), mask	57.8	84.0	66.9	33.8	74.1	80.7	67.0	90.4	76.9	46.2	79.7	85.7
FFL (no field), simple poly	61.1	87.4	71.2	35.1	74.5	82.3	64.7	89.4	74.1	41.7	77.9	85.7
FFL (with field), mask	57.7	83.8	66.3	33.8	73.8	81.0	68.1	91.0	77.7	47.5	80.0	86.7
FFL (with field), simple poly	61.7	87.6	**71.4**	35.7	74.9	83.0	65.4	89.8	74.6	42.5	78.6	85.8
FFL (with field), ACM poly [10]	61.3	87.4	70.6	33.9	75.1	83.1	64.9	89.4	73.9	41.2	78.7	85.9
PolyWorld (offset off)	58.7	86.9	64.5	31.8	80.1	85.9	71.7	92.6	79.9	47.4	85.7	94.0
PolyWorld (offset on)	**63.3**	**88.6**	70.5	**37.2**	**83.6**	**87.7**	**75.4**	**93.5**	**83.1**	**52.5**	**88.7**	**95.2**

Figure 13: Comaprison results oon CrowdAI test dataset by PolyWorld

Figure represents MS COCO results on the CrowdAI test dataset for all the building extraction and polygonization experiments. The results of PolyWorld are calculated discarding the correction offsets (offset off), and refining the vertex positions (offset on). FFL refers to the Frame Field Learning method. The results are computed with and without frame field estimation. “mask” refers to the pure segmentation produced by the model. “simple poly” refers to the Douglas–Peucker polygon simplification, and “ACM poly” refers to the Active Contour Model polygonization method [16]

Metric	Meaning
AP	Average Precision (overall) – higher is better. General performance measure combining precision and recall.
AP50	AP at IoU threshold 0.5 – more lenient match condition.
AP75	AP at IoU threshold 0.75 – stricter match condition.
APS / APM / APL	AP for small, medium, and large buildings, respectively.
AR	Average Recall (overall) – measures how well true objects are detected.
AR50 / AR75	AR at IoU thresholds 0.5 and 0.75.
ARS / ARM / ARL	AR for small, medium, and large objects, respectively.

Source Code : GitHub

Improved version , Re:PolyWorld (2023)

Following PolyWorld, Zorzi and Fraundorfer (2023)[17] introduced Re:PolyWorld, which is claimed to be an improved multi-stage version of the framework . Re:PolyWorld added a second refinement stage where an initial polygon prediction is further optimized and made even more regular by an additional GNN module.

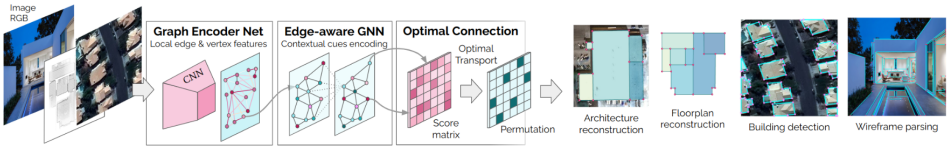

Figure 14: Re:PolyWorld Methodology

With these enhancements, Re:PolyWorld achieved new state-of-the-art scores on the CrowdAI dataset, improving both the precision and the shape quality of footprints. For example, it improved the mean intersection-over-union (IoU) and corner angle error metrics beyond what PolyWorld and a strong frame-field baseline had achieved.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_S	AR_M	AR_L
Mask R-CNN [7]	41.9	67.5	48.8	12.4	58.1	51.9	47.6	70.8	55.5	18.1	65.2	63.3
PANet [12]	50.7	73.9	62.6	19.8	68.5	65.8	54.4	74.5	65.2	21.8	73.5	75.0
PolyMapper [10]	55.7	86.0	65.1	30.7	68.5	58.4	62.1	88.6	71.4	39.4	75.6	75.4
FFL, simple poly [6]	61.7	87.6	71.4	35.7	74.9	83.0	65.4	89.8	74.6	42.5	78.6	85.8
FFL, ACM poly [6]	61.3	87.4	70.6	33.9	75.1	83.1	64.9	89.4	73.9	41.2	78.7	85.9
PolyWorld [30]	63.3	88.6	70.5	37.2	83.6	87.7	75.4	93.5	83.1	52.5	88.7	95.2
Re:PolyWorld	**67.2**	**89.8**	**75.8**	**42.9**	**85.3**	**89.4**	**78.6**	**94.1**	**86.7**	**58.3**	**90.3**	**96.2**

Figure 15: Benchmark dataset of Re:PolyWorld

he continued success of these GNN-based methods demonstrates the value of treating polygon formation as a graph problem (where deep networks ensure the graph forms nice cycles with desired properties) rather than a pixel-by-pixel segmentation problem

Transformer-Based Sequence Models : Pix2Poly

Very recently, researchers have applied transformers the sequence modeling architecture behind advances in NLP to polygon extraction. Pix2Poly [18] is an attention-based model that casts building footprint delineation as a sequence prediction problem, handled entirely by a transformer encoder-decoder.

The key idea is to avoid the multi-step detour that graph-based models take (e.g., detect vertices → match into polygon). Instead, Pix2Poly's transformer directly outputs an ordered list of vertex coordinates in sequence, one vertex after another, in a single forward pass. To do this, it discretizes continuous image coordinates into a sequence of tokens (similar to how one might tokenize words or subwords in language) and trains the network to emit the token sequence corresponding to the building outline.

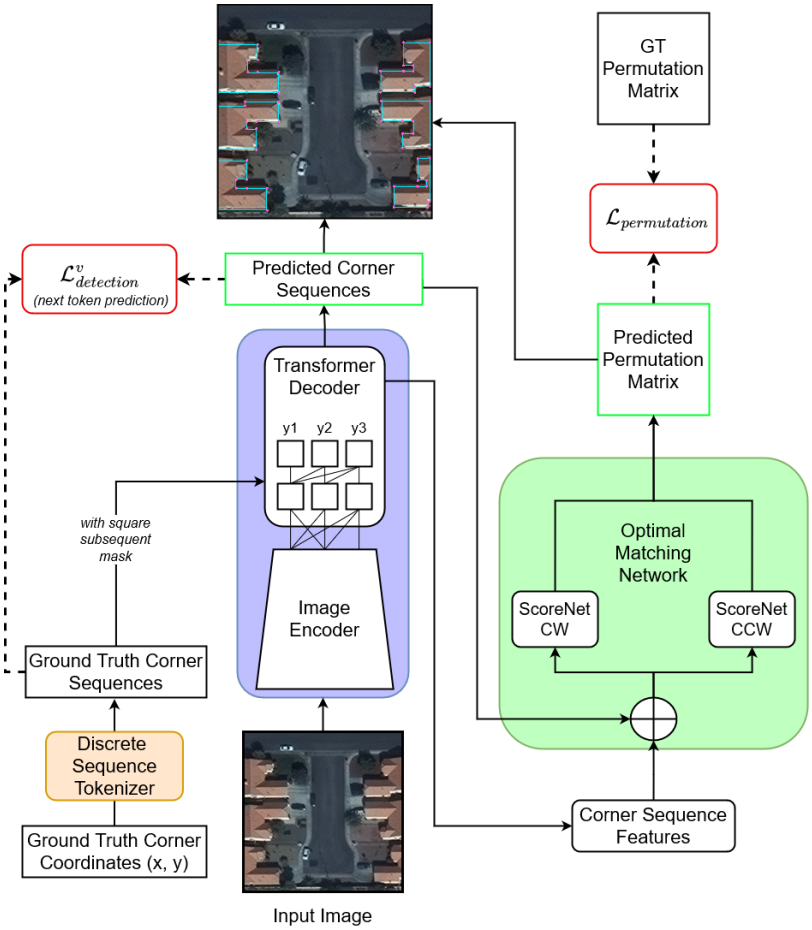

Figure 16: Overview of Pix2Poly Architecture

Because the transformer's self-attention can attend globally to the image, Pix2Poly can, in theory, capture the global shape of the building while placing each vertex. The authors highlight that it avoids certain bottlenecks of earlier methods: for example, it doesn't require a non-maxima suppression step to select vertices (which was non-differentiable in many prior pipelines), nor does it need a separate graph matching module, since the sequence inherently encodes the connectivity.

The entire model is differentiable end-to-end, making training more straightforward and cohesive. In their experiments, Pix2Poly achieved state-of-the-art results not only for building footprints but also for road network extraction, indicating the versatility of the approach.

Essentially, Pix2Poly represents the convergence of transformer-based detection with graph learning: it uses a transformer as a "vertex sequence detector" and still incorporates an optimal matching network (similar to PolyWorld's assignment module) to ensure the predicted sequence forms closed polygons. This model claimed to be less complex as compared to FLL , PolyWorld as it has total parameter count of (31.9M) [18]

Figure 17: Example of Pix2poly output

Source Code : Github

Other Noticable Advances

Alongside the above, there have been other notable modern approaches. PolyBuilding (2022) [1] introduced a similar concept of a “polygon transformer” that directly predicts vector representations of buildings. It emphasizes fully end-to-end training and shows that a transformer can outperform CNN+RNN hybrids on benchmark aerial image datasets.

Generative models have also been explored: for instance, RegGAN (2022) [19] used a generative adversarial network to refine building masks such that their boundaries look more like real building shapes. In RegGAN, a generator CNN outputs a building mask and a discriminator network critiques it, especially focusing on boundary regularity. This adversarial training leads to output masks with sharper, straighter edges than a standard segmentation network.

Similarly, another study proposed Poly-GAN (2023)[11] to post-process OpenStreetMap building footprints, adjusting vertices via a GAN to better align and orthogonalize them . These GAN-based approaches can be seen as learned versions of the old heuristic regularization rather than applying a Hough transform, they apply a discriminator that has learned what a “correct” building outline looks like and thus encourages the output to conform to those learned patterns.

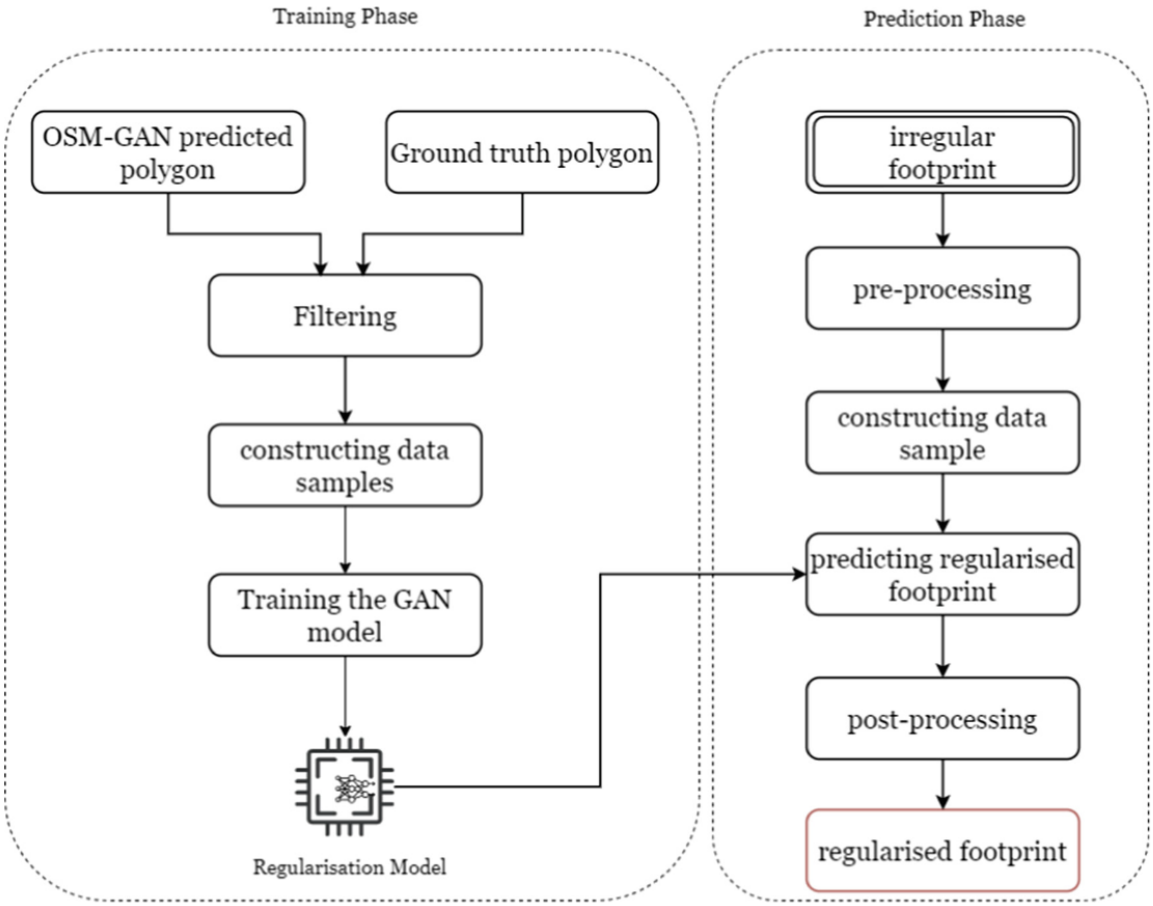

Figure 18: Schematic diagram of the polygon regularization process linking the Poly-GAN model training phase to the (predicted) building regularization phase [11]

Methodology

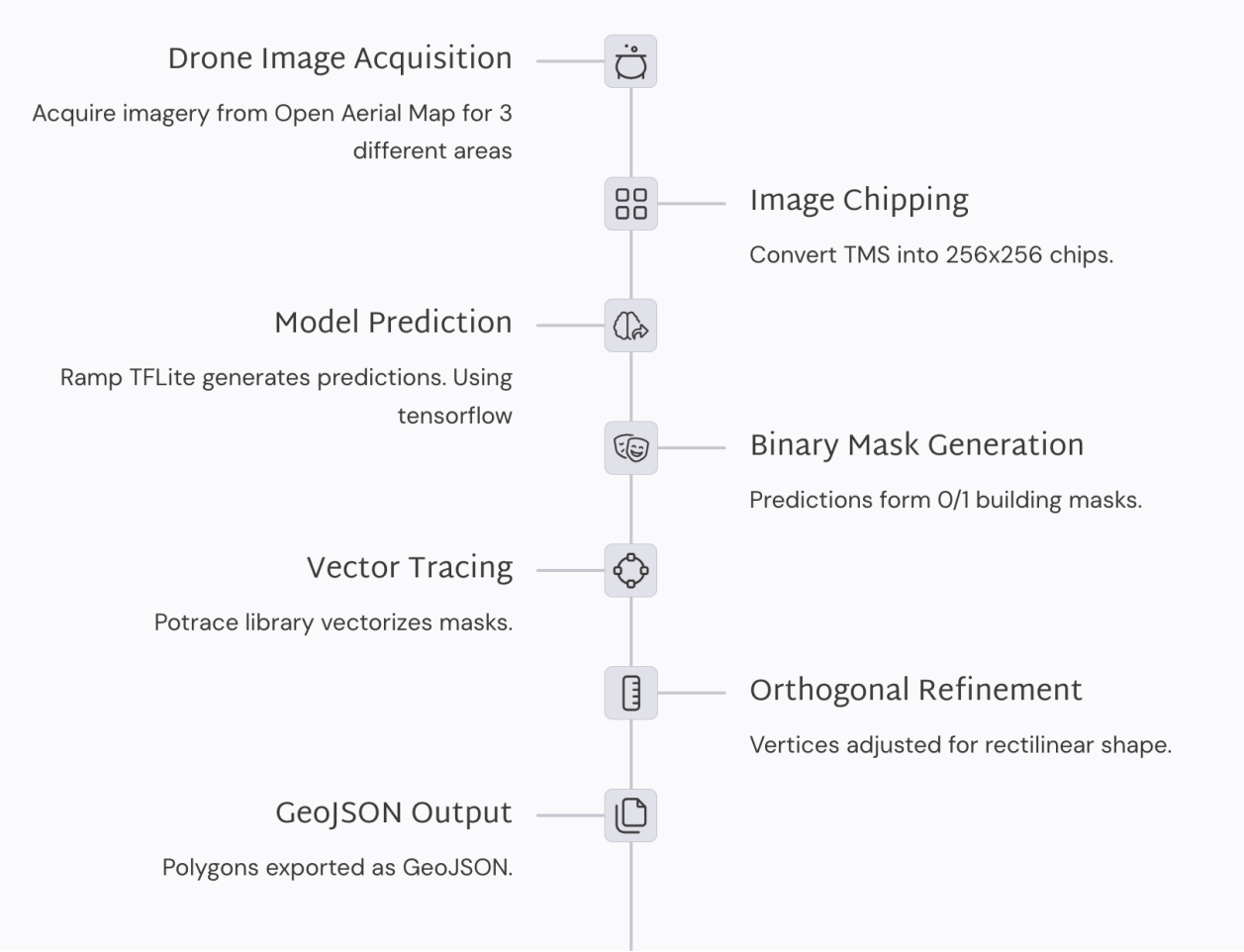

Figure 19: Methodology Utilized For the Comparison

In this study , Literature review was done by reading papers , going through their pros and cons and trying out the methhodologies discussed on the paper . For visual analysis two of the methods were picked up and only traditional methods were analyzed as part of the first initial research . In this comparision, the goal was to extract building footprints shape geometry from aerial imagery rather than evaluate the performance of the deep learning model. The workflow began by gathering high-resolution aerial images from OpenAerialMap, covering three different locations. These images were then divided into smaller tiles of 256×256 pixels to make them suitable for processing. A lightweight building segmentation model from the RAMP (Replicable AI for Microplanning), running on TensorFlow Lite, was used to predict the locations of buildings in these tiles. The output of the model was binary masks, where buildings were represented by white (1) pixels and background by black (0). These masks were then converted into vector outlines using the rasterio library. & this is where multiple algorithm was applied generate the building shapes,& finally building footprints were exported in GeoJSON format for visualization. Later on proposal is to do comparitive study with modern deep learning frameworks on the same area.

Discussion

While doing the literation review between traditional and deep learning methods following things were observed . This table is generated solely based on the interpretation from different papers and blogs cited on the references.

Comparison: Traditional vs. Deep Learning Methods

Accuracy and Performance

Aspect	Traditional Methods	Deep Learning Methods
Detection Accuracy	Moderate; struggles with small/faint buildings	High; CNNs and transformers achieve state-of-the-art IoU and recall
Processing Speed	Very fast (per building) on CPU	Slower per image but GPU-accelerated; parallelization possible
Scalability	Needs tuning for new regions	Scales to large areas
Example	Hough Transform, DP simplification	PolyWorld, Pix2Poly, Frame Field Learning

Flexibility and Generalization

Aspect	Traditional Methods	Deep Learning Methods
Adaptability	Manual reprogramming required	Retrainable and fine-tunable on new data
Shape Handling	Biased to rectilinear structures	Learns to detect irregular, curved, or complex forms
Data Sensitivity	Edge-based; poor in low contrast	Learns semantic cues (shadows, context)
Example	Thresholding, edge detectors	CNNs, Transformers trained on diverse imagery

Cartographic Quality

Aspect	Traditional Methods	Deep Learning Methods
Output Regularity	Hard constraints (e.g. 90° angles)	Learned regularity (polygon loss, angle constraints, GANs)
Visual Quality	Very clean, stylized	Near finer details as compared from hand-crafted results
Limitations	May snap overly aggressively	May allow some deviation; occasional noise
Example	Regularize Footprint tool	PolyWorld, Pix2Poly with angle loss, GAN refinement

GIS Integration

Aspect	Traditional Methods	Deep Learning Methods
Output Format	Vector-ready (Polygons)	Historically : Raster masks , Now outputs GeoJSON/Shapefiles directly
Workflow Fit	Compatible with legacy GIS	Integrated in QGIS, ArcGIS Pro via plugins, OpenCV
Post-Processing	Quite sophisticated	Often Minimal with latest pipelines
Example	Manual digitization, vector tools	Microsoft's global footprint pipeline, Google Open Buildings

Quality Control

Aspect	Traditional Methods	Deep Learning Methods

Aspect	Traditional Methods	Deep Learning Methods
Failure Visibility	Obvious errors, easy to flag	May generate plausible but wrong results
Correction	Manual re-runs or inspection	Hybrid review: DL + regularization + optional human validation
Robustness	Deterministic but brittle , Easy to explain	Robust to noise, generalizes well across geographies, Hard to explain

Quality is really subjective and it is hard to do qualitative analysis for this kind of project ! For sample comparison I picked few areas and compare two of the traditional methods. I picked three different areas, On algorithms : one is douglas pickler theorem another one is right angle optimization theorem which is defined here Both traditional approach to see how they perform practically on following datasets.

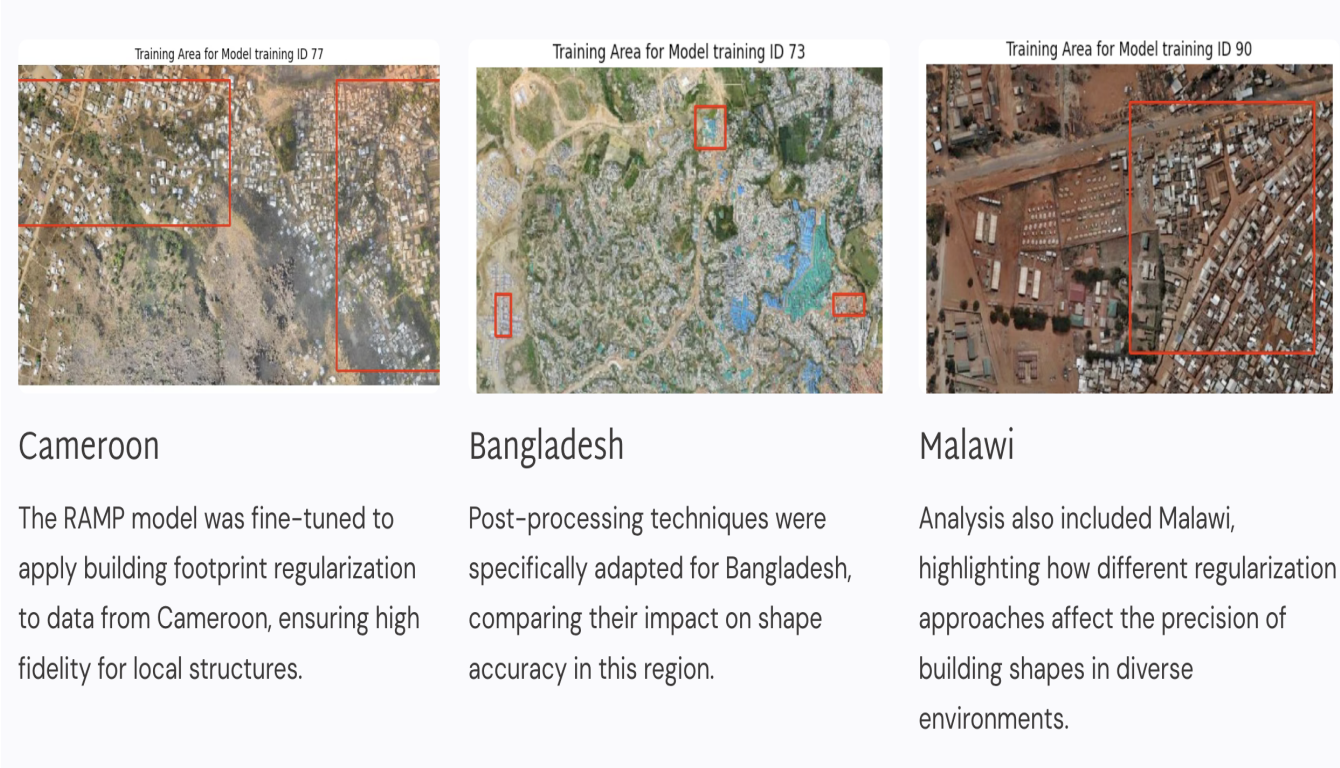

Figure 20: Example test areas

Experiment : Model Based Fitting

Figure 21: Cameroon

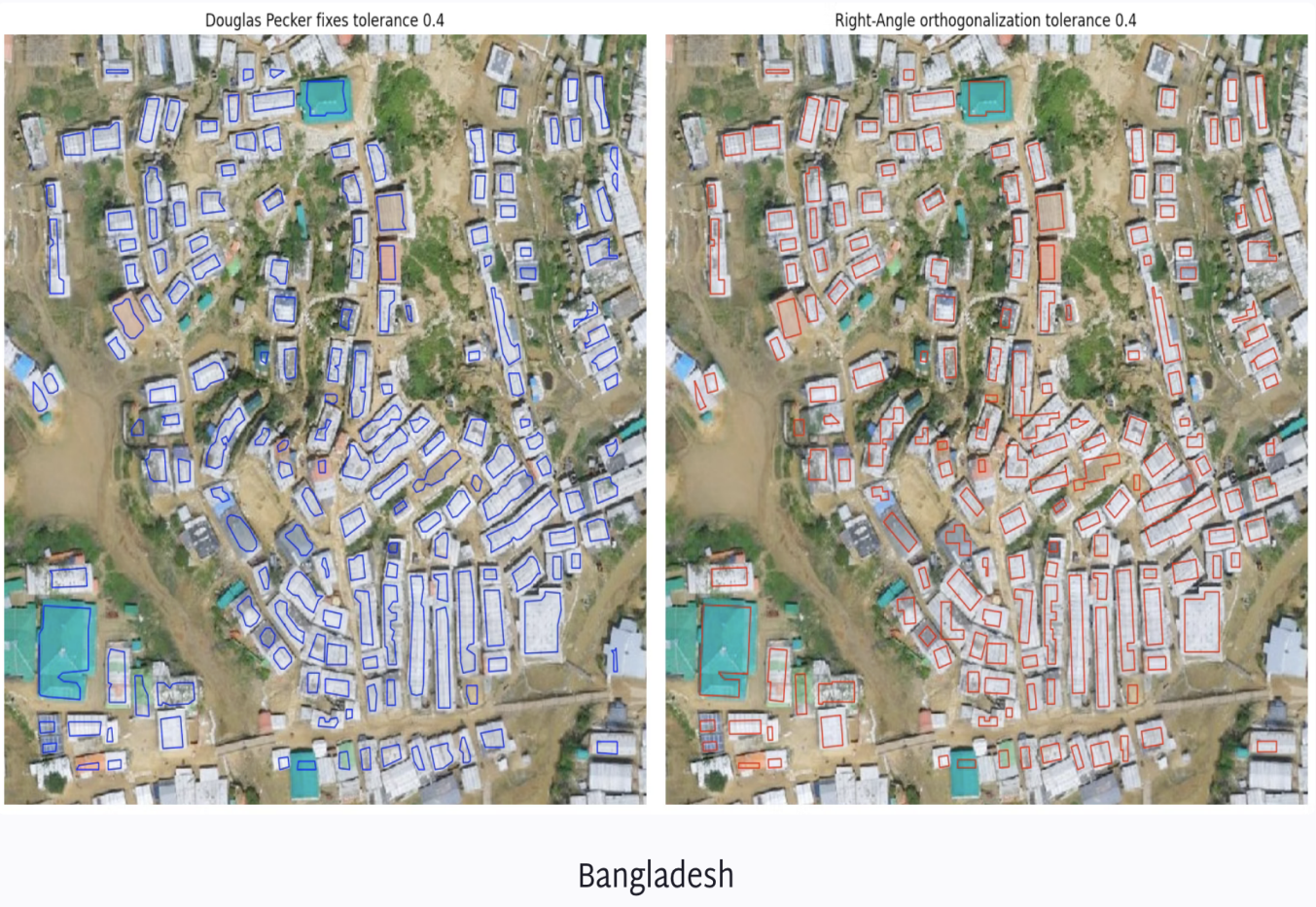

Figure 22: Bangladesh

Figure 23: Malawi

Right-angle optimization methods have shown slight superiority in certain contexts; however, the Douglas-Peucker algorithm remains widely relevant and is employed in numerous projects to date. The primary discussion point is whether an additional deep learning model is necessary for improved shape accuracy, or if feature extraction alone is sufficient. This determination is contingent upon the specific use case and the complexity of the algorithm employed. There is no straightforward solution to this problem.

The measurement of quality in this context is subjective and challenging. For some applications, a well-defined right-angled shape is considered optimal, while others prioritize a closer match to the background shape. The determination of cartographic quality is further complicated by the need for feature alignment. Even if a feature is accurately traced according to the imagery, its placement may not be contextually appropriate. For instance, a building near a road or a tall structure might not be mapped according to its rooftop; it could be slightly shifted or deliberately repositioned to better align with other features.

Conclusion

This review traces the evolution of building footprint regularization over the past three decades, from simple geometric rules to complex deep learning models. However, tracking this technical progress has revealed a more profound issue: the criteria for a “good” building footprint remain elusive. Even in OpenStreetmap itself people have different opinion which is good quality mapping and bad quality mapping interms of cartographic accuracy. A comparative analysis of different methods across three

areas yielded more questions than answers. Should the priority be perfect right angles, accurate representation of real building shapes, or aesthetic appeal on a map? Traditional methods excel at producing clean geometric shapes, while deep learning models are better at detecting complex buildings & slightly going towards understanding the context specially with the transformer based models. Deep learning might potentially solve the problem by applying different algorithms or cases per building rather than applying to the whole scene ; they can work on objects & rotating them to fit the context. However, neither approach fully addresses what humans actually want from these tools. While deep learning shows promise, it is not yet mature enough to provide a definitive solution.

It is important to acknowledge the limitations of this study. The computational resources required by different methods, training times, and runtime speeds were not evaluated. Comparitive analysis with deep learning model based approches can be taken forward on future research. Additionally, large-scale accuracy tests on standard datasets were not conducted. These omissions are not accidental; comprehensive testing would require significant computational resources and may not provide insights into what matters most in real-world applications. Current metrics of success, such as intersection over union scores used by some of the methods, overlook critical factors like whether a building polygon appears correct to human eyes. The field is still trying to justify with how to meaningfully measure cartographic quality.

While the field has made significant strides in automatically detecting buildings and handling complex shapes, this does not necessarily translate to better outputs for practical mapping work. For practitioners involved in projects like OpenStreetMap, the choice of method often depends on available tools, familiarity with the technology, and specific project goals. Building footprint regularization is as much about human judgment and aesthetics as it is about algorithms. Rather than seeking a one-size-fits-all solution, the focus should be on understanding the strengths and limitations of different approaches and providing users with the tools to make informed choices tailored to their specific needs.

References

1. **PolyWorld: Polygonal Building Extraction With Graph Neural Networks in Satellite Images**
Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, Friedrich Fraundorfer
(2022)
https://openaccess.thecvf.com/content/CVPR2022/html/Zorzi_PolyWorld_Polygonal_Building_Extraction_With_Graph_Neural_Networks_in_Satellite_CVPR_2022_paper.html

2. **Rectilinear Building Footprint Regularization Using Deep Learning**
Philipp Schuegraf, Zhixin Li, Jiaojiao Tian, Jie Shan, Ksenia Bittner
ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2024-06-10) https://doi.org/g9k82r
DOI: 10.5194/isprs-annals-x-2-2024-217-2024

3. **Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN**
Philipp Schuegraf, Ksenia Bittner
ISPRS International Journal of Geo-Information (2019-04-12) https://doi.org/ggwr8s
DOI: 10.3390/ijgi8040191

4. **Building Footprint Simplification Based on Hough Transform and Least Squares Adjustment**
Richard Guercke, Monika Sester
Proceedings of the 14th Workshop of the ICA Commission on Generalisation and Multiple Representation (2011-07)

5. **Use of the Hough transformation to detect lines and curves in pictures**
Richard O Duda, Peter E Hart
Communications of the ACM (1972-01) https://doi.org/b4t7tv
DOI: 10.1145/361237.361242

6. **Automated Building Footprint Extraction (Part 3): Model Architectures • Element 84** (2022-11-09) https://element84.com/software-engineering/automated-building-footprint-extraction-part-3-model-architectures/

7. **Algorithms for the reduction of the number of points required to represent a digitized line or its caricature**
David H Douglas, Thomas K Peucker
The Canadian Cartographer (1973)

8. **Hough Transform**
Surya Teja Karri
Medium (2019-09-27) https://medium.com/@st1739/hough-transform-287b2dac0c70

9. **Aggregation of LoD 1 building models as an optimization problem**
R Guercke, T Götzelmann, C Brenner, M Sester
ISPRS Journal of Photogrammetry and Remote Sensing (2011-03) https://doi.org/fbs43j
DOI: 10.1016/j.isprsjprs.2010.10.006

10. **AUTOMATIC EXTRACTION AND REGULARIZATION OF BUILDING OUTLINES FROM AIRBORNE LIDAR POINT CLOUDS**
Bastian Albers, Martin Kada, Andreas Wichmann
The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2016-06-09) https://doi.org/gcc7nx

DOI: 10.5194/isprs-archives-xli-b3-555-2016

11. **Poly-GAN: Regularizing Polygons with Generative Adversarial Networks**
Lasith Niroshan, James D Carswell
Lecture Notes in Computer Science (2023) https://doi.org/g9m2vh
DOI: 10.1007/978-3-031-34612-5_13

12. **Learning Deep Structured Active Contours End-to-End**
Lisa Zhang, Min Bai, Renjie Liao, Raquel Urtasun, Diego Marcos, Devis Tuia, Benjamin Kellenberger
2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018-06)
https://doi.org/ghj5xz
DOI: 10.1109/cvpr.2018.00925

13. **End to End Trainable Active Contours via Differentiable Rendering**
Shir Gur, Tal Shaharabany, Lior Wolf
arXiv (2019) https://doi.org/g9pzd8
DOI: 10.48550/arxiv.1912.00367

14. **End-to-End Trainable Deep Active Contour Models for Automated Image Segmentation: Delineating Buildings in Aerial Imagery**
Ali Hatamizadeh, Debleena Sengupta, Demetri Terzopoulos
Lecture Notes in Computer Science (2020) https://doi.org/g9pzd7
DOI: 10.1007/978-3-030-58610-2_43

15. **Polygonal Building Extraction by Frame Field Learning**
Nicolas Girard, Dmitriy Smirnov, Justin Solomon, Yuliya Tarabalka
2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021-06)
https://doi.org/gnt53r
DOI: 10.1109/cvpr46437.2021.00583

16. **PolyWorld: Polygonal Building Extraction with Graph Neural Networks in Satellite Images**
Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, Friedrich Fraundorfer
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022-06)
https://doi.org/gtmwjv
DOI: 10.1109/cvpr52688.2022.00189

17. **Re:PolyWorld - A Graph Neural Network for Polygonal Scene Parsing**
Stefano Zorzi, Friedrich Fraundorfer
2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023-10-01)
https://doi.org/g9mrpk
DOI: 10.1109/iccv51070.2023.01537

18. **Pix2Poly: A Sequence Prediction Method for End-to-end Polygonal Building Footprint Extraction from Remote Sensing Imagery** https://arxiv.org/html/2412.07899v1

19. **RegGAN: An End-to-End Network for Building Footprint Generation with Boundary Regularization**
Qingyu Li, Stefano Zorzi, Yilei Shi, Friedrich Fraundorfer, Xiao Xiang Zhu
Remote Sensing (2022-04-11) https://doi.org/gqhkzx
DOI: 10.3390/rs14081835