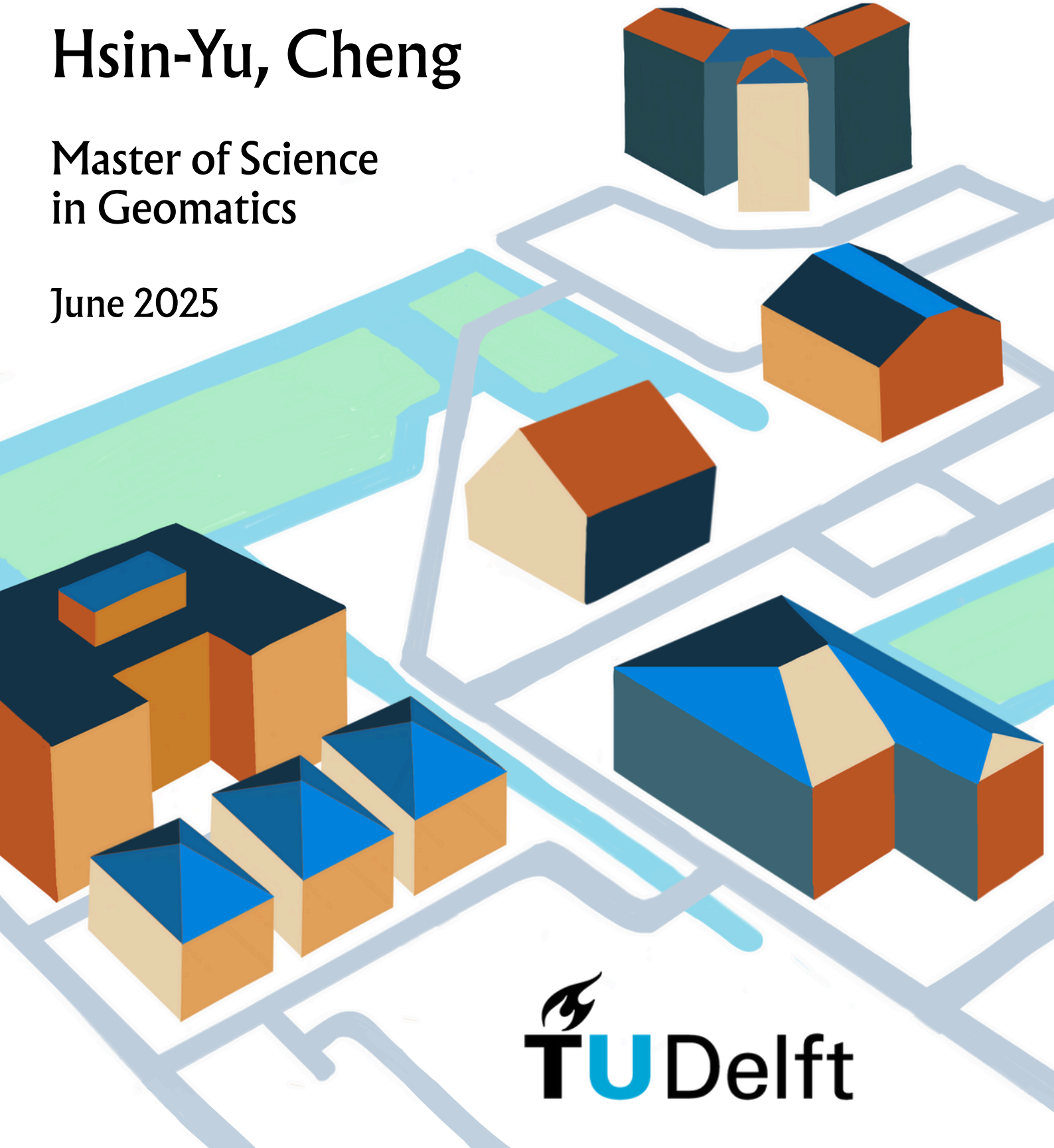
Roof Structure Extraction from Remote Sensing Images

Hsin-Yu, Cheng

Master of Science
in Geomatics

June 2025



TU Delft

MSc thesis in Geomatics

Roof Structure Extraction from Remote Sensing Images

Hsin-Yu, Cheng

June 2025

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of Master of Science in Geomatics

The work in this thesis was carried out in the:

3D geoinformation group
Delft University of Technology

Supervisors:	Dr. Liangliang Nan
	Dr. Weixiao Gao
Co-reader:	Dr. Seyran Khademi

Abstract

This thesis presents a method for extracting structured roof surfaces from remote sensing images. It achieved this by combining semantic segmentation with polygon-based refinement, which allows rooftop boundaries to be described more accurately using line and shape information. The method includes three main stages: (1) using an instance segmentation model to detect and classify rooftop areas; (2) generating polygonal candidates for plannar roof regions based on detected line features; and (3) optimizing label assignments through a Markov Random Field (MRF) model, which integrates prediction confidence with the spatial relationships between polygons. Experiments on benchmark datasets show that this approach improves the accuracy and consistency of rooftop segmentation while reducing incorrect detections. The system is modular and flexible, making it suitable for applications that require reliable roof structure analysis in urban environments.

Acknowledgements

My deepest gratitude goes to Dr. Liangliang Nan, for his early and continuous support throughout this thesis—academically and beyond. He provided valuable literature, reached out to original authors for code access, and offered opportunities to grow through workshops and presentations. These experiences helped me understand what it truly means to do research, not just to complete a thesis.

I am equally thankful to Dr. Weixiao Gao, whose insightful feedback and technical guidance shaped much of my methodology. From data collection and hardware setup to parameter tuning and debugging, he was always patient and precise in helping me identify and resolve what I often couldn't see.

Throughout this journey, I struggled to stay focused, tried too many directions, misunderstood concepts like MRF, and filled early drafts with excessive details while missing the big picture. Thank you both for repeatedly helping me refocus, rewrite, and refine. More importantly, thank you for making me realize that research is not about having my name in a book, but about identifying gaps, articulating motivation, and making meaningful contributions.

I also want to thank my family across the ocean in Taiwan, who reminded me every day to eat well and stay warm, never asking how many pages I'd written. Your care meant everything, especially in cold Dutch winters. Dad, thank you for making it possible for Mom and my sister to attend my final presentation. Sis, thank you for designing my thesis cover. And to my brother—thank you for being there when I needed someone to talk to.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background and Motivation

With the advancement of airborne LiDAR, UAV photogrammetry, and semantic 3D city models such as CityGML [Ledoux and Meijers, 2011; Biljecki et al., 2015], the accurate extraction of rooftop planar structures has become foundational for various smart city applications. These include solar potential estimation [Jochem et al., 2009; Nelson and Grubesic, 2020], urban-scale energy modeling [Peronato et al., 2017; Chen and Hong, 2018], and disaster management [Rezaeian and Gruen, 2011; Calantropio et al., 2021]. As urbanization accelerates, the demand for high-resolution rooftop data grows, highlighting the significance of robust extraction techniques in both research and practice.

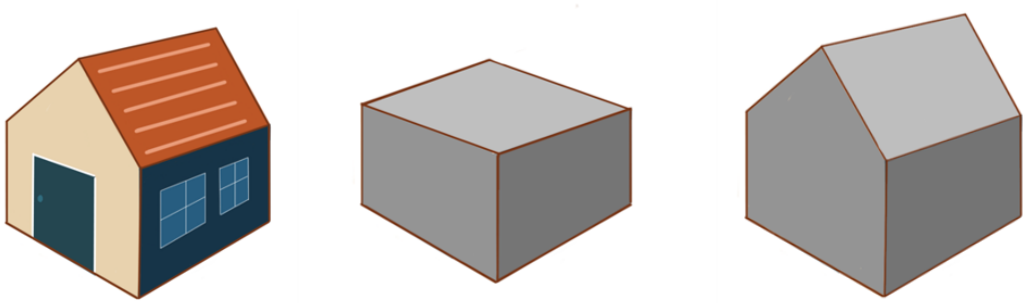
(a) Building Safety Monitoring (b) Green Roof Design (c) Drainage Simulation (d) Solar Potential Estimation

Figure 1.1.: Examples of rooftop applications benefiting from accurate roof structure extraction.

Building segmentation is a fundamental step in urban scene interpretation and 3D modeling. However, the resulting segmentation outputs often lack structural semantics and cannot reliably support downstream geometric reconstruction.

Although segmentation techniques based on LiDAR or imagery data enable automatic detection of building regions, they frequently produce arbitrary polygonal meshes that are difficult to edit and lack watertightness [Verma et al., 2006]. Mesh-based representations, while preserving geometric detail, tend to compromise semantic structure and compactness [Lafarge and Mallet, 2012]. In complex urban environments, occlusions and shadows further reduce segmentation accuracy, leading to broken or misclassified building outlines [Awrangjeb et al., 2014].

To overcome these limitations, recent research has increasingly focused on the extraction of *roof planar structures*, aiming to enhance both the geometric completeness and semantic clarity of building models. This task involves subdividing roof regions into geometrically coherent and semantically meaningful facets, laying the foundation for compact and interpretable 3D reconstruction.



(a) Reference Building (b) Segmentation Output (c) Structure Extraction

Figure 1.2.: Comparing building reconstruction strategies. (a) shows the reference building. (b) illustrates the 3D reconstruction outpupt obtained from segmentation masks alone, resulting in a coarse box model. (c) depicts our method using extracted roof structures, yielding a more accurate and semantically meaningful reconstruction.

1.2. Inspiration and Our Approach

Planar building surfaces, as fundamental geometric units in 3D scenes, not only exhibit clear spatial boundaries but also reflect inherent structural regularity and topological coherence. Recent advances in structured vision modeling suggest that explicit geometric cues—such as contours, edges, and line segments—can serve as effective intermediaries between low-level visual signals and high-level architectural understanding. Incorporating such primitives into attention-based frameworks has been shown to improve the global consistency of plane-related predictions, facilitating unified reasoning across segmentation, depth, and geometry [Tan et al., 2022; Yin et al., 2023].

However, in complex urban environments, image-based segmentation remains highly susceptible to visual disturbances—such as vegetation occlusion, deep shadows, and lighting variation [Awrangjeb et al., 2014]. These factors often result in fragmented boundaries, misaligned contours, or missing roof regions. Moreover, instance segmentation models may yield overlapping fragments or ambiguous labels, especially in dense cityscapes where complex morphologies are common [Yang et al., 2023; He et al., 2017]. In such cases, one roof surface might be incorrectly assigned to multiple predicted regions, undermining the topological integrity and usability of the output [Verma et al., 2006]. These limitations highlight that segmentation outputs, even when accurate at the pixel level, often lack the structural completeness and geometric consistency required for reliable 3D modeling.

These challenges motivate a shift in perspective: rather than treating segmentation as the final goal, we consider it an intermediate cue to be structurally refined. While existing methods can detect roof outlines and reconstruct LoD2 models from orthophotos and elevation data [Gao et al., 2024], typically emphasize geometric delineation rather than modeling inter-facet relationships. Likewise, multi-task models that jointly predict segmentation and plane parameters rarely enforce global consistency across roof structures [Yin et al., 2023]. We instead take inspiration from approaches that combine semantic reasoning with explicit geometry, using intermediate outputs as building blocks for structured reconstruction.

In particular, recent work has shown that polygonal surface selection can be framed as a global optimization problem using a Markov Random Field (MRF) formulation [Chen

et al., 2022]. By defining energy terms that penalize geometric redundancy and reward data consistency, these models produce watertight and topologically sound reconstructions through graph cut inference.

Building on this idea, we propose an MRF-based pipeline tailored to image-derived inputs. We begin by extracting structural line segments from orthophotos, which are polygonized to form candidate roof planes. Each candidate is then evaluated using instance segmentation predictions—specifically, confidence scores and per-pixel probability maps—to estimate its semantic plausibility. These form the unary potentials in our MRF, while pairwise terms capture spatial adjacency and consistency across candidates. Optimizing this energy yields a compact and topologically coherent polygonal roof structure, suitable for downstream 3D modeling tasks.

Rather than relying solely on pixel-wise segmentation, our method incorporates an intermediate polygonal representation that integrates semantic confidence with geometric structure. This leads to more consistent and interpretable roof plane assignments in complex urban environments.

1.3. Research Question

This work addresses the problem of extracting structured roof planar representations from a single satellite or aerial image. While deep segmentation networks can effectively delineate building regions at the pixel level, transforming such outputs into compact, topologically coherent roof structures remains a non-trivial challenge. Existing methods often struggle to enforce global consistency or preserve geometric simplicity, especially in complex urban environments.

Our main research question is:

> ***How can roof planar structure be extracted from a single satellite or aerial image using segmentation and optimization?***

To investigate this question, we focus on three sub-problems:

- **Polygonal Conversion:** How can structured polygonal representations be derived from pixel-based segmentation masks?
- **Plane Assignment:** How can fragmented roof parts be grouped and assigned to coherent roof planes while preserving topological integrity?
- **Efficiency and Compactness:** How can the extraction process be made more efficient while maintaining geometric precision and structural consistency?

Together, these questions guide the design of our image-driven reconstruction pipeline, which integrates semantic segmentation cues with geometric priors under a global optimization framework.

1.4. Thesis Organization

This thesis is structured into five chapters, progressively developing the motivation, framework, implementation, and evaluation of a geometry-aware approach for roof plane extraction from 2D imagery:

- **Chapter 1: Introduction** outlines the motivation for structured roof modeling in urban environments and identifies the limitations of existing segmentation methods. It introduces roof plane extraction as a more geometrically coherent alternative and presents the central research question along with an overview of the proposed pipeline.
- **Chapter 2: Related Work** reviews prior research in four main categories: building segmentation, instance-level roof modeling, roofline detection, and planar structure extraction. It compares existing approaches in terms of modeling strategies, output representations, and typical limitations, thereby motivating the proposed method.
- **Chapter 3: Methodology** introduces a three-stage pipeline: instance segmentation generates semantic masks, polygon proposal extracts geometric rooftop regions, and MRF optimization refines labels by integrating semantic and structural cues.
- **Chapter 4: Implementation, Results, and Discussion** details the dataset preparation, model configurations, and training procedures. It presents both qualitative and quantitative evaluation on benchmark datasets, compares results before and after MRF refinement, and analyzes typical failure cases. The chapter also discusses the observed strengths and limitations of the proposed approach, offering insight into generalizability and robustness across different rooftop structures.
- **Chapter 5: Conclusion and Future Work** summarizes the contributions of the proposed pipeline and reflects on its performance across diverse datasets. The chapter also outlines several future research directions aligned with each stage of the pipeline—ranging from enhanced backbone architectures, more robust polygon proposal mechanisms, to refined MRF modeling strategies—with the goal of further improving semantic accuracy and structural consistency in urban environments.

2. Related Work

In this chapter, we review recent works related to structured building modeling. We categorize these works into four types based on the level of detail and the nature of the output:

- **Building Segmentation** – methods that identify building regions in satellite images, aerial photographs, or point clouds. The output is typically a binary or semantic mask distinguishing building and non-building areas.
- **Building Instance Segmentation** – methods that separate individual buildings from one another. The goal is not just to detect buildings, but to distinguish between multiple adjacent structures. These methods assign a unique label to each building instance but do not necessarily provide accurate geometric outlines.
- **Roof Line Extraction** – methods that extract the geometric outlines or edges of rooftops, often as vectorized polygons or line segments. Unlike instance segmentation, these methods focus on reconstructing accurate 2D roof footprints or contours, regardless of whether building instances are distinguished. The emphasis is on geometric precision rather than object identity.
- **Roof Planar Structure Extraction** – methods that recover the internal 3D structure of rooftops, including multiple planar surfaces, their orientations, and spatial relationships. These methods aim to support detailed 3D model reconstruction, such as LoD2 or LoD2.2 representations.

This categorization helps clarify whether a method is designed for semantic understanding, geometric modeling, or full 3D reconstruction. The following sections review representative works in each category.

2.1. Building Segmentation

Semantic segmentation of buildings in overhead imagery is a fundamental step for downstream tasks such as instance-level analysis and 3D reconstruction. Most existing methods generate pixel-wise masks that identify building regions, but lack geometric structure or topological awareness, limiting their applicability in tasks that require accurate contours or shape regularity.

Progress in this area has been supported by large-scale benchmark datasets featuring diverse urban scenes. Among these, RoofVec [Hensel et al., 2021] offers instance-level rooftop annotations derived from imagery, and is used in this study as the benchmark dataset. While some datasets offer high-resolution footprints, their limited coverage and relatively uniform architectural styles restrict generalization [Demir et al., 2018]. Others introduce variation in imaging conditions to improve robustness, but performance remains affected by occlusion, shadow, and viewing angle distortions [Weir et al., 2019].

Common segmentation architectures rely on multi-scale feature fusion to enhance spatial precision and typically perform well across different datasets. Nonetheless, their outputs are raster-based, making them suboptimal for geometry-aware tasks such as contour extraction or vectorized reconstruction—especially in dense urban areas where building boundaries are often occluded or highly irregular [Chicchon et al., 2024].

(a) Input Image (b) Segmentation Output

Figure 2.1.: Example of semantic building segmentation. (a) shows an aerial input image, and (b) presents the corresponding segmentation mask output. While such pixel-wise predictions identify building regions, they lack geometric structure and are difficult to convert into accurate vector representations for 3D modeling.

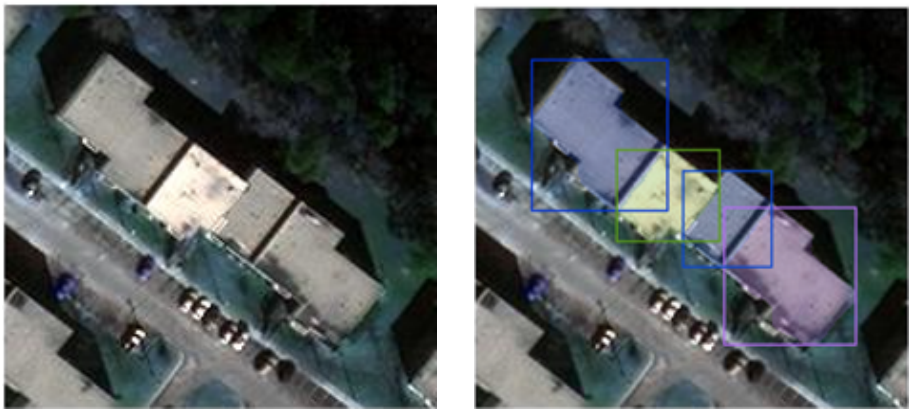
2.2. Building Instance Segmentation

Instance segmentation identifies and separates individual buildings within a scene, supporting object-level applications such as change detection and 3D reconstruction. Unlike semantic segmentation, this task must resolve adjacency and occlusion in dense urban layouts.

General-purpose frameworks have been adapted to this domain. These models perform well in natural image settings but often fail to respect geometric regularities in urban environments, resulting in irregular masks or merged buildings [He et al., 2017].

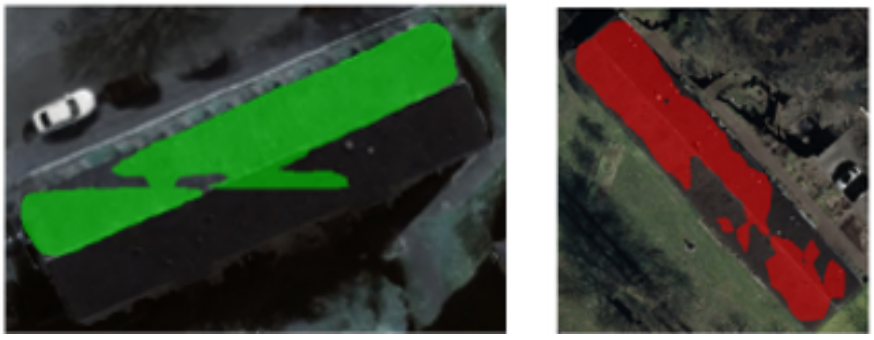
Recent approaches introduces a large-scale 3D dataset with fine-grained building instances. Their proposed method segments point clouds efficiently by learning point-level relations, avoiding clustering-based overhead [Yang et al., 2023].

Nonetheless, challenges remain, including structural ambiguity, noise sensitivity, and the lack of geometric constraints in model outputs. These limitations point to the need for segmentation approaches that integrate architectural priors and spatial regularization.



(a) Input Image (b) Instance Segmentation Output

Figure 2.2.: Example of building instance segmentation. (a) shows the aerial input image, and (b) displays the corresponding instance masks and bounding boxes. Compared to semantic segmentation, instance-level predictions must resolve close adjacency and preserve object boundaries.



(a) Blurry Boundaries and Polygon Distortion (b) Blurry Boundaries and Polygon Distortion

Figure 2.3.: Failure cases in instance segmentation. Both examples suffer from blurry boundaries and distorted polygon shapes, leading to inaccurate and geometrically inconsistent masks.

2.3. Roof Line Extraction

Roof line extraction focuses on recovering vectorized rooftop contours with high geometric fidelity, typically represented as structured polygons or planar graphs. This task supports applications such as cadastral mapping, topological modeling, and downstream 3D reconstruction, where pixel-wise segmentation is insufficient.

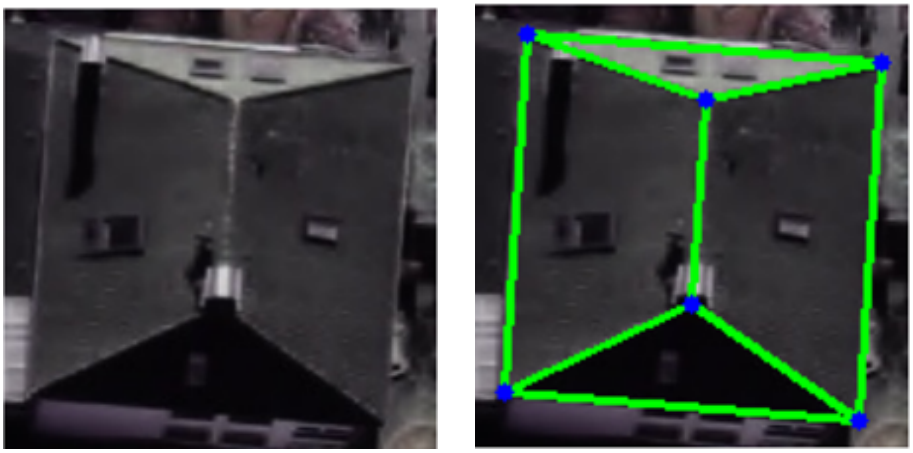
Approaches to this task can be broadly categorized into unsupervised refinement, mask-to-boundary alignment, and direct polygon prediction. Unsupervised methods operate with-

out training data, often using geometric priors and line detection to generate building outlines. These methods are flexible and data-efficient, but remain sensitive to texture loss, shading, and footprint ambiguity—especially in dense urban imagery [Gao et al., 2024].

A second line of work refines segmentation masks into vector-aligned contours by introducing structural constraints during post-processing. While this improves alignment accuracy, performance is tightly coupled to the quality of the initial mask predictions, and tends to degrade with occlusion or small-scale features [Xu et al., 2023].

Recent methods directly predict polygonal structures using deep learning. These include transformer-based models that detect and connect roof corners, graph neural networks that infer topological relations, and hybrid frameworks combining CNNs with optimization solvers. Although these approaches achieve better geometric consistency, they can still produce invalid shapes under noise or occlusion, and some incur significant computational costs [Chen et al., 2021; Zorzi et al., 2023; Nauata and Furukawa, 2020].

Overall, roof line extraction remains an open challenge due to the competing demands of structural regularity, scene complexity, and computational efficiency.

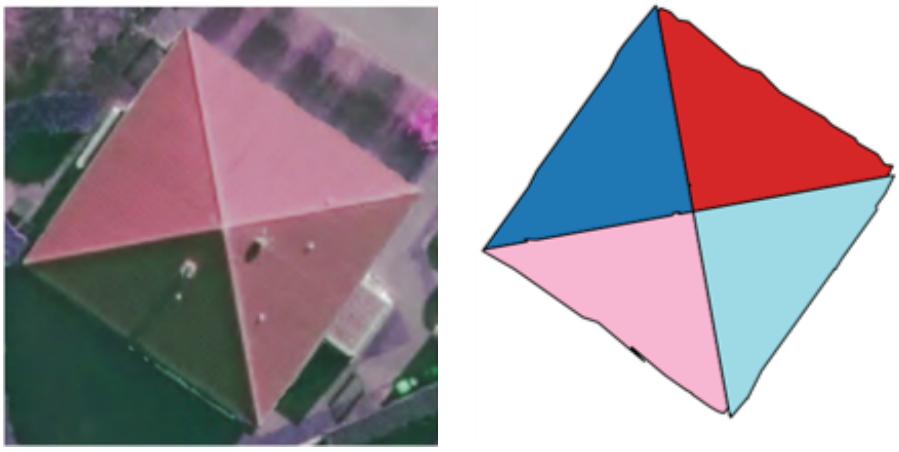


(a) Input Image (b) Roof Line Extraction Result

Figure 2.4.: Example of roof line extraction. (a) shows the raw aerial image, while (b) illustrates the rooftop outlines. The result preserves geometric structure and facilitates downstream modeling tasks.

2.4. Roof Planar Structure Extraction

Recovering the planar composition of rooftops—typically represented as sets of polygonal surfaces with geometric and topological consistency—is essential for producing LoD2-level 3D building models. Recent work in this domain can be broadly grouped by their input modality and modeling strategy: image-only approaches, image-height fusion methods, and learning-based geometric modeling.



(a) Input Image (b) Extracted Planar Structure

Figure 2.5.: Example of rooftop planar structure extraction. (a) shows an input aerial image, and (b) depicts the extracted roof facets represented as geometrically consistent polygonal surfaces. This structural representation forms the basis for LoD2-level 3D building modeling.

2.4.1. Image-only Methods

These approaches operate solely on RGB images or orthophotos to infer planar roof structures. Methods based on monocular cues can estimate roof sections and heights from a single image, but often produce systematic height errors and struggle with occlusions and generalization to unseen cities [Lussange et al., 2023]. Techniques using implicit fields model surface continuity in a latent space, yet may fail to distinguish roof and wall boundaries and are sensitive to input sparsity or complex geometry [Chen et al., 2022]. Graph-based optimization frameworks support interactive editing with geometric constraints, but typically require manual specification of roof topology and rely on computationally intensive post-optimization to ensure planarity [Ren et al., 2021].

2.4.2. Image + Height Fusion Methods

These methods leverage both RGB images and height information such as digital surface models (DSMs) or normalized DSMs (nDSMs) to improve roof reconstruction. By combining spectral and geometric cues, they aim to better separate rooftops from vegetation and capture building heights more accurately. Recent approaches fuse these modalities either early, at the input level, or late, through multi-branch neural networks [Schuegraf et al., 2023; Xu et al., 2024].

Despite improved performance over image-only systems, fusion-based approaches still face notable limitations. Raster-based predictions often result in broken or rounded roof lines, and the subsequent polygon and 3D model generation stages remain underdeveloped and require further research [Xu et al., 2024]. In addition, deep learning methods for DSM–RGB

fusion are still relatively immature, with limited architectural standardization and generalization ability [Xu et al., 2024]. The transition to clean vectorized models remains challenging, as these methods require careful post-processing to generate valid polygons and often struggle to distinguish buildings from spectrally similar objects such as roads or vegetation [Schuegraf et al., 2023]. Moreover, vertical structures are poorly reconstructed due to the near-nadir viewing angle of most airborne sensors, limiting the completeness of facade geometry [Bauchet et al., 2024]. Finally, iterative refinement frameworks that operate on top of initial reconstructions rely heavily on the quality of the starting model and are constrained by a limited set of heuristic actions, which restricts their flexibility in handling diverse architectural layouts [Zhang et al., 2021].

2.4.3. Learning-based Geometric Modeling

These approaches aim to learn structured 3D roof geometry representations directly from data, often targeting higher levels of abstraction than per-pixel or voxel outputs. Some methods represent geometry implicitly, learning continuous surface fields from image input, while others use explicit primitives such as polygons or rectangles, combined with graph-based neural architectures.

Implicit field methods do not rely on intermediate mesh or point cloud representations and instead encode surface geometry continuously. However, these methods often struggle to preserve sharp roof edges or discontinuities, require iso-surfacing as a post-processing step to obtain explicit geometry, and suffer from limited scalability due to high memory and computation costs [Chen et al., 2022].

Primitive-based methods decompose roofs into geometric units such as rectangles or polyhedra and model their arrangement and relations. Roof-GAN learns to generate roof primitives and their spatial relationships using adversarial training, offering diversity and realism in generated roof structures. Still, it assumes fixed wall height and symmetry, limits roof types to simple configurations, and is trained on a relatively small dataset [Qian et al., 2021]. Similarly, PolyGNN reconstructs polyhedral building models from point clouds by jointly optimizing geometry and connectivity through a graph neural network. While effective under ideal conditions, its performance depends on accurate primitive extraction and high-quality synthetic training data, potentially limiting generalization to noisy or real-world inputs [Chen et al., 2024].

3. Methodology

3.1. Overview

This chapter details the methodology for extracting and structurally refining rooftop instances from 2D aerial imagery through a multi-stage processing pipeline. The overall approach integrates instance-level semantic predictions with polygon-based geometric representations to produce rooftop segmentations with improved spatial coherence and structural integrity.

The process begins with instance segmentation applied to the input RGB image, producing soft semantic masks with pixel-wise confidence scores. In parallel, the RGB image is processed to generate candidate rooftop shapes in the form of polygon proposals, capturing the geometric boundaries of potential rooftop regions.

Subsequently, semantic probabilities from instance segmentation are aggregated over each polygonal region to produce per-polygon label probabilities. These probabilities are used to define unary potentials in a Markov Random Field (MRF) framework. Meanwhile, adjacency relationships among polygons are encoded as pairwise terms in the MRF based on their spatial connectivity.

The final step involves minimizing the total energy of the MRF, integrating both unary and pairwise terms, to infer the most probable label assignments for each polygon. This process allows the combination of fine-grained semantic information with structured geometric cues for improved segmentation reliability.

Figure 3.1 illustrates the full pipeline. The blue modules represent the independent input stages, which generate semantic predictions and polygon proposals. The orange modules denote the MRF-based refinement process that integrates both information streams to produce the final rooftop segmentation output.

The overall process can be summarized into three main stages:

- **Instance Segmentation:** Produces semantic masks with confidence scores for each rooftop object.
- **Polygon Proposal Generation:** Extracts oversegmented polygonal regions representing rooftop boundaries.
- **MRF Optimization:** Computes per-polygon unary and pairwise potentials using semantic and geometric cues. Then, minimizes the total MRF energy to assign instance labels to polygons.

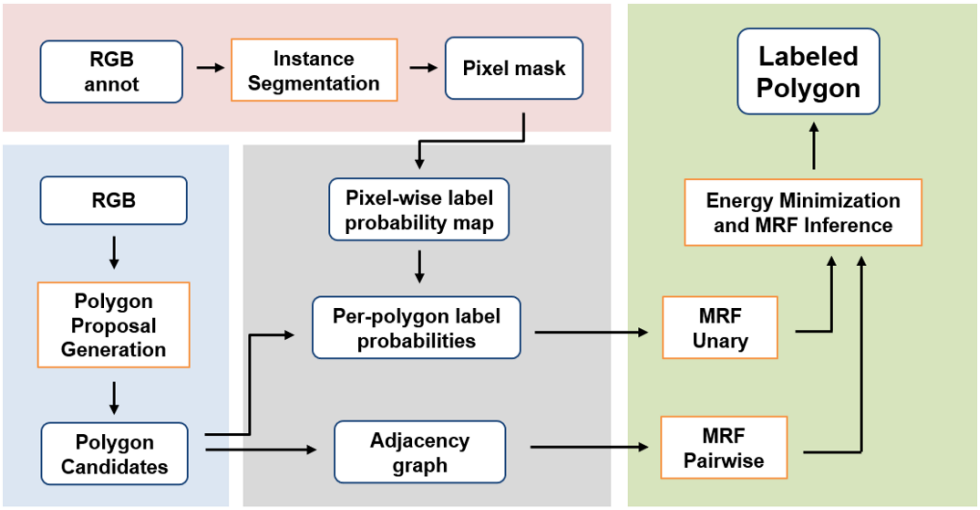


Figure 3.1.: An overview of our approach. Blue modules represent inputs and outputs, while orange modules indicate algorithmic components of the pipeline.

The following sections elaborate on each component of the pipeline in detail, highlighting the rationale behind their design and how they collectively enhance rooftop segmentation by improving structural coherence, instance separation, and geometric alignment.

3.2. Instance Segmentation

The first stage of the proposed pipeline focuses on instance segmentation. This step is responsible for detecting individual rooftop instances from aerial imagery, and it provides the initial semantic input used in later structural refinement.

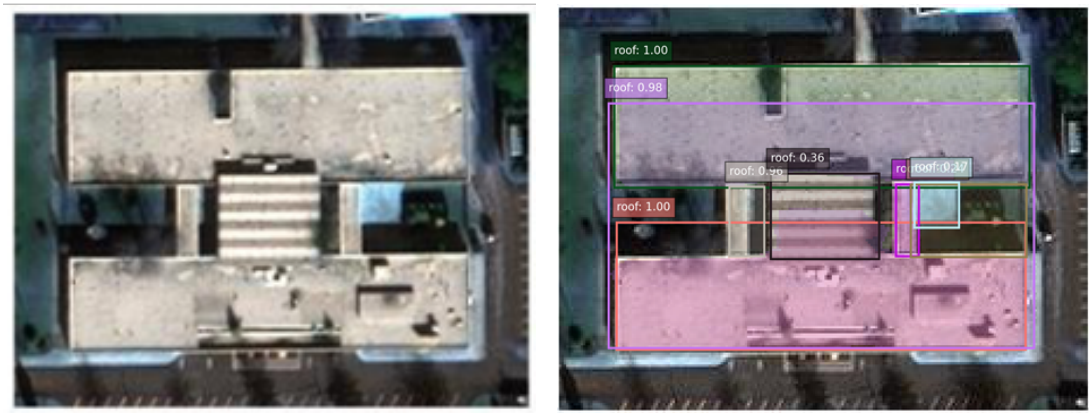
To achieve this, we employ a deep learning–based instance segmentation model that produces object-level semantic masks, enabling each rooftop to be identified as a distinct instance—an essential prerequisite for topologically aware post-processing.

The model extracts hierarchical visual features from the input image using a deep backbone network with multi-scale capability. These features are then used to predict instance-level masks for rooftop objects, producing high-resolution binary masks that delineate each rooftop as a separate entity.

The model outputs a set of instance-level masks, where each instance includes both a classification confidence score and a per-pixel probability map (soft mask). The confidence score reflects the likelihood that the predicted object belongs to the rooftop class, while the soft mask provides pixel-wise probabilities indicating how likely each pixel belongs to the predicted instance. These two outputs are later combined to form weighted semantic evidence for downstream inference.

Figure 3.2 shows an example of the model output, where rooftop instances are predicted as colored regions with associated confidence scores and bounding boxes. To further illustrate the prediction quality and potential limitations, Figure 3.3 visualizes a set of segmented instances, comparing them against ground truth. Despite the presence of multiple predictions,

only one region correctly overlaps with a true rooftop, emphasizing the need for additional geometric refinement in the next stage.



(a) Original aerial image. (b) Instance segmentation result.

Figure 3.2.: Instance segmentation output compared to the original RGB input. In (b), Each colored region represents a predicted rooftop instance, along with confidence scores and bounding boxes.

Figure 3.3.: Visualization of instance segmentation prediction. Among 8 predicted polygons, only Pred 2 is a true positive (TP), while the rest are false positives (FP).

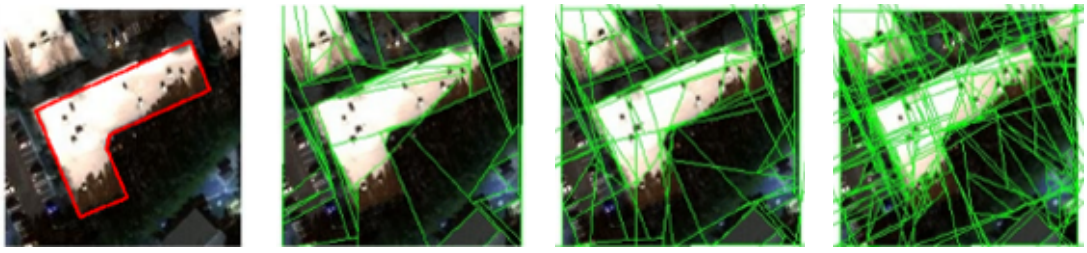
While instance segmentation provides valuable pixel-wise semantic cues, its output often suffers from noise and fragmentation near object boundaries. To overcome these limitations and introduce explicit geometric structure, we incorporate a complementary polygon proposal stage.

3.3. Polygon Proposal Generation

Instance segmentation models often suffer from imprecise object boundaries, fragmented regions, and overlapping predictions—limitations that hinder the extraction of topologically coherent rooftop structures. To ad dress these issues, we introduce a polygon proposal generation stage that constructs geometrically constrained regions directly from the RGB image, providing structured spatial units for subsequent probabilistic inference.

This stage is built upon the KIPPI algorithm (KInetic Polygonal Partitioning of Images) [Bauchet and Lafarge, 2018], which begins with dense line segment detection and proceeds with geometric regularization and kinetic-based partitioning. The method integrates local edge evidence with global shape priors such as parallelism and perpendicularity, producing boundary-aligned line networks that remain robust under occlusions, lighting variation, and urban clutter.

To ensure high recall of potential rooftop surfaces—particularly in complex or fragmented environments—we adopt an oversegmentation strategy that favors completeness over compactness. This results in a fine-grained set of candidate regions, which introduces redundancy but enhances downstream label flexibility. Figure 3.4 illustrates how increasing the segmentation granularity progressively decomposes the rooftop into finer polygonal regions, which enables better adaptation to structural complexity.



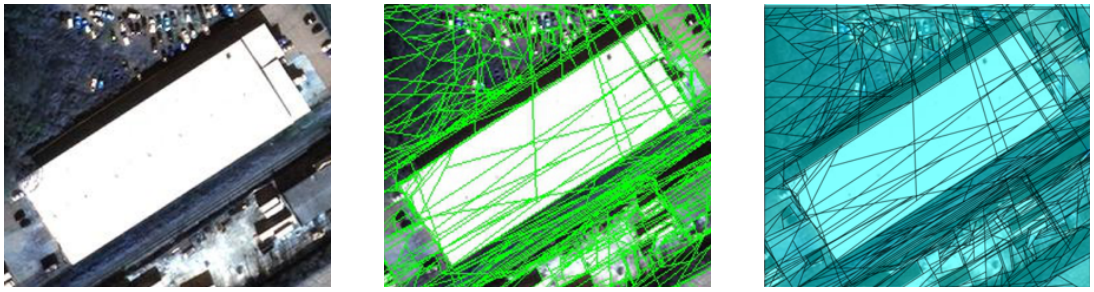
(a) Ground truth rooftop (b) Low segmentation level (c) Moderate segmentation level (d) High segmentation level

Figure 3.4.: Effect of oversegmentation granularity. Increasing segmentation levels introduce more polygon candidates, improving the likelihood of capturing fine rooftop boundaries, especially in cluttered or irregular regions.

Following line extraction, the segments are merged based on geometric connectivity to form a continuous rooftop line network. Closed polygonal regions are then automatically generated from this network, resulting in a set of candidate rooftop polygons. These polygons serve as the structural units for downstream MRF-based label inference.

Figure 3.5 visualizes the full polygon proposal pipeline, starting from raw roofline detection and ending with the construction of closed polygonal candidates. Implementation specifics, including parameter settings and performance considerations, are provided in Section 3.3.

With a set of oversegmented polygonal regions capturing geometric layout, the next step is to integrate these with semantic evidence from instance segmentation. This fusion is achieved through a probabilistic graphical model that assigns labels to each polygon in a globally consistent manner.



(a) Original image (b) Rooflines detected (c) Aggregated into polygons

Figure 3.5.: Illustration of the polygon proposal process. Rooflines are first detected from the input image and then aggregated to form closed polygonal regions for downstream inference.

3.4. MRF Optimization

This section presents a graph-based energy minimization framework for assigning instance labels to candidate rooftop polygons. Unlike deep learning–based instance segmentation models that output pixel-wise masks and probability maps, our method directly assigns discrete integer labels to each polygonal candidate. Here, label 0 denotes background, while labels $1 \dots N$ correspond to foreground rooftop instances. This explicit labeling scheme enhances the clarity of instance separation and ensures better geometric consistency, facilitating downstream structural processing and evaluation.

To achieve this, the problem is formulated as a Markov Random Field (MRF), where each polygon corresponds to a node in a graph, and edges represent geometric adjacency between polygons. The final label assignment is obtained by minimizing a global energy function composed of a unary term and a pairwise term. This function integrates confidence-weighted probability scores from model predictions with adjacency-based geometric information, balancing intra-region confidence with inter-region label coherence.

3.4.1. Problem Formulation

Given a set of polygon proposals extracted from an image, the objective is to assign a label $L_k \in \{0, 1, \dots, N\}$ to each polygon k, where non-zero labels represent foreground rooftop instances. The labeling problem is formulated as the minimization of the following energy function:

$$E(L) = \sum_k U_k(L_k) + \sum_{(i,j)\in E} \lambda \cdot w_{ij} \cdot V(L_i, L_j) \tag{3.1}$$

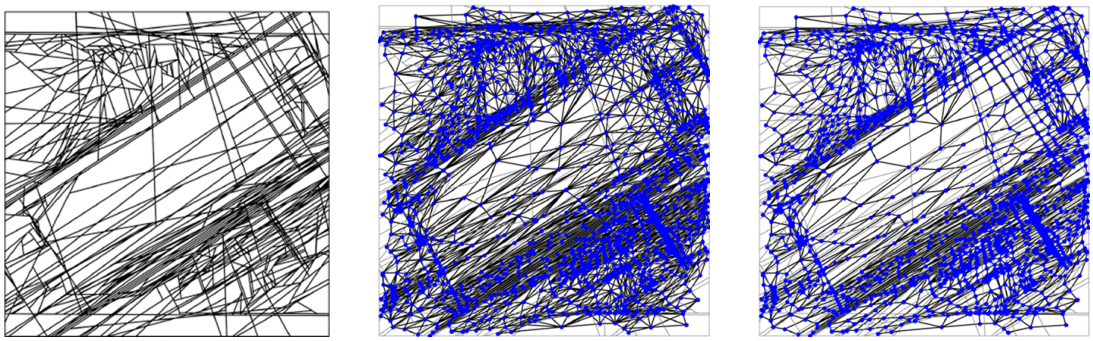
Here, $U_k(L_k)$ denotes the unary cost of assigning label L_k to polygon k, w_{ij} represents the geometric weight between adjacent polygons i and j, $V(a, b)$ is the label penalty function, and λ is a hyperparameter controlling the influence of the pairwise term.

3.4.2. Graph Construction

We construct an undirected graph $G = (V, E)$ from the set of polygons. Each node in V corresponds to a polygon, and an edge is added between polygons i and j if they satisfy a spatial adjacency criterion. Two adjacency types are considered:

- **Edge-based adjacency**: polygons share a non-zero boundary length.
- **Point-based adjacency**: polygons touch at a single vertex.

In this work, we adopt edge-based adjacency to better preserve geometric continuity and avoid spurious connections caused by point contacts. The edge weights w_{ij} are subsequently computed based on the shared boundary length l_{ij}. Details on this design choice and related settings are discussed further in Section 3.4.



(a) Raw Polygon Proposal (b) Adjacency by point (c) Adjacency by edge

Figure 3.6.: Different kinds of Adjacency graph

3.4.3. Unary Term

The unary term quantifies how likely a polygon corresponds to each semantic label, based on predictions from Mask R-CNN.

Confidence-weighted mask probability. For each predicted instance i, Mask R-CNN outputs a soft mask $M_i(x, y) \in [0, 1]$ and a classification confidence $s_i \in [0, 1]$. We define:

$$P_i(x, y) = s_i \cdot M_i(x, y) \quad \text{and} \quad P_{\text{bg}}(x, y) = 1 - \max_i P_i(x, y) \tag{3.2}$$

Per-polygon label probability. Each polygon k is rasterized to generate a binary mask Ω_k, and the average probability over each label j is computed as:

$$p_{k,j} = \frac{1}{|\Omega_k|} \sum_{(x,y)\in\Omega_k} P_j(x, y) \tag{3.3}$$

Unary Cost Transformation: To encourage the model to favor high-probability label assignments, we transform the label probability into a cost. Specifically, lower probabilities incur higher costs:

$$U_{k,j} = (1 - p_{k,j}) \cdot \alpha \tag{3.4}$$

Here, α is a scaling parameter that controls the penalty strength for low-confidence label assignments.

While the unary term captures how well each polygon aligns with semantic predictions, it alone may result in noisy or inconsistent labeling. To promote spatial coherence across adjacent polygons, we introduce a pairwise term that encodes neighborhood smoothness.

3.4.4. Pairwise Term

The pairwise term encourages spatial smoothness by penalizing label differences between adjacent polygons.

Edge weight normalization. For each edge $(i, j) \in E$, we compute a normalized weight based on the shared boundary length l_{ij}:

$$w_{ij} = f(l_{ij}) = \sqrt{l_{ij}/l_{\max}} \cdot \text{scale} + \text{offset} \tag{3.5}$$

We empirically use $l_{\max} = 80$ pixels, scale $= 5$, and offset $= 1$. This transformation mitigates the impact of extremely long edges due to the long-tailed distribution of l_{ij} values.

Label penalty matrix. To discourage inconsistent labels across adjacent polygons, we define a symmetric penalty:

$$V(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{otherwise} \end{cases} \tag{3.6}$$

Final pairwise cost. Combining both components, the full pairwise term becomes:

$$\text{PairwiseCost}_{ij}(a, b) = \lambda \cdot w_{ij} \cdot V(a, b) \tag{3.7}$$

The hyperparameter λ controls the overall strength of spatial regularization. Larger values of λ favor label smoothness; smaller values emphasize unary confidence.

Having defined both the unary and pairwise components of the energy function, we now turn to the optimization process used to infer the most likely label configuration.

3.4.5. Energy Minimization and Label Inference

Given the unary cost matrix $U \in \mathbb{R}^{N\times(N+1)}$ and the edge-weighted adjacency graph, the final label configuration $L = \{L_1, L_2, \ldots, L_N\}$ is obtained by minimizing the total energy:

$$L^* = \arg\min_L \left(\sum_k U_k(L_k) + \sum_{(i,j)} \lambda \cdot w_{ij} \cdot V(L_i, L_j) \right) \tag{3.8}$$

This multi-label optimization problem is solved using the α-expansion algorithm [Boykov et al., 2001], which efficiently minimizes the MRF energy by iteratively proposing large label moves while preserving approximate optimality. The resulting label assignment provides a globally consistent segmentation based on both semantic evidence and geometric adjacency.

After inference, the labeled polygons are converted into pixel-level masks for evaluation by rasterizing each polygon into a 2D canvas using its predicted label. This produces image-aligned segmentation maps that are directly comparable to ground truth annotations. These outputs are evaluated using standard semantic metrics (IoU, precision, recall) and structural consistency scores.

4. Implementation, Results, and Discussion

4.1. Implementation Details

4.1.1. Implementation Overview

To facilitate instance-level roof segmentation from 2D imagery, we propose a modular framework composed of three main stages: instance segmentation, polygon proposal generation, and MRF optimization. First, Mask R-CNN is employed to detect rooftop instances and produce soft masks. Then, the structural outlines of roofs are extracted using edge detection and converted into candidate polygons through the KIPPI algorithm. Finally, a Markov Random Field (MRF) model is constructed over these polygons to assign binary labels, integrating both semantic cues from the soft masks and spatial coherence from the polygon adjacency graph.

This chapter provides key implementation details of each module, including dataset preparation, model configuration, polygon generation parameters, and MRF inference.

4.1.2. Dataset Preparation

The **Cities** dataset [Nauata and Furukawa, 2020] contains 1,985 aerial images with a fixed resolution of 256×256 pixels. It features industrial and commercial buildings with flat yet complex rooftops. Many images include large white margins, leading to reduced effective rooftop areas.

The **RoofVec** dataset [Hensel et al., 2021] comprises 7,640 images of varying resolutions, primarily depicting residential buildings with sloped and relatively simpler roof structures.

Neither dataset is orthorectified. As part of preprocessing, we convert all rooftop annotations into polygon-based semantic masks, which serve as supervision for training the instance segmentation model. Following this, we randomly split each dataset into 60% for training, 20% for validation, and 20% for testing.

4.1.3. Instance Segmentation Model Training

Backbone Configuration We adopt Mask R-CNN [He et al., 2017] with a ResNet-50 backbone and a Feature Pyramid Network (FPN) architecture for instance-level rooftop segmentation. The output heads are configured to produce soft masks for each rooftop instance.

Training Procedure The model is trained using stochastic gradient descent (SGD) with a batch size of 4, momentum of 0.9, weight decay of 0.0005, and an initial learning rate of 0.005, which is reduced by a factor of 0.1 every 20 epochs. Early stopping is triggered after 15 epochs without improvement in validation IoU. All experiments are conducted on NVIDIA A100 GPUs. Loss and IoU metrics are logged at each epoch.

Validation Strategy Validation is performed on the 20% held-out set. Predicted masks are matched to ground truth instances using an IoU threshold of 0.5. A greedy matching strategy based on descending confidence scores is applied, with each ground truth instance matched at most once. The mean IoU across matched pairs is used for validation and model selection.

Training Curves Figures 4.1–4.3 illustrate the training dynamics. The learning rate follows a step-decay schedule. Training loss steadily decreases, and validation IoU improves until convergence, indicating effective training and minimal overfitting.

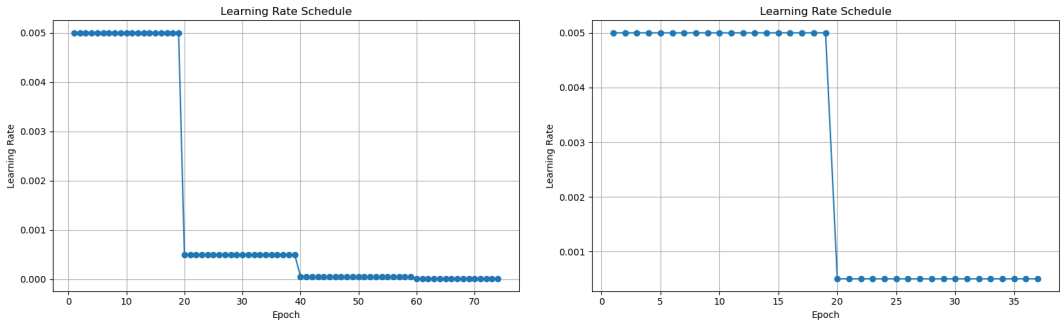


Figure 4.1.: Learning rate schedules during training for RoofVec (left) and Cities (right).

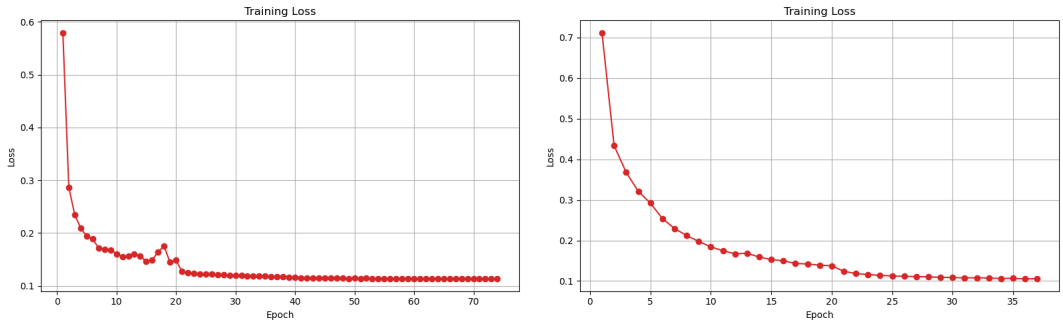


Figure 4.2.: Training loss curves over epochs for RoofVec (left) and Cities (right).

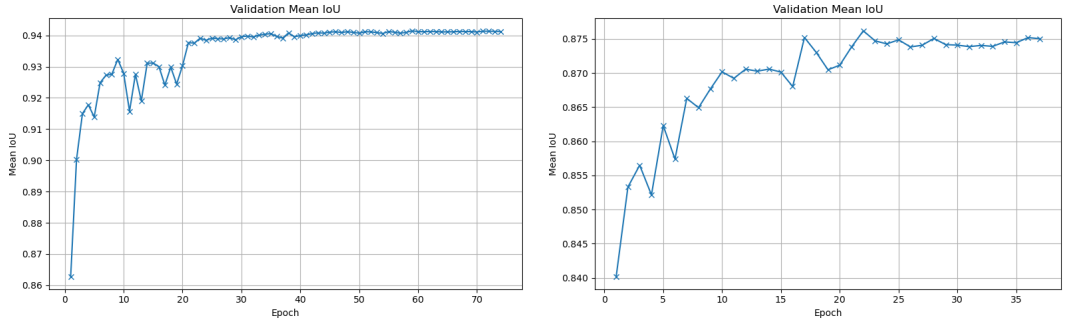


Figure 4.3.: Validation mean IoU over epochs for RoofVec (left) and Cities (right).

4.1.4. Polygon Proposal Generation

We use the line detection module from the KIPPI pipeline to extract structural line segments from RGB aerial images in the Cities dataset. These segments are then aggregated into polygonal rooftop candidates through a combination of intersection refinement and polygonization.

Parameter Configuration Two parameters significantly affect the granularity of the resulting polygon proposals: (1) `lsd_scale`, which controls the degree of Gaussian smoothing prior to line detection, and (2) `num_intersections`, which defines the number of iterations for resolving line intersections. We conduct a grid search over `lsd_scale` $\in \{0.8, 1.0, 1.2, 1.5, 1.8\}$ and `num_intersections` $\in \{1, 2, 5, 7, 9\}$.

Quantitative Analysis (10-Image Sample) Table 4.1 summarizes descriptive statistics for selected parameter combinations, computed over a sample of 10 randomly selected test images from the Cities dataset. Each cell reports the mean and standard deviation across these 10 images. As `num_intersections` increases, the number of polygons rises sharply, while average area and perimeter decrease—indicating finer spatial decomposition. Meanwhile, the average vertex count remains relatively stable, reflecting preserved structural complexity per shape.

Table 4.1.: Proposal statistics for selected parameter combinations (10 Cities test images).

LSD / Intersections	Polygon Count	Vertex Count	Area (pixels^2)	Perimeter (pixels)
0.8 / 1	64.1 ± 36.8	6.81 ± 2.26	1021.8 ± 2833.4	137.4 ± 114.5
1.2 / 2	200.3 ± 115.6	5.81 ± 1.40	324.0 ± 1029.7	68.9 ± 66.2
1.5 / 9	**1010.4 ± 636.5**	5.23 ± 1.02	**64.9 ± 319.1**	**27.4 ± 34.4**
1.8 / 9	620.4 ± 182.6	5.39 ± 1.18	102.6 ± 467.5	35.2 ± 41.5

Full results for all parameter combinations are listed in Appendix A.2.

Final Selection and Full Evaluation We adopt `lsd_scale` = 1.5 and `num_intersections` = 9 for all subsequent experiments. This configuration achieves a high number of proposals while preserving reasonable polygon complexity, making it well-suited for downstream MRF-based labeling.

Table 4.2 shows summary statistics computed over the full Cities test set using this final configuration. These values provide a more comprehensive view than the earlier 10-image sample.

Table 4.2.: Final polygon statistics (LSD = 1.5, Intersections = 9) over the entire Cities test set.

Metric	Mean	Std. Dev.	Median	Max
Polygons per image	1416.9	767.8	1277.0	4153
Polygon area (pixels^2)	46.3	244.5	7.18	20607.5
Polygon perimeter (pixels)	24.5	29.8	15.3	649.7
Vertex count per polygon	5.31	1.10	5.00	27

4.1.5. MRF Optimization

Graph Construction As introduced in the Section 3.4.2, we represent polygon proposals as an undirected graph where each polygon corresponds to a node. Among the adjacency strategies discussed, we implement only edge-based adjacency: an edge is created between two polygons if they share a non-zero boundary segment. The length of the shared boundary is computed and stored for each pair and later used to determine pairwise weights in the MRF formulation.

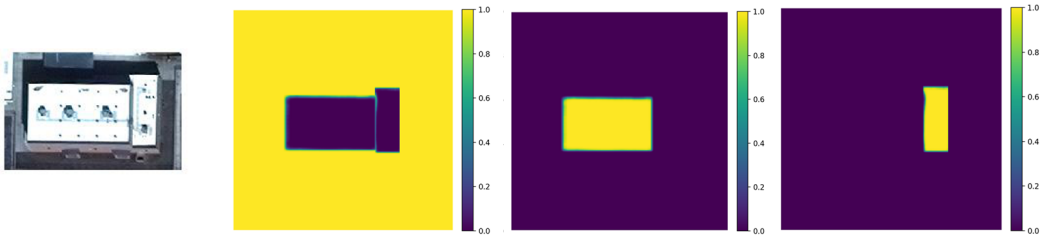
Unary Term Computation As introduced in Section 3.4.3, the unary term reflects how likely each polygon corresponds to semantic classes, based on predictions from Mask R-CNN. In practice, this is computed by first rasterizing each polygon onto the image plane and then averaging the soft mask probabilities within the region to obtain class-wise probabilities $p_{k,j}$. These values are subsequently transformed into energy costs using the inverse mapping $U_{k,j} = (1 - p_{k,j}) \cdot \alpha$, with $\alpha = 10$.

We fix $\alpha = 10$ across all experiments to provide a stable unary baseline. This allows us to systematically explore how the overall label assignment is affected by the strength of spatial regularization, which is controlled separately through the pairwise term.

Figure 4.4 visualizes the class-wise soft probability masks produced by Mask R-CNN for a sample input, which form the basis for downstream polygon-level aggregation. These probability maps serve as the foundation for computing polygon-level scores.

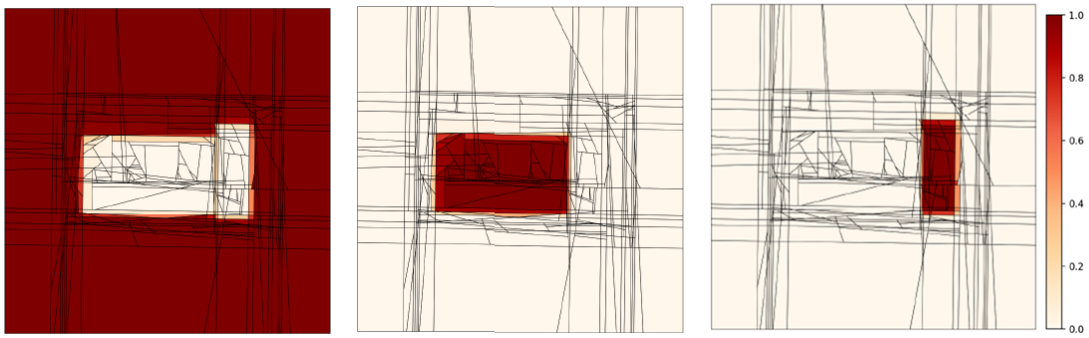
Figure 4.5 further illustrates the results of polygon-level aggregation. The top row shows the averaged per-polygon probabilities, while the bottom row displays the resulting unary cost matrix. Darker regions indicate higher energy penalties, typically corresponding to low-confidence predictions.

For completeness, Appendix A.3 provides a numerical example showing the per-polygon probability matrix and its transformed unary cost values, illustrating the exact input format for MRF optimization.



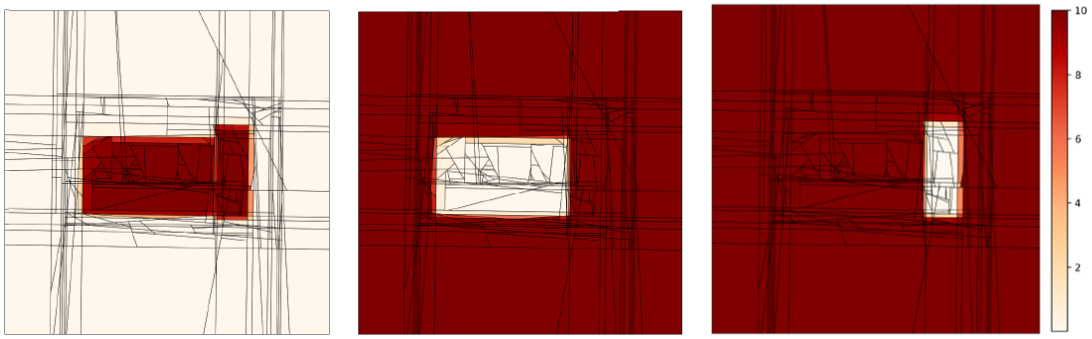
(a) Raw RBG (b) Class 0 (c) Class 1 (d) Class 2

Figure 4.4.: Raw pixel-level soft masks from Mask R-CNN. Values represent per-pixel probabilities for each class.



(a) Polygon Prob (raw) class 0 (b) Polygon Prob (raw) class 1 (c) Polygon Prob (raw) class 2

Figure 4.5.: Per-polygon class probability maps computed from instance segmentation outputs. Values represent the aggregated semantic likelihood for each polygon belonging to class 0, 1, or 2, respectively.



(a) Unary cost - class 0 (b) Unary cost - class 1 (c) Unary cost - class 2

Figure 4.6.: Corresponding unary cost maps derived from the polygon-level class probabilities. Higher values indicate lower confidence and contribute more to the MRF energy.

Pairwise Term Definition To encourage spatial smoothness in the final labeling, we define pairwise costs between adjacent polygons based on the length of their shared boundary. Let l_{ij} denote the edge length shared by polygon pair (i, j). A direct linear use of l_{ij} tends to overemphasize a small number of large connections, due to the long-tailed distribution of edge lengths observed in practice.

Figure 4.7 illustrates this distribution over the test set. Most polygon pairs share very short boundaries, while a few pairs have disproportionately long edges, which can dominate the energy minimization if not properly normalized.

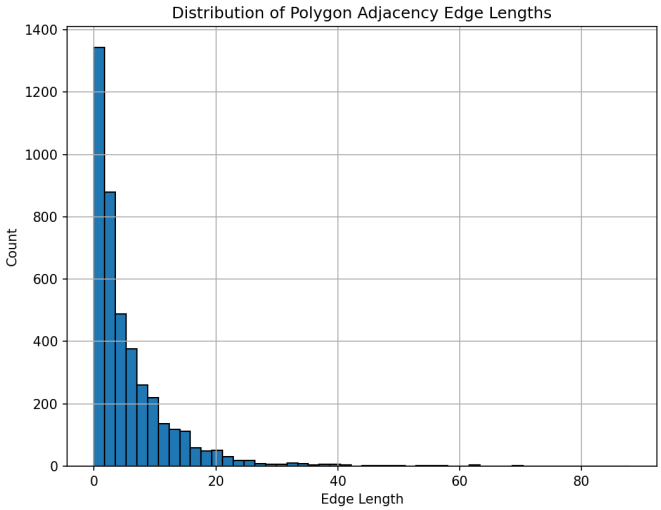


Figure 4.7.: Histogram of raw shared edge lengths.

To mitigate this, we apply a square root transformation and cap the values at 80 pixels. The normalized weight is computed as:

$$w_{ij} = \sqrt{l_{ij}/80} \cdot 5 + 1 \tag{4.1}$$

This transformation compresses the dynamic range while preserving the relative order of edge strengths. The square root reduces the influence of extreme long edges without nullifying their spatial importance.

The scaling factor ($\cdot 5$) and offset ($+1$) are chosen empirically to map the resulting weights into a numerically stable range, approximately $[1, 6]$. This aligns the magnitude of pairwise terms with the unary energy scale (controlled by $\alpha = 10$), ensuring balanced contributions from both semantic and geometric cues. The offset also prevents zero-weight edges, which would eliminate spatial regularization for weakly connected regions.

The final pairwise energy between two adjacent polygons becomes:

$$\text{Cost}_{ij}(a, b) = \lambda \cdot w_{ij} \cdot \mathbb{1}[a \neq b] \tag{4.2}$$

where λ is a smoothness hyperparameter, and $\mathbb{1}[a \neq b]$ indicates whether the polygons are assigned different labels. This encourages label consistency along long, continuous boundaries.

Inference and Rasterization The MRF energy function, consisting of unary and pairwise terms, is minimized to obtain a label assignment for all polygons. Each polygon is then rasterized back into the image plane using its assigned label, producing a dense mask for evaluation. These outputs are assessed using both semantic metrics (e.g., mean IoU) and structural criteria such as label consistency across adjacent regions.

Smoothness Parameter Sweep We explore the impact of the smoothness weight λ through a two-stage sweep. A coarse search over $\{0.001, 0.01, 0.1, 1, 10, 100\}$ identifies the general sensitivity range. A finer search within the transition region (e.g., $\lambda = 0.05$) is then conducted to refine the trade-off between over-smoothing and under-regularization. The best value is selected based on downstream performance on both semantic and structural metrics.

4.2. Result Analysis and Comparison

4.2.1. Quantitative Evaluation Metrics

We adopt standard instance-level evaluation metrics to quantify model performance: True Positives (TP), False Positives (FP), False Negatives (FN), Precision, Recall, and Mean Intersection over Union (IoU). These are defined as follows:

- **Precision:** the proportion of predicted instances that are true rooftops.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.3}$$

- **Recall:** the proportion of ground truth rooftops that are correctly predicted.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.4}$$

- **Mean IoU:** the average intersection-over-union score over all matched predictions.

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{4.5}$$

  where P is the predicted polygon and G is the corresponding ground truth. We compute IoU for each matched pair and report the average over all matches.

These metrics capture both semantic correctness and instance-level overlap quality, and are used consistently throughout this section.

We evaluate the quantitative performance of our method on two datasets: **Cities** and **RoofVec**. Comparisons are made across three stages of the pipeline: (1) raw instance predictions from Mask R-CNN, (2) MRF labeling using only unary probabilities, and (3) full MRF inference using unary and pairwise term.

Baseline Performance: Mask R-CNN We begin by reporting the instance segmentation performance of the raw Mask R-CNN model. Table 4.3 summarizes the metrics. On the Cities dataset, the precision is only around 0.59 due to many overlapping or redundant predictions. RoofVec performs better, but still has room for improvement.

Table 4.3.: Instance-level performance of Mask R-CNN on test sets.

Dataset	TP	FP	FN	Precision
Cities	870	600	197	0.5918
RoofVec	3338	356	48	0.9036

Ablation Study: Smoothness Weight (λ) in MRF To evaluate the impact of spatial regularization, we evaluate the MRF performance across a range of smoothness weights $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 10.0, 100.0\}$, with the unary scale α fixed at 10. Increasing λ strengthens spatial smoothness, which helps suppress false positives. However, overly large λ values cause the pairwise term to dominate, leading to excessive smoothing and foreground collapse. Since mean IoU depends on the number and quality of true positives, it does not necessarily improve as false positives decrease. We highlight in bold the configurations that best balance semantic accuracy and structural coherence. On the RoofVec dataset, $\lambda = 0.1$ achieves the highest mean IoU, while $\lambda = 0.5$ further reduces false positives with only a slight drop in IoU.

Ablation: Unary-Only vs. Full MRF To isolate the effect of pairwise terms, we compare the performance of the unary-only labeling with the full MRF output at $\lambda = 0.1$. Table 4.5 shows that using only the unary term provides reasonable segmentation, but adding the pairwise term improves overall label coherence. On Cities, mean IoU improves slightly, and false positives decrease. On RoofVec, the improvement is more significant — both precision and IoU increase, showing that pairwise consistency helps especially on clean, structured datasets.

Performance Summary Across all evaluations, the proposed framework consistently improves structural integrity and semantic accuracy over baseline Mask R-CNN outputs. The full MRF model, combining semantic confidence with spatial smoothness, offers a principled post-processing solution for instance segmentation refinement.

4.2.2. Comparative Analysis

We compare our method against several structured polygon prediction approaches using the region-level F_1-score, following an evaluation protocol where predicted and ground-truth polygons are matched based on an IoU threshold of 0.7 [Nauata and Furukawa, 2020].

Under this setting, our method achieves a region F_1-score of **66.1%**, outperforming previous graph- and sequence-based methods [Nauata and Furukawa, 2020; Zhang et al., 2021]. Although one recent fully-convolutional method [Chen et al., 2021] reports a higher F_1-score of 70.6%, our method maintains a simpler modular design and does not rely on edge supervision, transformer-based decoders, or joint primitive prediction.

Table 4.4.: MRF sweep results on the Cities and RoofVec datasets (unary scale = 10).

(a) Cities dataset

λ	TP	FP	FN	Precision	Recall	Mean IoU
0.001	749	319	200	0.7012	0.7893	0.8331
0.01	749	316	200	0.7035	0.7893	0.8331
0.05	749	313	200	0.7051	0.7893	0.8333
0.1	**748**	**306**	**201**	**0.7097**	**0.7882**	**0.8336**
0.5	738	277	211	0.7271	0.7778	0.8260
1.0	723	262	226	0.7340	0.7620	0.8179
10.0	373	84	576	0.8162	0.3937	0.5064
100.0	0	1	949	0.0000	0.0000	0.0000

(b) RoofVec dataset

λ	TP	FP	FN	Precision	Recall	Mean IoU
0.001	3265	111	90	0.9362	0.9731	0.9125
0.01	3264	109	91	0.9370	0.9729	0.9127
0.05	**3263**	**99**	**92**	**0.9705**	**0.9726**	**0.9130**
0.1	**3262**	**94**	**93**	**0.9720**	**0.9723**	**0.9130**
0.5	3261	73	94	0.9781	0.9720	0.9117
1.0	3248	70	107	0.9789	0.9681	0.9049
10.0	1863	96	1492	0.9510	0.5552	0.5568
100.0	14	69	3341	0.1683	0.0042	0.0063

Table 4.5.: Comparison of polygon labeling with and without MRF smoothing.

Dataset	**Method**	TP	FP	FN	Precision	Mean IoU
Cities	Unary only	749	319	200	0.7012	0.8331
	MRF ($\lambda = 0.1$)	748	**306**	201	**0.7097**	**0.8336**
RoofVec	Unary only	3265	111	90	0.9362	0.9125
	MRF ($\lambda = 0.1$)	**3262**	**94**	93	**0.9720**	**0.9130**

For completeness, we also report our performance under a relaxed IoU threshold of 0.5, where the region F_1-score increases to **74.6%**, highlighting robustness to small geometric discrepancies.

Table 4.6.: Region-level F_1 score (%) comparison with prior methods on roof structure extraction, evaluated at IoU threshold 0.5 and 0.7.

Method	IoU Threshold	Region F_1 (%)
Nauata and Furukawa [2020]	0.7	60.8
Zhang et al. [2021]	0.7	63.5
Chen et al. [2021]	0.7	**70.6**
Ours	0.7	**66.1**
Ours	0.5	**74.6**

In addition to the region-level F_1-score, we also compute the Macro Mean IoU between matched polygon pairs. Our method achieves 0.8286 at an IoU threshold of 0.7 and 0.8406 at 0.5, indicating high geometric consistency across predicted and ground-truth regions. While previous works do not report this metric, we believe it offers complementary insight into the alignment quality of predicted shapes.

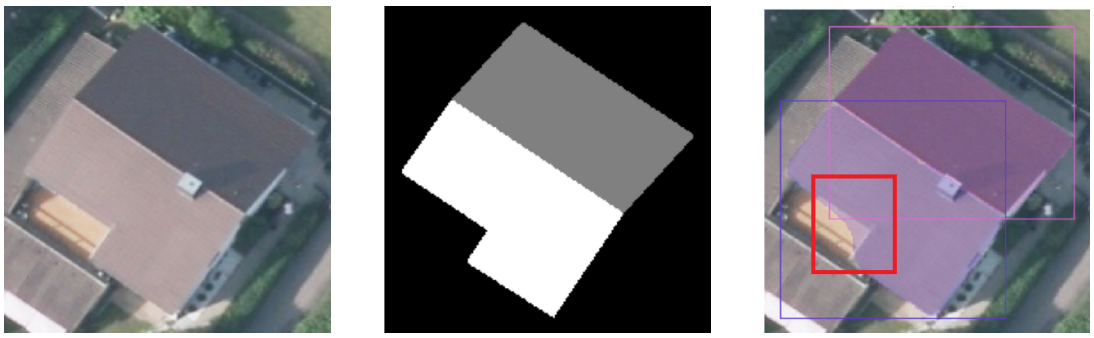
We further observe that under a stricter IoU threshold (0.7), false positives tend to increase, which we attribute to instance-level over-segmentation from the initial proposal stage. Specifically, a single ground-truth region is often fragmented into multiple predicted instances, each covering only part of the target shape. These redundant polygons dilute the match quality and lead to an inflated false positive count. A more precise or post-refined instance generation process could potentially reduce such errors and narrow the remaining performance gap.

4.2.3. Qualitative Visualization

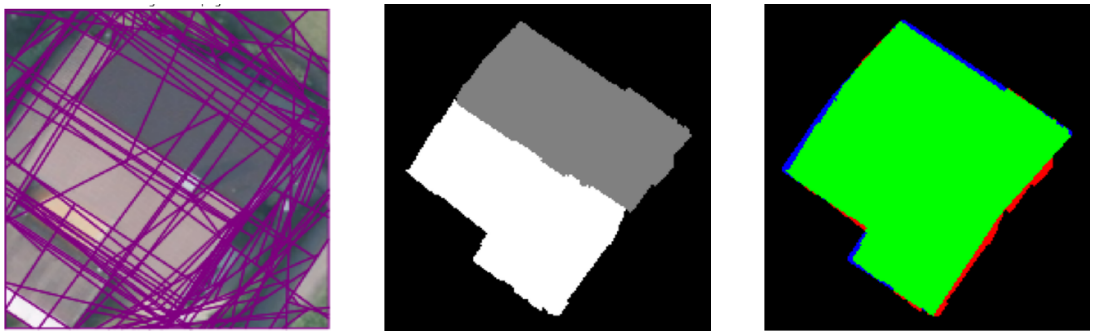
We present a series of qualitative examples to demonstrate the impact of our method on initial instance segmentation results. Each example highlights a specific issue commonly found in raw predictions and shows how our method addresses it effectively.

The following qualitative examples (Figures 4.8–4.11) illustrate how different components of our MRF model address common segmentation issues such as spurious fragments, over-fragmentation, internal boundary noise, and misalignment.

Example 1: Removing Spurious Fragments In this example, a small extra triangular fragment appears in the raw instance segment prediction (highlighted by the red box in Figure 4.8 (c)). Such fragments are typically caused by noisy or ambiguous boundaries. After applying the MRF refinement, the fragment is correctly removed, yielding a cleaner segmentation output and demonstrating the model's ability to suppress isolated, erroneous predictions.



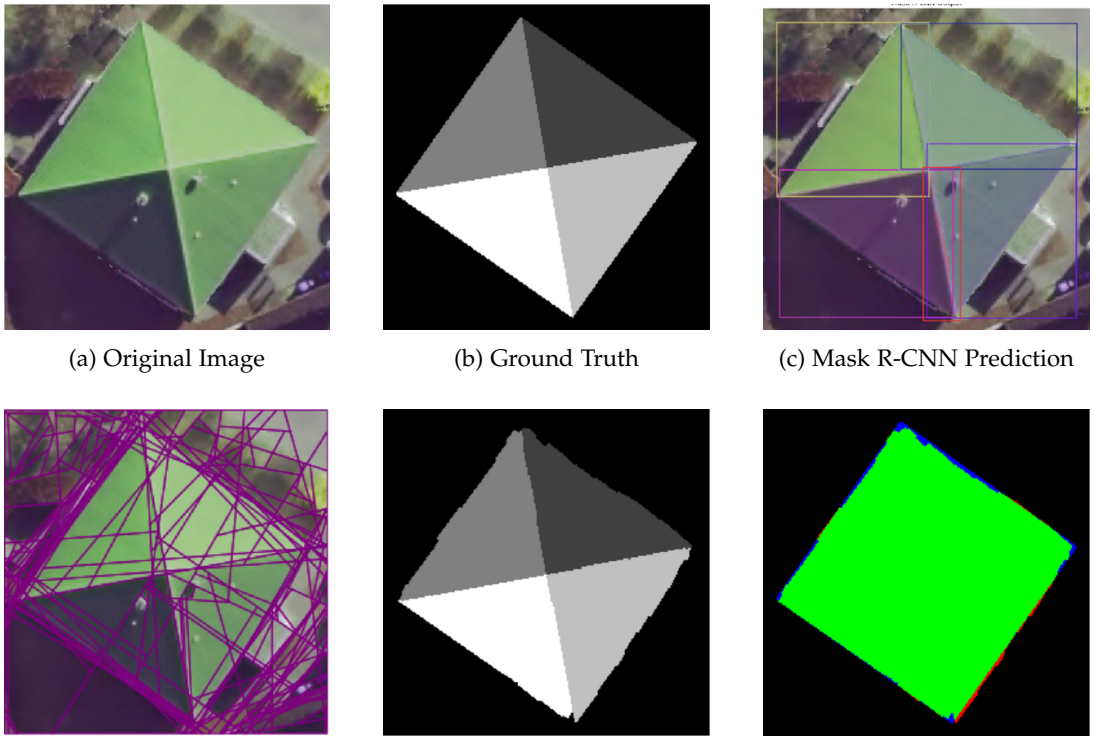
(a) Original Image (b) Ground Truth (c) Mask R-CNN Prediction



(d) Polygon Proposal (e) MRF Result (f) MRF Instance (TP: Green, FP: Red, FN: Blue)

Figure 4.8.: MRF post-processing removes spurious fragments and refines boundaries to produce cleaner segmentations.

Example 2: Merging Over-Segmented Instances The raw Mask R-CNN prediction exhibits over-fragmentation, producing six separate instances for what is in fact a single object (see Figure 4.9 (c)). This not only increases the perceived object count but also disrupts structural coherence. Our MRF refinement successfully merges these redundant segments into a unified instance mask.



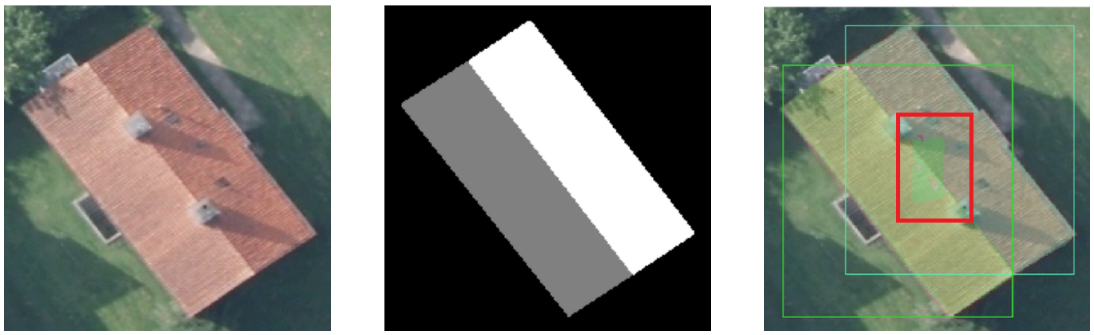
(a) Original Image (b) Ground Truth (c) Mask R-CNN Prediction

(d) Polygon Proposal (e) MRF Result (f) MRF Instance (TP: Green, FP: Red, FN: Blue)

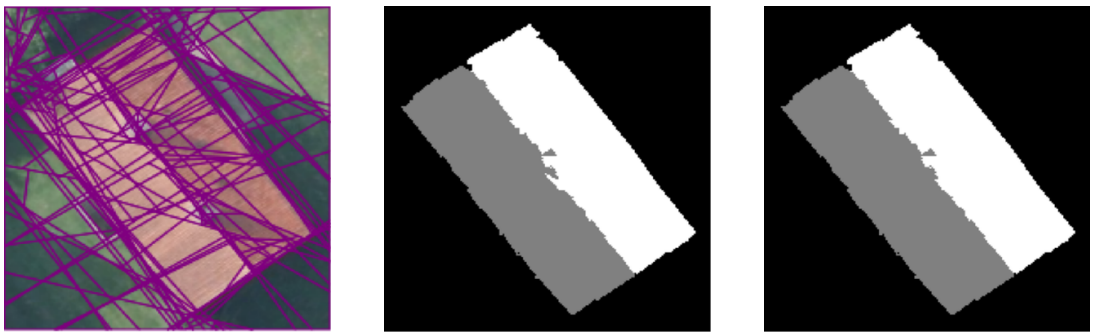
Figure 4.9.: MRF merges over-segmented instances into cohesive masks and reduces redundant predictions.

Example 3: Smoothing Internal Boundaries As seen in the red box in Figure 4.10 (c), the baseline prediction includes jagged internal boundaries within a single instance. These result in fragmented shapes and poor alignment. The unary-only MRF result in Figure 4.10 (e) improves the shape but retains some irregularities. The full MRF (Figure 4.10 (f)), with pairwise terms, produces smoother, more consistent boundaries.

Example 4: Correcting Alignment and Filling Gaps As shown in Figure 4.11 (c), the baseline prediction contains overlapping and misaligned masks. Unary-only refinement (Figure 4.11 (e)) introduces false holes due to limited spatial context. By incorporating the pairwise term (Figure 4.11 (f)), these holes are corrected and the masks are reshaped to better match the ground truth.

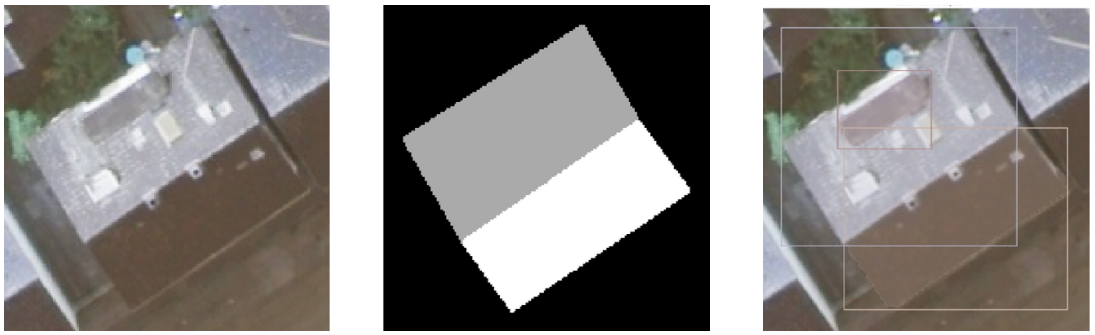


(a) Original Image (b) Ground Truth (c) Mask R-CNN Prediction



(d) Polygon Proposal (e) MRF Result (Unary Only) (f) MRF Result (Full)

Figure 4.10.: MRF smooths jagged internal boundaries using pairwise terms, resulting in more geometrically consistent masks.



(a) Original Image (b) Ground Truth (c) Mask R-CNN Prediction

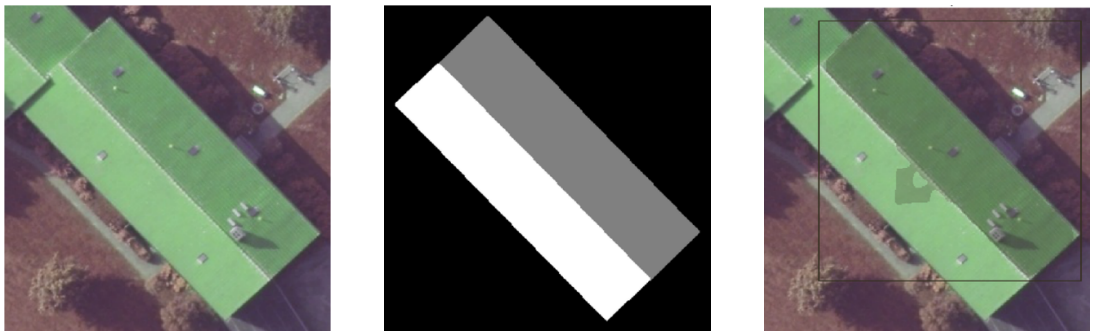
(d) Polygon Proposal (e) MRF Result (Unary Only) (f) MRF Result (Full)

Figure 4.11.: MRF improves alignment and fills false gaps by combining unary and pairwise terms.

4.2.4. Failure Cases

Despite the overall performance improvements, our framework exhibits several failure modes. We categorize and illustrate them as follows.

Example 1: Missed Detection from Mask R-CNN Figure 4.12 illustrates a fundamental limitation of our framework: when the base instance segmentation model completely misses a rooftop, subsequent post-processing cannot recover it. In this case, the Mask R-CNN output only captures the right-side rooftop, completely missing the left segment. As our method relies on unary probabilities projected onto polygon proposals, the absence of any initial detection results in irreversible false negatives.

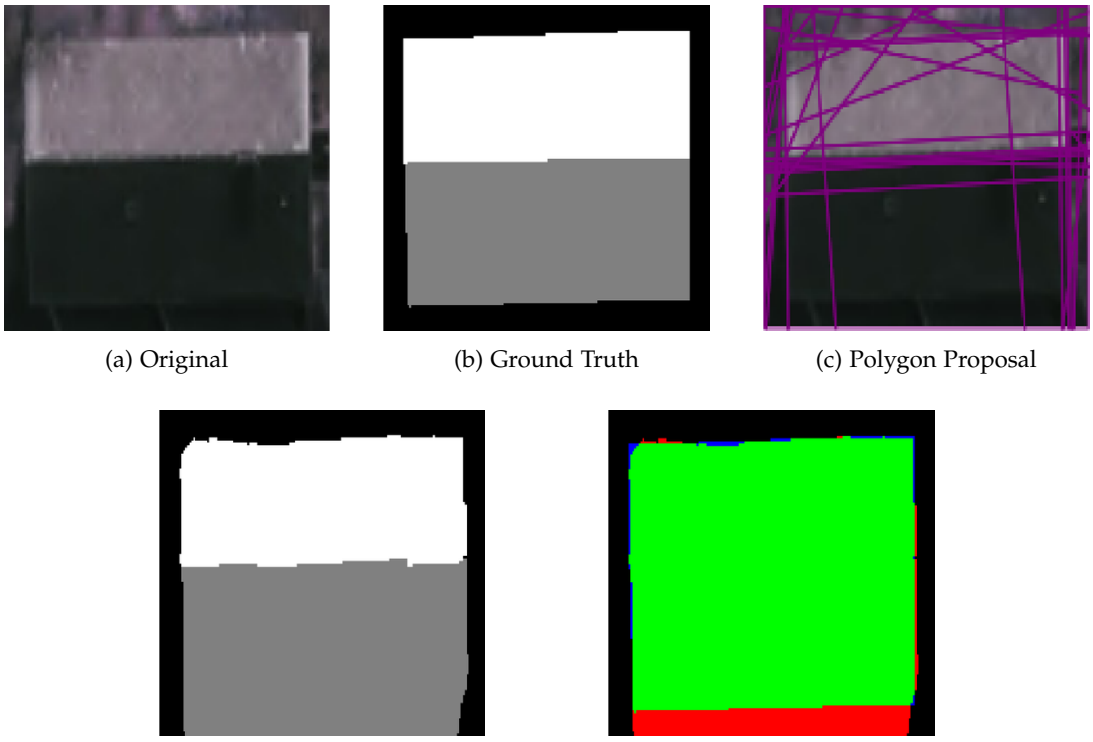


(a) Original image (b) Ground Truth (c) Mask R-CNN Prediction

(d) Polygon Proposal (e) MRF result (f) MRF Instance TP/FP/FN

Figure 4.12.: Failure Case 1: Missed rooftop on the left cannot be recovered without any initial detection.

Example 2: Shadowed Rooflines Lead to Missing Proposals Another failure pattern arises from weak polygon initialization. As shown in Figure 4.13, the bottom portion of the rooftop is heavily shadowed, leading to a lack of strong line segments in the initial proposal phase. Even under a dense intersection configuration, the bottom rooftop boundary remains poorly delineated.



(a) Original (b) Ground Truth (c) Polygon Proposal

(d) MRF Result (e) TP / FP / FN

Figure 4.13.: Failure Case 2: Shadow occlusion causes missing polygon edges in the proposal phase, leading to incorrect MRF segmentation.

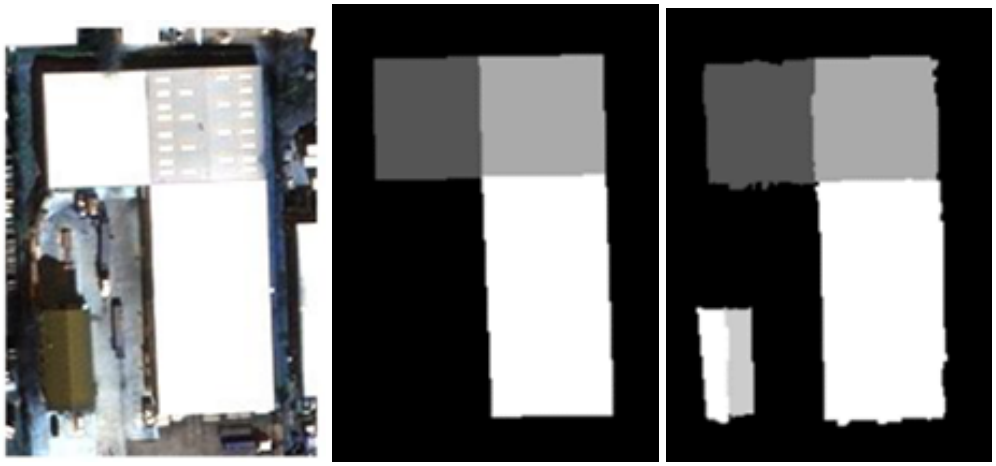
Example 3: Instance Over-Segmentation Limits Refinement As shown in Figure 4.14, Mask R-CNN initially predicts eight fragmented instances (N = 8), some of which are misaligned. The subsequent polygon matching and MRF refinement are limited by this fragmented input. Although GCO reduces label noise, the initial over-segmentation imposes a ceiling on final quality. This failure highlights the challenge of resolving complex or repetitive roof patterns when instance masks are highly fragmented.

(a) Original (b) Ground Truth (c) Polygon Proposal

(d) MRF Result (e) TP / FP / FN

Figure 4.14.: Failure Case 3: Over-segmented instance masks limit final quality, even with MRF refinement.

Example 4: Annotation Noise Causes Evaluation Errors Finally, Figure 4.15 demonstrates the influence of annotation quality on evaluation outcomes. The left-bottom rooftop is clearly visible in the input image and correctly predicted by the model, yet it is absent in the ground truth annotation. This results in the prediction being counted as a false positive, despite being semantically valid. Such incomplete labeling introduces ambiguity during both training and evaluation, underscoring the need for cleaner, human-verified annotations in urban-scale datasets.



(a) Original (b) GT (c) Our Result

Figure 4.15.: Failure case caused by incomplete ground-truth annotation. The rooftop in the lower-left corner is correctly segmented by our method (c) but is missing in the ground truth (b), leading to an apparent false positive during evaluation.

These failure cases highlight critical limitations of our pipeline—namely, its reliance on accurate instance proposals, coherent polygon extraction, and high-quality annotations. Future improvements could incorporate learned polygon refinement or uncertainty-aware training to address such limitations.

4.2.5. Application Demonstration

To illustrate the practical benefits of our proposed pipeline, we present a real-world demonstration of its applicability in 3D urban modeling. By combining our refined rooftop segmentations with height data, we can generate plausible 3D reconstructions of urban buildings. These outputs support downstream applications such as city planning, solar panel placement, and disaster response simulation.

Given the cleaned and instance-aware rooftop masks obtained through our MRF optimization process, we extrude each segmented region into a 3D polygon using estimated building heights. For demonstration purposes, we assign height values manually or infer them from LiDAR or DSM (Digital Surface Model) data when available. Each rooftop instance is represented as a planar polygon extruded vertically to form a prism.

To highlight the geometric expressiveness of our approach, we showcase reconstructions of several common roof types:

- **Gable roof**: Two sloped planes meeting along a central ridge.
- **Pyramidal hip roof**: Four sloped planes converging to a single apex.
- **Hip roof with ridge**: Four sloped planes meeting along a horizontal ridge line.
- **Flat roof**: A horizontal planar roof surface with vertical walls.

The following examples show 3D renderings of these rooftop structures from multiple viewing angles.

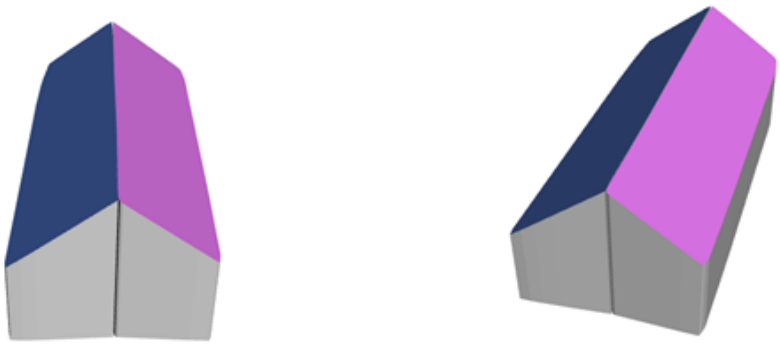


Figure 4.16.: Example 1: Gable roof reconstruction. Two sloped surfaces form a central ridge.

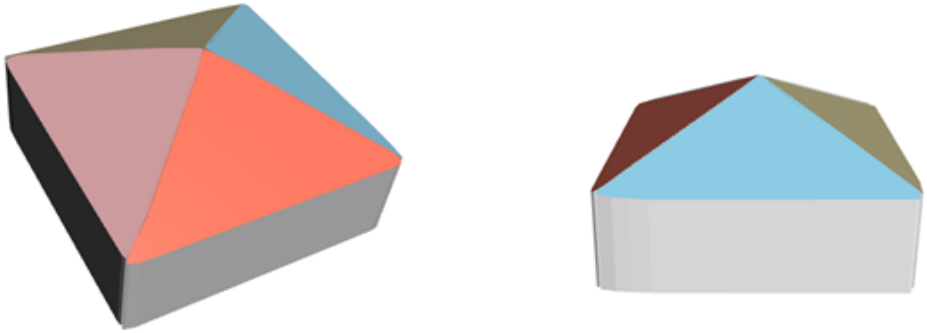


Figure 4.17.: Example 2: Pyramidal hip roof reconstruction. Four sloped planes meet at a single apex point.

These visualizations demonstrate that even with simple extrusion, our pipeline is capable of generating plausible and structurally diverse 3D rooftops. This enables scalable and lightweight reconstruction of urban environments using only 2D aerial imagery and minimal elevation data.

4.3. Discussion

This thesis presents a three-stage pipeline for rooftop segmentation that combines instance-level segmentation, polygon proposal generation, and MRF-based optimization. Unlike prior approaches—such as building segmentation, building instance segmentation, roof line

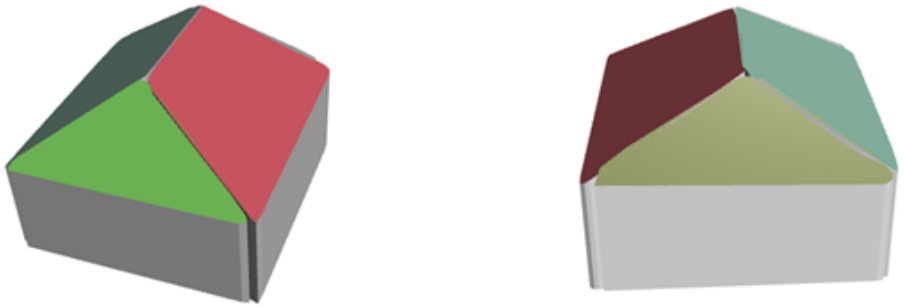


Figure 4.18.: Example 3: Hip roof with ridge. The sloped surfaces form a horizontal ridge at the top.

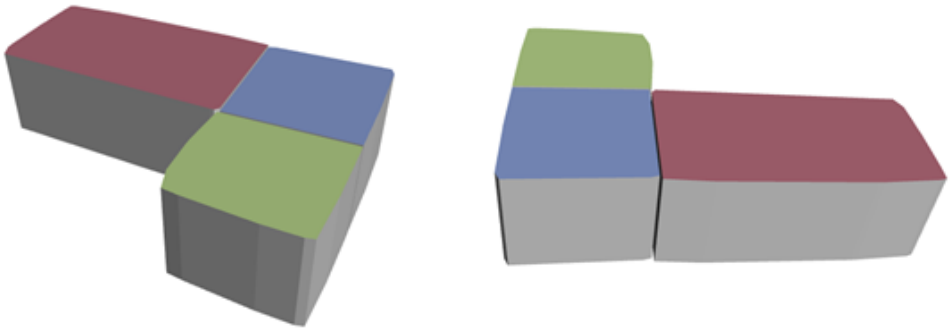


Figure 4.19.: Example 4: Flat roof structure. All roof faces are horizontal with vertical extrusion walls.

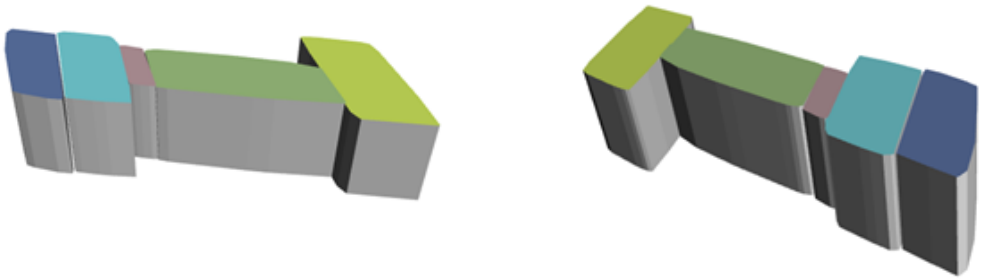


Figure 4.20.: Example 5: Another example of a flat roof building generated using our extrusion pipeline.

extraction, or roof planar structure extraction—our method integrates both geometric structure and semantic confidence into a unified, modular, and model-agnostic framework. This enables more consistent, interpretable, and robust results across diverse urban scenes.

A core motivation for this work lies in addressing the structural challenges commonly observed in rooftop segmentation. As discussed in Chapter 1, conventional instance segmentation models often suffer from visual ambiguities caused by occlusion, shadows, or background clutter. These effects lead to incomplete roof boundaries, broken masks, and inconsistent instance assignments—particularly in dense urban scenes where rooftops are tightly packed and visually similar.

Such issues are clearly reflected in the *Cities* dataset, where the raw predictions from Mask R-CNN frequently contain overlapping segments, fragmented regions, or missing rooftop parts. These errors result in low semantic precision (around 0.59) and undermine the geometric integrity needed for downstream applications. By introducing a polygon-based intermediate representation and applying MRF-based optimization, our pipeline effectively reduces redundancy, merges fragmented parts, and restores structurally plausible rooftop regions. As a result, both precision and intersection-over-union (IoU) scores improve significantly. In addition, experiments with different polygon proposal settings—such as line segment thresholds and intersection attempts—show that our selected configuration (LSD = 1.8, Intersections = 9) provides a good trade-off between coverage and geometric regularity, adapting well to occlusions and blurred edges.

In contrast to the noisy conditions of the *Cities* dataset, the *RoofVec* dataset features cleaner imagery and more consistent annotations. As a result, the baseline instance segmentation model already achieves strong performance, with F1 scores around 0.94. Nevertheless, our post-processing pipeline continues to provide measurable improvements—particularly in refining object boundaries, removing small redundant fragments, and enhancing the visual clarity of rooftop shapes. These results highlight an important characteristic of the proposed method: its benefits are not limited to low-quality or error-prone predictions. Even when the initial segmentation is already accurate, the integration of geometric reasoning and topological constraints can correct subtle inconsistencies and improve structural coherence. This

suggests that the pipeline generalizes well across data conditions, making it a practical and robust refinement tool for real-world urban modeling tasks.

A major strength of the proposed pipeline lies in its structured representation of rooftop regions using polygons. Unlike pixel-wise masks, which often contain noise or irregular shapes, polygons provide clear, closed boundaries that are well-suited for spatial analysis and vector-based modeling. This intermediate representation enables precise region adjacency modeling, supports 3D extrusion, and facilitates integration with GIS or urban planning systems. As demonstrated in Chapter 4.2.5, the ability to produce topologically coherent and interpretable outputs enables a direct transition from 2D segmentation to plausible 3D rooftop reconstruction.

In addition, the pipeline is fully modular and model-agnostic. It operates independently of the segmentation backbone and does not require retraining or task-specific tuning. This flexibility allows it to serve as a post-processing module on top of any instance segmentation network, such as ResNet, Swin Transformer, or Mask2Former. Compared to prior methods, the proposed framework also offers advantages in terms of automation and supervision requirements. For example, Xu et al. [2023] use multi-level supervision to directly predict polygon structures, but require extensive annotations and dedicated training. Ren et al. [2021] incorporate manual correction stages, limiting scalability. Zhang et al. [2021] apply geometric optimization but incur high computational cost. In contrast, our method operates without additional training, relies on lightweight post-processing, and delivers structured outputs efficiently.

Finally, the system is computationally efficient. The average runtime per image is approximately 3 seconds, with most of the time spent on polygon scoring and graph construction. Deep model inference accounts for only a small portion of the total processing time. This runtime profile supports the method's use as a lightweight refinement stage in large-scale urban analysis pipelines, where both throughput and geometric fidelity are important. Table 4.7 summarizes the average time spent in each step across 10 test images.

Table 4.7.: Average runtime per step (10 images, Cities dataset)

Step	**Time (s)**
Load image	0.1169
Model inference	0.2164
Generate pixel probability map	0.0005
Compute polygon probabilities	1.9052
Prepare unary	0.0000
Build graph	0.6838
GCO labeling	0.1035
Rasterize prediction	0.1056
Evaluate metrics	0.0027
Smoothness stats	0.0034

4.4. Limitation

Despite the overall effectiveness of our method, several limitations remain.

First, the performance of the system is still influenced by the quality of the original dataset annotations. As observed in our experiments, the *Cities* dataset contains inconsistent or coarse annotations, including under-segmentation and missing rooftop labels. These inconsistencies negatively impact evaluation metrics such as false positives and false negatives, and make the benefits of post-processing harder to assess reliably. In contrast, the *RoofVec* dataset offers more consistent and complete labels, which partially explains the higher precision and IoU observed in that setting.

Second, although our method is effective at correcting redundant or overlapping instance predictions from Mask R-CNN, it cannot fully recover from severely incorrect initial segmentations. If the base model fails to detect a rooftop entirely or misclassifies a large structure, no amount of geometric post-processing can compensate for the absence of semantic evidence. This underscores the importance of having a reasonably accurate base model in order for downstream refinement to be effective.

Third, the generation of polygon candidates is still susceptible to challenging visual conditions. Even after carefully tuning parameters such as the LSD threshold and the number of intersection attempts, issues such as tree shadows, strong occlusion, or low roof-to-background contrast can hinder reliable line detection. As discussed in Section 3.3, these conditions often lead to incomplete or distorted polygon proposals, particularly along visually ambiguous boundaries.

Moreover, tuning for dense polygon coverage introduces a trade-off of its own trade-offs: aggressive configurations may generate an excessive number of small, fragmented polygons, resulting in jagged boundaries and irregular shapes that complicate label assignment. On the other hand, conservative settings risk under-segmenting the scene, causing valid rooftop areas to be poorly captured or entirely missed. Striking a balance between proposal completeness and geometric regularity remains a difficult challenge, and significantly impacts the effectiveness of the downstream MRF labeling process.

Fourth, the MRF-based label refinement depends heavily on parameter tuning—especially the smoothness cost λ. As shown in our ablation experiments, different λ values can lead to trade-offs between semantic accuracy (true/false positive counts) and geometric consistency (mean IoU). Tuning this parameter requires careful balancing and extensive experimentation. A setting that improves label continuity may, in some cases, suppress fine-grained distinctions or introduce smoothing artifacts that reduce overall IoU.

These limitations suggest that while our framework is robust and generalizable, its performance is bounded by the quality of the underlying inputs—both in terms of data annotation and model prediction—as well as by the sensitivity of its geometric components to scene complexity and parameter settings.

5. Conclusion and Future Work

5.1. Conclusion

This thesis presents a three-stage framework for structure-aware rooftop segmentation, integrating deep instance prediction, polygon-based geometric abstraction, and MRF-based label regularization. Starting from Mask R-CNN outputs, we introduce a polygon proposal mechanism based on line detection and spatial partitioning, followed by graph-based optimization to refine labels across adjacent regions. The goal is to improve both semantic accuracy and geometric consistency in complex urban settings.

Experiments on two distinct datasets—*Cities* and *RoofVec*—demonstrate the effectiveness and generality of the approach. On the challenging *Cities* dataset, the method significantly reduces false positives and improves interpretability over baseline predictions. On the higher-quality *RoofVec*, it further enhances boundary precision and spatial coherence. Ablation studies validate each component's contribution, and comparisons with prior work highlight the benefits of structured, polygon-based outputs over conventional pixel-level masks.

Beyond performance, the framework is fully automated, modular, and model-agnostic. It requires no extra supervision or retraining, and can be applied to any instance segmentation output. This makes it especially suitable for practical use in remote sensing workflows, where label quality varies or geometric precision is essential.

In summary, the proposed pipeline offers a robust, scalable, and interpretable solution for rooftop segmentation. By bridging deep learning with geometric modeling, it produces GIS-ready results well-suited for urban analysis, 3D reconstruction, and large-scale deployment.

5.2. Future Work

Future extensions to this framework can be naturally aligned with its three-stage structure: instance segmentation, polygon proposal, and MRF-based optimization.

First, improvements to the instance segmentation stage can directly enhance the semantic foundation of the entire pipeline. While our approach is model-agnostic, future work could explore alternative backbones such as ResNeXt, Swin Transformer, or segmenter-based architectures to investigate how richer feature representations impact rooftop mask quality. Stronger initial predictions may reduce false positives and improve the coverage of small or ambiguous rooftops, thereby supporting more accurate downstream geometric refinement.

Second, the polygon proposal stage presents opportunities to further refine the balance between structural fidelity and robustness. In this work, we have already performed a systematic analysis of key parameters—such as `lsd_scale` and `num_intersections`—and evaluated their effects on polygon count, area, and complexity. However, we did not propagate each configuration through the full MRF pipeline to assess downstream segmentation performance. Future work could extend this analysis by connecting geometric proposal quality with end-task outcomes, helping to identify parameter regimes that optimize both structure and accuracy. Additionally, polygon filtering or merging operations may help suppress redundant or spurious candidates without sacrificing completeness.

Finally, the MRF optimization stage could be further explored by refining the definitions of unary and pairwise terms. Currently, unary scores are computed by averaging pixel-level probabilities within each polygon, which may underrepresent confidence in large or boundary-touching regions. Future work could investigate weighted integration schemes that place more emphasis on high-confidence or centrally located pixels to better capture semantic certainty. In terms of spatial regularization, alternative graph constructions—such as vertex-based adjacency or affinity-based edges—might offer improved structural awareness. Similarly, pairwise energy terms could be reformulated using logarithmic scaling or edge-aware weighting functions to enhance label propagation, especially in cluttered or occluded areas.

These directions aim to improve the precision, robustness, and geometric awareness of the pipeline, thereby enhancing its applicability to large-scale and structurally diverse urban environments.

A. Appendix

A.1. Distribution of Polygon Proposals

Figures A.1–A.3 illustrate the distribution of polygon proposal characteristics under the selected configuration (`lsd_scale` = 1.8, `num_intersections` = 9). The majority of polygons are compact and regular, though area and perimeter distributions exhibit long-tailed behavior, indicating the presence of a few large or elongated structures.

A.2. Extended Polygon Proposal Statistics

Table A.1 reports detailed statistics for each combination of LSD scale and number of intersection attempt used during polygon proposal generation. Each value is computed over 10 randomly sampled tiles from the Cities dataset.

Table A.1.: Proposal statistics under different LSD/intersection settings (Cities dataset). Each cell shows mean ± standard deviation across 10 tiles.

LSD / Intersections	Polygon Count	Vertex Count	Area (pixels^2)	Perimeter (pixels)
0.8 / 1	64.14 ± 36.77	6.81 ± 2.26	1021.76 ± 2833.45	137.42 ± 114.47
0.8 / 2	171.90 ± 97.62	5.59 ± 1.31	381.24 ± 1110.24	74.29 ± 69.75
1.0 / 1	82.43 ± 47.18	6.83 ± 2.34	795.05 ± 2518.28	120.38 ± 106.11
1.2 / 1	115.71 ± 71.09	6.85 ± 2.42	566.38 ± 2126.02	97.16 ± 92.49
1.2 / 2	200.30 ± 115.57	5.81 ± 1.40	323.95 ± 1029.65	68.88 ± 66.24
1.2 / 5	251.54 ± 148.84	5.79 ± 1.49	260.54 ± 912.67	60.71 ± 60.11
1.5 / 2	193.72 ± 108.39	5.57 ± 1.45	226.84 ± 826.44	56.90 ± 57.40
1.8 / 1	231.92 ± 195.17	6.08 ± 1.67	197.45 ± 821.54	51.19 ± 53.69
1.8 / 5	281.83 ± 155.77	5.38 ± 1.18	76.94 ± 387.47	29.66 ± 36.78
1.8 / 7	415.81 ± 159.75	5.18 ± 1.13	72.01 ± 330.93	28.61 ± 34.71
1.8 / 9	620.38 ± 182.61	5.39 ± 1.18	102.58 ± 467.53	35.23 ± 41.51
1.8 / 2	835.64 ± 518.27	5.29 ± 1.08	78.43 ± 375.87	30.33 ± 37.20
1.5 / 9	1010.38 ± 636.53	5.23 ± 1.02	64.86 ± 319.09	27.39 ± 34.44

A.3. Unary Cost Example

Table A.2 shows a sample of polygon-wise semantic probabilities and their corresponding unary costs, computed according to the transformation $U_{k,j} = (1 - p_{k,j}) \cdot \alpha$ with $\alpha = 10$.

Each row represents one polygon, and columns correspond to class labels (e.g., background, rooftop planar).

Table A.2.: Sample polygon-level class probabilities and transformed unary costs.

Polygon	**Class 0**	**Class 1**	**Class 2**
Probabilities	0.9313	0.0000	0.0687
Unary Costs	0.6867	10.0000	9.3133
Probabilities	0.6558	0.0000	0.3441
Unary Costs	3.4415	10.0000	6.5585
Probabilities	1.0000	0.0000	0.0000
Unary Costs	0.0000	10.0000	10.0000

This table corresponds to the intermediate output passed into the MRF optimization framework after soft mask aggregation and cost transformation.

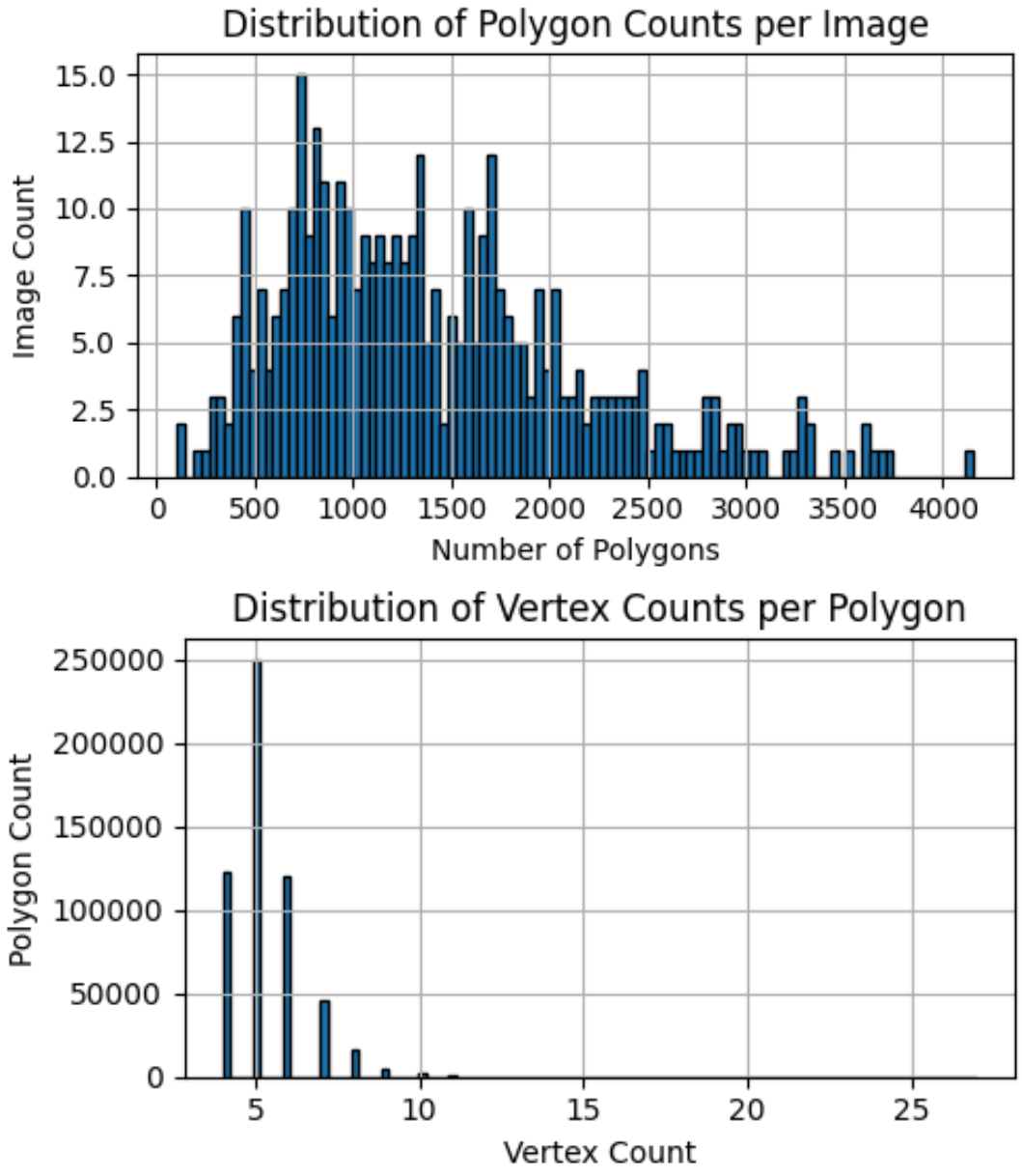


Figure A.1.: Top: Number of polygon proposals per image. Bottom: Vertex count distribution.

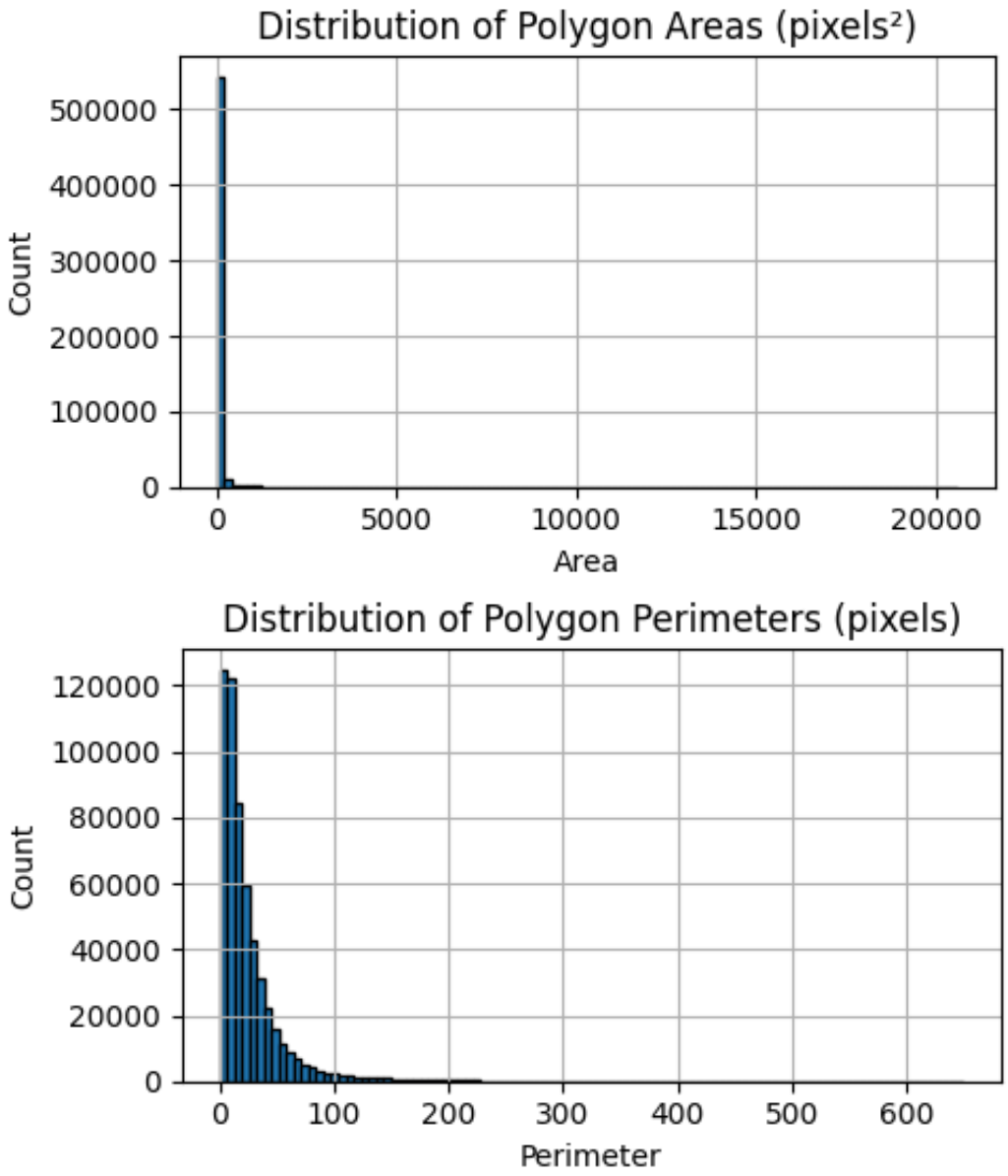


Figure A.2.: Distribution of polygon area (top) and perimeter (bottom). Most proposals are compact.

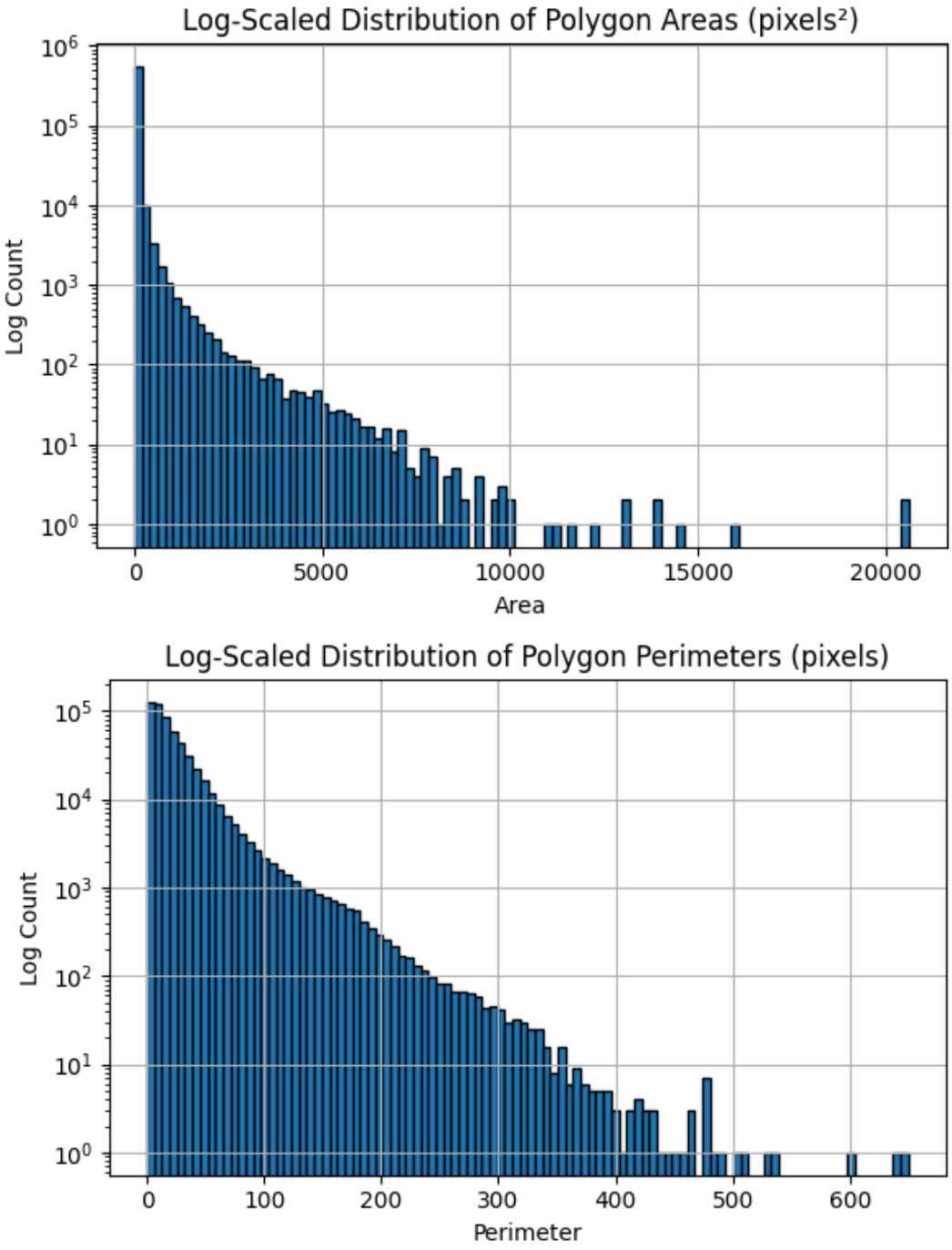


Figure A.3.: Log-scaled plots highlighting long-tailed behavior in area and perimeter distributions.

Bibliography

Awrangjeb, M., Lu, G., and Fraser, C. S. (2014). Automatic building extraction from lidar data covering complex urban scenes. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XL-3, pages 25–31. ISPRS.

Bauchet, J.-P. and Lafarge, F. (2018). Kippi: Kinetic polygonal partitioning of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bauchet, J.-P., Sulzer, R., Lafarge, F., and Tarabalka, Y. (2024). Simplicity: Reconstructing buildings with simple regularized 3d models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. USM3D Workshop.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., and Çöltekin, A. (2015). Applications of 3d city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Calantropio, A., Chiabrando, F., Codastefano, M., and Bourke, E. (2021). Deep learning for automatic building damage assessment: Application in post-disaster scenarios using uav data. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-1-2021, pages 113–120. Copernicus Publications.

Chen, Y., Chen, Y., Lin, W., Fang, Y., and Qin, J. (2021). Heat: Holistic edge attention transformer for structured building segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15945–15954.

Chen, Y. and Hong, T. (2018). Impacts of building geometry modeling methods on the simulation results of urban building energy models. *Applied Energy*, 215:478–488.

Chen, Z., Ledoux, H., Khademi, S., and Nan, L. (2022). Reconstructing compact building models from point clouds using deep implicit fields. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:58–73.

Chen, Z., Shi, Y., Nan, L., Xiong, Z., and Zhu, X. X. (2024). Polygnn: Polyhedron-based graph neural network for 3d building reconstruction from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:693–706.

Chicchon, M., Malinverni, E. S., Sanità, M., Pierdicca, R., Colosi, F., and León Trujillo, F. J. (2024). Building semantic segmentation using unet convolutional network on spacenet public data sets for monitoring surrounding area of chan chan (peru). *Geomatics and Environmental Engineering*, 18(3):25–44.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., and Raskar, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–181.

Gao, W., Peters, R., and Stoter, J. (2024). Unsupervised roofline extraction from true orthophotos for lod2 building model reconstruction. In Kolbe, T. H. et al., editors, *Recent Advances in 3D Geoinformation Science*, Lecture Notes in Geoinformation and Cartography, pages 425–436. Springer.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969.

Hensel, S., Goebbels, S., and Kada, M. (2021). Building roof vectorization with ppgnet. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLVI-4/W4-2021, pages 85–90.

Jochem, A., Höfle, B., Rutzinger, M., and Pfeifer, N. (2009). Automatic roof plane detection and analysis in airborne lidar point clouds for solar potential assessment. *Sensors*, 9(7):5241–5262.

Lafarge, F. and Mallet, C. (2012). Creating large-scale city models from 3d-point clouds: A robust approach with hybrid representation. *International Journal of Computer Vision*, 99(1):69–85.

Ledoux, H. and Meijers, M. (2011). Topologically consistent 3d city models obtained by extrusion. *International Journal of Geographical Information Science*, 25(4):557–574.

Lussange, J., Debeir, O., Burie, J.-C., and Ogier, J.-M. (2023). 3d detection of roof sections from a single satellite image and application to lod2-building reconstruction. *Remote Sensing*, 15(4):1152.

Nauata, N. and Furukawa, Y. (2020). Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. *arXiv preprint arXiv:1912.05135*.

Nelson, J. R. and Grubesic, T. H. (2020). The use of lidar versus unmanned aerial systems (uas) to assess rooftop solar energy potential. *Sustainable Cities and Society*, 61:102353.

Peronato, G., Kämpf, J. H., Rey, E., and Andersen, M. (2017). Integrating urban energy simulation in a parametric environment: a grasshopper interface for citysim. In *PLEA Conference Proceedings*.

Qian, Y., Zhang, H., and Furukawa, Y. (2021). Roof-gan: Learning to generate roof geometry and relations for residential houses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2857–2866.

Ren, J., Zhang, B., Wu, B., Huang, J., Fan, L., Ovsjanikov, M., and Wonka, P. (2021). Intuitive and efficient roof modeling for reconstruction and synthesis. *ACM Transactions on Graphics (TOG)*, 40(6):249:1–249:24.

Rezaeian, M. and Gruen, A. (2011). Automatic 3d building extraction from aerial and space images for earthquake risk management. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 5(1):77–96.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., and Bittner, K. (2023). Roof3d: A real and synthetic data collection for individual building roof plane and building sections detection. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume X-1/W1-2023, pages 971–978. ISPRS.

Tan, B., Xue, N., Bai, S., Wu, T., and Xia, G.-S. (2022). Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16614–16623.

Verma, V., Kumar, R., and Hsu, S. (2006). 3d building detection and modeling from aerial lidar data. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2213–2220. IEEE.

Weir, N., Lindenbaum, D., Bastidas, A., Van Etten, A., McPherson, S., Shermeyer, J., Kumar, V., and Tang, H. (2019). Spacenet mvoi: A multi-view overhead imagery dataset. *arXiv preprint arXiv:1905.07293*.

Xu, B., Xu, J., Xue, N., and Xia, G.-S. (2023). Hisup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:284–296.

Xu, Y., Jubanski, J., Bittner, K., and Siegert, F. (2024). Roof plane parsing towards lod-2.2 building reconstruction based on joint learning using remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 133:104096.

Yang, G., Xue, F., Zhang, Q., Xie, K., Fu, C.-W., and Huang, H. (2023). Urbanbis: A large-scale benchmark for fine-grained urban building instance segmentation. In *ACM SIGGRAPH Conference Proceedings*, pages 1–11. Association for Computing Machinery.

Yin, X., Wu, Z., Yang, J., Fu, J., Zhang, Z., Liu, Z., and Lu, H. (2023). Planerectr: Weakly-supervised 3d plane reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20483–20492.

Zhang, F., Xu, X., Nauata, N., and Furukawa, Y. (2021). Structured outdoor architecture reconstruction by exploration and classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Zorzi, S., Bazrafkan, S., Habenschuss, S., and Fraundorfer, F. (2023). Polyworld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12923–12932.

Colophon

This document was typeset using LaTeX, using the KOMA-Script class `scrbook`. The main font is Palatino.