ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XII-4/W1-2026
21st International 3D GeoInfo Conference 2026, 28–29 September 2026, Sofia, Bulgaria

Roof Structure Extraction from Remote Sensing Images

Hsin-Yu Cheng[1], Weixiao Gao[1,2], Liangliang Nan[1]

[1] Delft University of Technology, The Netherlands
[2] Wageningen University & Research, The Netherlands
hycheng@u.nus.edu, gaoweixiaocuhk@gmail.com, liangliang.nan@tudelft.nl

Keywords: Roof structure extraction, Instance segmentation, Polygon labeling, Markov Random Field, Remote sensing imagery

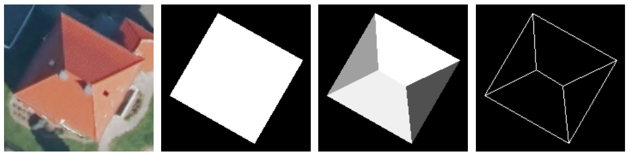
Abstract

Accurate roof structure extraction from aerial imagery is important for applications such as solar energy assessment, urban analysis, and 3D city modeling, where roof surfaces require geometrically reliable polygonal representations. Although instance segmentation methods can detect rooftop regions, their masks often contain redundant or overlapping predictions and irregular boundaries, making them difficult to convert into coherent roof surfaces. To address these limitations, we propose a polygon-level structured inference framework for roof-region extraction. Rather than directly using raster masks as final outputs or reconstructing roof polygons from local vector primitives, the proposed method generates over-segmented polygon candidates from line-based cues and aggregates confidence-weighted probabilities at the polygon level. Roof-region assignment is formulated as a Markov Random Field (MRF) optimization problem, where unary terms encode segmentation evidence and pairwise terms enforce spatial consistency. Experiments on the *Cities* and *RoofVec* datasets show that the proposed framework suppresses redundant and spurious predictions while improving region coherence. On RoofVec, false positives are reduced from 111 to 94 and precision improves from 96.7% to 97.2% compared with unary-only polygon labeling, while maintaining comparable mean IoU. These results show improved geometric consistency in roof extraction from 2D remote sensing imagery. The source code is publicly available at `https://github.com/appleadele/Roof-Structure-Extraction-from-Remote-Sensing-Images`.

1. Introduction

With the growing availability of LiDAR, photogrammetry, and semantic 3D city models such as CityGML (Biljecki et al., 2015), 3D urban data have been increasingly used in a wide range of applications. In particular, detailed building geometry and rooftop information are important for rooftop solar potential assessment (Jochem et al., 2009, Nelson and Grubesic, 2020), urban-scale building energy modeling (Peronato et al., 2017, Chen and Hong, 2018), and disaster-related urban inventory and damage assessment (Rezaeian and Gruen, 2011, Calantropio et al., 2021).

Building segmentation is a key step in urban scene interpretation and 3D modeling, and remote sensing imagery enables automatic detection of buildings and rooftop regions. As illustrated in Fig. 1, rooftop interpretation can be represented at different levels, including building-level segmentation, roof planar structure segmentation, and roof line extraction. However, segmentation outputs are typically represented as raster masks that do not explicitly encode structural information, often resulting in geometrically inconsistent boundaries such as broken or rounded rooflines that are difficult to use for vectorized reconstruction and structured roof modeling (Xu et al., 2024). These challenges become more pronounced in dense urban scenes, where closely spaced buildings, occlusions, and complex roof configurations make extraction more difficult (Awrangjeb et al., 2014, Chicchon et al., 2024). As a result, even accurate pixel-level predictions may fail to produce clean, consistent, and structurally complete roof representations. Beyond raster-only segmentation, recent studies have moved toward vectorized and structured roof representations, including rooflines, corner–edge graphs, planar graphs, and roof-plane polygons (Gao et al., 2024, Nauata and Furukawa, 2020, Chen et al., 2022a, Xu et al., 2024). These methods highlight the importance of explicit geometric primitives and connectivity for coherent roof modeling. At the same time, they also suggest that vectorized reconstruction remains sensitive to the reliability of local geometric evidence: incomplete rooflines, missed corners, fragmented or spurious edges, and weak roof boundaries may alter graph connectivity, prevent the formation of valid closed roof regions, or lead to over-partitioned roof structures. Such topological and geometric errors can propagate to downstream 3D building reconstruction by producing incorrect roof partitions, unstable surface topology, or inconsistent roof geometry.



(a) (b) (c) (d)

Figure 1. Illustration of different levels of rooftop interpretation based on ground-truth annotations. (a) Original image. (b) Building-level semantic segmentation. (c) Roof planar structure segmentation. (d) Roof line extraction result.

These limitations motivate a formulation of roof extraction as structured labeling over polygonal candidates, rather than as either raster-mask prediction or direct vector reconstruction from locally detected corners and edges. We first use line-based geometric cues to generate an over-segmented polygon partition, which provides a closed and non-overlapping hypothesis space for possible roof regions. Instance segmentation outputs are used only as semantic evidence for polygon labels, rather than as final masks. Inspired by MRF-based structured reconstruction that selects consistent surfaces from candidate partitions (Chen et al., 2022b), we optimize polygon labels globally using confidence-weighted semantic scores and polygon adjacency. In this way, roof topology is determined by selecting and merging regions within a shared partition, allowing overlapping masks, fragmented predictions, and local boundary artifacts to be reduced in a single structured inference step, without requiring a separate corner-, edge-, or topology-repair stage.

The main contributions of this work are threefold: (1) roof surface extraction is formulated as a structured labeling problem over an over-segmented polygon partition generated from line-based geometric cues, rather than as direct raster-mask prediction or direct vector reconstruction from local primitives; (2) an MRF energy is designed for polygon-level roof instance assignment, integrating confidence-weighted segmentation evidence with shared-boundary pairwise regularization; and (3) experiments on public roof extraction datasets show reduced redundant and spurious rooftop predictions and improved precision while maintaining competitive IoU performance.

2. Related Work

Related work relevant to rooftop interpretation from aerial or satellite imagery includes semantic and instance segmentation, roof boundary extraction, and planar roof structure reconstruction. In this section, we review segmentation-based and geometry-oriented approaches for roof modeling.

2.1 Semantic and Instance Segmentation

Semantic segmentation of buildings from aerial and satellite imagery has become a widely adopted approach for urban scene interpretation and downstream 3D modeling tasks. Existing approaches can be broadly categorized into semantic segmentation and instance segmentation methods. Semantic segmentation methods generate pixel-wise masks that distinguish building from non-building regions, enabling large-scale extraction of building footprints. Large-scale benchmark datasets have further supported the development and evaluation of building segmentation methods (Demir et al., 2018). However, segmentation outputs remain predominantly raster-based and do not explicitly encode geometric structure. Rasterized roof predictions may suffer from broken or irregular line segments, leading to rounded corners and limiting their direct use for vectorized reconstruction (Xu et al., 2024). In dense urban scenes, segmentation masks may also exhibit imprecise boundaries or merge adjacent buildings, especially under shadows or background clutter (Chicchon et al., 2024). As a result, even accurate pixel-level predictions are not directly suitable for geometry-aware tasks such as vectorized reconstruction or planar modeling.

Instance segmentation extends semantic segmentation by separating adjacent buildings into distinct object instances. General-purpose frameworks such as Mask R-CNN provide a representative formulation for this task by predicting an instance mask for each detected region (He et al., 2017). However, these methods still produce independent raster masks for each instance and do not explicitly model structural relationships among neighboring roof regions. Consequently, irregular or misaligned boundaries, redundant or overlapping predictions, and sensitivity to occlusion further limit the direct usability of segmentation outputs for structured roof reconstruction. These limitations motivate the development of approaches that move beyond pixel-level labeling toward geometry-aware and structurally consistent roof representations, particularly those that can explicitly model relationships between roof regions.

2.2 Roof Boundary and Planar Structure Extraction

Beyond pixel-wise segmentation, prior work has explored geometric rooftop representations, particularly vectorized roof boundaries and structured roof components. These efforts include methods based on geometric priors and line detection, segmentation-based contour refinement, and learning-based polygon or planar-graph prediction. Unsupervised methods rely on geometric priors and line detection to extract rooflines without requiring labeled training data (Gao et al., 2024). While such methods can provide useful geometric cues for roof reconstruction, the extracted lines may still be incomplete or fragmented for complex buildings, making additional structural reasoning necessary for forming coherent roof regions. Segmentation-to-polygon methods attempt to reduce the gap between raster masks and vector representations by incorporating additional geometric supervision, such as vertices, line segments, or boundary cues, into the learning process (Xu et al., 2023). Learning-based vector reconstruction methods aim to infer structured polygonal or planar-graph representations by detecting corners, junctions, edges, lines, regions, or other geometric primitives and reasoning over their connectivity through graph-based models, Transformer-based architectures, optimal-transport matching, or explicit optimization (Nauata and Furukawa, 2020, Chen et al., 2022a, Zorzi et al., 2022, Wang et al., 2025). While these approaches generally improve geometric regularity compared with raster-mask outputs, they may still be sensitive to missed corners or weak image cues, which can lead to incomplete or invalid structures for complex buildings; hybrid optimization pipelines may also introduce additional inference cost (Gao et al., 2024, Nauata and Furukawa, 2020).

Several studies further aim to recover planar roof structures for LoD2-level building reconstruction. Existing approaches include image-only methods that infer roof sections from a single satellite image and image–height fusion methods that combine spectral imagery with elevation data such as DSMs or nDSMs (Lussange et al., 2025, Xu et al., 2024). Image-only methods can recover roof geometry from a single satellite image by learning indirect 3D cues such as perspective distortion, facade deformation, shadows, and roof ridge appearance; however, inferring reliable height and planar structure from image evidence alone remains challenging, especially when such cues are weak, ambiguous, or affected by imaging conditions (Lussange et al., 2025). Fusion-based methods improve reconstruction accuracy by incorporating height information, yet raster-based predictions may still produce broken or irregular roof boundaries, making the transition to clean vectorized models challenging and often requiring additional connectivity reasoning or post-processing to obtain valid polygons (Xu et al., 2024). Related RGB–DSM studies on roof-plane and building-section segmentation further show that appearance ambiguity remains challenging, since RGB inputs may confuse buildings with roads or impervious surfaces, whereas DSM inputs may confuse rooftops with high vegetation (Schuegraf et al., 2023).

Despite these advances, many existing approaches still rely on local predictions or independently generated components and do not explicitly enforce global consistency across roof regions. As a result, integrating these predictions into coherent and structurally consistent roof models remains challenging. This motivates a region-level formulation in which semantic evidence and spatial adjacency are considered jointly over a shared polygonal partition.

https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-81-2026

3. Method

3.1 Overview

Rooftop extraction from aerial imagery often relies on raster-based segmentation outputs, which lack explicit geometric structure and are not well suited for region-level reasoning. To obtain geometrically consistent roof representations, we aim to perform rooftop labeling on structured regions rather than pixel grids. To this end, we propose a polygon-based framework for rooftop instance extraction, where labeling is performed on a set of polygonal regions generated from geometric cues instead of directly using raster segmentation masks as the final representation. Our formulation follows a candidate-generation-and-selection principle related to optimization-based 3D structured reconstruction, in which planar primitives are used to generate spatial cells and coherent building surfaces are extracted through global optimization (Chen et al., 2022b). Rather than reconstructing 3D surfaces from point-cloud-derived planar primitives and cell complexes, our method applies this principle to a 2D image-derived polygon partition, using MRF optimization to assign spatially consistent roof-region labels from instance-segmentation evidence.

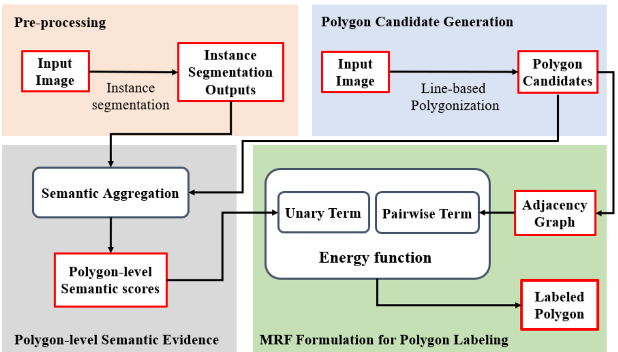


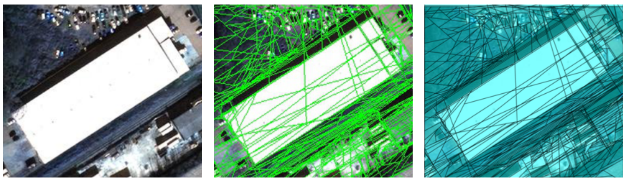
Figure 2. Overview of the proposed framework. Red boxes denote data representations, and blue boxes denote processing modules. Instance segmentation provides pixel-level semantic predictions, while line-based polygonization generates polygon candidates. Polygon-level semantic scores and geometric adjacency are integrated into an MRF for final labeling.

As illustrated in Figure 2, the framework consists of three main stages. First, an instance segmentation model is applied to the input image to produce rooftop predictions, including instance confidence scores and soft probability masks. These outputs provide pixel-level semantic evidence. Second, structural line cues are extracted from the image and polygonized to generate candidate regions. The resulting partition is intentionally over-segmented to preserve fine structural boundaries and provide flexible geometric units for subsequent inference. Third, the semantic predictions are aggregated within each polygon to obtain polygon-level semantic scores. These semantic scores define the unary terms of a Markov Random Field (MRF), while adjacency relationships between polygons define pairwise interactions. The final labeling is obtained by minimizing the MRF energy over the polygon graph. This formulation integrates semantic predictions with geometric structure, producing roof surface delineations that are more spatially consistent while preserving fine-scale segmentation cues.

3.2 Polygon Candidate Generation

Rooftop structures in aerial imagery often exhibit strong geometric regularities, such as straight edges and orthogonal boundaries, which are not explicitly captured by pixel-based representations. To enable geometry-aware and region-level inference beyond pixel-based representations, we partition the input image into polygonal regions that serve as candidate rooftop surfaces. The partition is generated from structural line cues extracted from the RGB image.

We adopt the KIPPI algorithm (Bauchet and Lafarge, 2018), which detects line segments and organizes them into a geometrically regularized line network. The method incorporates geometric constraints such as parallelism and orthogonality to produce boundary-aligned line structures under image noise, illumination variation, and partial occlusions. Closed polygonal regions are then obtained by polygonizing the resulting line network. To ensure that potential rooftop structures are not missed, the partition is intentionally oversegmented. This strategy favors high recall by generating a dense set of candidate polygons, allowing complex roof surfaces to be represented as combinations of multiple regions. Although this introduces redundancy, it provides flexible geometric units for the subsequent labeling stage. The resulting polygon set defines the structural representation used for subsequent inference. Each polygon corresponds to a node in the MRF graph, where semantic evidence and geometric adjacency are integrated to infer the final rooftop labels. The polygon candidate generation process is illustrated in Fig. 3.



(a) Input image (b) Detected rooflines (c) Polygon candidates

Figure 3. Illustration of polygon candidate generation. (a) Input image. (b) Detected rooflines, where green lines represent extracted line segments. (c) Polygon candidates generated from the line network, where blue regions indicate polygonal partitions and black lines denote polygon boundaries.

3.3 Polygon-level Semantic Evidence

To associate semantic information with the geometric representation, predictions from instance segmentation are aggregated over the polygon candidates obtained in the previous stage. We perform semantic aggregation to convert pixel-level predictions into region-level semantic evidence, which can be integrated into the subsequent MRF inference.

To integrate the detection confidence with the mask prediction, we define a confidence-weighted pixel probability map as

$$P_i(x, y) = s_i M_i(x, y) \tag{1}$$

where s_i denotes the confidence score of instance i, $M_i(x, y)$ represents the mask probability at pixel (x, y), and $P_i(x, y)$ is the resulting confidence-weighted pixel probability. The instance segmentation model provides these quantities, where each prediction consists of a confidence score s_i and a soft mask

$M_i(x, y)$ that indicates the likelihood of each pixel belonging to the predicted instance.

Given the polygon partition described in Section 3.2, the semantic score for polygon k with respect to instance i is obtained by averaging the pixel-wise probabilities within the polygon:

$$p_{k,i} = \frac{1}{|\Omega_k|} \sum_{(x,y)\in\Omega_k} P_i(x, y) \tag{2}$$

where Ω_k denotes the set of pixels contained in polygon k. For the background label, the pixel-level probability is computed as one minus the maximum probability among all detected rooftop instances, and is averaged within polygon k to obtain $p_{k,0}$. This aggregation step establishes the connection between raster-based semantic predictions and the polygonal representation used in the proposed framework. The resulting polygon-level semantic scores serve as unary evidence in the subsequent MRF formulation, while spatial relationships between polygons define the pairwise interactions.

3.4 MRF Formulation for Polygon Labeling

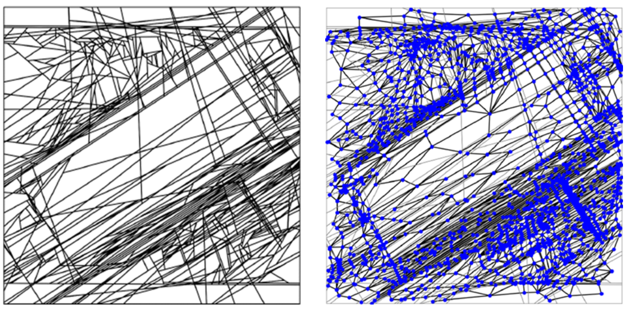
Given the polygon-level semantic scores obtained in Section 3.3, the final rooftop labeling is formulated as an energy minimization problem using a Markov Random Field (MRF). In this formulation, each polygon corresponds to a node in a graph, and the goal is to assign either the background label or a roof-instance label to each polygon while maintaining spatial consistency between neighboring regions.

Let $G = (V, E)$ denote an undirected graph constructed from the polygon partition, where each node $k \in V$ represents a polygonal region and an edge $(k, l) \in E$ connects two polygons that share a common boundary. The resulting adjacency graph is illustrated in Fig. 4.

Each polygon k is assigned a label L_k from the label set

$$L_k \in \{0, 1, \ldots, N\} \tag{3}$$

where 0 denotes the background class and $1, \ldots, N$ correspond to the detected rooftop instances obtained from the instance segmentation model.



(a) Polygon candidates (b) Adjacency graph

Figure 4. Illustration of adjacency graph construction. (a) Polygon candidates with black polygon boundaries. (b) Adjacency graph, where blue dots denote graph nodes, black lines denote adjacency connections, and gray lines indicate the underlying polygon boundaries.

The unary term measures the cost of assigning label i to polygon k, where $i = 0$ denotes the background label and $i = 1, \ldots, N$ denote the detected rooftop instances. The cost is based on the polygon-level semantic score $p_{k,i}$ defined in Section 3.3. A higher semantic score indicates stronger evidence that the polygon belongs to the corresponding label. The unary cost for all labels is defined as

$$U_{k,i} = \alpha(1 - p_{k,i}), \tag{4}$$

where α is a scaling factor controlling the influence of the unary term. In particular, the background cost is computed as $U_{k,0} = \alpha(1 - p_{k,0})$. This formulation assigns lower costs to labels with stronger semantic support, thereby encouraging each polygon to receive a well-supported label during optimization.

The pairwise term enforces spatial consistency between neighboring polygons by penalizing label discontinuities along shared boundaries. For two adjacent polygons k and l, the pairwise penalty is defined as

$$\psi(L_k, L_l) = \begin{cases} 0 & \text{if } L_k = L_l, \\ 1 & \text{if } L_k \neq L_l. \end{cases} \tag{5}$$

where L_k and L_l denote the labels assigned to polygons k and l, and $\psi(\cdot, \cdot)$ defines the pairwise penalty between neighboring labels. To account for geometric connectivity, the penalty is weighted by the length of the shared boundary between the polygons. Let w_{kl} denote a boundary-length-dependent weight for adjacent polygons k and l.

Combining the unary and pairwise terms, the total MRF energy is defined as

$$E(L) = \sum_k U_{k,L_k} + \lambda \sum_{(k,l)\in E} w_{kl}\psi(L_k, L_l), \tag{6}$$

where λ is a regularization parameter controlling the balance between semantic evidence and spatial smoothness.

The optimal labeling L^* is obtained by minimizing the total energy:

$$L^* = \arg\min_L E(L). \tag{7}$$

This formulation integrates semantic evidence derived from instance segmentation with geometric adjacency relationships between polygons, producing spatially consistent rooftop instance labels.

To solve the resulting multi-label optimization problem, we employ the α-expansion algorithm (Boykov et al., 2001), which efficiently approximates the global optimum of the MRF energy. The optimization iteratively performs expansion moves solved via graph-cut optimization until no further energy reduction can be achieved, resulting in a spatially consistent polygon labeling.

4. Experiments and Results

4.1 Experimental Setup

Experiments are conducted on two publicly available datasets: the *Vectorizing World Buildings* dataset (hereafter referred to as *Cities*) (Nauata and Furukawa, 2020), based on SpaceNet imagery (Van Etten et al., 2018), and the *RoofVec* dataset from the Vectorization-Roof-Data-Set project (Hensel et al., 2021). Cities contains 1,985 aerial images of size 256×256, mainly depicting industrial and commercial buildings with complex flat rooftops. RoofVec contains 7,640 images with varying resolutions and primarily features residential buildings with simpler sloped roofs. All rooftop annotations are converted into polygon-based semantic masks for training. Each dataset is randomly split into 60% training, 20% validation, and 20% testing sets.

We use Mask R-CNN (He et al., 2017) with a ResNet-50-FPN backbone as a standard and widely used baseline. Since the focus of this work is the proposed polygon-level structured inference framework rather than the design of a new segmentation network, we adopt a stable baseline to evaluate the contribution of the polygon-based labeling stage more clearly. The model is trained using stochastic gradient descent with a batch size of 4, momentum 0.9, weight decay 5×10^{-4}, and an initial learning rate of 0.005, which is reduced by a factor of 10 every 20 epochs.

Polygon candidates are generated using the line detection module from the KIPPI pipeline. After parameter exploration, we use `lsd_scale` $= 1.5$ and `num_intersections` $= 9$ to obtain dense polygon partitions. Unary costs are derived from aggregated soft mask probabilities with scaling factor $\alpha = 10$. Pairwise terms are weighted by a boundary-length-dependent weight:

$$w_{kl} = \sqrt{l_{kl}/80} \cdot 5 + 1. \tag{8}$$

The smoothness parameter λ is evaluated over $\{0.001, 0.05, 0.1, 1.0, 10.0\}$.

Predictions are matched to ground-truth instances using an IoU threshold of 0.5. For instance-level evaluation, we report True Positives (TP), False Positives (FP), False Negatives (FN), and Precision. For polygon-based evaluation, we additionally report mean IoU over matched instances. Recall is also reported in the smoothness-parameter ablation.

4.2 Results and Analysis

We first evaluate the raw Mask R-CNN predictions. As shown in Table 1, the baseline performs substantially better on RoofVec than on Cities. In particular, the precision on Cities is only 59.2%, indicating a large number of redundant or overlapping predictions in complex urban scenes. By contrast, the baseline reaches 90.4% precision on RoofVec, reflecting the cleaner imagery and simpler roof structures in that dataset. As also observed in Fig. 5(b), the baseline predictions exhibit multiple overlapping detections for the same rooftop structures, leading to redundant predictions.

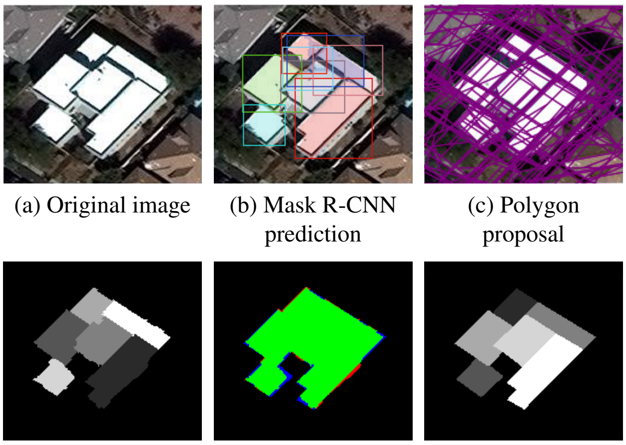
We then evaluate the polygon-based stages of the pipeline, including unary-only labeling and the full MRF model. Table 2 compares polygon labeling with and without the MRF pairwise term. The main effect of the MRF term is to reduce redundant and spurious polygon labels rather than to substantially increase mean IoU. Compared with unary-only labeling, the full MRF reduces false positives from 319 to 306 on Cities and from 111 to 94 on RoofVec, while improving precision from 70.1% to 71.0% and from 96.7% to 97.2%, respectively. The mean IoU remains similar because many correctly matched roof regions already overlap well with the ground truth after unary labeling; the pairwise term mainly improves spatial coherence by suppressing isolated or inconsistent polygon labels.

Data	TP	FP	FN	Precision
Cities	870	600	197	59.2%
RoofVec	3338	356	48	90.4%

Table 1. Instance-level performance of the Mask R-CNN baseline.

Data	Method	TP	FP	FN	Prec.	mIoU
Cities	Unary	749	319	200	70.1%	83.3%
Cities	MRF	748	306	201	71.0%	83.4%
RoofVec	Unary	3265	111	90	96.7%	91.3%
RoofVec	MRF	3262	94	93	97.2%	91.3%

Table 2. Comparison of polygon labeling with and without the MRF pairwise term ($\lambda = 0.1$).



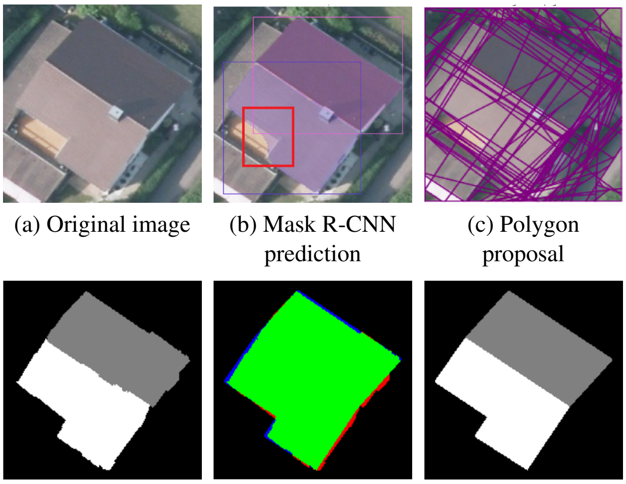
(a) Original image (b) Mask R-CNN prediction (c) Polygon proposal

(d) MRF result (e) TP (green), FP (red), FN (blue) (f) Ground truth

Figure 5. Example of redundant instance predictions in a multi-region roof. The proposed polygon-level framework suppresses inconsistent assignments and produces a more coherent roof-region labeling.

To further illustrate the type of errors reduced by the proposed polygon-level framework, Fig. 5 shows a representative example with multiple roof regions, some of which are affected by shadows. As shown in Fig. 5(b), the Mask R-CNN baseline produces eight instance predictions for six annotated roof regions, including overlapping and partially redundant predictions for the roof structure. The line-based polygon partition in Fig. 5(c) provides candidate regions for subsequent structured labeling. After MRF optimization, Fig. 5(d) shows a more coherent roof-region labeling, where inconsistent assignments are suppressed while the main roof components are preserved. The TP/FP/FN visualization in Fig. 5(e) is consistent with the reduction in false positives and the improvement in precision reported in Table 2.

In addition to suppressing redundant rooftop predictions,

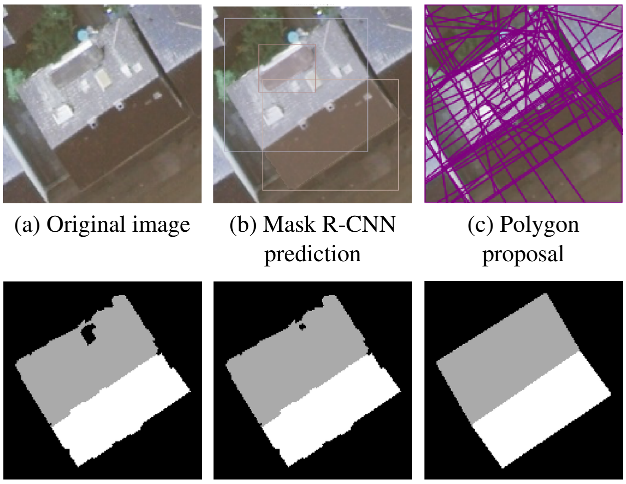


(a) Original image (b) Mask R-CNN prediction (c) Polygon proposal

(d) MRF result (e) TP (green), FP (red), FN (blue) (f) Ground truth

Figure 6. A representative example of boundary artifact removal. The Mask R-CNN baseline produces a small spurious protrusion near the roof boundary, highlighted by the red box, while polygon-level MRF optimization removes it and yields a cleaner, more regular boundary.

the proposed polygon-level framework also regularizes local boundary artifacts and irregular roof outlines. As shown in Fig. 6, the Mask R-CNN baseline produces a small spurious protrusion near the roof boundary, resulting in an irregular and curved outline. After MRF optimization, this boundary artifact is removed, yielding a cleaner segmentation result and a more regular boundary.



(a) Original image (b) Mask R-CNN prediction (c) Polygon proposal

(d) Unary only (e) Full MRF (f) Ground truth

Figure 7. Effect of the MRF pairwise term. Compared with unary-only labeling, the full MRF model suppresses fragments, fills small holes, and improves boundary consistency by enforcing spatial consistency between adjacent polygons.

To isolate the effect of the MRF pairwise term, Fig. 7 compares unary-only polygon labeling with the full MRF result. The Mask R-CNN baseline in Fig. 7(b) captures the overall rooftop structure, but the unary-only result in Fig. 7(d) exhibits irregular boundaries and small holes within the predicted region. After full MRF optimization in Fig. 7(e), these artifacts are reduced, resulting in smoother boundaries and more coherent rooftop regions. This example illustrates the role of the pairwise term: by enforcing spatial consistency across adjacent polygons, the MRF model reduces local boundary artifacts and fills small holes. Such region-level consistency is difficult to obtain when raster masks are used directly as independent final outputs, because adjacency relationships between neighboring regions are not explicitly optimized. Although the overall mIoU changes only slightly in Table 2, the qualitative comparison shows that the MRF term improves local spatial coherence.

For contextual comparison with prior roof structure extraction methods, we report region-level F_1 scores following the evaluation protocol used in previous polygon-based roof-structure extraction work (Nauata and Furukawa, 2020), where predicted and ground-truth polygons are matched at an IoU threshold of 0.7. Because many recent roof-structure extraction methods are evaluated on different benchmark datasets and also differ in output representations and evaluation protocols, we focus on reference results that are based on the same or closely related benchmark and are closest to the region-level polygon matching protocol used here. The reference results are taken from the original papers of the listed methods (Nauata and Furukawa, 2020, Zhang et al., 2021, Chen et al., 2022a), and should therefore be regarded as reference results rather than strictly comparable benchmarks.

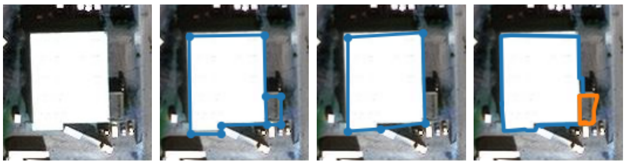
As shown in Table 3, the proposed method achieves a region-level F_1 score of 66.1%. Relative to the reported numbers in prior work, this value is above those of Nauata and Furukawa (2020) and Zhang et al. (2021), while remaining below the 70.6% reported by Chen et al. (2022a). Although these comparisons are not strictly controlled, they suggest that the proposed method is competitive while relying on a relatively simple modular pipeline without specialized polygon decoders or edge supervision. Under a relaxed IoU threshold of 0.5, our method achieves 74.6%, indicating robustness to small geometric discrepancies.

Method	IoU	F_1
Nauata and Furukawa (2020)	0.7	60.8%
Zhang et al. (2021)	0.7	63.5%
Chen et al. (2022a)	0.7	70.6%
Proposed method	0.7	**66.1%**
Proposed method (IoU = 0.5)	0.5	74.6%

Table 3. Reported region-level F_1 scores of previous roof structure extraction methods and the result of the proposed method. Since the experimental settings are not identical, these numbers are provided for reference only.

To provide a qualitative reference, we applied the publicly available pretrained model from Zhang et al. (2021) directly to our test images without retraining under our data split. Since the original method was developed and trained under a different experimental setup, the examples in Figures 8 and 9 are intended only as qualitative illustrations rather than a controlled comparison. In these examples, our method appears visually more consistent with the annotated roof structures.

In terms of computational efficiency, each image on the Cities dataset generates about 1400 polygon candidates on average, and the full pipeline takes approximately 3 seconds per image. Most of the time is spent on computing polygon probabilities and constructing the adjacency graph, while MRF optimization accounts for only a small fraction of the runtime. This suggests that the proposed framework can be used as a lightweight structured inference stage.



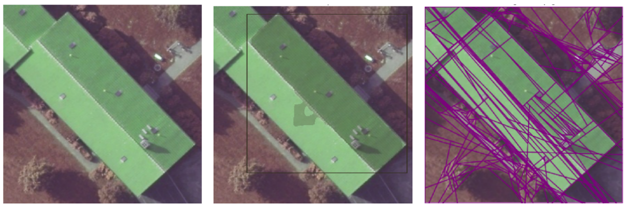
(a) Original image (b) Ground truth (c) Zhang et al. (d) Ours

Figure 8. Qualitative comparison with Zhang et al. (2021). Our method better preserves the small attached roof structure.

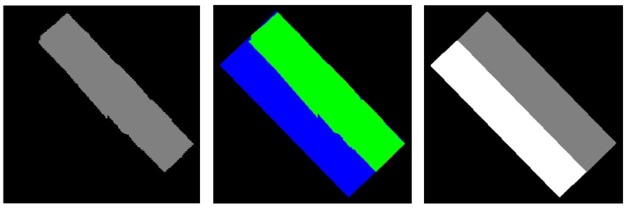
(a) Original image (b) Ground truth (c) Zhang et al. (d) Ours

Figure 9. Qualitative comparison with Zhang et al. (2021). Our method avoids implausible polygon connections and better matches the ground truth.

A representative failure case is shown in Fig. 10. When the base instance segmentation model completely misses a rooftop region, polygon labeling cannot recover it because the unary evidence is absent. Weak or incomplete line structures caused by shadows, occlusions, or low contrast may also lead to insufficient polygon proposals. Thus, the final performance still depends on the quality of both the initial segmentation and the polygon partition. The evaluation may also be affected by annotation quality, since incomplete or coarse rooftop labels can cause visually reasonable predictions to be counted as false positives. Future work will focus on stronger unary predictions, more robust polygon proposal generation, and improved instance-level annotations.



(a) Original image (b) Mask R-CNN prediction (c) Polygon proposal



(d) MRF result (e) TP (green), FP (red), FN (blue) (f) Ground truth

Figure 10. Failure Case: Missed rooftop on the left cannot be recovered without any initial detection.

4.3 Ablation on the Smoothness Parameter

We further analyze the influence of the smoothness parameter λ while fixing the unary scaling factor to $\alpha = 10$. As shown in Table 4, small to moderate values of λ improve spatial consistency by encouraging neighboring polygons to share the same label.

On the Cities dataset, the best trade-off is obtained at $\lambda = 0.1$, achieving the highest mean IoU and a reduction in false positives. On RoofVec, λ values between 0.05 and 0.1 yield the most stable results, with consistently high precision and IoU. In contrast, larger values (e.g., $\lambda \geq 10$) lead to excessive smoothing, causing a significant drop in recall and overall performance.

These results indicate that moderate spatial regularization improves segmentation coherence, while overly strong smoothness suppresses valid rooftop regions. Moreover, the performance remains relatively stable within a reasonable range of λ, suggesting that the proposed model is not overly sensitive to parameter tuning.

(a) Cities dataset						
λ	TP	FP	FN	Prec.	Rec.	mIoU
0.001	749	319	200	70.1%	78.9%	83.3%
0.05	749	313	200	70.5%	78.9%	83.3%
0.1	748	306	201	**71.0%**	78.8%	**83.4%**
1.0	723	262	226	73.4%	76.2%	81.8%
10.0	373	84	576	81.6%	39.4%	50.6%
(b) RoofVec dataset						
λ	TP	FP	FN	Prec.	Rec.	mIoU
0.001	3265	111	90	96.7%	97.3%	91.3%
0.05	3263	99	92	97.1%	97.3%	91.3%
0.1	3262	94	93	**97.2%**	97.2%	**91.3%**
1.0	3248	70	107	97.9%	96.8%	90.5%
10.0	1863	96	1492	95.1%	55.5%	55.7%

Table 4. Ablation study on the smoothness parameter λ for the Cities and RoofVec datasets ($\alpha = 10$).

5. Conclusions

We presented a polygon-level structured inference framework for roof structure extraction from remote sensing imagery. Rather than using raster instance masks as final outputs or assembling roof polygons from local geometric primitives, the method generates an over-segmented polygon partition from line-based cues and formulates roof-region assignment as a structured labeling problem. By combining confidence-weighted instance segmentation evidence with adjacency-based pairwise regularization in an MRF optimization framework, the method produces spatially more consistent roof-region labels.

Experiments on the *Cities* and *RoofVec* datasets show that the proposed framework reduces redundant and spurious rooftop predictions and improves precision while maintaining competitive IoU performance. Qualitative results further demonstrate that polygon-level structured inference can suppress inconsistent assignments, reduce local boundary artifacts, and improve region coherence compared with direct raster-mask outputs or unary-only polygon labeling.

The proposed framework depends on the quality of the initial segmentation and polygon partition. Future work will focus on improving unary predictions and developing more robust polygon generation for complex rooftop configurations.

References

Awrangjeb, M., Lu, G., Fraser, C. S., 2014. Automatic building extraction from LiDAR data covering complex urban scenes.

Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XL-3, 25–32. doi.org/10.5194/isprsarchives-XL-3-25-2014.

Bauchet, J.-P., Lafarge, F., 2018. KIPPI: KInetic polygonal partitioning of images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3146–3154. doi.org/10.1109/CVPR.2018.00332.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A., 2015. Applications of 3D city models: State of the art review. *ISPRS Int. J. Geo-Inf.*, 4(4), 2842–2889. doi.org/10.3390/ijgi4042842.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11), 1222–1239. doi.org/10.1109/34.969114.

Calantropio, A., Chiabrando, F., Codastefano, M., Bourke, E., 2021. Deep learning for automatic building damage assessment: application in post-disaster scenarios using UAV data. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-1-2021, 113–120. doi.org/10.5194/isprs-annals-V-1-2021-113-2021.

Chen, J., Qian, Y., Furukawa, Y., 2022a. HEAT: Holistic edge attention transformer for structured reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3866–3875. doi.org/10.1109/CVPR52688.2022.00384.

Chen, Y., Hong, T., 2018. Impacts of building geometry modeling methods on the simulation results of urban building energy models. *Applied Energy*, 215, 717–735. doi.org/10.1016/j.apenergy.2018.02.073.

Chen, Z., Ledoux, H., Khademi, S., Nan, L., 2022b. Reconstructing compact building models from point clouds using deep implicit fields. *ISPRS J. Photogramm. Remote Sens.*, 194, 58–73. doi.org/10.1016/j.isprsjprs.2022.09.017.

Chicchon, M., Malinverni, E. S., Sanità, M., Pierdicca, R., Colosi, F., Trujillo, F. J. L., 2024. Building semantic segmentation using UNet convolutional network on SpaceNet public data sets for monitoring surrounding area of Chan Chan (Peru). *Geomatics Environ. Eng.*, 18(3), 25–43. doi.org/10.7494/geom.2024.18.3.25.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. DeepGlobe 2018: A challenge to parse the earth through satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 172–181. doi.org/10.1109/CVPRW.2018.00031.

Gao, W., Peters, R., Stoter, J., 2024. Unsupervised roofline extraction from true orthophotos for LoD2 building model reconstruction. T. H. Kolbe, A. Donaubauer, C. Beil (eds), *Recent Advances in 3D Geoinformation Science: Proceedings of the 18th 3D GeoInfo Conference*, Lecture Notes in Geoinformation and Cartography, Springer, Cham, 425–436. doi.org/10.1007/978-3-031-43699-4_27.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969. doi.org/10.1109/ICCV.2017.322.

Hensel, S., Goebbels, S., Kada, M., 2021. Building roof vectorization with PPGNet. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-4/W4-2021, 85–90. doi.org/10.5194/isprs-archives-XLVI-4-W4-2021-85-2021.

Jochem, A., Höfle, B., Rutzinger, M., Pfeifer, N., 2009. Automatic Roof Plane Detection and Analysis in Airborne Lidar Point Clouds for Solar Potential Assessment. *Sensors*, 9(7), 5241–5262. doi.org/10.3390/s90705241.

Lussange, J., Yu, M., Tarabalka, Y., Lafarge, F., 2025. KIBS: 3D detection of planar roof sections from a single satellite image. *ISPRS J. Photogramm. Remote Sens.*, 220, 207–216. doi.org/10.1016/j.isprsjprs.2024.11.014.

Nauata, N., Furukawa, Y., 2020. Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. *European Conference on Computer Vision*, Springer, 711–726. doi.org/10.1007/978-3-030-58598-3_42.

Nelson, J. R., Grubesic, T. H., 2020. The use of LiDAR versus unmanned aerial systems (UAS) to assess rooftop solar energy potential. *Sustain. Cities Soc.*, 61, 102353. doi.org/10.1016/j.scs.2020.102353.

Peronato, G., Kämpf, J. H., Rey, E., Andersen, M., 2017. Integrating urban energy simulation in a parametric environment: a Grasshopper interface for CitySim. *Proceedings of PLEA 2017 Edinburgh: Design to Thrive*, Edinburgh, UK.

Rezaeian, M., Gruen, A., 2011. Automatic 3D building extraction from aerial and space images for earthquake risk management. *Georisk*, 5(1), 77–96. doi.org/10.1080/17499511003679998.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., Bittner, K., 2023. Roof3D: A real and synthetic data collection for individual building roof plane and building sections detection. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-1/W1-2023, 971–979. doi.org/10.5194/isprs-annals-X-1-W1-2023-971-2023.

Van Etten, A., Lindenbaum, D., Bacastow, T. M., 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv:1807.01232*.

Wang, J., Chen, G., Zhang, X., Wang, T., Tan, X., Yang, Q., Zhou, W., Zhu, K., 2025. RoofMapNet: Utilizing geometric primitives for depicting planar building roof structure from high-resolution remote sensing imagery. *Int. J. Appl. Earth Obs. Geoinf.*, 141, 104630. doi.org/10.1016/j.jag.2025.104630.

Xu, B., Xu, J., Xue, N., Xia, G.-S., 2023. HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision. *ISPRS J. Photogramm. Remote Sens.*, 198, 284–296. doi.org/10.1016/j.isprsjprs.2023.03.006.

Xu, Y., Jubanski, J., Bittner, K., Siegert, F., 2024. Roof plane parsing towards LoD-2.2 building reconstruction based on joint learning using remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.*, 133, 104096. doi.org/10.1016/j.jag.2024.104096.

Zhang, F., Xu, X., Nauata, N., Furukawa, Y., 2021. Structured outdoor architecture reconstruction by exploration and classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12427–12435. doi.org/10.1109/ICCV48922.2021.01220.

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1848–1857. doi.org/10.1109/CVPR52688.2022.00189.