RoofVIP Benchmark Dataset: 2D Roof Planar Polygons and Very High-Resolution Digital Orthophotos Pairs for Building Roof Reconstruction

Chaikal Amrullah, Daniel Panangian, Guneet Mutreja, Youssef Abdelhedi and Ksenia Bittner

German Aerospace Center (DLR),
Institute for Remote Sensing Technology (IMF), 82234 Wessling, Germany
(chaikal.amrullah, daniel.panagian, guneet.mutreja, youssef.abdelhedi, ksenia.bittner)@dlr.de

Keywords: Building Roof Reconstruction; Vector-Image Benchmark Dataset; Segmentation and Geometric Model Evaluation

Abstract

Accurate building roof modeling is fundamental to urban analytics, digital twins, and 3D city reconstruction. However, progress in deep learning–based reconstruction is constrained by the limited availability of diverse, high-resolution datasets with detailed geometric annotations. This study introduces ROOFVIP dataset, a large-scale benchmark featuring very high-resolution RGB orthophotos paired with 2D roof vectors that capture diverse urban morphologies across Munich, Germany. Following Level of Detail (LoD) 2 principles, ROOFVIP encompasses a broad range of roof geometries and architectural complexities, providing a robust foundation for evaluating both segmentation- and vectorization-based reconstruction methods. Two reconstruction paradigms are examined: a two-step segmentation-based approach (Cascade Mask R-CNN, Mask R-CNN, SOLOV2, YOLACT) and a one-step direct vector prediction approach (HEAT, PolyRoof). ImageNet-pretrained region-based models, particularly Mask R-CNN and Cascade Mask R-CNN, achieve the highest segmentation accuracy, effectively delineating complex roof boundaries while revealing challenges in small or irregular structures. Geometry-based models exhibit complementary strengths: HEAT prioritizes topological regularity, while PolyRoof emphasizes geometric precision. Although performance metrics are lower than those on simpler datasets such as HEAT and Roof Intuitive, ROOFVIP effectively exposes the challenges of geometric diversity and scale variation, serving as a rigorous benchmark for future research. The dataset includes predefined training, validation, and test splits, enabling consistent benchmarking across methods. By providing a challenging and diverse geometric landscape, ROOFVIP aims to advance geometry-aware deep learning approaches and support scalable, high-fidelity 3D urban modeling. The dataset is publicly available through the project page at `https://chaikalamrullah.github.io/RoofVIP/`.

1. Introduction

1.1 Background

Modeling buildings remains an open research challenge attracting broad attention across disciplines (Biljecki et al., 2016; Cheng et al., 2022; Kutzner et al., 2020; Luo et al., 2022; Qian et al., 2021; Schuegraf et al., 2023; Zorzi and Fraundorfer, 2023). Accurate and up-to-date building models are essential for cadastral mapping, digital twins, disaster management, and autonomous navigation (Atazadeh et al., 2021; Dosovitskiy et al., 2017; Fan et al., 2021; Khan et al., 2022), underscoring the need for detailed, reliable 3D representations in modern geospatial analytics.

However, generating such models remains resource-intensive. Traditional mono- and stereo-plotting require expert operators, while semi-automated methods relying on geometry, shadows, or LiDAR offer limited adaptability to complex urban morphologies (Braun et al., 1995; Rottensteiner and Briese, 2003; Sun and Salvaggio, 2013; Verma et al., 2006). Deep learning has since emerged as a scalable alternative (Amrullah et al., 2025; Qian et al., 2021; Schuegraf et al., 2023; Zorzi and Fraundorfer, 2023), but progress is constrained by the scarcity of diverse, high-resolution datasets with accurate geometric annotations. Existing datasets often lack the spatial detail and structural variability needed for generalization across cities.

Among building components, roofs play a pivotal role: they define geometry and architectural style while being clearly visible in very high-resolution (VHR) imagery. This strong visibility and geometric regularity make roofs ideal for learning-based vectorization, enabling models to infer structure directly from imagery.

To address data scarcity, we introduce the ROOFVIP dataset, a large-scale and high-quality benchmark comprising paired VHR RGB orthophotos and 2D roof vectors that capture diverse urban forms across Munich, Germany (Bayerische Vermessungsverwaltung, 2024b). ROOFVIP combines extensive manual verification with geometric and topological consistency checks to ensure label accuracy. It captures a broad spectrum of roof complexities and establishes a unified benchmark for evaluating segmentation- and vectorization-based deep learning models for roof reconstruction.

1.2 Related Study

Building roof modeling has been extensively investigated across multiple LoD, forming the foundation for automated building reconstruction. The ROOFVIP dataset adopts a 2 Dimension (2D) representation consistent with LoD 2.0, the coarsest variant of the LoD2 specification (Biljecki et al., 2016; Kutzner et al., 2020), characterized by simplified roof geometries that omit fine architectural details and rooftop superstructures.

Several existing datasets follow a similar level of geometric abstraction. The Holistic Edge Attention Transformer for Structured Reconstruction (HEAT) dataset, also referred to as Vectorizing World Buildings (VWB), was introduced in Holistic Edge

Attention Transformer for Structured Reconstruction (Chen et al., 2022; Corley et al., 2024). Derived from the SpaceNet 1 challenge, it comprises 2,001 buildings from 0.3 m Maxar WorldView RGB imagery (256×256 px tiles). The Roof Intuitive or Mesh Image Pair (MIP) dataset (Corley et al., 2024; Ren et al., 2021) contains 3,583 suburban residential roofs annotated with 2D vectors and synthetic 3 Dimension (3D) meshes. While HEAT primarily features simple rectangular industrial roofs, MIP captures more varied residential structures with triangular and irregular polygons. Both, however, are limited in spatial resolution, coverage, and geometric diversity. Examples of building shapes from both datasets are shown in Figure 1.

(a) HEAT dataset

(b) Roof Intuitive dataset

Figure 1. Examples of existing roof datasets.

Together, existing datasets provide roughly 5,500 vector-image pairs for 2D roof modeling research. As of November 3, 2025, other datasets, such as Zeitview Rooftop Geometry (ZRG) (Corley et al., 2024), reportedly contained approximately 22,000 pairs. However, these datasets were not publicly available. Both HEAT and MIP have supported deep learning studies in polygonal roof reconstruction and structural graph prediction (Amrullah et al., 2025; Cheng et al., 2022; Luo et al., 2022; Zorzi and Fraundorfer, 2023). **ROOFVIP** extends these foundations by providing approximately twice the sample size, finer geometric precision, and greater architectural diversity across mixed urban contexts in Munich, enabling more comprehensive benchmarking and generalizable evaluation for automated roof reconstruction.

2. Data Source Challenge

Obtaining a suitable dataset for roof vector reconstruction, particularly at higher levels of detail, remains a considerable challenge due to the dependence on the accuracy and consistency of publicly available geospatial data. The ROOFVIP dataset was developed to address this gap by providing paired vector data and Digital Orthophoto (DOP) imagery for building modeling and reconstruction research. The imagery originates from the Bavarian Open Geodata (Bayerische Vermessungsverwaltung, 2024a,b,c), an open-access initiative of the State of Bavaria, Germany, offering high-quality geospatial resources suitable for automated modeling studies. The following section describes the image data, with the vector component detailed later in this paper.

Figure 2. Example of building samples from study areas. The first row illustrates structures typical of the suburban region, while the second row presents examples from the inner-city area of Munich.

The DOP are orthorectified aerial photographs with a spatial resolution of 20 cm/pixel (DOP20) (Bayerische Vermessungsverwaltung, 2024c), covering two contrasting Munich regions: suburban west and the dense, mixed-use inner city (Figure 2). The suburban zone mainly comprises detached residential buildings and large industrial structures, while the inner city features compact historical blocks, mixed commercial–residential units, and institutional buildings typical of central European architecture. Figure 3 shows the geographic coverage of the study area.

As with most geospatial datasets, certain challenges arise from the data generation process itself. Although not explicitly documented, the DOP appear to have been produced through a rectification procedure that likely integrates auxiliary inputs such as Digital Surface Model (DSM) derived from Light Detection and Ranging (LiDAR) point clouds or stereo imagery. This rectification process can introduce geometric distortions in areas of abrupt elevation change, resulting in minor stretching or skewing along building edges and roof boundaries (see Figure 4a). Consequently, some roof outlines in the orthophotos may exhibit slight deviations from their true geometric form, as several other open datasets with the same level of detail also face similar flaws (Federal Office of Topography or swisstopo, 2024; Senatsverwaltung für Stadtentwicklung, Bauen und Wohnen Berlin, 2023).

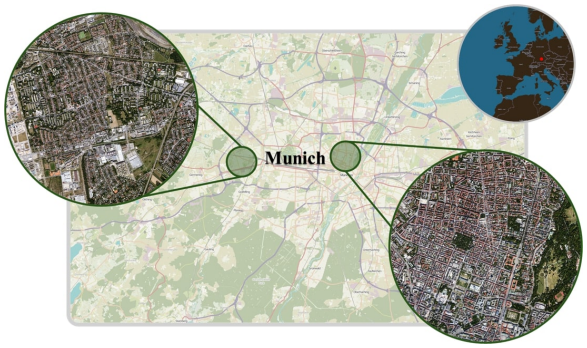
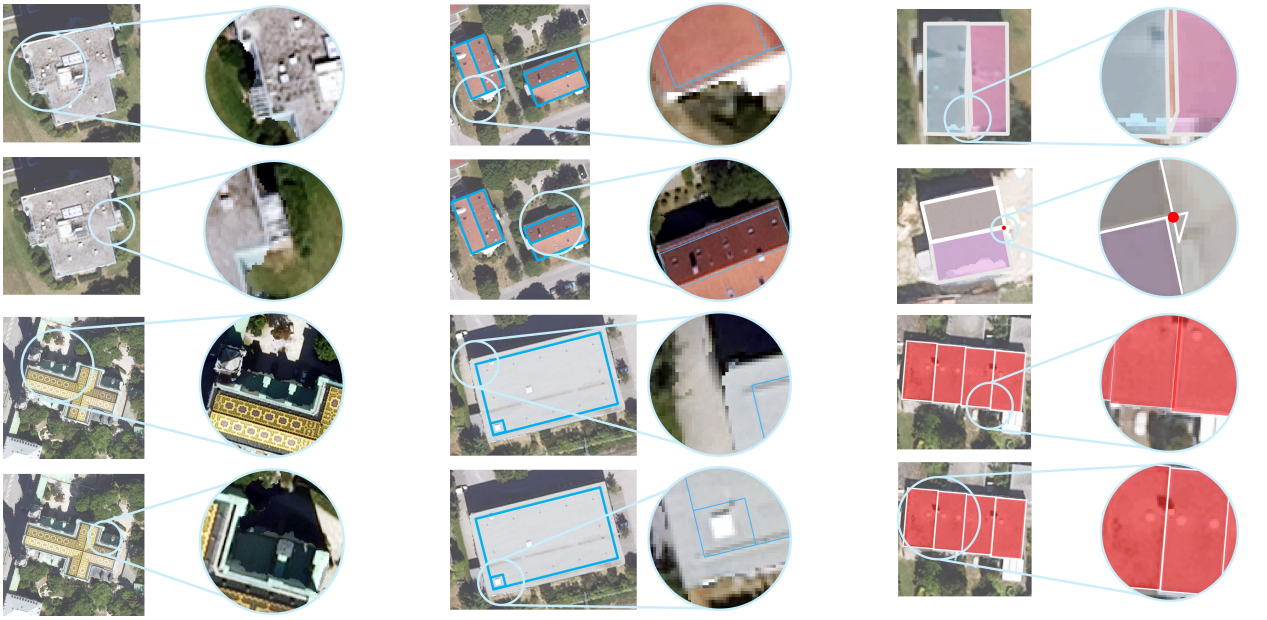


Figure 3. Study area: the metropolitan area of Munich, Germany. The base map for this figure is from Open Street Map (OpenStreetMap, 2025) while the DOP is from (Bayerische Vermessungsverwaltung, 2024c).

As with most aerial datasets, several challenges stem from the data generation process. The rectification of DOP imagery, which likely incorporates DSM data derived from LiDAR or stereo sources, can introduce minor geometric distortions near

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-207-2026



(a) Geometric distortions in the DOP caused by the rectification process, resulting in stretching or skewing along building edges and roof boundaries

(b) Inaccuracies in the original LoD2 roof planar vector data, including misplaced corners, misaligned edges, and over- or under-segmentation of roof structures

(c) Manual labeling topological inconsistencies, such as sliver or gap artifacts, self-intersecting delineation, overlap polygons, and multipart geometry structures

Figure 4. Overview of the data source challenge.

abrupt elevation changes, resulting in stretched or skewed roof edges (Figure 4a). Likewise, vector data from the Bavarian LoD2 3D Building Model (Bayerische Vermessungsverwaltung, 2024a) may contain errors such as misplaced corners, misaligned edges, and over- or under-segmented roof structures (Figure 4b).

Because these procedural artifacts affect both imagery and vector annotations, a manual re-annotation process was performed following LoD 2 specifications to ensure geometric and topological consistency. Such data imperfections, including skewing from rectification, vector mislabeling, and inconsistent segmentation granularity, represent the three primary challenges affecting model learning. These issues can propagate through the reconstruction pipeline, resulting in biased feature learning, incorrect geometric relationships, and reduced generalization performance in subsequent experiments (see Figure 4).

3. RoofVIP Dataset

3.1 Manual Roof Segment Labeling

Given the challenges inherent in both the vector and imagery sources, manual labeling was essential to ensure high-quality and consistent roof vector data. The process began with monoplotting the DOP imagery to extract 2D roof polygons. Due to local distortions in the DOP, roof corners and edges were carefully interpreted and regularized to preserve geometric accuracy. Roofs partially obscured by trees, shadows, or adjacent structures were manually completed based on visual cues, symmetry, and overall building shape (Figure 5).

To maintain consistency, a structured two-step review procedure was followed: an initial labeling (version 1) was refined by a second annotator (version 2), followed by a final validation from a third reviewer. This ensured uniform interpretation of roof structures and alignment with established LoD2 modeling principles. The labeling adhered to subclass 2.1 definitions (Biljecki et al., 2016; Kutzner et al., 2020), capturing primary roof planes and architectural elements larger than 2 m², such as balconies or roof extensions. While this approach minimizes inconsistencies, some degree of interpretative variability remains, representing a known limitation of the dataset.

Following annotation, the labeled segments belonging to the same building footprint were merged using a *unary union* operation. This procedure consolidates multiple roof parts into single building instances, simplifying instance management while preserving geometric hierarchy. However, as discussed in later section, it also introduces a few exceptionally large instances containing multiple roof sections, influencing the model’s learning behavior and complexity distribution.

3.2 Geometry and Topology Validation

After the monoplotting and merging process, geometric and topological validations were performed to ensure internal consistency. Each building is represented as a structured sequence of 2D coordinates forming closed roof polygons. Validation was conducted semi-automatically by examining polygon hierarchies and their spatial relationships.

The checks involved verifying geometric integrity, removing sliver polygons, snapping misaligned vertices, decomposing multipart geometries, and standardizing vertex orientation for consistent data formatting (Figure 4c). Special attention was given to resolving topological exceptions, such as nested roof parts or misrepresented holes (i.e., polygons defined as separate entities rather than reversed inner rings). These corrections ensured that all roof segments were geometrically valid, topologically coherent, and compatible with deep learning frameworks for subsequent model training.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-207-2026 | © Author(s) 2026. CC BY 4.0 License.

With labeling, merging, and validation completed, the finalized dataset was organized into a structured format in preprocessing step described in the following section.

Figure 5. Manual labeling examples. The white lines indicate interpreted roof edges, while the different colors represent individual roof planes resulting from the manual labeling process.

3.3 Preprocessing

After completing the semi-automatic geometry and topology validation, the next stage involves preprocessing the data to prepare it for the learning pipeline. This step includes cropping each individual building into a paired image and vector sample.

Individual buildings are defined based on their geometric groupings rather than administrative boundaries. In practice, planar roof structures that are physically connected are treated as a single building (see Figure 6). This grouping process is achieved through unary union operations on the roof polygons. As a result, the dataset comprises a total of 7,744 vector–image pairs.

An important characteristic of this dataset is that it preserves the Ground Sampling Distance (GSD) of the original imagery. The image size, however, varies depending on the building’s maximum width or height, with additional padding applied to maintain a square format. Only geometries that are fully contained within the image boundary are retained in the corresponding vector file. Each image–vector pair then undergoes a final quality inspection, combining visual assessment with automated topological validation.

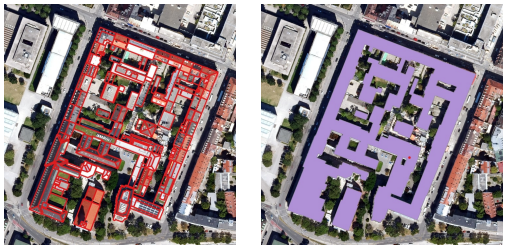


Figure 6. Single-building extraction and cropping process. Geometrically connected roof structures are grouped as individual buildings using unary union operations.

The vector coordinates are converted into the local coordinate system of each image. To ensure spatial traceability, a JavaScript Object Notation (JSON) metadata file accompanies each pair. This file contains the transformation parameters for converting predictions back to global coordinates, as well as details of the GSD and Coordinate Reference System (CRS) in European Petroleum Survey Group (EPSG) format. The overview of this study and the subsequent analysis, focuses primarily on single-building samples to enable a fair comparison with existing benchmark datasets.

3.4 Dataset Comparison

To assess the characteristics of the proposed dataset relative to established benchmarks such as HEAT and Roof Intuitive, a systematic quantitative comparison was conducted. The analysis begins with the computation of geometric descriptors for each building vector. These include basic geometric features such as vertex count and area, as well as higher-level shape attributes derived from geometric and topological properties. Graph-based metrics such as node centrality, path length, and connectivity were also extracted to describe roof-structure complexity.

The relevance of each feature was evaluated using Analysis of Variance (ANOVA) F-values and Mutual Information (MI) scores. Dimensionality reduction was then applied through Principal Component Analysi (PCA) to obtain a composite metric called the geometric complexity score. All scores were normalized to enable consistent comparison across datasets. Figure 7 illustrates the score distributions and representative samples, while Table 1 summarizes the statistical properties and the Earth Mover’s Distance (EMD) between distributions. Smaller EMD values indicate greater similarity between datasets.

A visual and statistical examination reveals distinct complexity profiles among datasets. The RoofVIP dataset exhibits a medium-to-high complexity range with two dominant peaks and a broad spread. Its distribution extends further toward higher complexity values than the others. This results in a higher variance and standard deviation, showing that RoofVIP captures diverse roof geometries and urban morphologies. Meanwhile, HEAT shows a similar general pattern but a narrower variation range, which agrees with its close EMD value to RoofVIP. The Roof Intuitive dataset demonstrates a slightly higher mean complexity, yet it spans a smaller portion of the total range and reflects a more homogeneous structure.

To ensure fair evaluation, the geometric complexity analysis was also applied internally to define the training, validation, and testing subsets. The datasets were split into 80-10-10% portions, respectively. Samples were binned according to their complexity scores so that each subset maintained a balanced representation of simple and complex building structures. This ensured that the training process remained consistent across varying levels of geometric detail, as shown in Figure 8.

4. Experiment

4.1 Two- and One-step Approach

Having established the dataset’s structure and quality, the next objective is to evaluate its capacity to support deep learning workflows for roof vector reconstruction. Specifically, this section assesses whether RoofVIP enables robust performance across established reconstruction paradigms and provides a foundation for model benchmarking and generalization studies.

Roof reconstruction methods can be broadly categorized into two paradigms: the two-step and the one-step approach. The

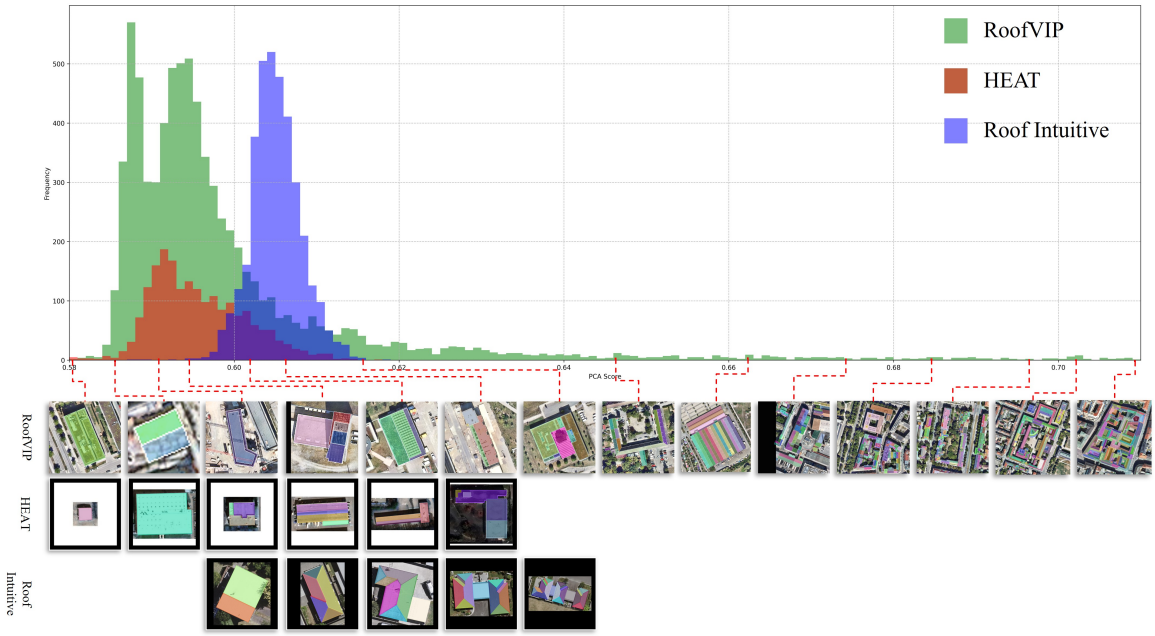


Figure 7. Comparison of the proposed ROOFVIP dataset with established benchmarks.

Dataset	Mean	Mode	Median	Variance ($\times 10^{-5}$)	Std. deviation
RoofVIP	60.14	58.75	59.40	80	2.84
HEAT	59.56	59.15	59.42	3	0.57
Roof Intuitive	60.50	60.45	60.44	1	0.03

(a) Statistical properties.

Dataset Name	Roof VIP	HEAT	Roof Intuitive
Roof VIP	0.0	0.00725	0.0144
HEAT	0.00725	0.0	0.00941
Roof Intuitive	0.0144	0.00941	0.0

(b) EMD.

Table 1. Quantitative comparison of dataset characteristics.

two-step process begins with pixel-level roof instance segmentation, or followed by post-processing for vectorization. In contrast, the one-step approach directly predicts roof primitives (points, edges, or polygons) from imagery, learning geometric relationships without an intermediate mask representation.

In this study, both paradigms are explored. The two-step approach applies instance segmentation using Cascade Mask Region-based Convolutional Neural Network (C Mask R-CNN) (Cai and Vasconcelos, 2018), Mask Region-based Convolutional Neural Network (Mask R-CNN) (He et al., 2017), Segmenting Objects by LOcation Version 2 (SOLOV2) (Wang et al., 2020), and You Only Look At CoefficienT (YOLACT) (Bolya et al., 2019) from the MMDetection framework (Chen et al., 2019). Region-based models (Mask R-CNN, C Mask R-CNN) employ hierarchical refinement to improve boundary delineation. In contrast, single-stage models (SOLOV2, YOLACT) prioritize computational efficiency and maintain balanced performance across scales.

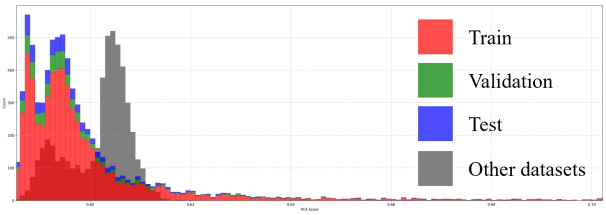


Figure 8. Distribution of building samples by geometric complexity of RoofVIP dataset. Each subset is binned to maintain balanced representation across training, validation, and testing splits.

The one-step paradigm is represented by HEAT (Chen et al., 2022) and PolyRoof (Amrullah et al., 2025). HEAT employs a Transformer-based architecture that encodes global context through a Vision Transformer and decodes it into roof primitives and their topological relationships using a holistic edge-attention mechanism. PolyRoof combines a convolutional neural network (CNN) backbone with a graph neural network (GNN) to refine edge connectivity and polygon topology. Its Area Segmentation Loss (Amrullah et al., 2025) balances roof-segment and building-instance reconstruction, promoting hierarchical understanding of roof geometry. Together, these models highlight trade-offs between segmentation precision, geometric fidelity, and structural coherence.

All models were trained using an adaptive stopping criterion. Training continued until performance plateaued and loss convergence was observed, with the assumption that no secondary plateau would occur. This approach aims to ensured dataset-driven convergence and minimized bias caused by architectural convergence differences. This evaluation framework enables

a balanced comparison between segmentation- and geometry-based reconstruction models, emphasizing how architectural design and learning formulation interact with the geometric complexity represented in ROOFVIP.

4.2 Metric Performance

To compare the performance of the one-step and two-step approaches, both segmentation-based and geometry-based metrics were employed.

For the two-step approach, standard segmentation metrics were used to evaluate mask quality prior to vectorization. The primary metric is the mean Average Precision (mean Average Precision (mAP)), which measures the average precision across all classes and detection thresholds. It is defined as follows:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{1}$$

Here, N denotes the total number of classes. For each class i, TP_i represents the number of *true positive* pixels correctly predicted as belonging to class i, and FP_i denotes the number of *false positive* pixels incorrectly assigned to class i.

Several mAP variants were also considered to capture performance under different evaluation settings. mAP_{50} and mAP_{75} correspond to mean Average Precision measured at Intersection over Union (IoU) thresholds of 0.50 and 0.75, respectively, reflecting lenient and strict localization criteria. The scale-based metrics mAP_s, mAP_m, and mAP_l follow the COCO convention and evaluate detection accuracy across small, medium, and large objects.

For geometry-based, performance was evaluated using metrics that directly assess the quality of vector outputs. These include *Positional Accuracy*, *Line Distance*, *Building Instance F1-score*, *Roof Segment F1-score*, *Reconstruction Score*, and *Count Completeness*. Collectively, these metrics capture geometric precision, completeness, and structural integrity of the reconstructed roof vectors. The formal definitions are provided below (Amrullah and Bittner, 2025; Amrullah et al., 2025).

Building and Roof F1-score: *Precision* and *Recall* are first computed for building and roof segment predictions, and the harmonic mean yields the F1-score:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where *Precision* is the ratio of correctly predicted elements to all predicted elements, and *Recall* is the ratio of correctly predicted elements to all ground truth elements.

Positional Accuracy: This is measured using the Root Mean Square Error (RMSE) between predicted and ground-truth points, p_{pred} and p_{gt}:

$$\text{RMSE}(p_{\text{pred}}, p_{\text{gt}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{i,\text{pred}} - p_{i,\text{gt}})^2} \tag{3}$$

where n is the number of points compared with illustration in Figure 9a.

Line Distance: The line-based discrepancy between predicted and ground-truth polygons is measured using a directed/undirected Hausdorff-type distance:

$$D_L(A,B) = \frac{1}{2}\Bigg[\max\Big(\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\Big) + \inf_{\alpha_a,\alpha_b}\max_{t\in[0,1]}\|\alpha_a(t)-\alpha_b(t)\|\Bigg] \tag{4}$$

where A and B denote the predicted and ground-truth roof polygons, α_a and α_b are their continuous curve parametrizations, and $\|\cdot\|$ is the Euclidean norm for inner and outer line as illustrates in Figure 9b.

Reconstruction Score: This composite score combines building and roof F1-scores to evaluate the overall segmentation quality (shown in Figure 9c):

$$\text{Reconstruction Score} = \frac{2 \cdot F_{1,\text{Building}} \cdot F_{1,\text{Roof}}}{F_{1,\text{Building}} + F_{1,\text{Roof}}} \tag{5}$$

Count Completeness: The completeness of the predicted roof primitives is assessed by comparing the number of predicted and ground truth elements:

$$\text{Completeness Score} = \max\left(0,\, 1 - \frac{|\text{count}_{\text{pred}} - \text{count}_{\text{gt}}|}{\text{count}_{\text{gt}}}\right) \tag{6}$$

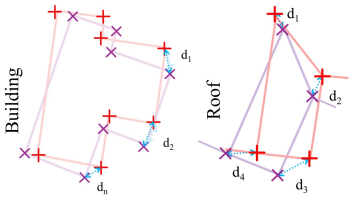
A value of 1 indicates perfect completeness, while 0 represents a completely missing or severely over/under-predicted instance (in Figure 9d).

The illustration of how the metric calculation is implemented on vector data is shown in Figure 9.

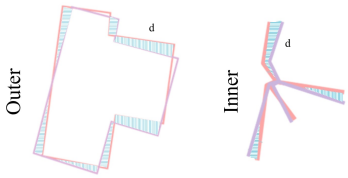
4.3 Result

This section presents the experimental results for the two- and one-step approach. The two-step approach started with instance segmentation process to single roof object class. All models use a ResNet50 (He et al., 2016) backbone to ensure consistent feature extraction and isolate architectural differences in performance. Other training hyper parameters, including batch size, learning rate, optimizer settings, and number of epochs, were kept identical across all experiments to ensure controlled comparison. In the first configuration, models were trained from scratch using only the ROOFVIP data to evaluate their ability to learn roof-specific features without external priors. In the second, ImageNet-pretrained weights (He et al., 2019; Russakovsky et al., 2015) were used to initialize the backbone, allowing analysis of the influence of transfer learning. The quantitative results of both experiments are summarized in Tables 2a and 2b.

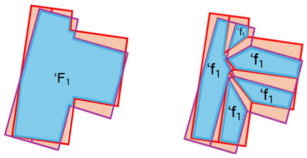
The performance comparison of geometry-based one-step models is summarized in Table 3. Quantitatively, PolyRoof achieves higher positional and line accuracy, indicating more precise geometric reconstruction and improved boundary alignment. Meanwhile, HEAT shows superior F1-scores in roof segmentation and overall geometry counting, demonstrating stronger topological regularity and global structural coherence due to its holistic edge attention design. Both models achieve comparable



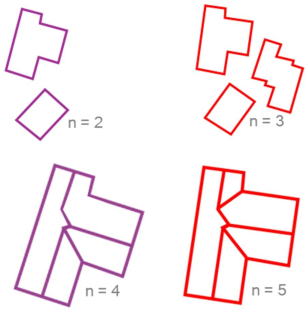
(a) Positional Accuracy



(b) Line Distance



(c) Reconstruction Score



(d) Count Completeness

Figure 9. Illustration of vector-based metric computation for roof reconstruction.

completeness in instance counts, with PolyRoof slightly outperforming in line and overall reconstruction metrics. Overall, these results highlight a trade-off between the spatial precision of PolyRoof and the structural regularity of HEAT, reflecting their differing design priorities in direct vector reconstruction.

5. Discussion and Limitation

Table 2 shows that ImageNet pretraining substantially improves segmentation performance across all models. When trained from scratch, SOLOV2 achieves the highest mAP (37.78%), followed by C Mask R-CNN (30.70%) and YOLACT (28.64%), indicating that single-stage models generalize better without prior knowledge. With pretrained initialization, Mask R-CNN attains the best overall mAP (58.06%), outperforming others across most IoU thresholds and scales, while SOLOV2 performs robustly on medium and large roofs. These results highlight the critical role of transfer learning in improving convergence and generalization, particularly for region-based architectures.

Figure 11 reinforces these findings: region-based models, especially Mask R-CNN, provide the most consistent roof delineations and effectively separate adjacent buildings. C Mask R-CNN performs similarly but occasionally over-segments, whereas SOLOV2 maintains good coverage yet merges nearby structures. In contrast, YOLACT struggles with small or low-contrast roofs.

Model	mAP (↑%)	mAP_{50} (↑%)	mAP_{75} (↑%)	mAP_s (↑%)	mAP_m (↑%)	mAP_l (↑%)
YOLACT	28.64	44.34	30.92	25.05	49.88	46.35
SOLOV2	**37.78**	**55.07**	**37.75**	27.54	46.92	**47.80**
Mask R-CNN	15.75	27.58	16.29	14.86	44.23	37.85
C Mask R-CNN	30.70	49.07	33.13	**29.71**	**50.02**	44.24

(a) Training from scratch.

Model	mAP (↑%)	mAP_{50} (↑%)	mAP_{75} (↑%)	mAP_s (↑%)	mAP_m (↑%)	mAP_l (↑%)
YOLACT	32.47	48.79	35.93	27.55	52.25	47.91
SOLOV2	47.23	68.07	52.75	37.83	63.78	61.69
Mask R-CNN	**58.06**	**76.93**	**65.70**	**43.72**	**72.89**	**75.09**
C Mask R-CNN	48.59	69.37	54.38	40.41	65.65	59.20

(b) Training with ImageNet pretraining.

Table 2. Performance comparison of instance segmentation with ResNet50 backbone in the two-step approach. The bold text indicates best performance while the underline text is for the second best.

An additional factor influencing performance is the dataset structure, where unary union preprocessing produces exceptionally large single-building samples with multiple roof sections, resulting in uneven geometric complexity. While this favors segmentation-based models, it poses challenges for geometry-driven one-step methods that rely on structured relations and permutation-based learning, often limiting convergence and generalization. As shown in Table 3 and Figure 11, PolyRoof achieves higher positional and line accuracy, whereas HEAT attains superior reconstruction scores and geometric counting. Qualitatively, HEAT generates smoother and more globally coherent structures, while PolyRoof captures finer geometric details. These results highlight complementary strengths: HEAT emphasizes topological regularity, whereas PolyRoof prioritizes geometric precision, reflecting a trade-off between structural coherence and detail fidelity.

Nevertheless, absolute performance remains lower than on simpler datasets such as HEAT and Roof Intuitive, likely due to the higher geometric variability in ROOFVIP. This limitation becomes more pronounced on large image patches, where increased scene complexity further degrades performance (Figure 10). Geometry-based models particularly struggle to pro-

Model	Positional Accuracy	Line Distance	Roof F1-score	Building F1-score	Reconstruction Score	Count		
						Point	Line	Polygon
	(↓ px)	(↓ px)	(↑%)	(↑%)	(↑%)	(↑%)	(↑%)	(↑%)
HEAT	2.04	14.71	**37.08**	22.32	27.86	**15.33**	13.07	**7.38**
PolyRoof	**1.89**	**12.92**	32.50	**31.43**	**31.95**	15.08	**15.92**	6.23

Table 3. Performance comparison of geometry-based models with one-step process. Bold text indicates the best performance.

duce complete and consistent polygon structures, often yielding fragmented or topologically incomplete outputs. In contrast, segmentation-based two-step models more reliably classify and separate roof instances, although they remain sensitive to image distortions, resulting in wavy or imprecise boundaries—an effect mitigated most effectively by Cascade Mask R-CNN (Figure 10 in crop image part). Notably, geometry-based methods, while less consistent overall, occasionally localize corner points with higher precision, highlighting a complementary advantage in fine structural reasoning.

6. Conclusion

This study introduced RoofVIP, a large-scale benchmark for automated roof reconstruction, combining high-resolution orthophotos with precisely annotated 2D roof vectors across diverse urban morphologies in Munich. Manual and semi-automatic validation ensures geometric and topological consistency, while statistical analysis shows broader complexity than datasets such as HEAT and Roof Intuitive, establishing RoofVIP as a robust benchmark for both segmentation- and vectorization-based methods.

Experiments show that ImageNet pretraining significantly improves segmentation-based two-step approaches, with Mask R-CNN and Cascade Mask R-CNN achieving the most reliable delineation of complex roofs. Geometry-based one-step models (HEAT, PolyRoof) exhibit complementary strengths: HEAT preserves structural coherence, while PolyRoof captures geometric precision, yet both struggle with generalization under high variability. These challenges intensify on large image patches and rectification distortions, where segmentation models may produce wavy boundaries and geometry-based methods often fail to generate complete, consistent polygons.

Overall, the results reveal a trade-off: segmentation methods provide more stable and interpretable outputs, whereas geometry-based approaches offer greater potential for fully vectorized, end-to-end modeling. Future work will extend ROOFVIP with richer metadata, higher geometric detail, and more diverse urban forms.

Credit Authorship Contribution Statement

Chaikal Amrullah: conceptualization; methodology; data curation; formal analysis; visualization; writing – original draft; investigation (geometric and segmentation models). Daniel Panangian: writing – review and editing; validation; investigation (geometric models). Guneet Mutreja: writing – review and editing; validation; investigation (segmentation models). Youssef Abdelhedi: validation; investigation (geometric models). Ksenia Bittner: writing – review and editing; supervision; project administration.

Acknowledgments

At the time of publication, Chaikal Amrullah (No. 57681552) was funded by a DLR–DAAD Research Fellowship to support his PhD studies.

References

Amrullah, C., Bittner, K., 2025. Graph roof reconstruction with synthetic data supervision from misaligned labels. *Proceedings of the DAGM German Conference on Pattern Recognition (GCPR)*, Freiburg, Germany.

Amrullah, C., Panangian, D., Bittner, K., 2025. Polyroof: Precision roof polygonization in urban residential buildings with graph neural networks. *Proceedings of the Joint Urban Remote Sensing Event (JURSE)*, 1–4.

Atazadeh, B., Halalkhor Mirkalaei, L., Olfat, H., Rajabifard, A., Shojaei, D., 2021. Integration of cadastral survey data into building information models. *Geo-spatial Information Science*, 24(3), 387–402.

Bayerische Vermessungsverwaltung, 2024a. 3d building model bavaria (lod2). `https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=lod2`. Accessed: 2025-09-29.

Bayerische Vermessungsverwaltung, 2024b. Bavarian open geodata portal. `https://geodaten.bayern.de/opengeodata/`. Accessed: 2025-09-29.

Bayerische Vermessungsverwaltung, 2024c. Digital orthophoto bavaria (dop20 rgb). `https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=dop20rgb`. Accessed: 2025-09-29.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59, 25–37.

Bolya, D., Zhou, C., Xiao, F., Lee, Y. J., 2019. Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9157–9166.

Braun, C., Kolbe, T. H., Lang, F., Schickler, W., Steinhage, V., Cremers, A. B., Förstner, W., Plümer, L., 1995. Models for photogrammetric building reconstruction. *Computers and Graphics*, 19(1), 109–118.

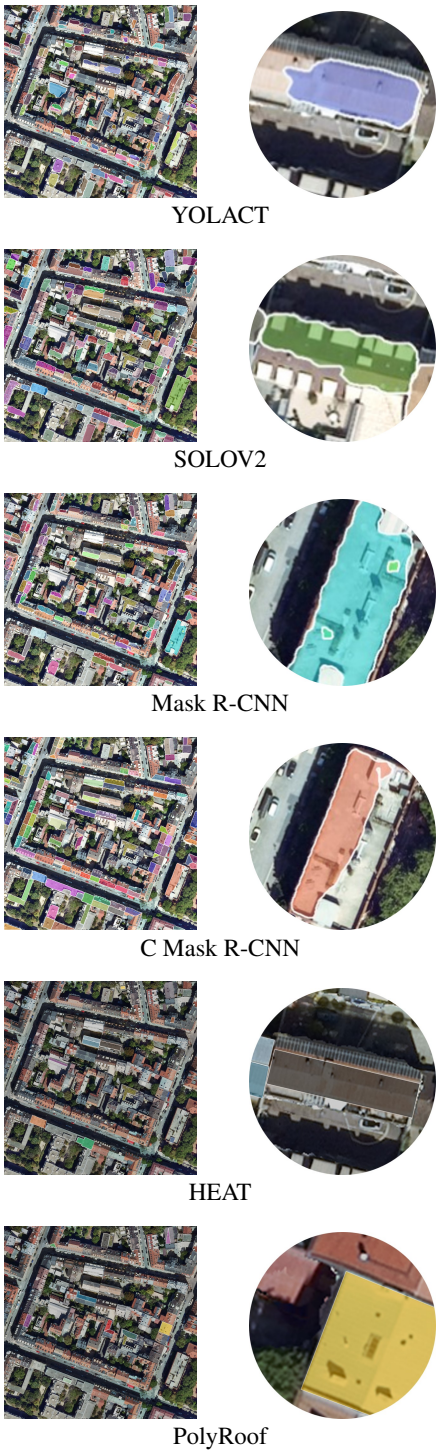


Figure 10. Qualitative comparison of roof reconstruction on a very high-resolution image patch (1338 × 1338 pixels). Each method presents the full prediction on the left and a cropped region on the right, highlighting local geometric details and rectification artifacts. Ground-truth roof planes are shown in distinct colors with white boundaries, and predictions are visualized in varying colors.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6154–6162.

Chen, J., Qian, Y., Furukawa, Y., 2022. Heat: Holistic edge attention transformer for structured reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3866–3875.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.

Cheng, A.-C., Li, X., Liu, S., Sun, M., Yang, M.-H., 2022. Autoregressive 3d shape generation via canonical mapping. *European Conference on Computer Vision*, Springer, 89–104.

Corley, I., Lwowski, J., Najafirad, P., 2024. Zrg: A dataset for multimodal 3d residential rooftop understanding. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4635–4643.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. Carla: An open urban driving simulator. *Conference on Robot Learning*, PMLR, 1–16.

Fan, C., Zhang, C., Yahja, A., Mostafavi, A., 2021. Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management*, 56, 102049.

Federal Office of Topography or swisstopo, 2024. Swissimage 10 cm: Digital color orthophotomosaic of switzerland. `https://www.swisstopo.admin.ch/en/orthoimage-swissimage-10`. Accessed: 2025-09-30.

He, K., Girshick, R., Dollár, P., 2019. Rethinking imagenet pretraining. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4918–4927.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Khan, A. H., Omar, S., Mushtary, N., Verma, R., Kumar, D., Alam, S., 2022. Digital twin and artificial intelligence incorporated with surrogate modeling for hybrid and sustainable energy systems. *Handbook of Smart Energy Systems*, Springer, 1–23.

Kutzner, T., Chaturvedi, K., Kolbe, T. H., 2020. CityGML 3.0: New functions open up new applications. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(1), 43–61.

Luo, Y., Ren, J., Zhe, X., Kang, D., Xu, Y., Wonka, P., Bao, L., 2022. Learning to construct 3d building wireframes from 3d line clouds. *arXiv preprint arXiv:2208.11948*.

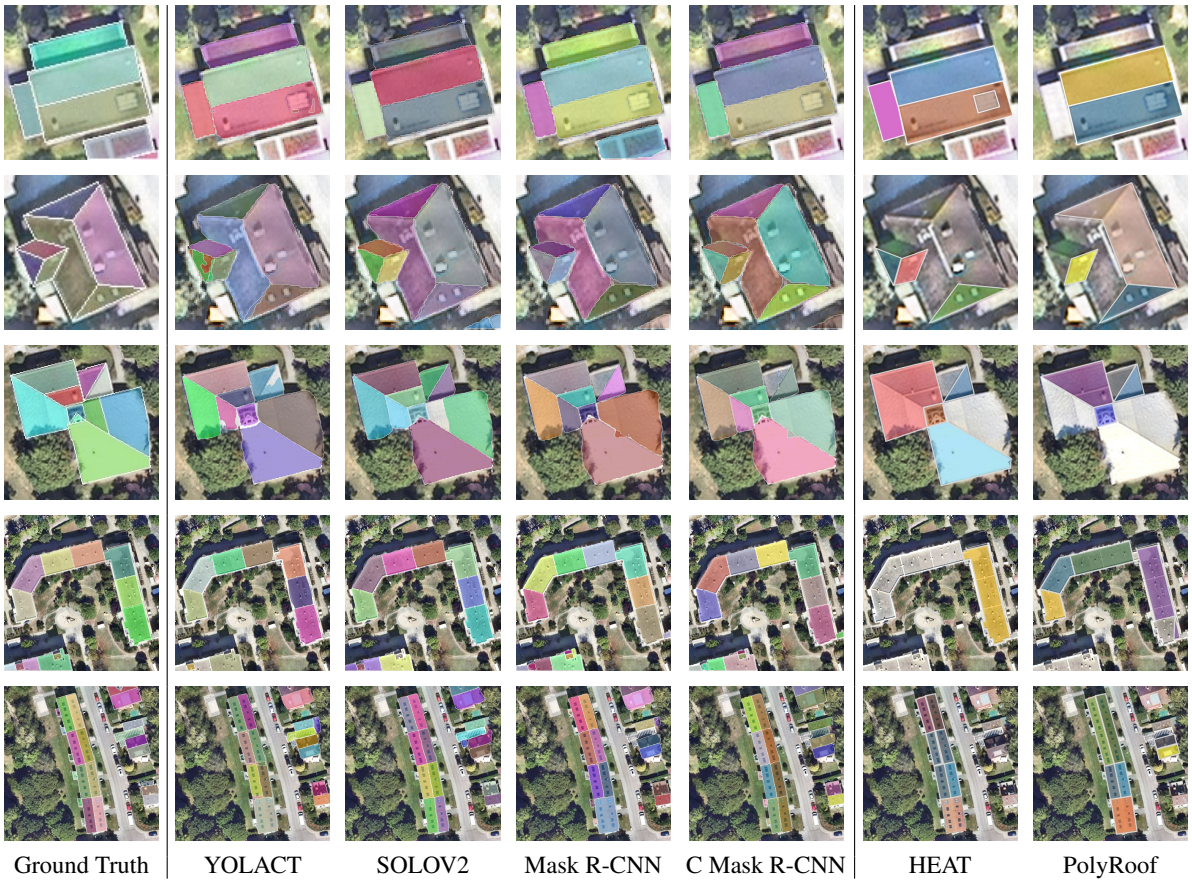


Figure 11. Qualitative comparison of instance segmentation results using ImageNet-pretrained models. Each row depicts urban scenes with varying building complexities, while each column shows outputs from different models. In the ground truth, individual roof planes are colored distinctly with white borders, whereas predictions are visualized in varying colors.

OpenStreetMap, 2025. Openstreetmap. `https://www.openstreetmap.org/#map=11/48.1569/11.5363`. Accessed: 2025-10-13.

Qian, Y., Zhang, H., Furukawa, Y., 2021. Roof-gan: Learning to generate roof geometry and relations for residential houses. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2796–2805.

Ren, J., Zhang, B., Wu, B., Huang, J., Fan, L., Ovsjanikov, M., Wonka, P., 2021. Intuitive and efficient roof modeling for reconstruction and synthesis. *arXiv preprint arXiv:2109.07683*.

Rottensteiner, F., Briese, C., 2003. Automatic generation of building models from LiDAR data and the integration of aerial images. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., Bittner, K., 2023. Roof3D: A real and synthetic data collection for individual building roof plane and building sections detection. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1, 971–979.

Senatsverwaltung für Stadtentwicklung, Bauen und Wohnen Berlin, 2023. Digitale farbige trueorthophotos 2023 (dop20rgbi). `https://daten.berlin.de/datensaetze/digitale-farbige-trueorthophotos-2023-dop20rgbi-3abe1323`. Accessed: 2025-09-30.

Sun, S., Salvaggio, C., 2013. Aerial 3D building detection and modeling from airborne LiDAR point clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(3), 1440–1449.

Verma, V., Kumar, R., Hsu, S., 2006. 3d building detection and modeling from aerial lidar data. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2, 2213–2220.

Wang, X., Zhang, R., Kong, T., Li, L., Shen, C., 2020. SOLOv2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, 33, 17721–17732.

Zorzi, S., Fraundorfer, F., 2023. Re: Polyworld – a graph neural network for polygonal scene parsing. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16762–16771.