Robust Joint Instance-Semantic Segmentation for Semantic Enrichment of 3D Roof Reconstruction from Noisy Labels

Valentina Schmidt [1,2,3], Martin Kada [2]
[1] State Mapping and Surveying Office Lower Saxony
[2] Institute of Geodesy and Geoinformation Science, Technische Universität Berlin
[3] Reimagine Spaces

Keywords: Roof structure analysis, Deep Learning for spatial data, training with noisy labels, semantic and instance segmentation

Abstract

We address the previously unstudied task of roof part instance segmentation in 3D building models, which provides fine-grained semantic and structural information beyond traditional roof surface segmentation. This paper presents a framework that leverages official LoD2 semantic building models and airborne LiDAR point clouds to automatically generate training data for joint semantic and instance segmentation of roof part instances. We introduce a multi-task ConvPoint-based network with bidirectional cross-attention modules for feature sharing, along with a two-stage noise-robust training pipeline designed to mitigate annotation noise and geometric complexity. Experiments on datasets derived from public 3D semantic building models demonstrate that our approach substantially improves segmentation quality under noisy, real-world conditions. The results highlight that progress in automated 3D building reconstruction depends not only on network design but critically on advanced training strategies that can exploit noisy, large-scale semantic building models, providing a reproducible methodology for harnessing public 3D city inventories.

1. Introduction

Three-dimensional (3D) building models are essential components of digital city representations, supporting applications in urban planning, environmental analysis and disaster management. Traditional reconstruction methods rely on manually crafted geometric rules and intensive human intervention, resulting in high costs and infrequent updates. Recent advances in deep learning (DL) have shown potential for automating 3D reconstruction; however, these approaches remain confined to research settings because large, accurately annotated datasets that capture real-world variability are scarce. Examples of DL-based 3D reconstruction include PolyGNN (Chen et al., 2024), which employs a graph-neural network for polyhedron-based reconstruction using a synthetic dataset supplemented with real-world data, and Point2Roof (Li et al., 2022), a two-stage graph-based approach trained on the relatively simple RoofN3D dataset (Wichmann et al., 2019). Although these methods achieve promising geometric results, they generalise poorly to real-world point clouds because of significant geometric and semantic domain gaps. Public 3D building models—such as CityGML LoD2 datasets—could provide large-scale annotations, yet their labels are inherently noisy owing to geometric simplification and limited semantic validation. Previous work (Wang et al., 2023) uses self-supervision to mitigate geometric annotation errors, but weakly supervised learning under semantic noise remains largely unexplored in this context. Consequently, a training strategy that is resilient to noisy labels is still missing. Semantic segmentation can enhance scene understanding by structuring urban data and guiding interpretation. Current methods treat buildings as monolithic entities—or extract roofs as planar primitives—thus missing higher-level semantics and failing to capture complex roof structures. The specific task of roof-part instance segmentation (RPI)—partitioning roofs into semantically meaningful components such as gables, hips or flats—has not been previously addressed. These finer-grained semantics can guide and constrain downstream geometric processing, support accurate physical simulations (e.g., wind loads and solar potential) and improve the interpretability of digital twins.

The task is challenging because roof structures often consist of adjacent, interleaved primitives with subtle geometric transitions, which complicates reliable segmentation and requires models capable of learning robust geometric and semantic patterns. To overcome these limitations, we propose a novel approach that jointly tackles semantic and instance segmentation by explicitly targeting detailed roof-structural components, termed Roof-Part Instances (RPIs). An RPI is the connected subset of roof-labelled points in a 3D point cloud corresponding to a canonical roof primitive—e.g., gable, hip or flat—from a fixed taxonomy; it terminates at geometric discontinuities such as ridges, valleys, eaves or height offsets. Our contribution comprises two tightly integrated advances: (i) Noise-resilient multi-stage training pipeline. We combine self-supervised pre-training, confidence-guided pseudo-labelling and selective relabelling to convert noisy, coarse CityGML data into effective supervision for RPI. (ii) an improved architecture for multi-task learning. A continuous-kernel ConvPoint-based (Boulch, 2020) encoder feeds separate semantic and instance decoders, which are coupled by multiscale bidirectional cross-attention modules. By exchanging information between the two decoder branches at multiple resolutions, these modules enhance the network's ability to learn robustly from training labels affected by geometric and semantic noise. Taken together, our noise-resilient pipeline and cross-attentive architecture turn large yet imperfect LoD2 repositories into reliable supervision for roof-part semantics, pushing the field markedly closer to fully automated, fine-grained 3D urban reconstruction at scale.

2. Related Work

2.1 Instance Segmentation for 3D Point Cloud Data

Various deep learning methods have been developed to tackle semantic and instance segmentation in 3D point cloud data, broadly categorized into proposal-based methods, proposal-free

methods, and multi-task learning approaches, each offering distinct strategies to address the challenges of this complex task.

2.1.1 Proposal-Based Methods

Instance segmentation in 3D point clouds is a core challenge in computer vision and geospatial analysis. Early approaches, inspired by 2D object detection, adopt a proposal-based (top-down) paradigm involving two stages: generation of coarse object proposals and subsequent refinement into instance masks. Examples include GSPN (Yi et al., 2019),which generates proposals through shape reconstruction, and methods that use voting schemes (Ding et al., 2020) or 3D anchor strategies (Boudjoghra et al., 2024). While effective in indoor scenes, these methods struggle in large-scale or structured outdoor environments such as roofscapes, where errors in the proposal stage and computational complexity are significant challenges.

2.1.2 Proposal-Free Methods

To address the limitations of proposal-based approaches, bottom-up methods have emerged that directly learn point-wise embeddings. These methods map each point to a high-dimensional space where embeddings from the same object instance are close, and those from different instances are well-separated. Instance labels are then derived via clustering algorithms such as mean-shift or DBSCAN (Zhao & Tao, 2020). Panoptic-PolarNet (Zhou et al., 2021) utilizes polar Bird's Eye View representations for real-time LiDAR panoptic segmentation, incorporating class-agnostic clustering and adversarial pruning to handle occluded roof instances in urban scenes effectively. Proposal-free approaches provide greater flexibility in scenes with complex geometries and densely packed structures, which are commonly found in roof data. However, their performance depends on embeddings that encode both local geometry and broader semantic context.

2.1.3 Multi-Task Learning for Joint Semantic and Instance Segmentation

Given the intrinsic relationship between semantic and instance segmentation, where each instance belongs to a specific class, recent work often frames the problem as multi-task learning (MTL). These methods employ a shared encoder with task-specific decoders for semantic labeling and instance embedding. ASIS (X. Wang et al., 2019) introduced inter-task communication modules to enhance mutual learning. This approach has proven effective in leveraging the synergy between semantic and instance segmentation tasks. Several efforts have adapted general-purpose frameworks to roof segmentation from airborne laser scanning (ALS) data. RoofNet (Zhang & Fan, 2022) builds upon ASIS, tailoring it to sparse ALS point clouds and leveraging cross-branch feature fusion to capture planar structures more effectively. (L. Li et al., 2024) further enhance the segmentation of adjacent roof planes by introducing a boundary-aware three-branch network. In addition to semantic and instance branches, an offset-prediction branch shifts points toward their instance centers in Euclidean space, and boundary points are treated separately to minimize their disruptive impact on clustering. These innovations reflect a broader shift toward architectures that integrate geometric priors. Nonetheless, current models largely remain focused on roof surface segmentation, lacking the capacity to identify fine-grained, semantically meaningful roof components.

To learn to discriminate between different types of roof surfaces, it is sufficient for most cases for a neural network to learn geometric regularities, such as height, surface normals, or planarity. For roof-part segmentation, however, these low-level geometric cues are insufficient. Canonical roof primitives, such as gables, hips, and dormers, often exhibit smooth transitions and similar geometric features, occur in close spatial proximity, and lack distinct separating boundaries. This challenge relates to the geometric shortcut phenomenon described by (Wu et al., 2025), where 3D models tend to collapse onto low-level geometric signals rather than learning semantic structure. Thus, roof-part instance segmentation demands expressive feature learning, deeper semantic-instance interaction, and training paradigms that resist shortcut-driven collapse, which is essential for precise, semantically enriched 3D roof analysis in digital twins and automated urban modeling.

2.1.4 Automatic Training Data Derivation from 3D Building Models

Recent advances in generating training data for semantic and instance segmentation of roof structures primarily leverage publicly available 3D city models and LiDAR point clouds. (Faltermeier et al., 2023) introduced a large-scale, automated method for generating datasets based on semantic CityGML models, addressing data scarcity in roof segment orientation classification tasks. Their approach utilizes roof polygons extracted from official LOD2 building models, projecting them onto orthophotos to generate pixel-wise semantic labels for roof orientation classes. While this method expands the dataset size and diversity, the segmentation granularity remains coarse, as it is restricted to entire roof segments without instance-level differentiation. Similarly, (Kong & Fan, 2024) proposed an automated method for generating datasets for roof segmentation directly from open-source LoD2 building models, emphasizing flexibility by allowing user-defined point densities and noise levels. Their dataset, NRW3D, leverages real geographic context to enhance model generalization but focuses primarily on semantic labeling without addressing detailed hierarchical modeling or roof type diversity. (R. Wang et al., 2023) created the Building3D dataset, comprising aerial LiDAR point clouds and corresponding mesh and wireframe models across urban-scale areas, designed to facilitate urban modeling research. This dataset emphasizes structural complexity and geometric fidelity. However, it does not explicitly integrate semantic labeling related to roof orientation or canonical roof shapes, limiting its direct applicability to detailed semantic segmentation tasks. Our work differentiates itself by introducing a methodology that integrates geometric and semantic information from publicly available CityGML-based LOD2 building models with LiDAR point clouds to generate high-quality annotations for simultaneous semantic and instance segmentation of roof part instances (RPIs). By explicitly exploiting hierarchical structures inherent to CityGML models, including building parts, roof surfaces, and the detailed roof type attributes included by these models, we provide finer-grained annotations at the instance level. Moreover, our approach explicitly addresses granularity mismatches and semantic ambiguities in the input data through targeted refinement strategies, aiming to produce a robust training dataset suited for deep learning methods.

3. Workflow for Automated Dataset Derivation and Point-Level Annotation

A roof part instance (RPI) is defined in this work as the connected subset of roof-labeled points in the 3D point cloud that corresponds uniquely to a canonical roof primitive a fixed taxonomy which currently includes eight roof classes: flat, shed, gable, hip, jerkinhead, pyramid, hip-gable, mansard roof. Each RPI terminates at geometric discontinuities such as ridges, valleys, eaves, or height offsets.

The main data source comprises official German Level-of-Detail 2 (LoD2) 3D building models delivered in the AdV-CityGML profile. Compliance with the "Product and Quality Standard for 3D Building Models" defined by the German Working Committee of the Surveying Authorities (AdV) ensures nationwide consistency in geometric representation and semantic annotation. According to this standard, a building is decomposed into building part objects whenever it contains two or more discrete volumetric bodies; contrasting roof forms frequently coincide with such divisions but are not a prerequisite. In LoD2, each building part is annotated with the roofType attribute, selected from the current AdV code list of sixteen roof-type classes, while the parent building carries this attribute only when no parts exist. This geometric–semantic hierarchy underpins the derivation of individual roof-section instances in our workflow. Additional data sources include airborne laser-scanning (ALS) point clouds in LAS format and pre-segmented 2D building-instance polygons generated via deep-learning methods from aerial imagery and elevation data.

Preprocessing: CityGML tiles are converted to CityJSON for efficient parsing. We extract semantic and geometric information, isolating individual buildings, their hierarchical subdivisions into building parts, and corresponding roof surfaces. Each extracted building and its associated surfaces—including roof, wall, and ground surfaces—are stored along with semantic roof-type labels and geometric parameters.

RPI Candidate Generation and Refinement:
A key component of the methodology is the generation and refinement of Roof Part Instance (RPI) candidates for precise annotations. Initially, ground surfaces from LoD2 data are used as preliminary polygons to define the horizontal extent of roof instances. Due to inherent geometric simplifications in LoD2 models, these polygons often require adjustment. Refinement integrates detailed planar roof surfaces from LoD2 data, matching them to candidates using a 50% overlap threshold to optimize accuracy and computational efficiency. Matched roof surfaces are merged into unified polygons, enhancing delineation by capturing geometric discontinuities like ridges and eaves. Further, attributes such as slope, surface normal orientation, and height differentials aid in identifying continuous roof structures across adjacent building parts. The process is exemplified in Figure 1, where the top panel displays the initial ground surface polygons outlined in orange alongside the individual roof surface polygons highlighted in red. The leftmost building illustrates a common challenge: the ground polygons fail to align with actual roof part transitions, resulting in inaccurate and fragmented delineations. In contrast, the central building serves as a successful example of refinement, where the integration of detailed roof surface geometries produces polygons that closely correspond to the actual roof structure, demonstrating improved accuracy. However, the rightmost building reveals a limitation of the approach, as neither the ground polygons nor the roof surface geometries adequately capture the transition between distinct roof parts, indicating a failure to resolve complex roof morphology.
The bottom panel presents the final automatically generated polygons for roof part instances (RPIs) in magenta, after the refinement process. While these refined polygons generally improve the representation of roof parts, several errors remain visible. These include missed transitions, where distinct roof parts are not separated, and misaligned boundaries, where polygon edges do not precisely follow roof discontinuities. Such errors underscore the inherent challenges and limitations of fully automated geometric and semantic annotation workflows, particularly in handling complex or irregular roof configurations. This visual comparison highlights the necessity for further methodological improvements to enhance delineation fidelity and semantic accuracy in roof part instance extraction.
Semantic labels are assigned based on LoD2 attributes, with a selective sampling strategy that prioritizes rare roof types and limits the overrepresentation of common types, such as flat and gable roofs, to ensure dataset diversity.

Semantic Labeling, Patch Extraction, and Annotation: In the concluding phase, point cloud patches are extracted from input tiles using Building Instance Polygons (BIPs) derived from aerial imagery and ALS data. Patches are refined to maintain spatial continuity across tile boundaries. For annotation, refined RPI polygons are projected onto the horizontal plane, and LiDAR point coordinates are assessed against these boundaries. Points within polygons are assigned unique RPI identifiers and corresponding roof type labels, while others are labeled as background. This workflow creates a comprehensive dataset for training algorithms in the semantic classification of diverse roof structures.

Figure 1: Top: Ground polygons (orange) and roof surface polygons (red). Bottom: Final roof part instance polygons (magenta) after refinement.

https://doi.org/10.5194/isprs-archives-XLVIII-4-W15-2025-141-2025

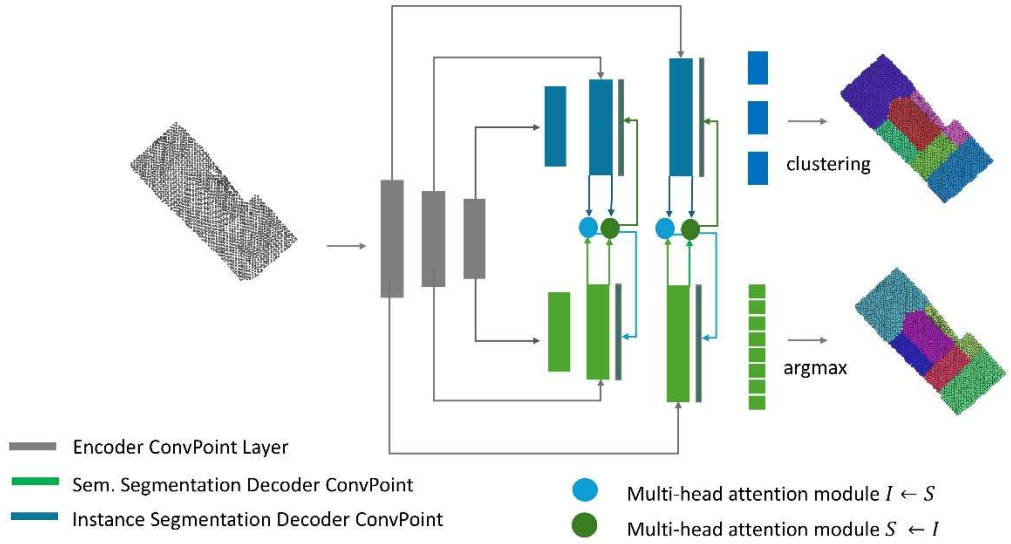

Figure 2: Architecture of the proposed bidirectional multi-task network. The encoder extracts features from the input point cloud, which are then processed by separate semantic segmentation (green) and instance segmentation (blue) decoder branches.

4. Method

The objective of this work is to develop a robust deep learning framework for roof-part instance segmentation (RPI) from airborne LiDAR point clouds. Specifically, the goal is to assign each point in a 3D point cloud patch—potentially containing complex or multiple adjacent roofs—both a semantic label (indicating roof-part type) and an instance label (a unique identifier for each contiguous roof part), even in the presence of noisy training annotations. while handling noisy annotations in the training data.

Let an input point cloud patch be denoted as: $P = (x_i, a_i)_{i=1}^{N}$ where $x_i \in R^3$ represents the 3D coordinates of point i, $a_i \in R^d$ denotes optional attributes such as intensity. Each point is annotated with:

- A semantic label $y_i^s \in \mathcal{C}$ where $\mathcal{C} = c_1, c_2, \dots, c_K$ is the set of semantic roof categories.
- An instance label $y_i^I \in 1,2, \dots, R$, where R is the number of roof-part instances in the patch

Due to the automatic nature of annotation derived from LoD2 3D building models, the observed labels $(\tilde{y}_i^S, \tilde{y}_i^I)$ may be corrupted versions of the true labels (y_i^s, y_i^I) i.e., they can be incorrect or inconsistent with the true underlying roof structure.
The core task is to learn a function f_θ parameterized by neural network weights θ, mapping the input point cloud to both semantic and instance predictions: $f_\theta: P \rightarrow \{(\hat{y}_i^S, \hat{y}_i^I)\}_{i=1}^{N}$ (1)
where $\hat{y}_i^S$ and $\hat{y}_i^I$, are the predicted semantic and instance label for point i in a input cloud patch including M points.

4.1 Multi-Scale Attention Network for Robust Roof-Part Instance Segmentation

Building upon recent advances in joint semantic and instance segmentation of 3D point clouds we propose a novel architecture designed explicitly for segmenting roof building parts in airborne laser scanning (ALS) point cloud patches. While prior works have demonstrated the effectiveness of multi-task learning and feature interaction mechanisms, significant challenges remain in delineating complex roof geometries, resolving ambiguous boundaries, and efficiently processing large-scale LiDAR data with potentially noisy annotations.
The network operates on a point cloud patch with a fix number of points and produces two types of output: per-point semantic logits $s \in \mathbb{R}^{NxC}$, where C is the number of semantic classes and instance embedding maps $E \in \mathbb{R}^{NxD}$, where D is the dimensionality of the learned embedding space.
The network backbone utilizes ConvPoint layers, which perform continuous convolutions directly on raw point clouds. This approach maintains permutation and translation invariance, while avoiding the information loss commonly associated with voxelization or quantization. As a result, the network is able to capture both fine-grained geometric features and broader contextual information. The encoder is structured as a sequence of down-sampling stages, each halving the number of points and increasing the feature dimensionality. Two dedicated decoder branches—one for semantic segmentation and one for instance segmentation—mirror the encoder by progressively restoring spatial resolution through inverse neighborhood aggregation. Skip connections link corresponding encoder and decoder stages, enabling the retention of high-resolution geometric details throughout the network. Figure 1 illustrates the overall architecture. This architecture is inspired by recent advances in point-based segmentation, emphasizing the value of multi-scale feature extraction and careful preservation of spatial structure.
To enable effective information exchange between semantic and instance segmentation tasks, the network employs two parallel decoder streams: one for semantic segmentation (S) and one for instance embedding I .At intermediary and top decoder level, we introduce a pair of multi-head attention ($MHAtt$) modules (Vaswani et al., 2017) , facilitating bidirectional, multi-scale feature interaction between the two streams:
$Z^{I \leftarrow S} = MHAtt(Q_I, K_S, V_S),\ Z^{S \leftarrow I} = MHAtt(Q_S, K_I, V_I)$ (2)
where: H_s and H_I denote the semantic and instance decoder features at the current layer, respectively. $Q_s = K_s = H_s$ are the queries, keys, and values from the semantic decoder. $Q_I = K_I = H_I$ are the queries, keys, and values from the instance decoder.

The updated decoder features are then computed as: $H_I \leftarrow H_I + Z^{I \leftarrow S}$, $H_s \leftarrow H_s + Z^{S \leftarrow I}$ (3)
This bidirectional attention is applied at multiple decoder scales, allowing the network to dynamically aggregate contextual information at different levels of abstraction. The semantic-to-instance path helps the instance decoder distinguish geometrically similar roof parts of different types, while the instance-to-semantic path improves the delineation of semantic boundaries between adjacent roof parts.

4.2 Noise-Robust Training Pipeline for Joint Semantic and Roof-Part Instance Segmentation

Learning from LoD2-derived roof labels presents unique challenges due to frequent misclassifications and per-point label corruption. To address this, we propose a three-stage training pipeline that (1) learns geometric priors without reliance on labels, (2) progressively denoises supervision during early training, and (3) guards against late memorization of noise.
Given a training set $D = \{(P_i, \tilde{y}_i^S, \tilde{y}_i^I)\}_{i=1}^M$ (4), where $\tilde{y}_i^S$, $\tilde{y}_i^I$ noisy labels, the network is trained to minimize a robust multi-task loss:
$\mathcal{L}(\theta) = \lambda_s \cdot \mathcal{L}_{sem}(\hat{y}_i^S, \tilde{y}_i^S) + \lambda_I \cdot \mathcal{L}_{inst}(\hat{y}_i^I, \tilde{y}_i^I)$ (5) where: $\mathcal{L}_{sem}$ is a noise-robust semantic segmentation loss function (e.g., Generalized Cross Entropy), $\mathcal{L}_{inst}$ is a discriminative loss, for instance embedding, encouraging points from the same instance to be close in embedding space and different instances to be well separated λ_s and λ_I are weighing factors. The discriminative loss (Wang 2019) acts on embeddings e_i:
$\mathcal{L}_{disc} = \frac{1}{M}\sum_{m=1}^{M}\left[\alpha \sum_{i \epsilon I_m} \|e_i - \mu_m\|^2 + \beta \sum_{n \neq m} max(0, \delta - \|\mu_m - \mu_n\|)^2\right]$ (5) where μ_m is the mean embedding of instance m.

4.2.1. Geometry-Centred Pre-Training

In order to equip the encoder with robust geometric priors, we utilize a self-supervised occlusion-completion pretext task (Wang et al., 2021). In each input patch, 60% of the points are occluded. The occluded portions are selected using farthest-point sampling. A decoder is trained to reconstruct the coordinates of the occluded points. As self-supervised loss we use Chamfer Distance (CD) $\mathcal{L} = CD(P_{mask,}\hat{P})$ (6) which calculates a symmetric distance between two sets of points. This produces weights θ_{geom} based on the learned features space and is agnostic to label noise.

4.2.2 Noise-Aware Warm-Up with Bootstrapping

Initializing with the encoder weights with θ_{geom}, we train the network with the multi-task loss (eq. 5), where for the semantic component we Generalized Cross-Entropy (Zhang and Sabuncu, 2018) $GCE(p_t, q) = \frac{1-p_t^q}{q}), 0 < q \leq 1$, (7) where p_t denotes the predicted probability assigned to the true class. Setting $q = 1$ recovers the standard cross-entropy loss, while values of $q < 1$ increases robustness to label noise.
To further mitigate noise, we apply soft bootstrapping (Reed et al., 2015) combining model prediction confidences and the noisy labels: $\hat{y}_i^S = (1-\beta)p + \beta\tilde{y}_i^S$ (7) . During this phase, we exploit the "memorization effect" (H. Zhang et al., 2024) deep networks first learn clean patterns before fitting noise. By monitoring prediction confidence, we dynamically refine the label set—replacing noisy labels with high-confidence predictions when they remain stable across epochs. This targeted pseudo-labeling reduces the effective noise rate and improves the reliability of supervision for semantic tasks.
The resulting hybrid loss function can be formally expressed as:
$\mathcal{L}_{sem} = (\hat{y}^S, p, \beta, q) = \frac{1-(\hat{y}^S \cdot p)^q}{q}$ (8) where $\hat{y}^S$ represents the soft bootstrap-adjusted labels, p represents the predicted probabilities from the model, parameters β and q control label adjustment and robustness, respectively. A β value close to 1 emphasizes reliance on the original annotations, whereas a smaller value increasingly trusts the model's predictions, effectively self-correcting mislabeled data points during training. The combined approach aims to benefit simultaneously from the self-correcting nature of bootstrap targets and the inherent noise robustness of GCE loss, thus enhancing the segmentation model's resilience to noisy annotations.

4.2.3 Stable Long-Run Optimization

During the final training stage, we employ beta-mixture re-weighting to mitigate the influence of noisy labels. This method fits a two-component beta mixture model to the distribution of per-sample training losses, distinguishing clean and noisy samples and dynamically adjusting their contributions to the total loss.
During training, we model the distribution of per-sample losses l_{i} as a two-component beta mixture:
$\text{p}(l_{\text{i}}) = \pi_1 \text{Beta}(l_{\text{i}};\, \alpha_1, \beta_1) + \pi_2 \text{Beta}(l_{\text{i}};\, \alpha_2, \beta_{21})$, (9) where π_1 and π_2 are the mixing coefficients and $\alpha_1, \beta_1, \alpha_2, \beta_2$ are the beta distribution parameters for clean and noisy samples, respectively.
The posterior probability that a sample is clean, w_{i} is used to weight its loss: $\text{w}_{\text{i}} = \frac{\pi_1 \text{Beta}(l_{\text{i}};\, \alpha_1, \beta_1)}{\text{p}(l_{\text{i}})}$ (10). are the beta distribution parameters for clean and noisy samples, respectively. The final loss is then a weighted sum: $\mathcal{L} = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}} \text{w}_{\text{i}} l_{\text{i}}$. (11). For a comprehensive description of the estimation and EM algorithm, we refer the reader to the original work by Arazo et al. (2019).

4.3 Evaluation Metrics

We evaluate our joint semantic and instance segmentation results using four complementary metrics: mean Intersection-over-Union (mIoU), mean coverage (mCov) and Panoptic Quality (PQ). We evaluate our joint semantic and instance segmentation results using four complementary metrics: mean Intersection-over-Union (mIoU), mean coverage (mCov) and Panoptic Quality (PQ). Mean Intersection-over-Union evaluates the semantic segmentation performance by averaging the IoU_c across all classes. $\text{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c}$ (9) where, TP_c, FP_c, and FN_c are the true positive, false positive, and false negative counts for class c. Mean coverage assesses instance segmentation quality by averaging the best intersection-over-union (IoU) between each ground-truth instance and its most overlapping predicted instance: $mCov(G, P) = \frac{1}{|\text{G}|}\sum_{\text{g}\in\text{G}} \max_{\text{p}\in\text{P}} \text{IOU(g, p)}$ (10), where G is the set of ground-truth instances and P s the set of predicted instances. Panoptic Quality (PQ) jointly measures semantic and instance segmentation performance by combining segmentation and recognition quality: $\text{PQ} = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP|} \cdot \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$ (11) where p, g are matched predicted and ground-truth segments.

4.4 Datasets

The RoofN3D dataset consists of point cloud data from New York City, providing geometric and semantic labels for building reconstruction tasks. It includes only simple, standard roof types—primarily flat, gabled, hipped, and pyramid roofs. For full details, readers are referred to Wichmann et al. (2019).

The SemRoof dataset covers 337 km² in Lower Saxony, Germany, and is based on official LoD2 building models, airborne LiDAR point clouds (≥4 points/m², ≤0.30 m horizontal and ≤0.15 m vertical accuracy). Approximately 75% of the data is from urban regions. LoD2 models are generated via building footprint intersection and model-driven shape recognition (~70% semantic roof type accuracy). The dataset comprises 42,690 point cloud patches (84,588 building parts), with 70% of buildings exhibiting complex roofs (1–11 segments) and eight semantic roof types. To enhance quality, single-segment standard roofs were excluded, patch areas were limited to 35–450 m², and points below the 5th height percentile were removed to reduce noise.

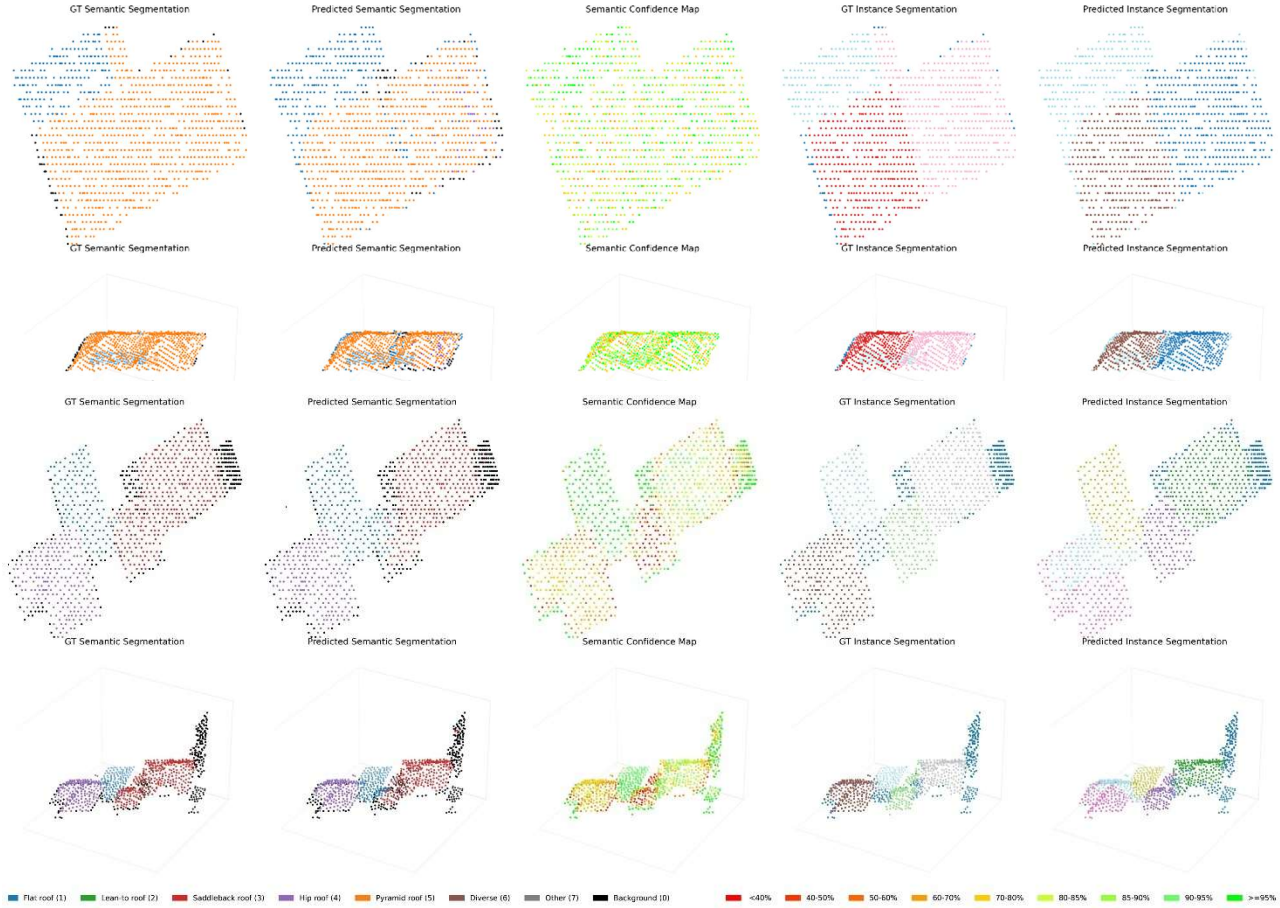

Figure 3: Illustration of segmentation and instance prediction results for two example cases. The figure includes: (a) noisy ground truth for segmentation (SemRoof dataset), (b) predicted segmentation, (c) confidence map for segmentation, (d) ground truth for instance segmentation

5. Experiments

The network is built on ConvPoint layers, each with 16 convolutional centers and an initial feature dimension of 48, which is doubled in the second layer and maintained thereafter. Both the encoder and each decoder branch comprise three ConvPoint layers, forming a three-level hierarchy with progressive 4:1 downsampling at each stage. Each input patch consists of 256 input points for RoofN3D data and 1,024 points for SemRoof data. Batch normalization and ReLU activation follow each convolutional layer. Bidirectional cross-attention modules operate on 96-dimensional feature spaces, enabling multi-scale interaction between the semantic and instance decoders. Skip connections and bottleneck layers are used to preserve spatial detail while controlling feature dimensionality.

The semantic decoder outputs per-point class logits for 12 semantic roof-part classes, which are converted to labels via argmax. The instance decoder outputs per-point embedding vectors of dimension 5. In post-processing, instance predictions are obtained by applying mean shift clustering with a bandwidth of 0.8 to the embedding space.

Training is performed using the AdamW optimizer with an initial learning rate of 0.001 and weight decay of 0.01, with a batch size of 1. The multi-task loss combines a semantic segmentation loss and a discriminative instance embedding loss, weighted by a factor of 5. Dropout regularization ($p = 0.5$) is applied throughout the network. Data augmentation strategies—including point jittering, random z-axis rotations, and small random occlusions—are used during training.

6. Results

On the RoofN3D dataset, our network demonstrates rapid convergence, achieving stable and robust results in fewer than 50 training iterations with the Adam optimizer and cross-entropy loss. To systematically evaluate the impact of feature knowledge transfer between the semantic and instance segmentation decoders, we conducted experiments under four scenarios: (1) no feature transfer, (2) transfer from the instance decoder to the semantic decoder (S ← I), (3) transfer from the semantic decoder to the instance decoder (I ← S), and (4) bidirectional transfer

(S ← I, I ← S). The quantitative results for each transfer scenario are summarized in Table 1.

	mCov	mIoU	PQ
Baseline	0.87	0.89	0.93
(S ← I)	0.91	0.93	0.95
(S ← I, I ← S).	0.92	0.94	0.96

Table 1:Performance metrics for different feature transfer scenarios on the RoofN3D dataset.

The baseline configuration without feature transfer achieves strong performance, indicating that simple geometric cues suffice for planar roof surfaces. Knowledge transfer from the instance to semantic branch (S ← I) improves both semantic and instance segmentation, supporting findings that cross-task feature sharing is beneficial. Transferring features from semantic to instance branch (I ← S) has little effect, likely because semantic classes align closely with geometric structure. Bidirectional transfer maintains the gains seen with instance-to-semantic transfer. The network performs consistently across roof types, with most remaining errors attributable to ground truth inaccuracies rather than model limitations.

The SemRoof dataset poses greater challenges due to higher annotation noise and increased roof-part complexity. Conventional training quickly led to stagnating metrics and overfitting, highlighting the need for robust strategies. To address this, we implemented a noise-robust training pipeline: the encoder is first initialized via self-supervised geometric pre-training, followed by a noise-aware warm-up using generalized cross-entropy loss (β=0.7β=0.7) and soft bootstrapping for 36 epochs. Once the loss plateaued, high-confidence predictions (over 90%) were used to replace likely corrupted ground truths, generating pseudo-labels for further optimization. The network was then retrained with these pseudo-labels for 120 epochs at a fixed learning rate of 0.001.
This protocol led to substantial improvements in both semantic and instance segmentation, as shown in Table 2. Many remaining errors are linked to ambiguous or incorrect roof-part boundaries, underscoring the need for higher-quality ground truth.

Training Stage	mCov	mIoU	PQ
Stage 1	0.60	0.55	0.68
Stage 3	0.76	0.67	0.78

Table 2: Performance metrics for the two training stages on the SemRoof dataset.

6.1 Discussion

This work demonstrates not only the feasibility of leveraging existing semantic 3D building models as training data for fine-grained roof-part interpretation, but—more importantly—the potential of noise-robust training strategies to advance deep learning for 3D semantic building reconstruction. Our results show that, under controlled conditions and with relatively clean datasets such as RoofN3D, the proposed bidirectional multi-task ConvPoint network achieves rapid convergence and high segmentation accuracy. Cross-attention modules, particularly with knowledge transfer from the instance to the semantic decoder, further enhance performance at challenging roof boundaries, confirming the value of multi-task feature sharing.
However, our experiments on the SemRoof dataset highlight the challenges of applying such models to real-world, automatically labeled data. Despite a systematic pipeline for dataset preparation, SemRoof exhibits substantial annotation noise, especially at roof-part boundaries and in complex structures. This noise has a clear negative impact on model performance, as seen in rapid overfitting and stagnating metrics with conventional training. Our multi-stage, noise-robust training protocol, which includes self-supervised pre-training, generalized cross-entropy, bootstrapping, and pseudo-labeling, substantially improves results; however, absolute performance remains constrained by the quality of the underlying labels. Furthermore, the manually verified evaluation set contains relatively simple cases, which may inflate reported metrics and limit generalizability to more complex scenarios.
Looking ahead, several directions are essential. Expanding datasets to include 3D building models generated by a variety of reconstruction algorithms will help disentangle model errors from those inherent to specific pipelines. Including more complex roof structures in training, validation, and test sets is necessary for robust evaluation and generalization. Our experiments were limited to patches of sloped roof surfaces smaller than 450 m² and 1,024 points per patch; adapting the approach for larger and more variable patches is essential for scalability. Incorporating a weak supervision strategy, where at least a portion of the training data is manually verified, would improve label quality and model reliability, supporting more accurate benchmarking.
While our two-step training protocol yields promising results, it remains complex and resource-intensive. Future work should explore ways to simplify and streamline the training process, reducing computational demands without compromising segmentation accuracy. Additionally, we note that architectural parameters were not exhaustively tuned, and no ablation studies were conducted in this direction; further optimization of these parameters may yield additional performance gains.
In summary, our findings underscore that the key to progress in 3D semantic building reconstruction lies not only in network design but in the effective exploitation of noisy, large-scale semantic building models through advanced training methodologies. This paradigm enables the field to move beyond the limitations of small, manually curated datasets and accelerate innovation in automated urban modeling.

References

Arazo, E. *et al.* (2019) 'Unsupervised label noise modeling and loss correction', *36th International Conference on Machine Learning, ICML 2019*, 2019-June, pp. 465–474.

Boudjoghra, M. E. A. *et al.* (2024) 'Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation', pp. 1–13. Available at: http://arxiv.org/abs/2406.02548.

Boulch, A. (2020) 'ConvPoint: Continuous convolutions for point cloud processing', *Computers and Graphics (Pergamon)*, 88, pp. 24–34. doi: 10.1016/j.cag.2020.02.005.

Chen, Z. *et al.* (2023) 'PolyGNN: Polyhedron-based Graph Neural Network for 3D Building Reconstruction from Point Clouds', *ISPRS Journal of Photogrammetry and Remote Sensing*. Elsevier B.V., 218(PA), pp. 693–706. doi: 10.1016/j.isprsjprs.2024.09.031.

Comaniciu, D. and Meer, P. (2002) 'Mean Shift: A Robust Approach Toward Feature Space Analysis', *IEEE Trans. Pattern Anal. Mach. Intell.* Washington, DC, USA: IEEE Computer Society, 24(5), pp. 603–619. doi: 10.1109/34.1000236.

Ding, Z., Han, X. and Niethammer, M. (2020) 'Votenet +: An Improved Deep Learning Label Fusion Method for Multi-Atlas Segmentation', *Proceedings - International Symposium on Biomedical Imaging*, 2020-April, pp. 363–367. doi: 10.1109/ISBI45749.2020.9098493.

Engelmann, F. *et al.* (2020) '3D-MPA: Multi proposal aggregation for 3D semantic instance segmentation', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9028–9037. doi: 10.1109/CVPR42600.2020.00905.

Faltermeier, F. L. *et al.* (2023) 'Improving Semantic Segmentation of Roof Segments Using Large-Scale Datasets Derived from 3D City Models and High-Resolution Aerial Imagery', *Remote Sensing*, 15(7). doi: 10.3390/rs15071931.

Fontana, M., Spratling, M. and Shi, M. (2024) 'When Multitask Learning Meets Partial Supervision: A Computer Vision Review', *Proceedings of the IEEE*, 112(6), pp. 516–543. doi: 10.1109/JPROC.2024.3435012.

Kong, G. and Fan, H. (2024) 'Automatic Generation of 3-D Roof Training Dataset for Building Roof Segmentation From ALS Point Clouds', *IEEE Transactions on Geoscience and Remote Sensing*, 62, pp. 1–12. Available at: https://api.semanticscholar.org/CorpusID:274451989.

Li, L. *et al.* (2022) 'Point2Roof: End-to-end 3D building roof modeling from airborne LiDAR point clouds', *ISPRS Journal of Photogrammetry and Remote Sensing*, 193, pp. 17–28. doi: https://doi.org/10.1016/j.isprsjprs.2022.08.027.

Li, L. *et al.* (2024) 'A boundary-aware point clustering approach in Euclidean and embedding spaces for roof plane segmentation', *ISPRS Journal of Photogrammetry and Remote Sensing*, 218, pp. 518–530. doi: https://doi.org/10.1016/j.isprsjprs.2024.09.030.

Reed, S. E. *et al.* (2015) 'Training deep neural networks on noisy labels with bootstrapping', *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, pp. 1–11.

Vaswani, A. *et al.* (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), pp. 5999–6009.

Wang, R., Huang, S. and Yang, H. (2023) 'Building3D: An Urban-Scale Dataset and Benchmarks for Learning Roof Structures from Point Clouds', *Proceedings of the IEEE International Conference on Computer Vision*, pp. 20019–20029. doi: 10.1109/ICCV51070.2023.01837.

Wang, X. *et al.* (2019) 'Associatively segmenting instances and semantics in point clouds', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, pp. 4091–4100. doi: 10.1109/CVPR.2019.00422.

Wichmann, A. *et al.* (2019) 'RoofN3D: A database for 3d building reconstruction with deep learning', *Photogrammetric Engineering and Remote Sensing*, 85(6), pp. 435–443. doi: 10.14358/PERS.85.6.435.

Wu, X. *et al.* (2025) 'Sonata: Self-Supervised Learning of Reliable Point Representations'. Available at: http://arxiv.org/abs/2503.16429.

Yang, H., Huang, S. and Wang, R. (2024) 'a Method for Roof Wireframe Reconstruction Based on Self-Supervised Pretraining', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10(2), pp. 239–246. doi: 10.5194/isprs-annals-X-2-2024-239-2024.

Yi, L. *et al.* (2019) 'GSPN: Generative shape proposal network for 3D instance segmentation in point cloud', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, pp. 3942–3951. doi: 10.1109/CVPR.2019.00407.

Zhang, C. and Fan, H. (2022) 'An Improved Multi-Task Pointwise Network for Segmentation of Building Roofs in Airborne Laser Scanning Point Clouds', *Photogrammetric Record*, 37(179), pp. 260–284. doi: 10.1111/phor.12420.

Zhang, H. *et al.* (2024) 'Point cloud self-supervised learning for machining feature recognition', *Journal of Manufacturing Systems*, 77, pp. 78–95. doi: https://doi.org/10.1016/j.jmsy.2024.08.029.

Zhang, Z. and Sabuncu, M. R. (2018) 'Generalized cross entropy loss for training deep neural networks with noisy labels', *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS), pp. 8778–8788.

Zhao, L. and Tao, W. (2020) 'JSNet: Joint instance and semantic segmentation of 3D point clouds', *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 12951–12958. doi: 10.1609/aaai.v34i07.6994.

Zhou, Z., Zhang, Y. and Foroosh, H. (2021) 'Panoptic-Polarnet: Proposal-free LiDAR Point Cloud Panoptic Segmentation', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 13189–13198. doi: 10.1109/CVPR46437.2021.01299.