RefineNet: a Confidence-aware Deep Online Learning Framework to Refine Real-world Point Cloud Semantic Segmentation

Sharath Chandra Madanu[1*], Shenglan Du[1*], Jantien Stoter[1], Daan van der Heide[1,2]

[1] 3D Geoinformation Group, Delft University of Technology, Julianalaan 134, Delft, NL
[2] Rijkswaterstaat, Derde Werelddreef 1, Delft, NL
madanusharathchandra@gmail.com, {shenglan.du, J.E.Stoter, D.H.vanderHeide-1}@tudelft.nl

Keywords: Point cloud semantic segmentation, Real-world data, Deep learning, Online learning

Abstract

Accurate interpretation and segmentation of 3D point clouds in real-world urban environments is a critical challenge in geospatial analysis, particularly due to the complexity of real-world scenes, inevitable data uncertainties, and potential annotation errors. This paper proposes a confidence-aware deep learning framework to refine the segmentation accuracy of real-world point cloud data. By incorporating multi-source information, such as aerial imagery, and embedding geospatial prior knowledge, this framework models data uncertainty through point-wise confidence scores. Besides, we design an iterative online learning strategy, allowing the network to improve both its predictions and the quality of training labels. Extensive experiments on large-scale airborne laser-scanned data demonstrate that our framework effectively enhances training data by reducing label noise and improving annotation quality, which leads to more robust, generalizable model performance. Our source code is publicly available at https://github.com/AutumnMoon00/RefineNet.

1. Introduction

Point cloud is a critical geospatial information source supporting a wide range of applications such as urban planning, ecological monitoring, and environmental analysis. Analyzing and interpreting such 3D data is increasingly driving innovation not only in academic research, but also in industry practice (Biljecki et al., 2015). For example, large-scale airborne laser-scanned data can capture high-resolution elevation and topographic information, which serve as a crucial input for the creation of Digital Surface Models (DSMs), local urban 3D reconstruction, and water resource management.

However, in practice, real-world point cloud datasets usually exhibit inevitable data quality issues, such as noise, outliers, and annotation inconsistencies. These errors can compromise the reliability of data interpretation, propagate through subsequent analyses, and, as a result, degrade the quality of final geospatial products.

In recent years, deep learning has significantly revolutionized the classification, interpretation, and analysis of 3D point cloud data (Qi et al., 2017a, Qi et al., 2017b, Thomas et al., 2019, Zhao et al., 2021). Despite their advancements, deep learning methods remain vulnerable to data quality issues: The model's performance heavily depends on the quality of the input data. High-quality inputs lead to accurate predictions, while noisy data can greatly downgrade the performance. Several data-efficient approaches have been introduced to address these issues. However, many of them rely on complex architectures, such as Generative Adversarial Networks (GANs) (Li et al., 2021), and require extensive training cycles. This underscores the need for a more straightforward yet practical approach to enhance model performance, particularly in correcting misclassification errors of real-world point clouds.

In this paper, we propose a confidence-aware learning framework, RefineNet, that integrates a confidence-based training label updating strategy to refine point cloud semantic segmentation of real-world datasets that may contain low-quality annotations. By measuring local semantic consistency and incorporating geospatial knowledge priors from auxiliary imagery, we assign a point-wise confidence score to assess the reliability of each training point label. Then we use the confidence scores to dynamically guide network training: High-confidence samples are prioritized for learning, and low-confidence samples are iteratively refined based on the network predictions. This strategy can effectively enhance both the quality of the training data and the segmentation performance of the model.

Overall, our proposed RefineNet aims to achieve accurate semantic segmentation of real-world scenes under imperfect data conditions, such as annotation errors and noise commonly appearing in real-world point clouds. The core innovation lies in the incorporation of confidence awareness and an iterative online learning strategy that jointly improves semantic segmentation and training label quality. Compared to existing studies, our proposed RefineNet provides a simple yet effective solution without the increased computational complexity or more sophisticated architectures.

We summarize our contributions in two folds:

- We introduce a confidence estimation method that assesses the reliability of existing semantic labels by leveraging local semantic consistency and incorporating geospatial priors from supplementary data sources;
- We propose an online learning framework with a label refinement mechanism, which dynamically selects high-confidence samples for training and iteratively updates labels of low-quality samples. In this way, both the performance and robustness of the deep learning model are effectively enhanced.

[*] Equal contribution. Authors are ordered alphabetically.

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-3-2026-179-2026

2. Related work

Recent advancements in deep learning have greatly revolutionized 3D classification and segmentation tasks. Existing deep learning approaches directly applied to 3D point clouds can be categorized into MLP-based (Qi et al., 2017a, Qi et al., 2017b, Qian et al., 2022), Convolution-based (Li et al., 2018, Thomas et al., 2019), and point transformers (Zhao et al., 2021, Lai et al., 2022, Wu et al., 2024). However, these methods are fully supervised and depend on high-quality annotated data. In this paper, we focus on data-efficient techniques that are designed to cope with data with limited or low-quality annotations.

2.1 Transfer learning

Transfer learning is a group of techniques that leverage knowledge acquired from a source dataset to enhance classification performance in a target domain. Tobin et al. (Tobin et al., 2017) proposed a domain randomization technique to transfer knowledge learned from 2D simulated images to real-world object detection tasks. It was later extended to 3D point cloud data (Wu et al., 2023), where synthetic point clouds are generated to improve segmentation performance on real-world scenes. Xiao et al. (Xiao et al., 2022) introduced SynLiDAR, a large-scale synthetic LiDAR dataset collected from diverse virtual environments. Then, a point-cloud translator was developed to mitigate the domain discrepancy between synthetic and real-world data. Biehler et al (Biehler et al., 2023) proposed PLURAL, a co-training framework that leverages contrastive instance alignment to bridge domain gaps and improve generalization.

2.2 Semi- and weakly supervised learning

This line of methods performs 3D classification and segmentation using fewer point labels to train deep neural networks. Wei et al. (Wei et al., 2020) introduced a multi-path region mining strategy combined with the class activation mapping technique (Zhou et al., 2016) to generate pseudo point-level labels, which are then used to train the segmentation network. Hu et al. (Hu et al., 2022) proposed the Semantic Query Network that implicitly augments sparse supervision signals by querying and summarizing features from neighboring points. Contrarily, Pan et al. (Pan et al., 2024) developed a label recommendation network, explicitly learning to provide recommendations for points to be labeled. A number of studies have also explored self-supervised pre-training techniques to fine-tune networks on the target 3D dataset with limited annotations (Hou et al., 2021, Sharma and Kaul, 2020, Zhang et al., 2021).

Furthermore, under the domain of semi-supervised learning, several studies have investigated active learning (Settles, 2012) and self-training (Amini et al., 2025) to effectively utilize both labeled and unlabeled data for training. Shi et al. (Shi et al., 2021) introduced an active learning method based on superpoint set selection to optimize the model under limited annotation budgets. Wang et al. (Wang and Yao, 2022) designed an online pseudo-labeling framework with a semantic consistency constraint, which provides additional supervisory signals to improve the robustness of point cloud semantic segmentation. Li et al. (Li et al., 2021) used unlabeled point samples and a pseudo-labeling strategy for training. However, it trains a separate GAN architecture to pick more reliable label predictions from unlabeled point clouds, which can be computationally costly.

Our approach shares conceptual similarities with the work of Li et al. (Li et al., 2021). We also incorporate pseudo-labeling and online learning techniques. However, we differ from this study in two key aspects: First, we explicitly leverage geospatial knowledge priors derived from auxiliary data sources, i.e., aerial imagery, to efficiently assess the reliability of training point labels. Second, our framework is unified, lightweight, and end-to-end trainable. We don’t require the integration of multiple networks and extensive training cycles.

3. Method

Our objective is to develop an online learning framework to refine the semantic segmentation of real-world point cloud datasets, which often contain noise, outliers, and annotation artifacts. By leveraging local semantic consistency and integrating geospatial priors, we estimate point-wise confidence scores to assess the reliability of each training point label. Then, we use these scores to guide the training, enabling the network to prioritize more trustworthy labels and simultaneously refining low-confidence annotations. Furthermore, due to the significant imbalance across semantic categories, we use a class-balanced loss function to supervise the network.

3.1 Confidence measurement

Confidence quantifies the reliability of a point’s label. Lower values suggest reduced trust in the correctness of the current classification, and higher values represent greater certainty. We measure point-wise confidence scores following two steps:

Computing local semantic consistency. We calculate the percentage of neighboring points that share the same semantic label as a given point, assuming that spatially proximate points tend to exhibit similar semantic characteristics. For each point i, we locate its spherical neighborhood within a radius of r. The local semantic consistency is computed as an initial confidence score c_i, following:

$$c_i = \begin{cases} \frac{N_j}{N_i} & \text{if } N_i \geq N_{min} \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where N_i is the total number of neighboring points, N_j is the number of neighboring points that share the same class label as the point i, and N_{min} is a user-defined threshold to ensure a minimum neighborhood density. We empirically set $N_{min} = 5$. In subsequent Section 4.4, we evaluate the impact of varying N_{min} and find that a value of 5 achieves an effective trade-off between neighborhood density and robustness to noise.

Incorporating geospatial priors. We further refine point-wise confidence scores for specific urban object categories by incorporating geospatial prior knowledge from auxiliary data sources, such as aerial images. A particular focus is given to buildings, a dominant class in real-world urban scenes. Specifically, in airborne laser-scanned data, building facades often have sparse point densities and tend to exhibit low confidence scores. Therefore, it is important to improve the confidence estimation for these regions.

We first extract building footprints from open-source DSMs and aerial ortho-imagery. We binarize the DSM map based on a height threshold ϵ_h, i.e., $\epsilon_h = 2$, then convert the binary map into polygonal shapes. Morphological operations, such as erosion and dilation, are used to smooth polygon boundaries,

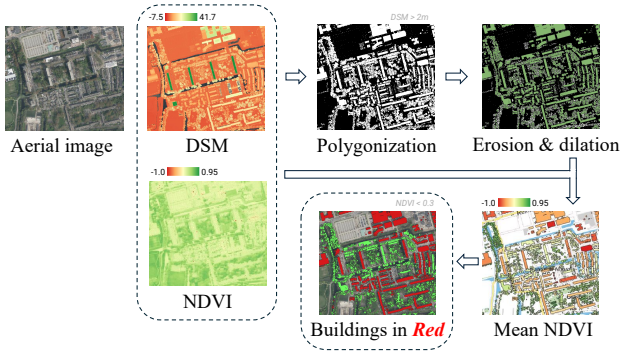

Figure 1. Buildings footprint extraction from open-source aerial imagery and DSMs. The mean NDVI is computed as the average NDVI value of all pixels within each extracted polygon.

remove small and noisy polygons, and retain topological consistency. The resultant polygons mostly contain urban buildings and vegetation. We further filter out vegetation polygons by computing the mean Normalized Difference Vegetation Index (NDVI). Polygons with mean NDVI values below a predefined threshold ϵ_v (i.e., $\epsilon_v = 0.3$) are considered building footprints. 3D points are then projected onto 2D building footprints. The confidence scores of the points corresponding to buildings are increased to 1.0 to emphasize facade regions. Figure 1 illustrates the overall building footprint extraction process.

3.2 Online learning

The input to our network is a real-world point cloud with pre-computed point-wise confidence scores, which may contain label annotation errors. For point-wise feature encoding, we adopt KP-Conv (Thomas et al., 2019) as the backbone network. While our approach is compatible with various backbone architectures such as MLP-based networks (Qi et al., 2017a, Qi et al., 2017b, Qian et al., 2022) and transformer networks (Zhao et al., 2021, Lai et al., 2022, Wu et al., 2024), we choose KP-Conv given its balance between computational efficiency and feature representational effectiveness. Moreover, in this paper, our focus is to investigate the impact of online learning strategies on refining real-world point cloud semantic segmentation, rather than to identify the optimal deep learning architecture.

We start the network training using only point samples with high confidence scores. The model is then used to make predictions for all points and give per-point class probability estimates. Predictions with high probabilities overwrite the original annotation labels, forming new labels referred to as pseudo-labels. These pseudo-labels are integrated into the training set for the network training in the next iteration. In this way, our proposed online learning strategy enables the network to iteratively correct mislabeled samples and simultaneously improve the point-wise confidence scores. Algorithm 1 presents the detailed steps of the online learning module. Overall, the online learning strategy yields two outcomes: (i) a refined and cleaned point cloud dataset with a significantly reduced number of low-confidence samples; (ii) a robustly trained network capable of accurate semantic segmentation, even in the presence of noisy or unreliable labels.

3.3 Class-balanced supervision

Class imbalance is a common issue in urban scene interpretation tasks, where a few dominant categories significantly outnumber the rest. For example, most points in urban environments belong to *building*, *road*, and *vegetation*, while the point

Algorithm 1: Online learning on point clouds

Input: input point cloud X and confidence map C
Output: trained model $f^{(e)}$ and updated dataset $X_{\overline{U}}$
Initialization: $e \leftarrow 0$, $X_{\overline{U}} \leftarrow \emptyset$, hyperparameters c_1, c_2, e_1, e_2, $c_1 < c_2$, $e_1 < e_2$
Segragate X into over- and under- confident point sets X_O and X_U,
$X_O \leftarrow \{\mathbf{x_i} \in X \mid c_i \geq c_1\}$
$X_U \leftarrow \{\mathbf{x_i} \in X \mid c_i < c_1\}$
repeat
  Train $f^{(e)}$ on $X_O \cup X_{\overline{U}}$
  if $e \geq e_1$ **then**
    $\Pi_e \leftarrow \{\mathbf{x_i} \in X_U,\ \Phi_y(\mathbf{x_i}, f^{(e)})\}$
      // $\Phi_y(\cdot,\cdot)$ denotes pseudo-labeling
    For Π_e, obtain the softmax probability map P
    $X_e \leftarrow \{\mathbf{x_i} \mid (\mathbf{x_i}, \widetilde{y_i}) \in \Pi_e \wedge p_i \geq c_2\}$
    $X_{\overline{U}} \leftarrow X_{\overline{U}} \cup X_e$
    $X_U \leftarrow X_U \setminus X_e$
    $e \leftarrow e + 1$
  end
until $e \geq e_2$ *or* $X_U = \emptyset$;

count of other classes is substantially lower. This imbalance limits the model's ability to learn discriminative features across all classes, since the network loss is overly exposed to only a few categories (Lin et al., 2017). To address the class imbalance issue, we employ a weighted cross-entropy loss function as follows:

$$L = -\sum_{i=1}^{N} w_k \log p_i^k, \tag{2}$$

where N is the total number of points, k is the Ground Truth (GT) semantic label of the i^{th} point, p_i^k is the predicted probability of the i^{th} point belonging to its GT category that can be obtained from the network softmax layer. w_k is the weight of the class k and is computed based on the inverse of its relative frequency in the dataset, i.e.,

$$w_k = \sqrt[3]{\frac{N_{max}}{N_k}}. \tag{3}$$

where N_k is the number of points of the class k, and N_{max} represents the point count of the class with the highest frequency.

4. Experiments

4.1 Dataset and implementation details

We use AHN (AHN, 2022), a nationwide airborne laser-scanned point cloud dataset collected in the Netherlands. This dataset is open-sourced and has been extensively used in both research and industrial applications. Specifically, we use the AHN4 version. AHN has six semantic classes: *ground*, *building*, *water*, *civil structure*, *high-tension*, and *other*, with vegetation currently categorized under the *other* class. The raw data contains xyz and intensity information. Due to its massive volume, we partition the dataset into smaller tiles, each measuring 0.25 × 0.3125 km, with a 10-meter overlap between adjacent tiles to mitigate edge effects. In our experiments, 52 tiles are used for training and 8 tiles for testing. Table 1 summarizes the distribution of point counts per class in the training and testing sets, where we have observed a significant class imbalance.

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-3-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

	Total	other	ground	building	water	high-tension	civil structure
Train	402.611M	102.866M	215.135M	66.535M	16.900M	0.048M	1.126M
Test	71.071M	17.606M	37.245M	5.804M	10.382M	0.001M	0.034M

Table 1. Distribution of point counts across all the classes in the training and testing sets.

Method	OA(%)	**mIoU(%)**	other	ground	building	water	high-tension	civil structure
Baseline	94.8	63.8	**86.6**	94.8	73.5	98.1	27.4	2.6
+ Online	**95.1**	**65.0**	85.4	94.8	**75.4**	**98.4**	**30.4**	**5.7**

Table 2. Segmentation results achieved on AHN dataset, using height and intensity as input features. OA (%), mIoU (%), and per-category IoU scores are reported. We average scores over three training runs to account for network performance variations.

Method	OA(%)	**mIoU(%)**	other	ground	building	water	high tension	civil structure
Baseline	**94.8**	**66.9**↑	**87.2**↑	**94.5**↓	**75.8**↑	**96.5**↓	**44.5**↑	**2.7**↑
+ Online	93.9↓	61.5↓	85.3↓	94.2↓	66.0↓	95.4↓	25.6↓	2.4↓

Table 3. Segmentation results achieved using height, intensity, and supplementary color information as input features. OA (%), mIoU (%), and per-category IoU scores are reported. We average scores over three training runs to account for network performance variations. We use ↑ to indicate performance improvements compared to the same model trained using only height and intensity features (Table 2), and ↓ for performance degradations.

We adopt KP-Conv (Thomas et al., 2019) as the backbone, which uses a voxel subsampling strategy to reduce the number of input points. We set the voxelization grid size as 20cm following common practice (Hu et al., 2021), and the number of kernels as 15. The network is trained for 300 epochs with a batch size of 6 and an initial learning rate of 0.01. For hyperparameters in Algorithm 1, we set $c_1 = 0.9, c_2 = 0.99, e_1 = 150, e_2 = 300$. Confidence thresholds (i.e., c) are set to be high to ensure that only the most reliable point samples can be used for model training. Besides, we set the warm-up period (i.e., e_1) to 150 epochs to allow the network to reach a partially converged state before introducing the online learning mechanism.

4.2 Quantitative results

To validate the effectiveness of our proposed confidence-aware framework for refining point cloud semantic segmentation, we compare its performance against the baseline network (Thomas et al., 2019). We train both models using a class-balanced loss function. Besides, we apply the same hyperparameters and training configurations to both networks to ensure a fair comparison. The segmentation performance is evaluated using standard metrics, including Overall Accuracy (OA), mean Intersection over Union (mIoU), and per-category IoU scores. Due to the high variations in network training, we repeatedly run each model three times and report the average scores.

Table 2 presents the segmentation performance of the baseline and its counterpart network, which is augmented with the online learning mechanism. Both networks use height and intensity as input features. The enhanced model demonstrates stronger performance, achieving gains of 0.3% in OA and 1.2% in mIoU. Among the six categories, five show superior or comparable performance, revealing that our method effectively facilitates network feature learning by prioritizing high-confidence data samples for training. Specifically, we achieve notable improvements in two minority classes, *high-tension* and *civil structure*, demonstrating that our method is more capable of handling underrepresented classes. However, we observe a performance decline of 1.2% mIoU for the *other* class. This is likely attributed to the fact that the *other* class is dominated by vegetation points. It tends to receive lower confidence scores due to the sparse and irregular spatial distributions of trees. Therefore, fewer point samples from the *other* category are incorporated during training, and the segmentation performance decreases.

The raw AHN data only contains xyz and intensity values. However, it has been post-processed and enriched with rgb colors by aligning the original point clouds with aerial imagery (GeoTiles, 2023). To further analyze the impact of this supplementary color information on the network's performance, we incorporate colors as additional input features. Table 3 reports the performance results for both the baseline and the network enhanced with our proposed online learning mechanism, trained using the combined height, intensity, and color features.

When incorporating supplementary color features, our online-learning enhanced network performs worse than the baseline. While adding colors leads to an overall improvement in the baseline network's performance, it results in a significant performance degradation for the online learning network among all categories. This observation illustrates one of the major limitations of our method: It strongly relies on the quality of input features. The colors are derived from aerial imagery, which may contain artifacts due to potential data occlusions or misalignments. The online learning approach is particularly sensitive to such artifacts or errors, as it selectively utilizes only a subset of high-confidence points for training rather than leveraging the entire input data. Thus, these wrong visual cues from supplementary colors can significantly influence the network learning and degrade the overall performance.

4.3 Qualitative results

Figure 2 presents our qualitative results achieved on the AHN dataset, with the model trained using the height and intensity features. Our method consistently outperforms the baseline when trained with height and intensity features by reducing segmentation errors. It can also detect urban objects, such as grasslands and bridges, with improved geometric completeness. This performance gain is attributed to the fact that our method selects only the high-confidence samples to participate in training, which enhances both the robustness and discriminative capacity of the learned feature representations. Note that real-world datasets often contain annotation inaccuracies. For example, in the third row of Figure 2, the dataset annotations incorrectly label all building points as *other*. While standard evaluation against this ground truth would consider these predictions incorrect, our method successfully classifies the *building* points based on their true geometries, showing its potential to correct misclassifications in practical, real-world applications.

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-3-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

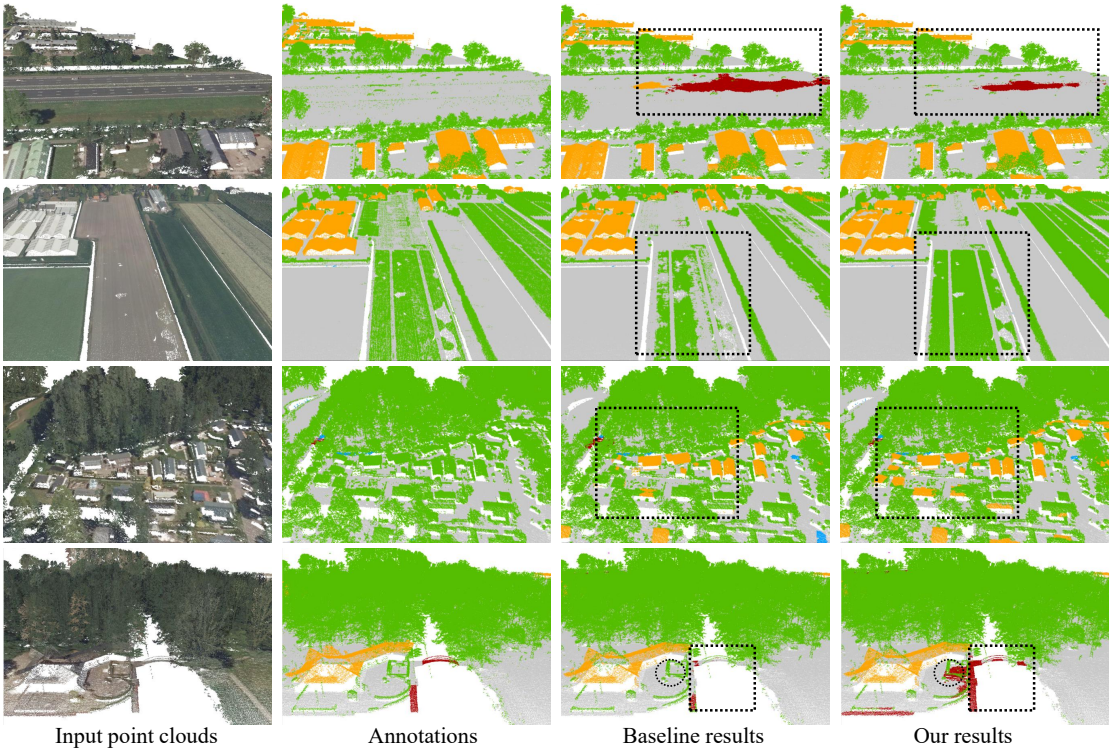

Figure 2. Qualitative results achieved on the AHN dataset, using height and intensity as input features. We use the following color schemes to render the scenes: for *building*, for *ground*, for *high-tension*, for *water*, for *civil structure*, and for *other*. Since AHN is a real-world dataset, its annotations may contain noise and errors. Our improvements are highlighted with black-dotted boxes. Minor segmentation noise in some local regions is marked by black-dotted circles.

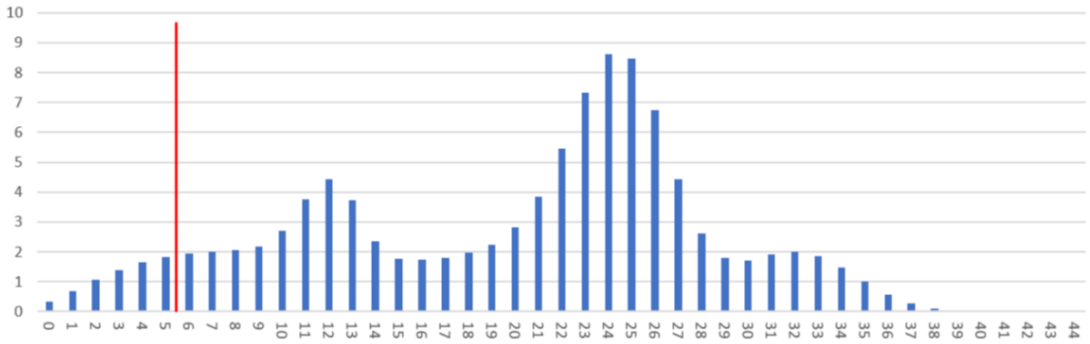

Figure 3. Histogram of the distribution of neighborhood point counts. The y-axis represents the point number expressed in millions.

4.4 Point density analysis

When computing local semantic consistency (Equation 1), we set $N_{min} = 5$ to ensure a minimum neighborhood density. In this section, we provide point cloud density analysis to empirically validate our hyperparameter choice.

Figure 3 shows the histogram of the point count distribution within a spherical neighborhood of 0.5m, obtained from the training set. It is observed that the neighboring point counts vary significantly. While most neighborhoods contain between 15 and 30 points, a great proportion still falls within the point range between 1 and 10, meaning that relatively sparse local regions are common in the dataset. However, neighborhoods with fewer than 5 points are rare in the dataset, which are also more susceptible to noise. Setting N_{min} too high may exclude many valid points and prevent us from gaining informative local context. On the contrary, setting N_{min} too low leads to overestimation of confidence scores for sparse and potentially noisy regions. Therefore, we set $N_{min} = 5$ for a balanced choice. It can effectively filter out extremely sparse, unreliable neighborhoods, while still preserving sufficient coverage to capture meaningful local contexts.

4.5 Label refinement on training data

As revealed in Section 4.2, given good input features, our proposed online learning framework can effectively improve the robustness and generalizability of the network. Besides, a natural byproduct of our approach is a cleaned training dataset with annotations progressively refined by pseudo-labels. This improvement in the quality of the training data is attributed to the proposed online learning mechanism (Algorithm 1). When the network makes high-confidence predictions on the set X_U, these new predictions serve as pseudo-labels and gradually replace the original labels. Through the iterative process of updating both the network parameters and the training labels, our method can mitigate label noise, reduce the influence of outliers, and correct annotation errors in the training set.

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-3-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

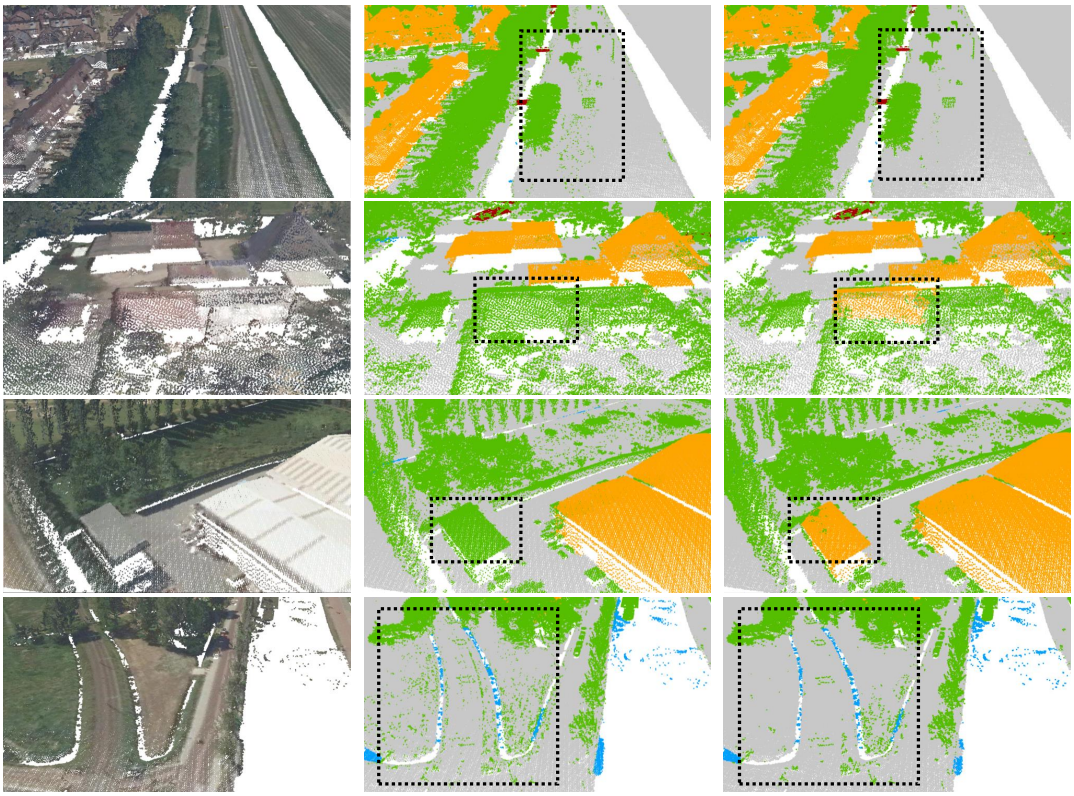

Figure 4. Label updates and refinements of the training point cloud data using the proposed online learning strategy. We use the same color schemes as in Figure 2 to render the scenes.

Figure 4 shows several representative examples of label refinements in the training set. In the raw data, *ground* points are often annotated inconsistently, exhibiting label noise. Also, many *building* points are erroneously labeled as *other*. Our online learning framework can effectively mitigate the noise in the *ground* points and correct misannotations in the *building* category, improving the overall quality of the dataset.

4.6 Limitations

Our primary research goal is to address 3D semantic segmentation of real-world point clouds under imperfect training data conditions, meaning that fully supervised semantic segmentation on ideally annotated datasets is out of our research scope. Based on this goal, we propose a confidence-aware online learning framework, RefineNet, that effectively enhances the segmentation performance in real-world scenes and shows potential for refining raw data labels with inevitable annotation errors, noise, and inconsistencies.

However, RefineNet still suffers from several limitations. It is highly sensitive to the quality of input features, as only a subset of the data is allowed to participate in training. When trained with reliable input features, our method outperforms the baseline. Nevertheless, its performance decreases when supplementary colors with potentially wrong visual cues are used in training. Another limitation lies in the estimation of point-wise confidence scores. We only leverage the geospatial priors with a specific focus on building-class points. Extending this strategy to incorporate prior knowledge for other urban objects can further enhance the robustness and overall performance. Besides, real-world datasets have inherent annotation errors. Our experimental evaluation may lack accuracy, as the available GT labels cannot be assumed to represent definitive ground truth.

5. Conclusions

We have proposed a confidence-aware online learning framework to explicitly address the challenges of real-world point cloud interpretation, specifically considering label noise, outliers, and annotation errors. Our framework integrates local semantic consistency measures and geospatial priors to assess the confidence level of existing annotations. These confidence scores are used to guide the online learning, where the network prioritizes high-confidence samples for training and iteratively refines the annotations of low-confidence points. Our approach yields a robustly trained segmentation model and a cleaned point cloud dataset with enhanced annotation quality. Experimental results on real-world datasets have validated the applicability of our approach. However, its performance remains sensitive to the quality of the input features. Our framework is directly applicable to real-world point cloud segmentation tasks. It can be used to denoise data, correct prominent annotation errors, and enhance overall data quality. It can also be integrated into semi-automated annotation pipelines to facilitate the annotation, updating, and iterative refinement of point cloud datasets.

In future work, we will focus on enhancing confidence estimation by integrating more enriched priors. We also consider leveraging the recent SAM technique (Kirillov et al., 2023) to extract high-fidelity object footprints directly from imagery to improve geospatial priors. Another potential direction is to incorporate synthetic point clouds to better handle the class imbalance issue. For example, by augmenting the dataset with samples of minor urban objects such as high-tension lines and civil structures, we can train the network on a balanced distribution to learn more robust and generalizable features.

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-3-2026-179-2026

References

AHN, 2022. Actueel hoogtebestand nederland. https://www.ahn.nl/. Accessed: 2025-09-01.

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., Maximov, Y., 2025. Self-training: A survey. *Neurocomputing*, 616, 128904.

Biehler, M., Sun, Y., Kode, S., Li, J., Shi, J., 2023. PLURAL: 3D point cloud transfer learning via contrastive learning with augmentations. *IEEE Transactions on Automation Science and Engineering*, 21(4), 7550–7561.

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

GeoTiles, 2023. Geotiles. https://geotiles.citg.tudelft.nl/. Accessed: 2025-09-01.

Hou, J., Graham, B., Nießner, M., Xie, S., 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15587–15597.

Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A., 2022. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. *European Conference on Computer Vision*, Springer, 600–619.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2021. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4977–4987.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *IEEE/CVF International Conference on Computer Vision*, IEEE, 4015–4026.

Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J., 2022. Stratified transformer for 3d point cloud segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8500–8509.

Li, H., Sun, Z., Wu, Y., Song, Y., 2021. Semi-supervised point cloud segmentation using self-training with label confidence prediction. *Neurocomputing*, 437, 227–237.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. *Advances in Neural Information Processing Systems*, 31. https://arxiv.org/abs/1801.07791.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *IEEE International Conference on Computer Vision*, 2980–2988.

Pan, Z., Zhang, N., Gao, W., Liu, S., Li, G., 2024. Less is more: Label recommendation for weakly supervised point cloud semantic segmentation. *AAAI Conference on Artificial Intelligence*, 38number 5, 4397–4405.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30. https://arxiv.org/abs/1706.02413.

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B., 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35, 23192–23204. https://arxiv.org/abs/2206.04670.

Settles, B., 2012. *Active learning*. Springer International Publishing.

Sharma, C., Kaul, M., 2020. Self-supervised few-shot learning on point clouds. *Advances in Neural Information Processing Systems*, 33, 7212–7221. https://arxiv.org/abs/2009.14168.

Shi, X., Xu, X., Chen, K., Cai, L., Foo, C. S., Jia, K., 2021. Label-efficient point cloud semantic segmentation: An active learning approach. *arXiv preprint*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *IEEE/CVF International Conference on Computer Vision*, 6411–6420.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., 2017. Domain randomization for transferring deep neural networks from simulation to the real world. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 23–30.

Wang, P., Yao, W., 2022. A new weakly supervised approach for ALS point cloud semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 237–254.

Wei, J., Lin, G., Yap, K.-H., Hung, T.-Y., Xie, L., 2020. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4384–4393.

Wu, C., Bi, X., Pfrommer, J., Cebulla, A., Mangold, S., Beyerer, J., 2023. Sim2real transfer learning for point cloud segmentation: An industrial application case on autonomous disassembly. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 4531–4540.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point transformer v3: Simpler faster stronger. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4851.

Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S., 2022. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. *AAAI Conference on Artificial Intelligence*, 36number 3, 2795–2803.

Zhang, Z., Girdhar, R., Joulin, A., Misra, I., 2021. Self-supervised pretraining of 3d features on any point-cloud. *IEEE/CVF International Conference on Computer Vision*, 10252–10263.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.