

Article

Attention-Enhanced MinkUNet for Label-Efficient Segmentation of Transmission Line LiDAR Point Clouds

Yijiang Wu ¹, Jianfeng Huang ²  and Yuxuan Lei ^{2,*}

¹ Zhongshan Power Supply Bureau, Guangdong Power Grid Co., Ltd., Zhongshan 528405, China; sjbwj888@163.com

² Beidou Research Institute, South China Normal University, Foshan 528225, China; huangjf9@m.scnu.edu.cn

* Correspondence: 2023025211@m.scnu.edu.cn

Abstract

Routine inspections of transmission lines are essential for maintaining the reliability of the power grid. Airborne LiDAR technology provides detailed 3D corridor data for automated hazard detection, such as vegetation encroachment and structural anomalies. However, manually analyzing large point clouds is inefficient, and current segmentation methods struggle with scene complexity, scale variation, and the high cost of annotation. In this study, we present a label-efficient segmentation method built on MinkUNet, a sparse voxel convolutional network enhanced with self-attention modules in its encoder–decoder for better spatial reasoning over corridor objects (e.g., trees, buildings, towers). To further handle structural diversity and class imbalance, we adopt task-specific data augmentations and focal loss. A multi-stage pseudo-labeling strategy is then employed to enable effective cross-scene generalization with minimal labeled data. We validate our method on three real-world transmission line datasets. On the Foshan dataset, it achieves a mean Intersection over Union (mIoU) of 0.740 with an inference time of 1.31 s. Cross-scene tests at two other locations, Shumuyuan and Langwang Village, yield mIoUs of 0.762 and 0.757, respectively. These results confirm robust performance even with limited annotations. Overall, our findings demonstrate the practicality of our approach for routine power line inspections, enabling reliable hazard detection with minimal annotation effort.

Keywords: transmission lines; LiDAR point clouds; semantic segmentation; MinkUNet; power-grid inspections; label-efficient; unmanned aerial vehicle

1. Introduction

Transmission lines are essential components of power systems, responsible for delivering electricity over long distances. Their safe operation directly impacts urban production, daily life, industrial activities, and regional energy management stability [1]. Over time, factors such as tree growth, building expansion, terrain obstructions, and extreme weather can introduce potential risks to these transmission corridors. Among these, tree encroachment is one of the most prevalent and critical engineering challenges faced during the inspection, operation, and maintenance of transmission lines. Traditionally, transmission line inspections have relied on manual field observations and periodic record-keeping. However, this method is heavily influenced by terrain conditions, inspection viewpoints, operational safety, and labor costs, making it challenging to meet the growing needs for large-scale, high-frequency, and detailed inspections [2].



Academic Editor: Mariusz Jasiński

Received: 12 June 2026

Revised: 1 July 2026

Accepted: 1 July 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In recent years, the integration of unmanned aerial vehicles (UAVs) and LiDAR sensors has offered a promising solution for inspecting transmission lines [3]. UAV LiDAR technology can quickly gather three-dimensional spatial information about transmission lines and their surrounding environments, even in complex terrain and vegetation. This technology produces point cloud data that includes various attributes, such as spatial coordinates, echo intensity, and color information. Unlike two-dimensional images, point clouds can provide detailed descriptions of power line sag, tower structures, canopy height, and three-dimensional distances between objects. This feature allows for more accurate calculations of tree-encroachment distances, risk assessments, and defect localization [4]. Nevertheless, point-cloud data are unordered, unstructured, density-imbalanced, and large in scale. One of the primary obstacles in intelligent transmission line inspection is the automatic categorization of towers, power lines, trees, buildings, background elements, and other features from extensive point cloud datasets.

Point-cloud semantic segmentation assigns a semantic class label to each point and is a crucial task for understanding three-dimensional scenes [5]. Current methods for point-cloud semantic segmentation can be broadly classified into three categories: point-based, multi-view/projection-based, and voxel-based methods. Point-based methods, such as PointNet [5], PointNet++ [6], and KPConv [7], utilize raw point sets or local neighborhood point sets as input. These methods effectively preserve the original geometric structures and spatial distributions of point clouds. However, they often require processes such as neighborhood search, sampling, grouping, and point-level feature aggregation, which can lead to high computational overhead and reduced efficiency in large-scale transmission-line point-cloud scenarios. Multi-view or projection-based methods, on the other hand, convert three-dimensional point clouds into two-dimensional views, spherical projection maps, or range images, enabling feature extraction and semantic prediction with established two-dimensional convolutional networks [8,9]. While these methods tend to be more computationally efficient and compatible with two-dimensional networks, the projection process can disrupt three-dimensional spatial relationships, potentially leading to information loss, especially in small objects or occluded regions.

In contrast, voxel-based methods convert unordered point clouds into regular voxels or sparse voxel grids and utilize three-dimensional or sparse convolutions to extract features. Sparse convolutional networks are particularly beneficial because they eliminate redundant computations on many empty voxels. This significantly enhances the efficiency of processing large-scale point clouds while still effectively representing three-dimensional structures. For these reasons, the sparse-convolution-based MinkUNet [10] framework is well-suited for the semantic segmentation of transmission-line point clouds.

MinkUNet can efficiently process large-scale point clouds; however, the object categories in transmission-line scenes exhibit significant differences in scale and spatial distribution. For instance, power lines are typically slender, continuous linear structures or may have local discontinuities. Towers are made up of complex metal components and prominent vertical structures. Tree point clouds have irregular shapes and are often closely intertwined with the background terrain in terms of height. In some scenes, the building category may contain only a few samples, making it easy to confuse buildings with the background. While voxelization can enhance computational efficiency, it may compress fine objects and local structural information, thereby affecting the segmentation accuracy of power lines, tower details, and building boundaries. Consequently, relying solely on a conventional voxel-convolutional architecture is insufficient to effectively capture long-range spatial dependencies among different objects in transmission-line scenes.

Moreover, existing methods for semantic segmentation of transmission-line point clouds mainly rely on a fully supervised learning approach, which requires accurate

labels for each point. Given the large scale of these point clouds and the complex boundaries between categories that demand expert annotation, point-wise labeling is both time-consuming and labor-intensive. Additionally, significant distribution shifts across different regions, voltage levels, and geomorphic conditions further complicate this process. When transferring a model to a new scene, it often necessitates the reconstruction of a fully annotated dataset and retraining, which severely limits the large-scale application of these algorithms in power-grid inspections. Therefore, reducing annotation costs while maintaining segmentation performance and improving the adaptability of models to new scenes is an urgent problem in transmission-line point-cloud segmentation.

To address the above challenges, this paper proposes a semantic segmentation method for transmission line LiDAR point clouds, based on an improved MinkUNet architecture and a label-efficient learning strategy. By balancing segmentation accuracy, inference efficiency, and annotation cost, our method enables practical applications such as early tree-encroachment warning, 3D corridor modeling, and automated component damage detection for intelligent power-line inspection and digital O&M. The main contributions of this work are summarized as follows:

- An improved MinkUNet framework utilizing sparse voxel convolutions has been developed to address the large spatial extent, significant scale variations, and elongated structural confounders found in transmission line point clouds. This framework enhances point-level recognition accuracy for critical objects while maintaining high inference efficiency.
- To improve segmentation for small-sample, fine-grained categories—such as power lines, towers, and buildings—self-attention enhancement modules have been integrated into important layers of the MinkUNet encoder and decoder. Additionally, focal loss and geometric data augmentation have been employed to strengthen the representation of hard-to-classify and minority-class samples.
- A label-efficient scene adaptation strategy has been proposed to reduce the high costs of per-point annotation and enhance cross-scene generalization. This strategy combines transfer learning, data augmentation, and a multi-stage threshold-based pseudo-labeling mechanism to leverage unlabeled point clouds from target scenes effectively.
- Experiments conducted on real-world transmission line point cloud datasets from Foshan, Shumuyuan (Zhongshan), and Langwang Village (Zhongshan) demonstrate that the proposed method achieves competitive performance, especially in terms of overall mean Intersection over Union (mIoU), power line segmentation accuracy, and inference efficiency.

2. Related Work

2.1. Point-Cloud Semantic Segmentation Methods

Point-cloud semantic segmentation is an important research topic in three-dimensional vision. Early methods commonly relied on handcrafted geometric features, local neighborhood statistics, or traditional machine-learning classifiers for point-wise classification. With the development of deep learning, PointNet was the first to directly process unordered point sets using shared multilayer perceptrons and a symmetric aggregation function, establishing a basic paradigm for end-to-end point-cloud learning [5]. PointNet++ further introduced hierarchical sampling, local neighborhood grouping, and multi-scale feature aggregation, enhancing the ability to model local geometric structures [6]. However, when processing wide-area and high-density transmission-line point clouds, point-based methods typically face the challenges of high neighborhood-search cost, low batch-processing efficiency, and insufficient large-scale contextual modeling.

Multi-view methods project three-dimensional point clouds onto multiple two-dimensional planes or spherical images, utilizing established two-dimensional convolutional networks for semantic segmentation. While these methods capitalize on the strong representation capabilities of two-dimensional vision models, the projection process can lead to issues such as occlusion, overlap, and loss of geometric information. This method is particularly inadequate for capturing the three-dimensional spatial relationships of elongated structures like power lines and towers. On the other hand, voxel-based methods convert point clouds into three-dimensional grids and employ three-dimensional convolutions for feature learning. Traditional dense voxel convolutions incur significant computational costs, especially in high-resolution scenes. In contrast, sparse convolutions enhance processing efficiency for large-scale point clouds by focusing convolution operations only on non-empty voxels. The Minkowski Engine and its notable network, MinkUNet, successfully strike a balance between accuracy and efficiency in both indoor and outdoor three-dimensional scene segmentation [10].

2.2. Intelligent Analysis of Transmission-Line Point Clouds

Intelligent analysis of LiDAR point clouds forms a critical technical foundation for automated inspection and maintenance of transmission line corridors [11,12]. This process typically encompasses a sequence of key tasks, including 3D scene reconstruction, detection and extraction of key infrastructure components like towers and power lines, classification of ground objects, measurement of safety clearances, and comprehensive risk assessment. By transitioning from manual inspection to automated, data-driven analysis, this approach significantly enhances the efficiency, accuracy, and safety of grid corridor management.

Transmission line scenes exhibit distinct engineering characteristics, characterized by complex spatial and topological relationships among various objects. For instance, power lines extend along the line direction while sagging, and towers are connected to these power lines, featuring intricate structures. Additionally, trees and buildings surround the corridor, and the terrain may have heights comparable to that of the surrounding vegetation. Traditional geometry-based methods typically rely on features such as height, linearity, curvature, connectivity, and spatial distance for object extraction. These methods have been successfully employed for tasks like power line extraction, tower localization, and detecting risks from tree encroachment [13,14]. However, they often depend on manually designed features and candidate-region screening, as well as scene-prior constraints. Consequently, their robustness can be compromised in complex terrain, under occlusion, with varying point density, and in the presence of noise [12]. In contrast, deep learning methods can learn more robust contextual features in a data-driven manner. However, their effectiveness is heavily reliant on the scale and quality of the training annotations, and they still face challenges with cross-scene generalization [1,12].

In the semantic segmentation of transmission-line point clouds, the model must accurately identify the background and trees, which typically comprise large portions of the data, while also effectively distinguishing smaller yet critically important elements such as power lines, towers, and buildings. Existing research indicates that ultra-high-voltage transmission-line point clouds exhibit large data volumes, sparse point distributions, and imbalanced category distributions. Specifically, the number of points associated with power lines, ground wires, towers, and buildings is often significantly lower than that for ground or vegetation classes [1]. Power line point clouds tend to be slender and have relatively small sample sizes, while buildings may appear in very limited numbers in certain datasets. This situation makes the model vulnerable to class imbalance. Additionally, transmission-line point cloud scenes cover extensive areas; therefore, using computationally intensive point-level networks may not satisfy the efficiency requirements for practical inspections.

Thus, it is essential to consider accuracy, efficiency, and engineering deployability simultaneously when designing semantic segmentation models for point clouds.

2.3. Attention Mechanisms and Sparse Feature Enhancement

Attention mechanisms adaptively compute correlation weights among features, thereby enhancing the representation of key regions and important objects in visual tasks [15]. Self-attention was originally widely used in sequence modeling. Its core idea is to compute dynamic weights based on the similarity among queries, keys, and values, so as to capture long-range dependencies [16]. In three-dimensional point-cloud tasks, attention mechanisms can be used for local neighborhood feature aggregation, global contextual modeling, and category-discriminative enhancement [17]. For transmission-line scenes, there are explicit spatial correlations between power lines and towers, and the relative positional relationship between trees and power lines is closely related to tree-encroachment risk. Therefore, introducing self-attention modules into sparse voxel features of MinkUNet can compensate for the limited receptive field of local convolutions and improve the modeling of discontinuous elongated objects.

2.4. Label-Efficient Learning and Scene Generalization

Label-efficient learning focuses on minimizing models' reliance on extensive, accurately annotated datasets. Common approaches include weakly supervised [18], semi-supervised [19], and active learning [20]. Weakly supervised learning enables models to be trained with incomplete, imprecise, or inaccurate annotations [21]. Semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data, leveraging strategies such as pseudo-labeling, self-training, and consistency regularization to improve training [22]. Recent advances in point cloud segmentation have adapted existing paradigms to 3D scenes. PointMatch, proposed by Wu et al. [23], employs a consistency-training framework that leverages both inherent data information and sparse labels, achieving strong robustness even under label scarcity. Yao et al. [24] designed Dual-Teacher, a dual-teacher-guided contrastive learning approach that filters low-quality pseudo-labels through cross-validation and refines them with a correction module. Yin et al. [25] introduced WSSS-ST, which integrates a self-training mechanism to harness large-scale unlabeled point clouds for industrial applications. These methods demonstrate the effectiveness of consistency regularization and pseudo-label refinement for label-efficient point cloud segmentation.

In this work, we implement a straightforward yet effective multi-stage threshold-based pseudo-labeling strategy specifically designed for transmission line point clouds. We recognize that the advanced methods mentioned earlier present valuable alternatives for future improvements. Our approach is driven by the unique characteristics of transmission line environments, where data distributions can vary significantly across different regions, terrains, and voltage levels. Applying a model trained on one scene directly to another can often lead to reduced performance. To tackle this issue, transfer learning can utilize general structural knowledge from the source scene, while pseudo-labeling methods can further leverage the distributional information found in unlabeled data from the target scene. Consequently, combining transfer learning with data augmentation and iterative pseudo-label optimization offers an effective and practical strategy for minimizing cross-scene adaptation challenges in the inspection of transmission line point clouds.

In summary, current research on the semantic segmentation of transmission-line point clouds encounters three main challenges. First, while point-level networks can preserve original geometric details, they tend to incur high computational overhead in large-scale transmission-line scenarios, making them unsuitable for rapid engineering inspections.

Second, although voxelized sparse convolutional networks offer efficiency benefits, voxel quantization and local convolution may compromise the representation of fine objects like power lines, tower components, and building boundaries. Third, most existing methods depend on having a sufficiently labeled dataset and often require reannotation and retraining across different regions, geomorphic types, and transmission lines, limiting their large-scale applicability. Consequently, it is crucial to develop a semantic segmentation method for transmission-line point clouds that integrates efficient inference, fine-grained structural representation, and low-annotation cross-scene adaptation.

3. Method

3.1. Problem Definition and Overall Framework

Given a set of transmission-line LiDAR point clouds $\mathcal{P} = \{p_i\}_{i=1}^N$, where the i -th point p_i contains three-dimensional coordinates $\mathbf{x}_i = (x_i, y_i, z_i)$ and optional attribute features $\mathbf{a}_i$, such as RGB color, echo intensity, or height-normalization information, the objective of point-cloud semantic segmentation is to learn a mapping function f_θ that predicts a class label for each point:

$$\hat{y}_i = f_\theta(\mathbf{x}_i, \mathbf{a}_i), \quad \hat{y}_i \in \{1, 2, \dots, C\}, \quad (1)$$

where C denotes the number of classes. In the experiments of this paper, the categories mainly include towers, power lines, background, trees, and buildings.

To address this semantic segmentation task in the context of transmission-line inspection, we design a comprehensive pipeline consisting of three main stages (Figure 1): data acquisition and preprocessing, an enhanced MinkUNet for accurate scene parsing, and a label-efficient strategy for cross-scene generalization.

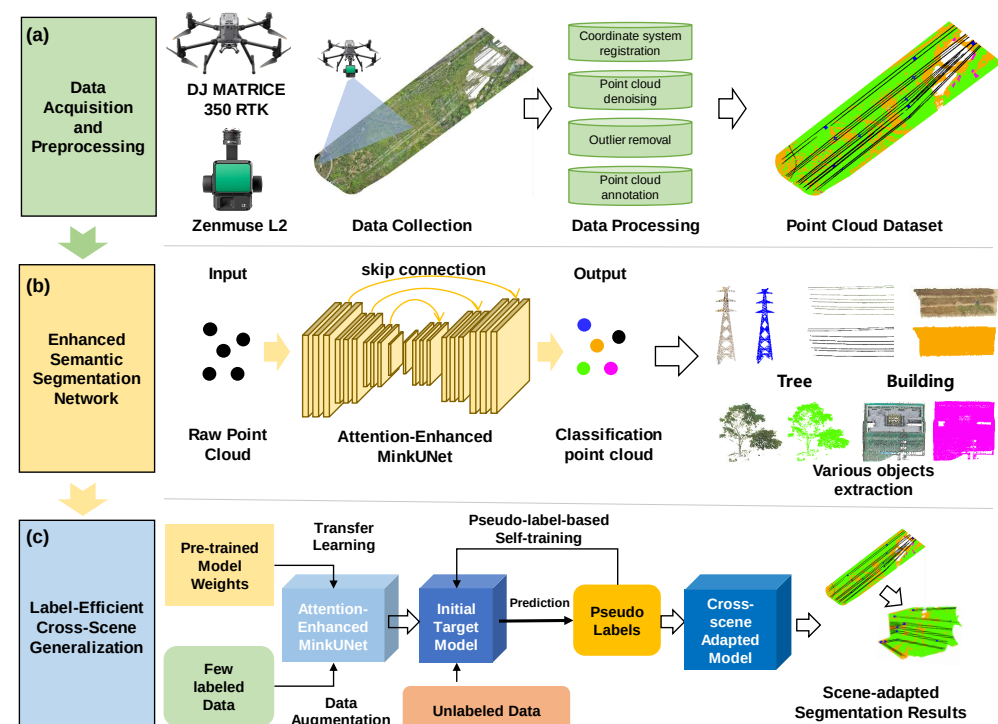


Figure 1. Overall framework of the proposed method. (a) Data Acquisition and Preprocessing; (b) Enhanced Semantic Segmentation Network (Attention-Enhanced MinkUNet); (c) Label-Efficient Cross-Scene Generalization via Pseudo-label Self-training.

- **Data Acquisition and Preprocessing.** The process begins with the collection of high-precision aerial LiDAR data using a DJI MATRICE 350 RTK drone (DJI, Shenzhen,

China), which is equipped with a Zenmuse L2 laser scanner (DJI, Shenzhen, China). The raw point clouds are then processed to create a unified dataset. This includes registering the coordinate system for georeferencing, removing noise to eliminate outliers, and manually annotating the data to generate ground-truth labels for supervised training. Detailed descriptions of the collected transmission-line data, specifications of the UAV-LiDAR platform, and the preprocessing procedures can be found in Section 4.1.

- **Enhanced Semantic Segmentation Network.** For fine-grained point cloud segmentation, we introduce an improved MinkUNet architecture. The raw point cloud is voxelized into a sparse tensor (voxel size 0.05 m), which is then fed into the network. Our key innovation lies in embedding self-attention modules within both the encoder and decoder stages, effectively capturing long-range topological dependencies between spatially dispersed objects (such as transmission towers, power lines, and background vegetation). The full architecture, including the attention mechanism and loss functions, is detailed in Sections 3.2–3.4.
- **Label-Efficient Cross-Scene Generalization.** To reduce reliance on extensive manual labeling in new environments, we adopt a semi-supervised adaptation strategy. The source-trained model weights are first transferred to the target scene as initialization. We then leverage a pseudo-labeling mechanism: the model predicts reliable labels for the abundant unlabeled target data, which are subsequently reused to retrain the model. This iterative refinement allows the network to efficiently adapt to the specific characteristics of the new scene, even with only a limited number of annotated samples. The complete adaptation procedure is described in Section 3.5.

3.2. MinkUNet Backbone Based on Sparse Voxel Convolution

MinkUNet adopts an encoder–decoder architecture and uses skip connections to integrate shallow spatial details with deep semantic information [10]. Different from dense three-dimensional convolutions, sparse convolutions are performed only at non-empty voxel locations, thereby significantly reducing the computational and memory costs of large-scale point-cloud processing [10,26]. Given a voxel size v , the original point coordinate $\mathbf{x}_i$ is quantized into a voxel coordinate:

$$\mathbf{q}_i = \left\lfloor \frac{\mathbf{x}_i}{v} \right\rfloor. \quad (2)$$

Point features within the same voxel are transformed into voxel features by averaging, random sampling, or feature aggregation. The sparse tensor can be represented as

$$\mathcal{S} = \{(\mathbf{q}_j, \mathbf{f}_j)\}_{j=1}^M, \quad (3)$$

where M is the number of non-empty voxels and $\mathbf{f}_j$ denotes the voxel feature.

The encoder consists of multiple sparse convolutional layers, batch normalization, ReLU activation, and residual blocks. As network depth increases, the feature-space resolution is gradually reduced and the semantic representation is progressively strengthened. The decoder gradually restores spatial resolution through sparse transposed convolutions and concatenates the decoded features with the corresponding encoder features at the same scale. Skip connections can transfer shallow local geometric structures, object boundaries, and spatial-position information to the decoding stage, alleviating the detail loss caused by voxelization and downsampling. The overall architecture of the proposed improved MinkUNet is shown in Figure 2. The network is built upon a sparse voxel convolutional encoder–decoder and introduces self-attention enhancement modules into key feature layers to improve the feature representation of elongated structures and minority-category objects in transmission-line point clouds.

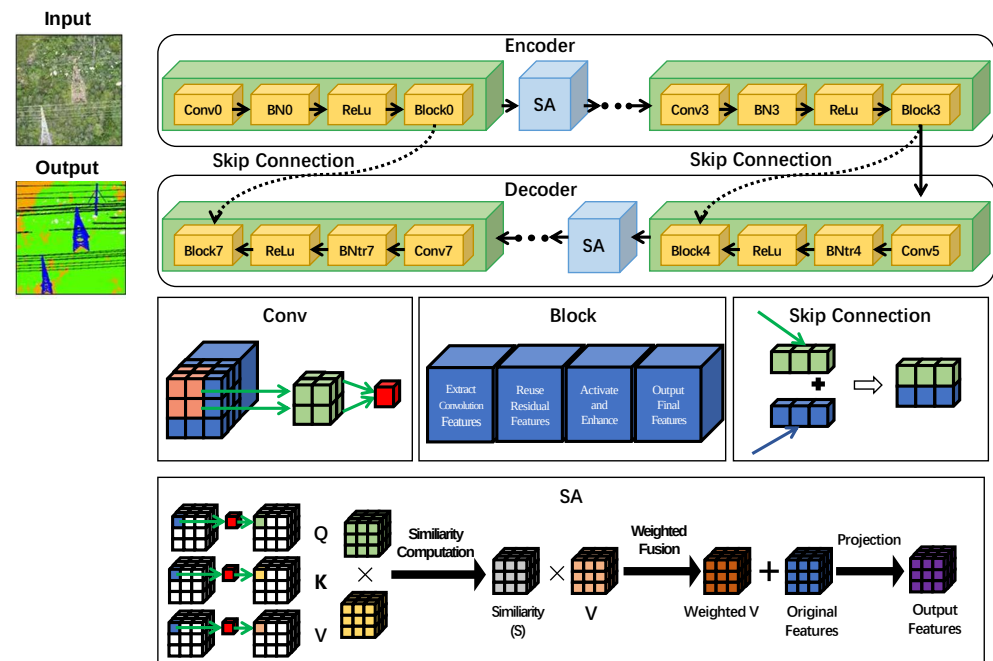


Figure 2. Schematic illustration of the improved MinkUNet architecture.

3.3. Self-Attention Enhancement Module

Conventional sparse convolutions mainly rely on local neighborhoods for feature aggregation, and their ability to model long-range dependencies and discontinuous structures is limited [16,27]. Power lines in transmission-line point clouds usually exhibit long-range linear distributions; towers and power lines have explicit spatial topological relationships; and trees, buildings, and corridor background are highly prone to local confusion. Therefore, relying only on local sparse convolutions is insufficient for fully modeling spatial correlations among different objects.

To address this issue, this paper inserts self-attention enhancement modules after the multi-scale output features of the MinkUNet encoder and before the fusion of decoder skip connections. The attention modules on the encoder side are used to enhance cross-region contextual dependencies, while those on the decoder side recalibrate key structural responses before the fusion of shallow geometric details and deep semantic features, thereby improving the discriminative ability for fine-grained categories such as power lines, towers, and building edges.

Let the input sparse feature of a given layer be $\mathbf{F} \in \mathbb{R}^{M \times d}$, where M denotes the number of non-empty voxels and d is the channel dimension. The self-attention module first generates query, key, and value features through a 1×1 sparse convolution or a linear mapping:

$$\mathbf{Q} = \mathbf{F}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}\mathbf{W}_V. \quad (4)$$

For the i -th center voxel, the attention weights are computed within its local neighborhood $\mathcal{N}(i)$ as

$$\alpha_{ij} = \frac{\exp(\mathbf{Q}_i^\top \mathbf{K}_j / \sqrt{d})}{\sum_{l \in \mathcal{N}(i)} \exp(\mathbf{Q}_i^\top \mathbf{K}_l / \sqrt{d})}, \quad j \in \mathcal{N}(i). \quad (5)$$

The neighboring value features are then aggregated with the learned weights:

$$\mathbf{O}_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{V}_j. \quad (6)$$

Finally, the self-attention output is passed through a projection layer, Dropout, and a residual connection to obtain the enhanced feature:

$$\mathbf{F}' = \mathbf{F} + \lambda \cdot \phi(\mathbf{O}), \quad (7)$$

where $\phi(\cdot)$ denotes projection and nonlinear transformation, and λ is a feature-scaling coefficient. This module can adaptively strengthen key structural features according to neighborhood feature correlations, enabling objects such as power lines and towers to obtain more discriminative representations in complex backgrounds.

3.4. Data Augmentation and Loss Function

To improve the robustness of the model to different acquisition angles, spatial positions, and local geometric variations, we apply data augmentation to the raw point cloud before voxelization. Specifically, the following strategies are used:

1. Translation augmentation: A random three-dimensional translation vector is generated and applied to the entire point clouds, expanding the distribution of objects under different spatial positions.
2. Rotation augmentation: Random rotation around the Z axis is performed to simulate viewpoint variations caused by UAVs acquisition from different flight headings.
3. Combined augmentation: Translation and rotation are randomly combined so that the model can learn stable feature representations under both position and angular variations.

To address the class imbalance problem in transmission-line point clouds, focal loss is adopted as the main supervised loss. For the predicted probability p_t corresponding to the ground-truth class, focal loss is defined as

$$\mathcal{L}_{\text{FL}}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (8)$$

where α_t is the class-balancing factor and γ is the focusing parameter. When a sample is easy to classify, $(1 - p_t)^\gamma$ approaches zero, reducing the contribution of easy samples to the loss. When a sample is difficult to classify, this factor retains a large weight, encouraging the model to focus more on hard samples and minority categories.

Adam is used as the optimizer, with the initial learning rate set to 1×10^{-3} and dynamically decayed using a cosine annealing strategy. To alleviate overfitting, Dropout regularization is added after sparse convolutional layers with a dropout rate of 0.2, and an L2 weight decay of 1×10^{-4} is used in the optimizer.

3.5. Label-Efficient Scene Generalization Strategy

Practical transmission-line inspection often involves different regions, voltage levels, terrains, and geomorphological conditions. If each new scene requires complete point-wise annotation and full retraining, the engineering cost will be high. This paper further proposes a label-efficient scene generalization strategy that achieves target-scene adaptation through transfer learning, data augmentation, and iterative pseudo-label optimization. The detailed training procedure is shown in Algorithm 1.

Let the labeled source-scene dataset be $\mathcal{D}_s = \{(\mathcal{P}_s^i, \mathcal{Y}_s^i)\}_{i=1}^{n_s}$, while the target scene contains only a small labeled dataset $\mathcal{D}_t^l$ and a large unlabeled dataset $\mathcal{D}_t^u$. First, model parameters θ_s are obtained by training on the source scene and are used as the initialization parameters for the target-scene model:

$$\theta_t^0 \leftarrow \theta_s. \quad (9)$$

Algorithm 1 Training procedure for label-efficient scene adaptation

Require: Source-scene labeled data $\mathcal{D}_s$, a small amount of target-scene labeled data $\mathcal{D}_t^l$, and target-scene unlabeled data $\mathcal{D}_t^u$
Ensure: Target-scene model parameters θ_t

- 1: Train the improved MinkUNet on the source-scene data $\mathcal{D}_s$ to obtain source model parameters θ_s
- 2: Initialize the target model parameters: $\theta_t \leftarrow \theta_s$
- 3: Perform initial fine-tuning using the target-scene labeled data $\mathcal{D}_t^l$
- 4: **for** the k -th training stage **do**
- 5: Use the current model to predict the unlabeled data $\mathcal{D}_t^u$
- 6: Select high-confidence points according to the confidence threshold τ_k and generate pseudo-labels
- 7: Jointly train the model using real labeled samples and pseudo-labeled samples
- 8: Apply consistency constraints to different augmented views of the unlabeled samples
- 9: Update the model parameters and adjust the threshold τ_k
- 10: **end for**
- 11: **return** θ_t

Then, the model is fine-tuned using the small amount of labeled target-scene data to obtain the initial target model f_{θ_t} . For an unlabeled point cloud $\mathcal{P}_u$, the model outputs the class probability $\mathbf{P}_u = f_{\theta_t}(\mathcal{P}_u)$. The pseudo-label is determined by the class with the maximum probability:

$$\hat{y}_{u,i} = \arg \max_c \mathbf{P}_{u,i}^c, \quad \hat{m}_{u,i} = \mathbb{I} \left[\max_c \mathbf{P}_{u,i}^c > \tau_k \right], \quad (10)$$

where τ_k denotes the confidence threshold at the k -th training stage, and $\hat{m}_{u,i}$ is the validity mask of the pseudo-label. Only high-confidence points participate in pseudo-label supervision, thereby reducing the negative effect of incorrect pseudo-labels on model training.

The total loss for target-scene training is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_u \mathcal{L}_{\text{pl}} + \lambda_c \mathcal{L}_{\text{con}}, \quad (11)$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss on real labeled samples, $\mathcal{L}_{\text{pl}}$ is the supervised loss on pseudo-labeled samples, and $\mathcal{L}_{\text{con}}$ is the consistency regularization term. The pseudo-label loss is defined as

$$\mathcal{L}_{\text{pl}} = -\frac{1}{\sum_i \hat{m}_{u,i}} \sum_i \hat{m}_{u,i} \log \mathbf{P}_{u,i}^{\hat{y}_{u,i}}. \quad (12)$$

Consistency regularization constrains the predictions of the same unlabeled sample under different augmentations to remain consistent:

$$\mathcal{L}_{\text{con}} = \frac{1}{N_u} \sum_{i=1}^{N_u} \|f_{\theta}(\mathcal{A}_1(\mathcal{P}_u))_i - f_{\theta}(\mathcal{A}_2(\mathcal{P}_u))_i\|_2^2, \quad (13)$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ denote two random augmentations. Through the iterative process of pseudo-label generation, screening, retraining, and threshold updating, the model can gradually exploit the structural information contained in unlabeled target-scene data.

The training process of the label-efficient adaptation strategy consists of three phases. First, supervised pre-training is conducted using a small set of real labeled samples from the target scene for 20 epochs, stabilizing after approximately 15 epochs. Next, the first pseudo-labeling stage applies a high confidence threshold of $\tau_1 = 0.8$ to select low-noise pseudo-labels for 40 epochs of iterative learning. Finally, as model accuracy improves, the threshold is lowered to $\tau_2 = 0.6$ for another 40 epochs, allowing more pseudo-labeled samples (potentially with some noise) to enrich target-domain feature learning. The total

training epochs across all stages is set to 100 for each dataset, with both pseudo-labeling stages stabilizing after approximately 30 epochs.

4. Experimental Results and Analysis

4.1. Datasets and Experimental Setup

Three transmission-line LiDAR point cloud datasets are used in the experiments. Figure 3 illustrates their spatial distribution and typical terrain characteristics. The Foshan dataset covers a flat urban plain in the Heyuebei area, representing typical urban transmission corridors in plain areas. The Shumuyuan dataset, located in Zhongshan City, features hilly terrain with steep slopes (25° – 35°) and narrow ravines, typical of mountain forest corridors. The Langwang Village dataset, also situated in Zhongshan City, covers flat rural terrain with elevation differences less than 3 meters, representing plains rural corridors.

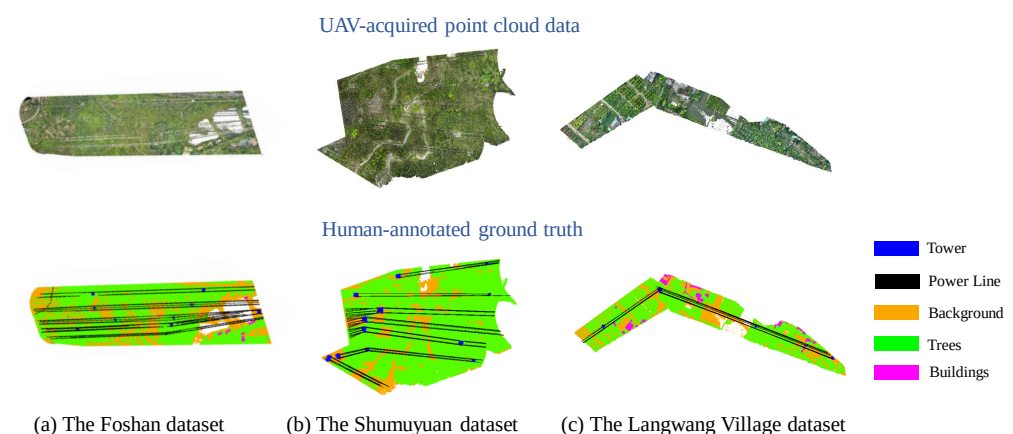


Figure 3. Overview of the UAV-acquired transmission line point cloud datasets and their corresponding human-annotated ground truth.

The Foshan dataset is used to evaluate single-scene segmentation accuracy and efficiency, with categories including towers, power lines, trees, buildings, and background. The Shumuyuan dataset is used to verify cross-scene generalization across categories of towers, power lines, trees, and background. The Langwang Village dataset further examines model adaptability using the same category set as Foshan dataset.

For cross-scene experiments on the Shumuyuan dataset, Line 1 serves as labeled data. Among the remaining lines, one is used for testing and the others as unlabeled training, with different test lines for cross-validation. On the Langwang Village dataset, one segment is labeled, another is unlabeled training, and a third is for testing, assessing adaptability to unseen sections of the same corridor.

Our point cloud datasets were acquired using a DJI Matrice 300 RTK (DJI, Shenzhen, China) quadcopter equipped with a Zenmuse L2 LiDAR module. Flight parameters were adjusted according to terrain and vegetation density, with a flight altitude of 80–120 m above ground, speed of 10 m/s, route overlap of 60%, and a target point density of ≥ 100 points/m². Coverage focused on the line corridor, including a 30–50 m buffer on each side and tower foundation areas. To ensure data quality, missions were scheduled under favorable weather conditions (no rain, wind < level 5). Detailed platform specifications are listed in Table 1.

Raw point clouds were processed using DJI Terra (v5.2.8, DJI, Shenzhen, China). The workflow included the following steps: (1) importing multi-flight raw data and setting the output coordinates to a unified reference system; (2) optimizing point cloud accuracy through GPS/IMU trajectory fusion and iterative closest point (ICP)-based multi-strip registration, assisted by ground control points for absolute georeferencing; (3) applying

multi-level denoising, which combines radius-based outlier removal and statistical filtering to suppress isolated noise while preserving fine geometric features of wires and branches; (4) filling in holes in occluded areas by using moving least squares surface fitting and manually supplementing points to repair missing wire segments and canopy gaps; and (5) converting the format to standardized LAS files. The cleaned point clouds were then manually labeled using CloudCompare (v2.13.2) to generate ground-truth annotations for six categories. This preprocessing chain ensures high-quality data suitable for subsequent segmentation tasks.

Table 1. Main system parameters of the UAV platform and LiDAR sensor.

Parameter	Value
UAV model	DJI MATRICE 350 RTK
LiDAR model	Zenmuse L2
Flight altitude (AGL)	80–120 m
Laser pulse repetition rate	240 kHz
Maximum number of returns supported	5
Point cloud data rate	Single return: max 240,000 pts/s Multiple returns: max 1,200,000 pts/s
Laser wavelength	905 nm
Ranging accuracy/point cloud density	0.04 m/ ≥ 100 points/m ²
Field of view (FOV)	Horizontal 70°, vertical 75° (non-repetitive)

4.2. Evaluation Metrics

Overall accuracy (OA), class-wise intersection over union (IoU), and mean intersection over union (mIoU) are used as the main evaluation metrics. For class c , IoU is defined as

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \quad (14)$$

where TP_c , FP_c , and FN_c denote the true positives, false positives, and false negatives of class c , respectively. mIoU is defined as the average IoU over all classes:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c. \quad (15)$$

OA is defined as the ratio of correctly classified points to all points:

$$\text{OA} = \frac{\sum_{c=1}^C \text{TP}_c}{N}. \quad (16)$$

4.3. Segmentation Results on the Foshan Transmission-Line Dataset

To evaluate the segmentation performance of the improved MinkUNet, the proposed method is compared with the original MinkUNet, KPConv, PointNet, and PointNet++. PointNet and PointNet++ are representative point-based networks [5,6], KPConv is a point-cloud segmentation method based on kernel point convolution [7], and MinkUNet [10] is the baseline version of the model improved in this paper. Figures 4 and 5 show the global segmentation results and the close-up views of different methods on the Foshan transmission-line dataset, respectively.

The visualization results show that the proposed method, MinkUNet, and KPConv can generally segment the overall region of the Foshan transmission-line scene, whereas PointNet and PointNet++ exhibit obvious omissions and category confusion in large-scale scenes. This indicates that, for transmission-line point clouds with long-range spans and complex spatial structures, methods that directly process point sets have limitations in

global contextual modeling and large-scale data processing. Compared with the baseline MinkUNet, the proposed method performs better on categories such as towers, power lines, and buildings, demonstrating that the self-attention modules and data augmentation strategies effectively enhance the model's discriminative ability for small objects and complex local structures.

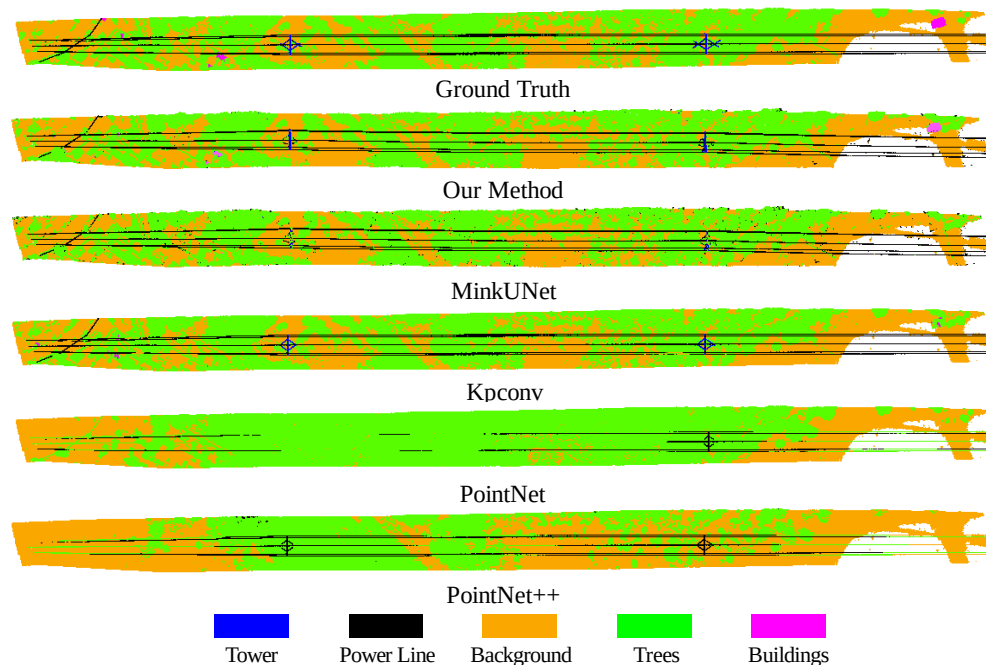


Figure 4. Global segmentation results of different methods on the Foshan transmission-line point-cloud dataset.

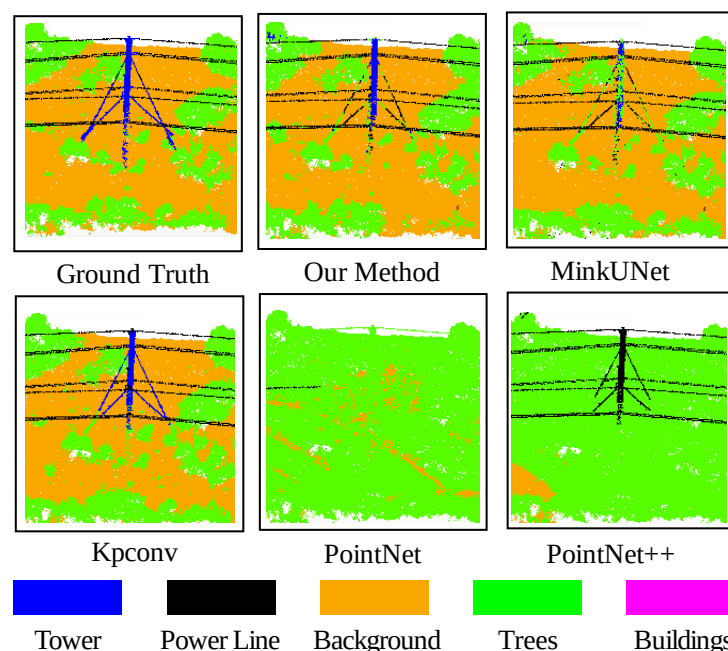


Figure 5. Local segmentation results on the Foshan transmission-line point-cloud dataset.

Table 2 presents the testing times for various methods. Our method achieves a testing time of only 1.31 s, which is significantly lower than that of KPConv, PointNet, and PointNet++. This finding suggests that the sparse-voxel-convolution-based MinkUNet framework maintains high efficiency when processing large-scale transmission-line point

clouds. Additionally, the self-attention modules introduced in this study do not compromise the overall inference efficiency, demonstrating promising potential for practical engineering applications.

Table 2. Comparison of testing time for different methods on the Foshan transmission-line dataset.

Segmentation Method	Testing Time (s)
Proposed method	1.31
KPConv	35.40
PointNet	300.01
PointNet++	660.07

Table 3 compares the quantitative metrics for various methods. The proposed method achieves a mIoU of 0.740, outperforming others in power line IoU and building IoU. This indicates a strong ability to recognize elongated structures and minority categories in transmission-line scenes. While KPConv excels in OA and performs well on tower IoU, background IoU, and tree IoU, it has a significantly longer inference time than the proposed method. This prolonged processing time makes it challenging to maintain efficiency in large-scale inspection scenarios. Compared with the baseline MinkUNet, the proposed method improves mIoU from 0.573 to 0.740 and increases the building IoU from 0.006 to 0.564. These improvements demonstrate that incorporating self-attention mechanisms and geometric data augmentation effectively enhances segmentation performance for minority categories and complex boundary cases.

Table 3. Comparison of evaluation metrics for different methods on the Foshan transmission-line dataset.

Segmentation Method	OA	Tower IoU	Power Line IoU	Background IoU	Tree IoU	Building IoU	mIoU
Proposed method	0.911	0.540	0.936	0.863	0.800	0.564	0.740
MinkUNet	0.906	0.306	0.903	0.857	0.793	0.006	0.573
KPConv	0.913	0.681	0.907	0.869	0.806	0.090	0.671
PointNet	0.726	0.000	0.261	0.595	0.554	0.000	0.282
PointNet++	0.746	0.000	0.477	0.687	0.441	0.000	0.321

4.4. Ablation Study

To analyze the independent and joint contributions of data augmentation (DA) and the self-attention module (SA), ablation experiments are conducted on the Foshan transmission-line dataset. Figure 6 shows the corresponding visualization results, and Table 4 reports the mIoU results under different strategy combinations.

Table 4. Quantitative evaluation of ablation studies on the Foshan transmission-line dataset.

DA	SA	mIoU
×	×	0.573
×	✓	0.583
✓	×	0.731
✓	✓	0.740

The results show that when only the self-attention module is added, mIoU increases from 0.573 to 0.583, indicating that this module can enhance spatial-dependency modeling without changing the overall structure of the backbone network. When only data augmentation is added, mIoU increases to 0.731, showing a more substantial improvement.

This suggests that transmission-line point clouds are sensitive to acquisition angle, spatial position, and line-direction variations, and geometric augmentation can significantly expand the spatial distribution of training samples. Furthermore, when DA and SA are used simultaneously, mIoU increases to 0.740, indicating that the two components are complementary. Data augmentation mainly improves robustness to spatial perturbations, while the self-attention module further strengthens the contextual representation of key objects such as power lines, towers, and buildings. These results verify the effectiveness of the proposed model in transmission-line point-cloud scenes.

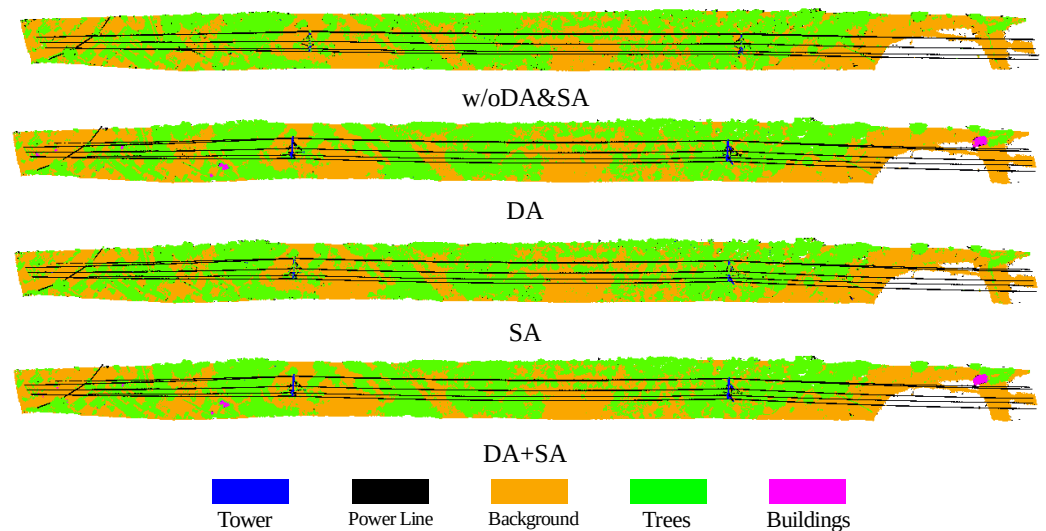


Figure 6. Visual comparison of ablation studies on the Foshan transmission-line dataset.

4.5. Results of the Label-Efficient Strategy on the Shumuyuan Dataset

For cross-scene experiments on the Shumuyuan dataset, Line 1 serves as labeled data. Among the remaining lines, one is used for testing and the others as unlabeled training, with different test lines for cross-validation. Table 5 presents the results of various strategy combinations applied to Line 4.

Table 5. Quantitative results of different label-efficient strategies on Line 4 of the Shumuyuan dataset.

Strategy	OA	Tower IoU	Power Line IoU	Background IoU	Tree IoU	mIoU
Supervision with labeled data only	0.784	0.197	0.886	0.531	0.712	0.581
Pseudo-labeling	0.773	0.019	0.116	0.581	0.690	0.351
TL + pseudo-labeling	0.781	0.336	0.889	0.528	0.708	0.615
TL + multi-stage threshold pseudo-labeling	0.781	0.337	0.892	0.531	0.709	0.617
TL + DA + pseudo-labeling	0.809	0.801	0.934	0.556	0.741	0.758
TL + DA + multi-stage threshold pseudo-labeling	0.812	0.802	0.938	0.564	0.745	0.762

As illustrated in Table 5, the basic pseudo-labeling strategy achieves a mIoU of only 0.351 when pretrained weights from the source scene and data-augmentation constraints are not applied. In this scenario, the IoU for the power line category decreases significantly, underscoring that the initial predictions from the target scene model are not sufficiently reliable. Low-quality pseudo-labels can be increasingly amplified during iterative training, leading to the accumulation of erroneous supervision. However, when transfer learning is introduced, the mIoU improves to 0.615, indicating that weights from the source scene model serve as an effective initialization for the target scene, leading to a notable enhancement in the quality of the pseudo-labels. Subsequently, with data augmentation, the mIoU further increases to 0.758. This underscores the importance of data augmentation to ensure model robustness, especially in settings with limited labeled data.

When transfer learning, data augmentation, and multi-stage threshold-based pseudo-labeling techniques are combined, the mIoU reaches 0.762. This marks an improvement of 0.181 over the 0.581 obtained by training on only a small set of labeled data from the target scene. This outcome demonstrates that the proposed label-efficient strategy effectively leverages the distributional information in unlabeled point clouds from the target scene, using a limited number of labeled samples, thereby significantly enhancing cross-scene adaptation performance.

Figure 7 shows the visualization results of different strategies in a local region. It can be observed that using pseudo-labeling alone easily causes tower omissions and power line category confusion, whereas transfer learning and data augmentation can significantly restore tower structures and power line continuity. Multi-stage threshold-based pseudo-labeling further filters out low-confidence predictions, making the model more stable during iterative training.

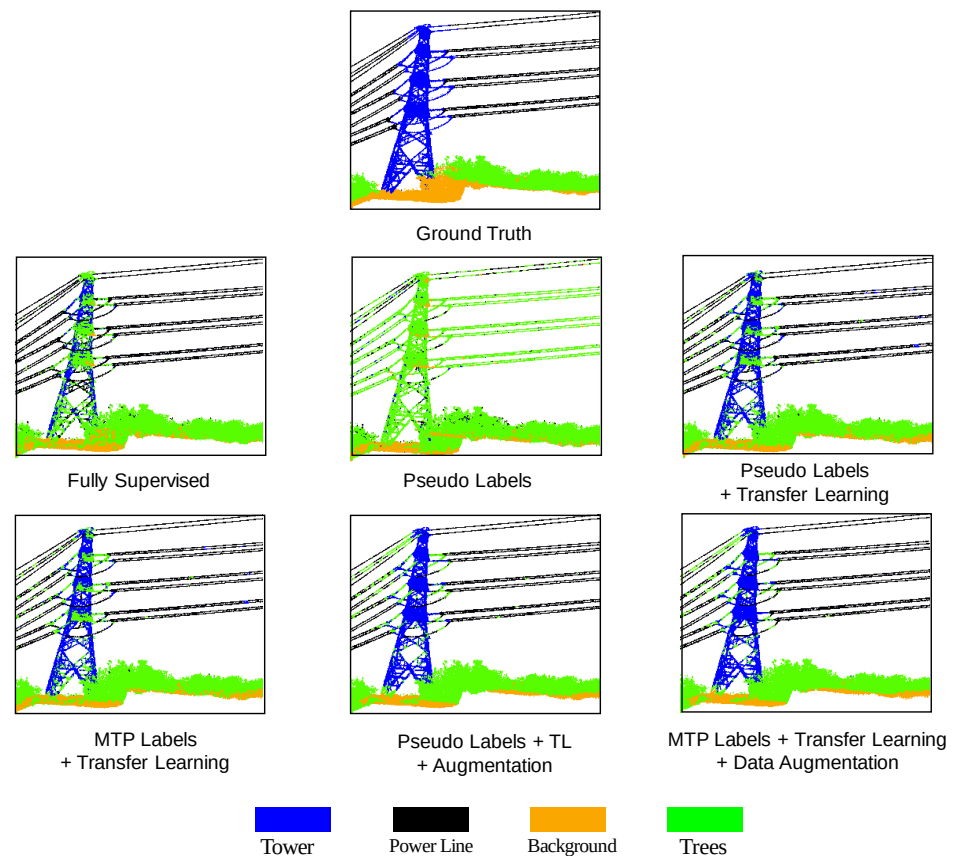


Figure 7. Local segmentation results of different strategies on the Shumuyuan dataset.

Furthermore, Figures 8–10 show the global segmentation results of Lines 2–3, Lines 4–5, and Lines 6–7 in the Shumuyuan dataset, respectively, while Figure 11 presents detailed segmentation results in local regions. These visualization results show that the proposed method can relatively completely extract key objects such as power lines and towers in different line scenes. Nevertheless, certain category confusion remains in regions with local occlusion, point-density variation, and complex background terrain.

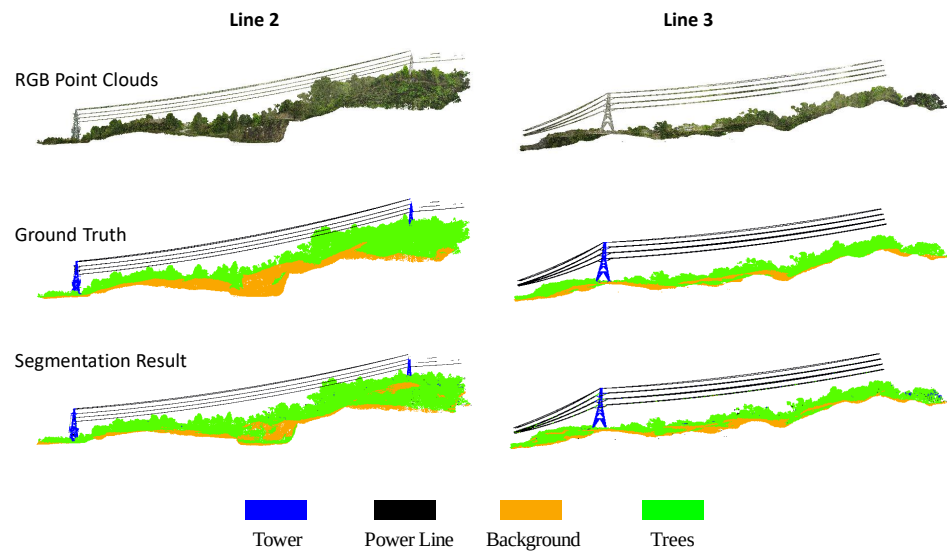


Figure 8. Global segmentation results of Lines 2 and 3 in the Shumuyuan dataset.

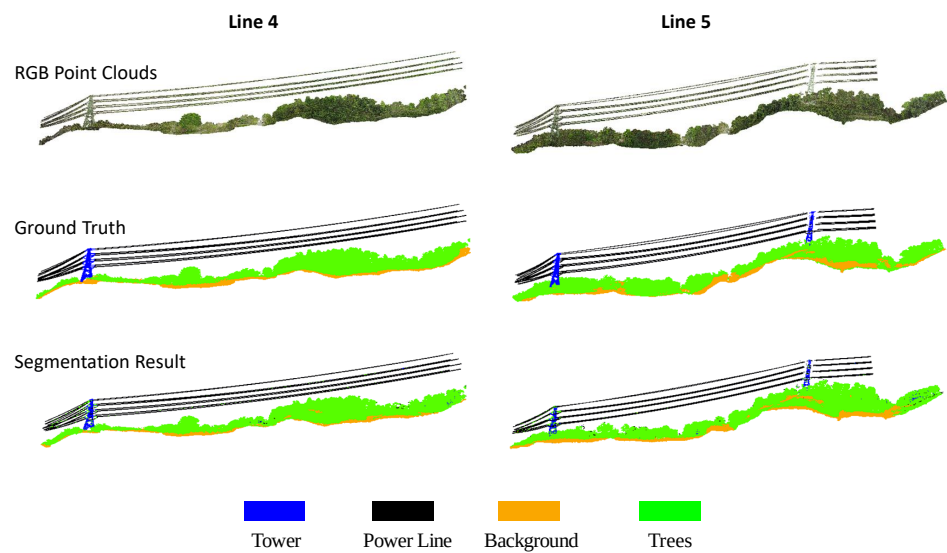


Figure 9. Global segmentation results of Lines 4 and 5 in the Shumuyuan dataset.

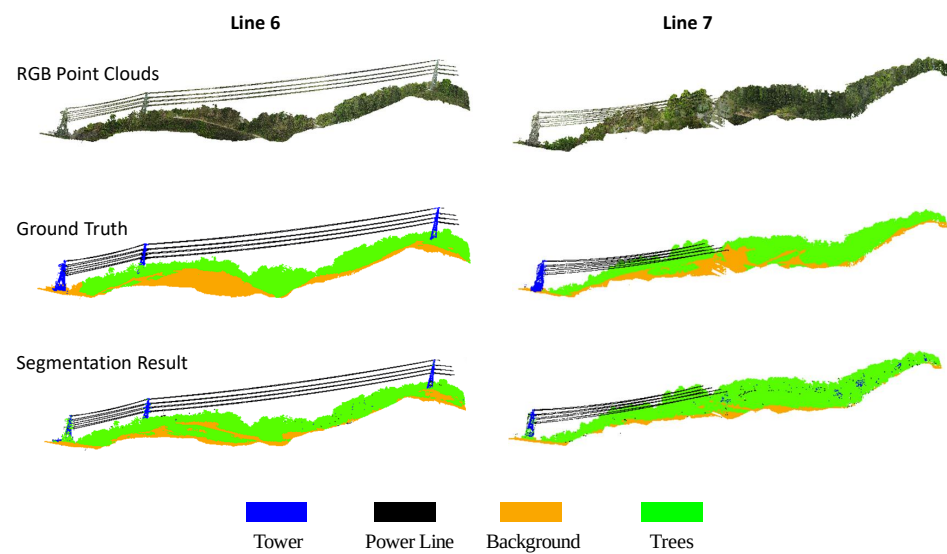


Figure 10. Global segmentation results of Lines 6 and 7 in the Shumuyuan dataset.

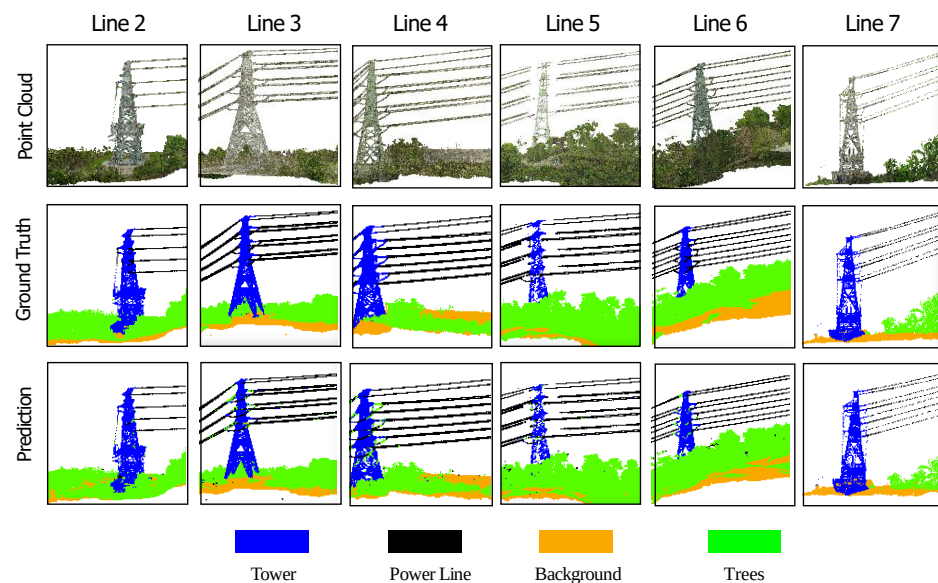


Figure 11. Local segmentation results on the Shumuyuan dataset.

Table 6 reports the segmentation results for different transmission lines in the Shumuyuan dataset. Overall, the proposed method achieves relatively stable segmentation performance across different test lines. Among them, Line 4 achieves the highest mIoU of 0.762, followed by Lines 2 (0.753), 3 (0.746), 5 (0.694), 6 (0.642), and 7 (0.599), yielding a mean mIoU of 0.699 ± 0.066 across the six test lines. Lines 6 and 7 obtain relatively lower mIoUs, which may be caused by more complex local tower structures, more severe tree occlusion, differences in point-density distribution, and similar heights between background terrain and vegetation, leading to more frequent confusion between towers and background. Nevertheless, the power line category maintains high IoU across all lines, indicating that the proposed method has strong cross-line generalization ability for the core linear objects of transmission lines.

Table 6. Segmentation results for different transmission lines in the Shumuyuan dataset.

Line ID	OA	Tower IoU	Power Line IoU	Background IoU	Tree IoU	mIoU
2	0.806	0.820	0.959	0.470	0.763	0.753
3	0.820	0.796	0.910	0.506	0.773	0.746
4	0.812	0.802	0.938	0.564	0.745	0.762
5	0.822	0.639	0.924	0.422	0.791	0.694
6	0.796	0.482	0.910	0.418	0.760	0.642
7	0.739	0.384	0.935	0.391	0.686	0.599

4.6. Results of the Label-Efficient Strategy on the Langwang Village Dataset

To further verify the adaptability of the proposed method to different transmission-line scenes, experiments are conducted on the Langwang Village dataset. Table 7 reports the quantitative metrics, while Figure 12 shows the RGB point cloud, ground-truth labels, and prediction results.

Table 7. Quantitative results on the Langwang Village dataset.

Line	OA	Tower	Power Line	Background	Tree	Building	mIoU
1	0.938	0.579	0.965	0.886	0.901	0.453	0.757

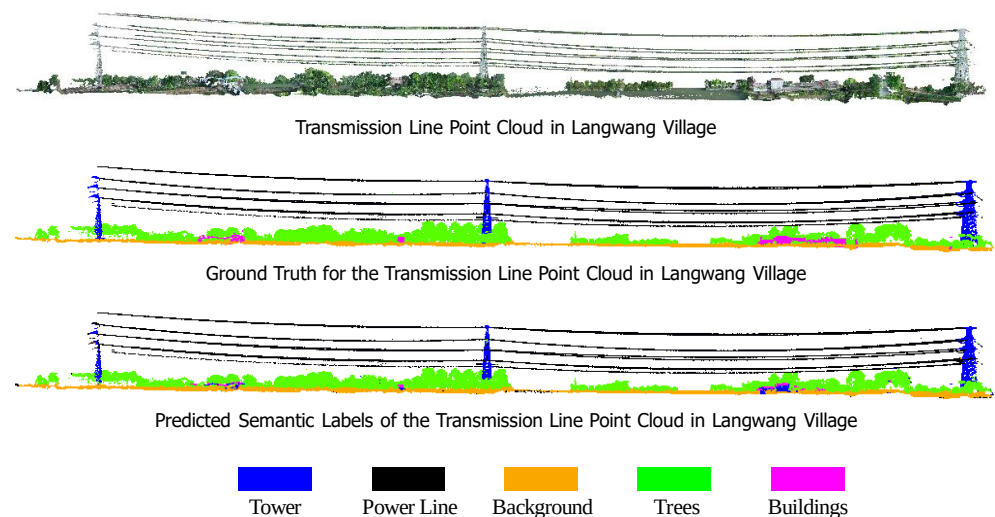


Figure 12. Segmentation results on the Langwang Village dataset.

The results show that the proposed method achieves an OA of 0.938 and an mIoU of 0.757 on the Langwang Village dataset. The IoUs of power line, background, and tree classes reach 0.965, 0.886, and 0.901, respectively, indicating that the model can recognize major structural objects and large-area environmental categories. In contrast, the IoUs of tower and building classes are 0.579 and 0.453, respectively, leaving room for further improvement. This may be related to the complex structure of tower point clouds, the small number of building samples, and local occlusion. Overall, the Langwang Village experiment verifies the adaptability of the label-efficient strategy in different real scenes.

5. Discussion

5.1. Balance Between Accuracy and Efficiency

Transmission-line point clouds typically cover large areas and contain a vast number of points, encompassing objects with significantly different spatial forms, such as power lines, towers, trees, buildings, and terrain. While networks that process points directly can preserve original point-level geometric information, they often require frequent sampling, grouping, and neighborhood searches in large-scale scenes. This results in high inference overhead. In contrast, sparse voxel convolution effectively reduces redundant computation in wide-area point-cloud processing by performing convolution only on non-empty voxels. This approach provides an efficient foundation for engineering deployment. Experimental results indicate that the proposed method achieves a mIoU of 0.740 on the Foshan dataset, with a testing time of only 1.31 s. This demonstrates that the sparse voxel framework can strike a favorable balance between efficiency and accuracy.

Additionally, although the baseline MinkUNet exhibits high inference efficiency, its ability to represent small or minority categories, such as power lines and buildings, is limited. The self-attention modules introduced in this paper can adaptively model feature correlations within local neighborhoods. They enhance the spatial contextual relationships between power lines and towers, as well as between trees and the surrounding corridor background. Consequently, the proposed method maintains the computational advantages of sparse convolution while improving segmentation accuracy for complex categories.

Beyond architectural choices, voxel size is another critical factor that affects sparse convolution performance. To further validate our configuration, we conducted a quantitative ablation study to assess the sensitivity to voxel size and its effect on segmentation performance. Consistent with the original MinkUNet, our improved model adopts a default voxel size of 0.05 m, achieving a mIoU of 0.740. We also tested voxel sizes of 0.03 m and

0.1 m. Reducing the voxel size to 0.03 m results in a mIoU of 0.732, which is slightly lower than the default setting, and it incurs a significant increase in computational cost due to the higher number of voxels. On the other hand, increasing the voxel size to 0.1 m yields a mIoU of 0.695, indicating a notable decline in accuracy due to the loss of fine geometric details. These results confirm that 0.05 m offers the best trade-off between segmentation accuracy and computational.

5.2. Mechanism of the Label-Efficient Strategy

Label-efficient experiments indicate that relying solely on pseudo-labeling does not guarantee improved performance in the target environment and may even lead to a decline in performance due to noise in the pseudo-labels. Transfer learning offers a more reliable initial decision boundary for the target scene, while data augmentation enhances the model's robustness to spatial variations and scene changes. Additionally, multi-stage, threshold-based pseudo-labeling minimizes erroneous supervision by applying confidence screening. Together, these three components allow the model to gradually integrate distributional information from unlabeled data, even when labeled data is limited. This finding suggests that, for cross-scene transmission-line point cloud adaptation, pseudo-labeling strategies should be combined with stable initialization and effective data perturbation techniques to leverage the potential of unlabeled data fully.

5.3. Limitations and Future Work

Despite the promising results, several limitations remain. First, the Building IoU is relatively low, ranging from 0.45 to 0.56. This is mainly due to the sparse and uneven distribution of building samples in the Foshan dataset, which is dominated by towers and dense vegetation. Although our method shows significant improvement over the baselines, the limited representativeness of the training data for buildings constrains further progress. Second, the most frequent classification errors occur under specific conditions. For instance, on steep terrain (e.g., Shumuyuan Lines 6–7), background points are often misclassified as trees due to significant elevation variations. In scenes with dense vegetation and similar height profiles (e.g., Langwang Village), buildings are often mistaken for power towers. These errors are further aggravated by the source-domain model's moderate performance on minority categories. Third, the greatest performance drops are concentrated in three scenarios: regions with steep slopes, areas with dense canopy cover, and categories with limited training samples. Such concentration of errors highlights the model's sensitivity to topographic complexity and data imbalance. Fourth, the cross-validation results on the Shumuyuan dataset (Table 6) are based on single-run evaluations due to the iterative pseudo-labeling pipeline. Although the per-fold variation (std = 0.066) reflects inter-scene variability, repeated runs with different random seeds would further improve statistical reliability. We plan to address this limitation in future work.

To address these limitations, future work will focus on both data-level and algorithmic improvements. At the data level, we plan to gather additional transmission-line point clouds across diverse building layouts and integrate co-registered RGB imagery to complement geometric features. Targeted augmentation for challenging scenarios, such as steep slopes, heavy occlusion, and sparse buildings, will also be explored. At the algorithmic level, beyond class-aware sampling, building-specific attention modules, and few-shot augmentation for minority categories, we aim to explore several advanced semi-supervised paradigms. These include dual-teacher guided contrastive learning [24] to filter low-quality pseudo-labels, and uncertainty-aware pseudo-label filtering to mitigate noise propagation.

6. Conclusions

This paper addresses challenges in transmission-line LiDAR point cloud segmentation: insufficient efficiency for large-scale scenes, difficulty in segmenting elongated objects, class imbalance, and high annotation costs for cross-scene adaptation. We propose a method combining a self-attention-enhanced MinkUNet with a label-efficient adaptation strategy. Building on MinkUNet, we insert self-attention modules into the encoder and decoder to improve spatial correlation modeling among power lines, towers, buildings, and background. Geometric data augmentation, focal loss, and regularized training are employed to improve recognition of minority categories. For cross-scene adaptation, we design a label-efficient strategy integrating transfer learning, data augmentation, and multi-stage threshold-based pseudo-labeling, enabling effective training with limited labeled data and abundant unlabeled data.

Experimental results demonstrate clear advantages over existing techniques. On the Foshan dataset, our method achieves a mIoU of 0.740 with an inference time of only 1.31 s, outperforming point-based networks (KPCnv, PointNet, PointNet++) in efficiency. Compared to baseline MinkUNet, mIoU improves by 16.7% and Building IoU by 93.8%, validating the effectiveness of self-attention and geometric augmentation for minority categories. Unlike conventional methods requiring extensive annotations per scene, our approach achieves superior transferability with minimal labeled data. Combining transfer learning, data augmentation, and multi-stage pseudo-labeling yields 0.762 mIoU on Shumuyuan and 0.757 on Langwang Village, surpassing the 0.581 obtained by training solely on limited target-scene data. These results confirm that the model effectively leverages source-domain knowledge and unlabeled target data, reducing labeling needs in large-scale deployments.

Future work will extend this framework to broader applications, including drone-based power line inspection, LiDAR-enabled 3D corridor mapping, real-time tree encroachment warning, and automated component damage detection. Technical improvements will focus on few-shot augmentation under class imbalance, cross-span topological relationship modeling, and dynamic pseudo-label threshold updating to further enhance robustness and generalization in complex scenes.

Author Contributions: Conceptualization, Y.W. and J.H.; methodology, J.H. and Y.L.; validation, Y.L.; formal analysis, Y.L.; investigation, Y.L.; resources, Y.W.; data curation, Y.W. and J.H.; writing—original draft preparation, J.H. and Y.L.; writing—review and editing, Y.L. and Y.W.; visualization, Y.L.; supervision, Y.L. and Y.W.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of China Southern Power Grid (CSG) (Project No. 032000KC23110037/GDKJXM20231237). The funding was administered by Zhongshan Power Supply Bureau of CSG, Guangdong.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are contained within the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge Cheng Zhang, Zhicong Dong, and Hongfa Li for their valuable assistance with data acquisition, field coordination, and experimental support at Zhongshan Power Supply Bureau. Their contributions were carried out under a formal collaboration agreement between the participating institutions.

Conflicts of Interest: Author Yijiang Wu was employed by the company Guangdong Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yu, H.; Wang, Z.; Zhou, Q.; Ma, Y.; Wang, Z.; Liu, H.; Ran, C.; Wang, S.; Zhou, X.; Zhang, X. Deep-learning-based semantic segmentation approach for point clouds of extra-high-voltage transmission lines. *Remote Sens.* **2023**, *15*, 2371. [\[CrossRef\]](#)
2. He, M.; Qin, L.; Deng, X.; Zhou, S.; Liu, H.; Liu, K. Transmission line segmentation solutions for uav aerial photography based on improved unet. *Drones* **2023**, *7*, 274. [\[CrossRef\]](#)
3. Cao, D.; Wang, C.; Liu, H.; Zhang, S.; Du, M.; Wang, P.; Nie, S.; Cheng, S. An automatic method for powerline extraction from als point cloud of powerline corridors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 10803–10829. [\[CrossRef\]](#)
4. Chen, C.; Jin, A.; Yang, B.; Ma, R.; Sun, S.; Wang, Z.; Zong, Z.; Zhang, F. DCPLD-Net: A diffusion coupled convolution neural network for real-time power transmission lines detection from UAV-Borne LiDAR data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102960. [\[CrossRef\]](#)
5. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660. [\[CrossRef\]](#)
6. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114. [\[CrossRef\]](#)
7. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 6411–6420. [\[CrossRef\]](#)
8. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953. [\[CrossRef\]](#)
9. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud. In Proceedings of the IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 1887–1893. [\[CrossRef\]](#)
10. Choy, C.; Gwak, J.; Savarese, S. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084. [\[CrossRef\]](#)
11. Matikainen, L.; Lehtomäki, M.; Ahokas, E.; Hyypä, J.; Karjalainen, M.; Jaakkola, A.; Kukko, A.; Heinonen, T. Remote sensing methods for power line corridor surveys. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 10–31. [\[CrossRef\]](#)
12. Shen, Y.; Huang, J.; Wang, J.; Jiang, J.; Li, J.; Ferreira, V. A review and future directions of techniques for extracting powerlines and pylons from LiDAR point clouds. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *132*, 104056. [\[CrossRef\]](#)
13. Zhang, R.; Yang, B.; Xiao, W.; Liang, F.; Liu, Y.; Wang, Z. Automatic extraction of high-voltage power transmission objects from UAV lidar point clouds. *Remote Sens.* **2019**, *11*, 2600. [\[CrossRef\]](#)
14. Xi, S.; Zhang, Z.; Niu, Y.; Li, H.; Zhang, Q. Power line extraction and tree risk detection based on airborne LiDAR. *Sensors* **2023**, *23*, 8233. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803. [\[CrossRef\]](#)
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [\[CrossRef\]](#)
17. Wang, J.; Cui, Y.; Guo, D.; Li, J.; Liu, Q.; Shen, C. Pointattn: You only need attention for point cloud completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5472–5480. [\[CrossRef\]](#)
18. Xu, X.; Lee, G.H. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13706–13715. [\[CrossRef\]](#)
19. Kong, L.; Ren, J.; Pan, L.; Liu, Z. Lasermix for semi-supervised lidar semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21705–21715. [\[CrossRef\]](#)
20. Shao, F.; Luo, Y.; Liu, P.; Chen, J.; Yang, Y.; Lu, Y.; Xiao, J. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 2575–2585. [\[CrossRef\]](#)
21. Zhou, Z.H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [\[CrossRef\]](#)

22. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1196. [[CrossRef](#)]
23. Wu, Y.; Yan, Z.; Cai, S.; Li, G.; Han, X.; Cui, S. Pointmatch: A consistency training framework for weakly supervised semantic segmentation of 3d point clouds. *Comput. Graph.* **2023**, *116*, 427–436. [[CrossRef](#)]
24. Yao, B.; Xiao, H.; Zhuang, J.; Peng, C. Weakly supervised learning for point cloud semantic segmentation with dual teacher. *IEEE Robot. Autom. Lett.* **2023**, *8*, 6347–6354. [[CrossRef](#)]
25. Yin, C.; Yang, B.; Cheng, J.C.; Gan, V.J.; Wang, B.; Yang, J. Label-efficient semantic segmentation of large-scale industrial point clouds using weakly supervised learning. *Autom. Constr.* **2023**, *148*, 104757. [[CrossRef](#)]
26. Graham, B.; Engelcke, M.; van der Maaten, L. 3D semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9224–9232. [[CrossRef](#)]
27. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.S.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16259–16268.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.