remote sensing

MDPI

Article

Semantic-Aware Cross-Modal Transfer for UAV-LiDAR Individual Tree Segmentation

Fuyang Zhou [1,2], Haiqing He [1,2,*], Ting Chen [3], Tao Zhang [1,2], Minglu Yang [1,2], Ye Yuan [4] and Jiahao Liu [5]

[1] School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China; fuyzhou@ecut.edu.cn (F.Z.); 2024100038@ecut.edu.cn (T.Z.); 2024120608@ecut.edu.cn (M.Y.)
[2] Jiangxi Key Laboratory of Watershed Ecological Process and Information, East China University of Technology, Nanchang 330013, China
[3] School of Water Resources and Environmental Engineering, East China University of Technology, Nanchang 330013, China; ct_201607@ecut.edu.cn
[4] Shenzhen DJI Innovations Technology Co., Ltd., Shenzhen 518057, China; 2019213817@ecut.edu.cn
[5] Jiangxi Helicopter Co., Ltd., Jingdezhen 333036, China; 2024120575@ecut.edu.cn
* Correspondence: hehaiqing@ecut.edu.cn

Abstract

Cross-modal semantic segmentation of individual tree LiDAR point clouds is critical for accurately characterizing tree attributes, quantifying ecological interactions, and estimating carbon storage. However, in forest environments, this task faces key challenges such as high annotation costs and poor cross-domain generalization. To address these issues, this study proposes a cross-modal semantic transfer framework tailored for individual tree point cloud segmentation in forested scenes. Leveraging co-registered UAV-acquired RGB imagery and LiDAR data, we construct a technical pipeline of "2D semantic inference—3D spatial mapping—cross-modal fusion" to enable annotation-free semantic parsing of 3D individual trees. Specifically, we first introduce a novel Multi-Source Feature Fusion Network (MSFFNet) to achieve accurate instance-level segmentation of individual trees in the 2D image domain. Subsequently, we develop a hierarchical two-stage registration strategy to effectively align dense matched point clouds (MPC) generated from UAV imagery with LiDAR point clouds. On this basis, we propose a probabilistic cross-modal semantic transfer model that builds a semantic probability field through multi-view projection and the expectation–maximization algorithm. By integrating geometric features and semantic confidence, the model establishes semantic correspondences between 2D pixels and 3D points, thereby achieving spatially consistent semantic label mapping. This facilitates the transfer of semantic annotations from the 2D image domain to the 3D point cloud domain. The proposed method is evaluated on two forest datasets. The results demonstrate that the proposed individual tree instance segmentation approach achieves the highest performance, with an IoU of 87.60%, compared to state-of-the-art methods such as Mask R-CNN, SOLOV2, and Mask2Former. Furthermore, the cross-modal semantic label transfer framework significantly outperforms existing mainstream methods in individual tree point cloud semantic segmentation across complex forest scenarios.

Keywords: airborne LiDAR; UAV images; instance segmentation; point cloud semantic segmentation; 2D-3D semantic mapping

check for updates

Academic Editors: Baoxin Hu and Linhai Jing

Received: 15 July 2025
Revised: 10 August 2025
Accepted: 12 August 2025
Published: 13 August 2025

Citation: Zhou, F.; He, H.; Chen, T.; Zhang, T.; Yang, M.; Yuan, Y.; Liu, J. Semantic-Aware Cross-Modal Transfer for UAV-LiDAR Individual Tree Segmentation. *Remote Sens.* **2025**, *17*, 2805. https://doi.org/10.3390/ rs17162805

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Semantic segmentation of individual tree-scale light detection and ranging (LiDAR) point clouds is a fundamental technique for precision forest management, analysis of ecological interaction mechanisms, and forest carbon stock assessment [1]. Traditional forestry surveys rely heavily on manual visual interpretation or ground-based measurements, which are labor-intensive, time-consuming, and inadequate for covering large and complex terrains [2]. In recent years, the rapid advancement of unmanned aerial vehicle (UAV)-based multi-sensor technologies has opened up new opportunities for forest resource surveys. By synchronously acquiring high-resolution RGB imagery and LiDAR point clouds, researchers are now able to capture forest structural information with greater completeness and detail [3,4]. However, existing supervised point cloud segmentation approaches remain heavily dependent on labor-intensive annotated datasets. The high cost of annotation and limited generalization across different forest environments constrain their applicability in dynamic forest monitoring [5]. Meanwhile, 2D image segmentation methods based on deep learning have demonstrated strong transferability across domains [6–8]. Nonetheless, efficiently transferring 2D semantic labels to the 3D point cloud domain remains a significant open challenge.

Current individual tree segmentation methods can be broadly categorized into traditional geometry-based algorithms and deep learning-based models. Traditional approaches, such as the watershed algorithm based on canopy height models (CHMs) [9,10] and voxel-based region growing methods [11,12], do not require annotated data. However, they are highly sensitive to canopy overlap and morphological diversity, often resulting in over-segmentation or under-segmentation. With the advent of deep learning, PointNet [13] pioneered end-to-end point cloud classification, but its global feature extraction mechanism overlooks local geometric relationships, leading to frequent misclassifications in dense canopy segmentation tasks. PointNet++ [14] addressed this limitation by hierarchically aggregating local neighborhood features, yet its supervised training paradigm still demands large amounts of labeled data and remains sensitive to variations in point cloud density. RandLA-Net [15] improves computational efficiency through random sampling and local feature enhancement, but its recall rate under dense canopies falls below 40%, which is insufficient for high-precision forestry applications. Moreover, supervised models exhibit limited generalization across forest types (e.g., coniferous vs. broadleaf forests), with significant performance degradation when applied to novel environments [16]. This issue is particularly critical in dynamically changing forest ecosystems.

Recent studies have explored multimodal fusion to enhance segmentation robustness. For instance, the FuseSeg framework [17] integrates RGB and LiDAR data via early-stage feature fusion, while the PMF (Prompt-Based Multimodal Fusion) method [18] employs a dual-stream network to align multimodal features within the camera coordinate system. However, these approaches are primarily designed for rigid urban objects and struggle to handle the flexible deformation and canopy occlusion commonly observed in forested environments. Individual tree segmentation in forest scenes exhibits distinct characteristics. Unlike the regular structures of rigid urban objects, forest canopies vary significantly across tree species—for example, conical spruces versus umbrella-shaped oaks—necessitating strong model generalization capabilities [19]. In addition, multi-layered canopies lead to sparse point cloud representations of understory trees, as LiDAR pulse penetration in coniferous forests is typically limited to 30–50%, further complicating data processing. Dynamic factors such as seasonal variations in leaf area also pose challenges to model robustness, requiring traditional supervised methods to be frequently retrained to adapt to shifts in data distribution [16].

To reduce the reliance on manual annotations, extensive research has explored unsupervised and cross-modal transfer techniques [20–23]. For example, the Segment Anything Model (SAM) [21] achieves zero-shot generalization through prompt-based mechanisms. However, when applied directly to LiDAR point clouds, it suffers from fragmented segmentation results due to the lack of spatial continuity in the 3D domain. Wang et al. [22] proposed an unsupervised method for forest point cloud instance segmentation, which achieved promising performance on benchmark datasets. However, it still struggles to effectively interpret LiDAR point cloud information in large-scale and complex forested areas. Weakly supervised methods such as SFL-Net (Slight Filter Learning Network) [23] leverage sparse annotations to guide model training, yet still require at least 20% of labeled data, making it difficult to completely eliminate human intervention.

The widespread adoption of UAV-based multi-sensor systems (LiDAR + RGB) has opened new avenues for cross-modal semantic transfer. RGB imagery provides rich color and texture cues that assist in species-level tree discrimination [24], while LiDAR point clouds offer precise 3D geometric information [25]. However, multimodal fusion presents several significant challenges. First, dense matching point clouds generated from UAV imagery (structure from motion point clouds) often exhibit scale discrepancies and coordinate offsets when compared to LiDAR point clouds, with direct registration errors reaching the meter level. Second, when 2D semantic labels from single-view images are projected into 3D space, canopy occlusion leads to incomplete label coverage. Finally, existing fusion frameworks (e.g., FuseSeg [17]) must concurrently process high-resolution imagery and large-scale point clouds, resulting in high computational complexity that hampers real-time performance—an essential requirement for operational forestry monitoring. To tackle this issue, this paper proposes a cross-modal semantic transfer framework for individual tree point cloud segmentation. By leveraging the synergistic analysis of UAV imagery and LiDAR data, the framework aims to provide a low-cost yet high-accuracy solution for large-scale forest monitoring.

To address the aforementioned challenges, this study proposes a cross-modal semantic transfer framework for individual tree point cloud segmentation in forest environments. The innovation lies in the integration of multimodal data advantages with optimized computational efficiency. First, we introduce a novel model, Multi-Source Feature Fusion Network (MSFFNet), which performs instance-level segmentation of individual trees based on UAV imagery. Second, a two-stage registration strategy is designed: a coarse alignment between dense matching point clouds and LiDAR data is achieved using the Two-stage Consensus Filtering (TCF) algorithm, followed by local refinement through the Iterative Closest Point (ICP) algorithm, effectively resolving spatial inconsistencies across data sources. Third, a semantic probability field is constructed by integrating multi-view geometry with the Expectation-Maximization (EM) algorithm, addressing inconsistencies in tree instance masks across different views (e.g., over-segmentation or misclassification). Finally, based on the semantic probability field and 3D point cloud, semantic association between 2D pixels and 3D points is established by jointly considering point cloud geometric features and semantic confidence. This enables the transfer of semantic labels from the 2D image domain to the 3D point cloud domain, completing the cross-modal semantic label mapping process.

The main contributions of this study are summarized as follows:

(1) This study presents a cross-modal semantic transfer framework for forest scenes, which leverages UAV imagery and LiDAR data to transfer semantic information from the 2D image domain to the 3D point cloud domain. This approach addresses the critical limitation of supervised point cloud segmentation methods that rely heavily on manual annotations.

(2) This study develops a label fusion algorithm based on multi-view spatial consistency, which effectively resolves semantic mapping conflicts caused by inconsistent instance masks of the same tree across multiple views. The proposed method ensures semantic consistency across views and preserves global intra-class geometric coherence.
(3) To evaluate the performance of the proposed cross-modal semantic transfer framework for individual tree point cloud segmentation, we conducted comprehensive experiments on two UAV-LiDAR datasets, demonstrating its effectiveness in comparison with several state-of-the-art methods.

The structure of this paper is organized as follows: Section 2 describes the data collection and preprocessing procedures. Section 3 presents the proposed method in detail, along with the experimental evaluation metrics. Section 4 reports and analyzes the experimental results of the proposed framework. Finally, Section 5 concludes the paper with a summary of the main findings.

2. Study Area and Data Pre-Processing

This section provides an overview of the study area and the data processing workflow, which mainly includes (1) the geographical location and environmental characteristics of the study area; (2) the acquisition and preprocessing of UAV imagery and LiDAR data; (3) the construction of ground truth data for accuracy validation.

2.1. Study Area

To validate the effectiveness of the proposed individual tree point cloud segmentation method for forest environments, we selected a representative study area in the southern part of Ganzhou, Jiangxi Province, China (115°30′–116°30′E, 25°00′–26°00′N), which consists of typical commercial forests. This region lies within the core construction zone of the national forest reserve and features a subtropical monsoon humid climate, with an average annual temperature of 15.9 °C and an average annual precipitation of 1548 mm. The terrain is notably undulating, with an average elevation of 478 m, and the vegetation coverage exceeds 92%. The forest community is dominated by Cunninghamia lanceolata, Pinus massoniana, and Phyllostachys edulis, forming a mixed coniferous-broadleaf and bamboo forest ecosystem. The vegetation exhibits a typical vertical stratification, consisting of arborous, shrub, and herbaceous layers. In this area, we collected high-resolution UAV-based RGB imagery and airborne LiDAR point cloud data, covering approximately 2 km^2. To evaluate the performance of the proposed method, we selected two representative subregions within the study area for experiments. The sizes of these two regions are approximately 40 m × 107 m and 30 m × 70 m, respectively. The geographic location and general characteristics of the study area are illustrated in Figure 1.

2.2. Data Acquisition and Processing

In this study, remote sensing data acquisition was conducted in August 2024 using the DJI Matrice 350 RTK UAV platform (DJI Innovations, Shenzhen, China), equipped with the Zenmuse L2 integrated multi-sensor system. This advanced system comprises a solid-state LiDAR module (wavelength: 905 nm; pulse repetition frequency: 240 kHz), a 20-megapixel RGB imaging unit, a high-precision inertial navigation system, and a GNSS positioning module, enabling synchronized acquisition of multi-source data. The system offers a wide field of view (70° horizontal, 75° vertical) with a maximum detection range of 450 m. At a distance of 200 m, it achieves a vertical accuracy of 4 cm and a horizontal accuracy of 5 cm, with an average point cloud density of 150 points/m^2, fully meeting the requirements of high-resolution 3D mapping. The RGB camera provides high-resolution aerial imagery, serving as the foundational data for individual tree segmentation. During each flight, both

LiDAR point clouds and high-resolution images of the study area were simultaneously acquired. Considering the complex terrain of the experimental site, terrain-following flight mode was adopted. The UAV flew at an altitude of approximately 100 m with a cruising speed of 10 m/s, capturing 194 high-definition orthophotos (5280 × 3956 pixels) and collecting approximately 123 million LiDAR points, resulting in a multimodal 3D forest dataset. The mission parameters are detailed in Table 1.

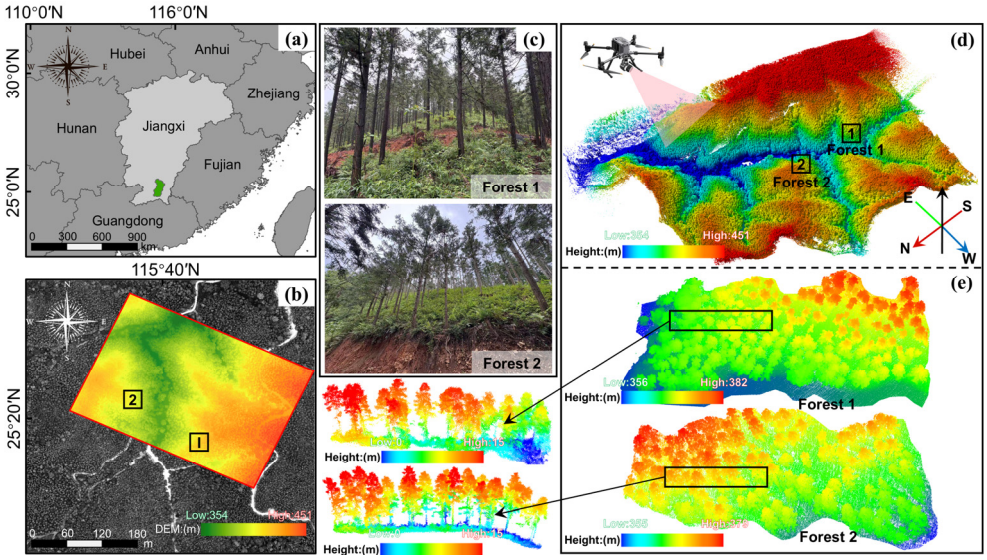


Figure 1. Study area. (**a**) Geographic location of the study site. (**b**) Selected forested experimental region. (**c**) Actual forest environment within the experimental plots. (**d**) Airborne LiDAR point cloud data acquired over the study area. (**e**) The two experimental plots used for evaluating the proposed method.

Table 1. UAV mission parameterization.

UAV (DJI Matrice 350 RTK)		Sensor (Zenmuse L2)	
Flight altitude	100 m	Maximum number of echoes supported	5
Flight speed	10 m/s	Point cloud data rate	1.2 million points/s
Forward overlap	70%	LiDAR ranging accuracy	2 cm @ 150 m
Side overlap	80%	Image size	5280 × 3956 pixels
Flight mode	terrain-following flight	Spatial resolution	3 cm/pixel

The LiDAR point cloud data were first preprocessed using DJI Terra (Version 4.5.0) [26] software to obtain optimized, high-precision point clouds. To separate non-ground points from the airborne laser scanning (ALS) data, the Cloth Simulation Filtering (CSF) algorithm [27] was employed for accurate classification of ground and non-ground points. The preprocessed non-ground point cloud was then used to generate a high-resolution CHM via Kriging spatial interpolation, effectively representing the three-dimensional structure of the forest canopy. Simultaneously, optical imagery was processed using DJI Terra's photogrammetry workflow. By integrating internal camera calibration parameters and lens distortion correction models, high-precision aerial triangulation was performed, resulting in a digital orthophoto map (DOM) with a ground resolution of 3 cm/pixel, as well as dense matched point clouds (MPC) data. It is worth noting that the image-derived point clouds were optimized using adaptive radius filtering in CloudCompare software (Version 2.13.0) [28], ensuring consistency in noise levels with the LiDAR point clouds.

2.3. Dataset Construction

Construction of the Individual Tree Crown Segmentation Dataset: To evaluate the effectiveness of the proposed individual tree instance segmentation method, we constructed a high-quality validation dataset. Using the Labelme annotation tool, manual labeling was performed on UAV-acquired RGB imagery, resulting in approximately 1364 tree instance masks in 2D. To ensure annotation accuracy, the crown delineation process integrated both the CHM and LiDAR point cloud data. The high-resolution RGB imagery served as the primary reference, providing clear visual cues for identifying and outlining tree crown boundaries. For individuals with ambiguous boundaries or overlapping crowns, additional information from CHM-derived tree height maps and LiDAR-based 3D cross-sectional profiles was incorporated to support more precise boundary determination. This annotation strategy, which fuses 2D and 3D information, significantly improves crown boundary accuracy and provides reliable ground-truth labels for training and evaluating the instance segmentation model—especially under dense vegetation or mixed forest conditions. The study area was divided into two subregions (as shown in Figure 1), which were used for cross-modal semantic mapping between 2D semantic instances and 3D point clouds. It should be noted that these two subregions were not involved in the training and validation stages of the MSFFNet but were reserved for model testing. We used DOM and CHM data outside of these two subregions for the training and validation of the MSFFNet. The training and validation sets cover approximately 80% of the labeled trees (about 1200), while the test set includes the remaining 20% (approximately 267 trees).

After completing the crown annotations, we constructed the original input dataset for individual tree crown instance segmentation by cropping the annotated images and their corresponding masks into 256 × 256 pixels patches with a stride of 128 pixels. Given the limited number of original samples, three data augmentation techniques were applied to enhance model robustness and generalization capability: geometric transformations [29], contrast adjustment, and Mixup-based augmentation [30]. These strategies effectively increased data diversity, reducing the risk of model overfitting and improving adaptability across different network architectures. After applying all three augmentation methods to each sub-dataset, a total of 7239 image patches were generated, covering 5824 distinct tree instances. Among them, the training, validation, and test sets consist of 5548, 686, and 878 images, respectively. The complete processing pipeline for individual tree crown instance segmentation is illustrated in Figure 2.

Construction of Individual Tree Point Cloud Segmentation Dataset: To support the evaluation of individual tree segmentation models in the 3D point cloud domain, we further constructed a dedicated dataset for individual tree point cloud segmentation. The annotation process was conducted using CloudCompare [28] (version 2.13.0) through an interactive labeling strategy, where each preprocessed ALS point cloud was manually segmented tree by tree. During annotation, 3D point clouds were jointly validated with their corresponding orthophotos. Using the "Segment" tool in CloudCompare, each tree's point cloud was carefully delineated with precise boundary definition. For quality control, a two-stage review mechanism was adopted: initial segmentation was performed by trained annotators, followed by expert validation by experienced forestry professionals through 3D visualization, ensuring label consistency and semantic accuracy. The resulting verification dataset contains 267 individually annotated tree samples from two representative species: Chinese fir (72%) and Pinus massoniana (28%). It is noteworthy that these 267 trees are distributed across two separate experimental sites (see Figure 1e), with 131 and 136 trees in each region, respectively. The two areas exhibit significant differences in terrain elevation, leading to varying degrees of point cloud overlap between adjacent trees (see Figure 1c). This creates substantial challenges for individual tree segmentation. At the same

time, the sample distribution provides an ideal basis for evaluating the adaptability and generalization capability of segmentation methods under diverse forest stand structures.

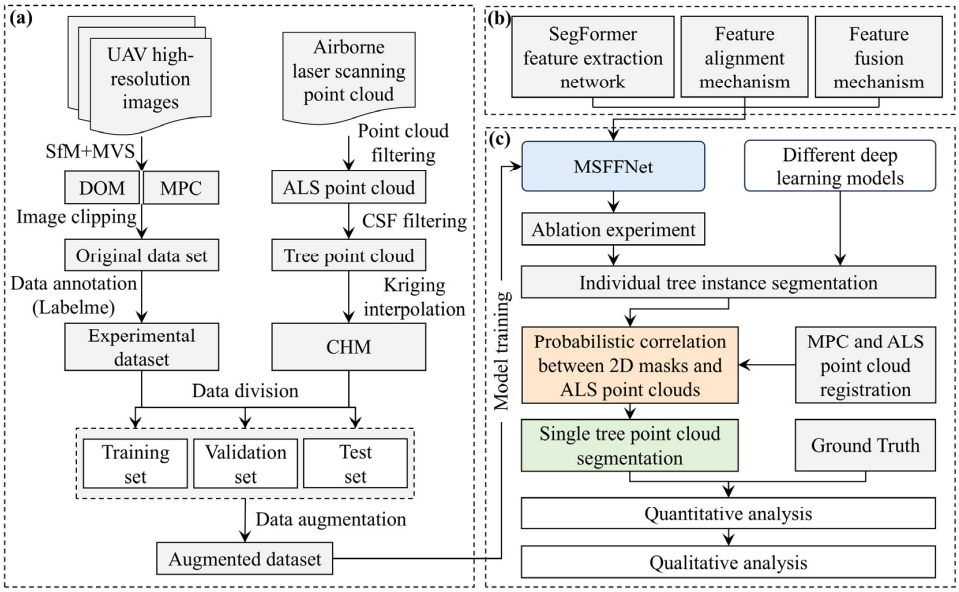


Figure 2. Pipeline of the proposed method. (**a**) Dataset construction process. (**b**) Core modules of the tree crown instance segmentation model. (**c**) Model training, inference, and individual tree point cloud segmentation process.

3. Proposed Method

For a forest scene with LiDAR point clouds and UAV imagery, our objective is to automatically transfer 2D semantic labels—obtained from a pretrained image segmentation network—into the 3D point cloud domain. Figure 2 illustrates the overall framework of the proposed method. We present a cross-modal label transfer approach tailored for individual tree semantic segmentation in forest environments. The framework takes RGB imagery and LiDAR point clouds as input and predicts instance-level tree crown semantic labels using the proposed segmentation model. The following subsections provide a detailed description of each module within the framework. As shown in Figure 2, the proposed method consists of three main stages: (1) 2D instance segmentation of individual tree crowns; (2) registration between aerial imagery and LiDAR point clouds; and (3) cross-modal semantic transfer based on a semantic probability field.

3.1. 2D Instance Segmentation of Individual Tree Canopies

Existing 2D instance segmentation networks [31–33] perform well on general datasets (such as MS-COCO [34] and PASCAL VOC [35]), but struggle to adapt to the unique complexity of forest scenes, for example, significant scale differences, dense forest spatial heterogeneity, and complex background interference caused by terrain undulations and canopy interlacing. To address this challenge, this study proposes a novel instance segmentation network named MSFFNet, which generates high-precision single-tree canopy segmentation masks through a multi-scale feature interaction fusion mechanism. To support model training and validation, we constructed a scene-adaptive annotated dataset (see Section 2.2), which is based on real forest imagery collected by a drone remote sensing platform and manually annotated by professionals to achieve precise annotation of single-tree canopy instances across the entire study area.

The architecture of MSFFNet is illustrated in Figure 3. It comprises three primary stages: feature extraction, region of interest (ROI) generation, and instance mask prediction. In the feature extraction stage, MSFFNet emphasizes deep interaction among

multi-source data to enhance hierarchical feature integration. Specifically, MSFFNet adopts a dual-branch parallel design that interactively fuses complementary information from multimodal inputs, with each branch dedicated to capturing modality-specific features. To address the computational challenges posed by high-resolution UAV imagery, both branches employ SegFormer as the backbone network. This design leverages sequence reduction mechanisms and implicit positional encoding to model long-range contextual dependencies while significantly reducing parameter overhead. It not only overcomes the inherent computational bottlenecks of Transformer-based architectures in handling high-resolution inputs but also mitigates the sensitivity of conventional methods to image resolution variability. Its efficient feature extraction aligns well with the multiscale structural characteristics of forest canopies, offering an ideal solution for individual tree segmentation in complex terrains. Considering the substantial geometric and semantic differences between RGB images and CHM data, a multi-level feature fusion (MLFF) module is introduced between the two branches. This module facilitates deep cross-modal feature fusion and improves the model's capability in crown detection and boundary delineation by integrating hierarchical representations across modalities. In the second stage, the fused multi-level feature maps are input into a region proposal network (RPN) to generate initial ROIs. In the final stage, deformable ROI alignment is applied to adaptively match each candidate region with the shared feature maps, enabling precise extraction of enhanced crown representations. The architecture simultaneously decodes three key outputs: crown class confidence, bounding box geometry, and pixel-wise instance masks—thus achieving joint optimization of geometric structure and semantic understanding.

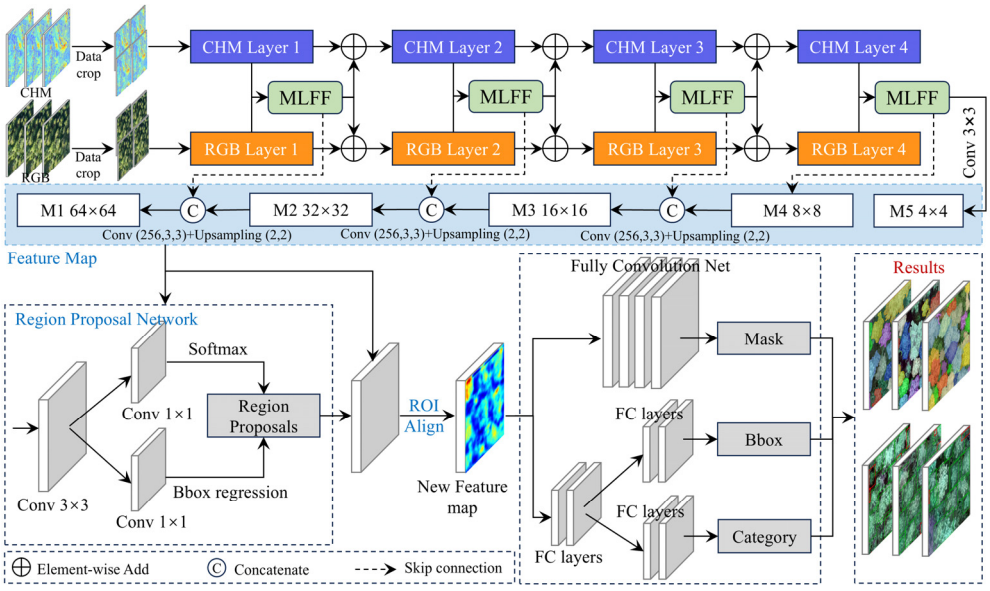


Figure 3. Overall architecture of MSFFNet.

As illustrated in Figure 4, the MLFF module integrates the texture and spectral information from RGB imagery with the spatial topological features from the CHM. While RGB data provides rich visual and appearance cues of the canopy, the CHM offers precise geometric structural representations—together forming a complementary multi-source data advantage. Deep fusion of these heterogeneous features enables the network to generate more accurate instance masks of individual tree crowns by improving crown boundary discrimination and reducing false positives and missed detections. As shown in Figure 4, the MLFF module consists of two main components: feature alignment and feature interaction. The feature interaction component further includes channel feature interaction (CFI) and

spatial feature interaction (SFI), enabling effective cross-modal information exchange and enhancing the discriminative power of the network.

To mitigate the impact of spatial misalignment between multi-branch and multi-level feature maps during feature interaction, we designed a Feature Alignment Module (FAM), as shown in Figure 4a. This module employs deformable convolution to achieve cross-branch spatial consistency correction, thereby providing higher-quality and more robust fused features for subsequent channel-wise and spatial feature interactions. Specifically, the feature maps from the dual branches are independently fed into deformable convolution blocks—each consisting of a fixed 3×3 deformable convolution kernel, batch normalization, and ReLU activation—for adaptive spatial alignment. The aligned feature maps are then concatenated and passed through a 3×3 convolution layer to predict a semantic offset field, which dynamically refines the spatial correspondence between modalities. This mechanism forms a robust foundation for downstream cross-modal feature fusion and interaction.

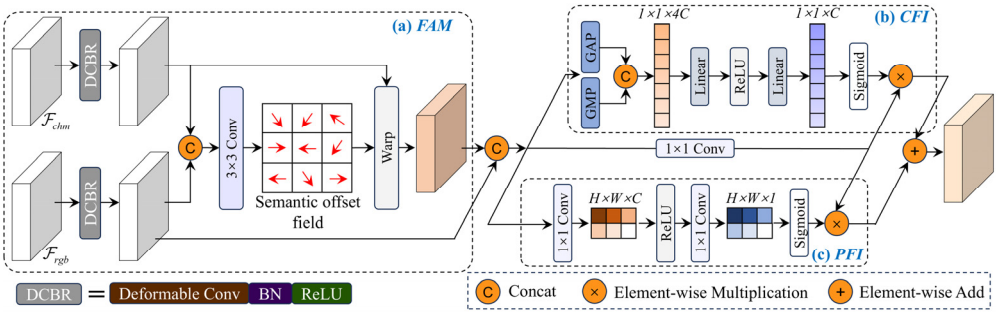


Figure 4. Architecture of the multi-level feature fusion (MLFF) module.

To achieve deep interaction between geometric and attribute features from multimodal data, the fused features output by the FAM are further processed by the CFI and SFI mechanisms. These modules allow multiple data streams within the network to emphasize each other's complementary information while aligning feature responses across modalities, thereby reducing the semantic and distributional discrepancies between them. As illustrated in Figure 4b, the CFI module consists of global average pooling, global max pooling, and a multi-layer perceptron (MLP). Specifically, the two input feature maps undergo both global average pooling and global max pooling to obtain tensors $U \in \mathbb{R}^{1\times1\times C}$ and $W \in \mathbb{R}^{1\times1\times C}$, which respectively capture the compressed global spatial information across each channel and enhance the channel-wise feature response. The resulting four feature vectors are concatenated and passed through an MLP to promote deep interaction and fusion of multimodal feature representations. As shown in Figure 4c, the SFI module comprises a convolutional layer followed by a sigmoid activation function. After the channel-wise interaction, the two aligned feature maps are concatenated and passed through a 1×1 convolution layer to generate a projection tensor $V \in \mathbb{R}^{H\times W\times1}$, which strengthens spatial responses across the fused feature maps and facilitates spatial-level interaction of multimodal features.

3.2. Aerial Imagery and LiDAR Point Cloud Registration

Due to sensor system errors (such as camera exposure delay, calibration residuals, and insufficient GPS/IMU observation quality control), there is typically a significant positional deviation between directly georeferenced aerial imagery and ALS data [36]. This study designed a two-stage 3D-3D registration strategy to ensure the accuracy of cross-modal semantic transfer. First, an MPC is generated from the UAV imagery, followed by noise filtering to remove outliers and obtain an optimized 3D point set. Then, the TCF algorithm

is employed to achieve coarse alignment between the MPC and ALS data. Finally, an improved ICP algorithm is applied to perform sub-pixel-level fine registration, addressing spatial inconsistencies across modalities and enhancing the spatial coupling precision between imagery and LiDAR point clouds. This provides a geometrically consistent foundation for subsequent semantic label transfer. The proposed registration framework consists of two main stages: coarse registration and fine registration.

(1) Coarse registration of point clouds based on the TCF algorithm

The Random Sample Consensus (RANSAC) algorithm is a widely used method for coarse point cloud registration [37]. Its core principle involves iteratively selecting a minimal set of randomly sampled point correspondences—typically three non-collinear pairs—to estimate a candidate rigid transformation consisting of a rotation matrix $\boldsymbol{R}$ and a translation vector $\boldsymbol{T}$. An inlier verification mechanism is then applied by evaluating whether the Euclidean distances between the transformed source points and the target points fall below a predefined threshold. The main advantage of RANSAC lies in its robustness to outliers in feature matching. By leveraging random sampling and consistency checking, the algorithm can tolerate more than 50% incorrect correspondences while still estimating a globally optimal coarse transformation. This provides a reliable initialization for subsequent fine registration, helping to avoid convergence to local minima and improving overall registration success rates.

To address the computational inefficiency and mismatch risks associated with conventional RANSAC algorithms in point cloud registration—especially under the presence of outliers arising from acquisition discrepancies between MPC and ALS data—this study introduces TCF approach to optimize coarse registration [38], thereby providing a high-quality initialization for subsequent fine alignment. Traditional RANSAC relies on purely random sampling during its hypothesis generation phase, resulting in exponentially increasing iterations and limited robustness under non-rigid deformation, as its distance-based inlier verification lacks geometric adaptability. In contrast, the proposed TCF method enhances both efficiency and accuracy through a dimension-reduction strategy, as formulated in Equation (1). In the first stage, a length-constrained single-point sampling scheme is employed to eliminate large-scale outliers. In the second stage, a two-point sampling strategy evaluates angular consistency to retain only high-confidence correspondences. Finally, the transformation parameters—including the rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{T}$—are robustly estimated using a scale-adaptive Cauchy Iterative Reweighted Least Squares (IRLS) solver. The workflow of the TCF algorithm is shown in Figure 5.

$$N_{ransac} = \left\lceil \frac{\log(1-\lambda)}{\log(1-(1-\vartheta)^{s})} \right\rceil \tag{1}$$

where s denotes the sampling dimension, ϑ denotes the proportion of outliers, N_{ransac} denotes the number of iterations, and λ denotes the success probability, which is generally 0.99.

(2) Point cloud fine registration based on the ICP algorithm

With the initial transformation parameters estimated by the TCF algorithm, the MPC and ALS point clouds can be roughly aligned. However, due to differences in features and scale between MPC and ALS point clouds, obtaining high-precision registration results still faces several challenges: (1) MPC and ALS feature distributions are uneven, with systematic biases in local regions (especially tree trunks, understory vegetation, and the ground), leading to misalignment in local point cloud registration; (2) Scale uncertainty (scale drift) in MPC causes minor scale differences between MPC and ALS; (3) Residual

point cloud noise and outliers from TCF, as well as registration errors caused by differences in sensor perspectives and occlusions.

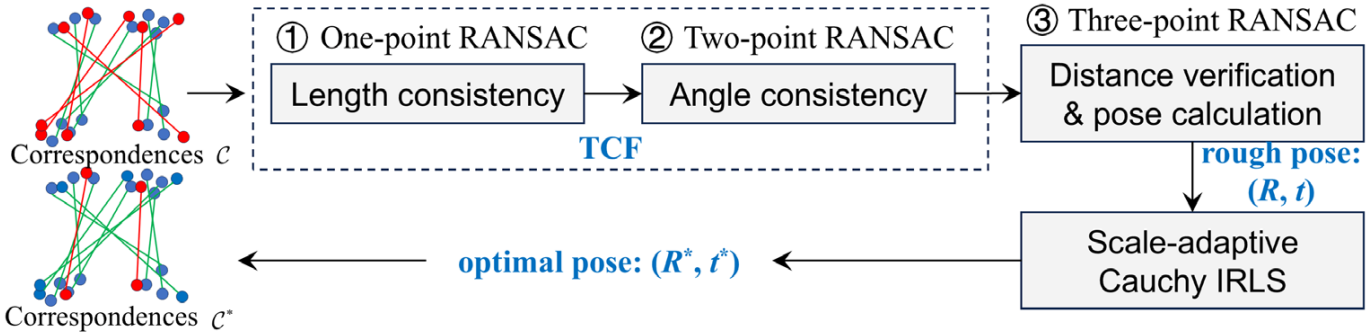


Figure 5. Registration pipeline between MPC and ALS point clouds. The red points in the figure represent mismatches, while the blue points represent correct matches.

To address local deformations and subtle residual deviations after coarse registration, and to further refine the rigid transformation parameters, this study employs the ICP algorithm to optimize the alignment between MPC and ALS point clouds. The ICP algorithm iteratively estimates the optimal rigid transformation by minimizing the Euclidean distance between corresponding point pairs, considering six degrees of freedom (three translations and three rotations), as defined by the objective function in Equation (2). During each iteration, the algorithm dynamically suppresses or removes outliers and searches for nearest neighbors across the entire scene, thereby improving the global spatial consistency. This process results in a registration outcome with significantly enhanced accuracy and robustness.

$$\text{argmin}\{\frac{1}{N}\sum_{i=1}^{N}\|q_i - Rp_i - t\|^2\} \tag{2}$$

where N represents the number of corresponding point pairs, and q and p represent the corresponding points in the two sets of point clouds.

The results of coarse and fine registration of point clouds are shown in Figure 6. After coarse alignment using the RANSAC algorithm, the MPC and ALS point clouds were roughly aligned; however, local deviations remained, as highlighted in the zoomed-in views. Subsequent application of the TCF algorithm produced significantly more accurate registration. As shown in Figure 6b, both global and local views indicate no noticeable misalignment between the MPC and ALS point clouds. In key regions such as tree canopies, understory trunks, and ground surfaces, the proposed two-stage registration strategy effectively eliminated prominent local deviations. These results demonstrate that the proposed approach achieves accurate alignment between MPC and ALS point clouds in complex forest environments, exhibiting high registration precision.

3.3. Cross-Modal Semantic Transfer

In Section 3.1, the proposed MSFFNet model is used for instance segmentation of tree crowns from multi-view UAV imagery. However, in multi-view images, multiple instance masks from different viewpoints may correspond to the same physical tree crown. To accurately segment individual trees in the 3D point cloud, it is essential to identify 2D masks that belong to the same tree and associate them with their corresponding regions in 3D space. This task is challenging due to segmentation errors in the 2D domain—such as tree crowns being correctly segmented in some views but over-segmented or falsely detected in others. Examples of such inconsistencies are illustrated in Figure 7.

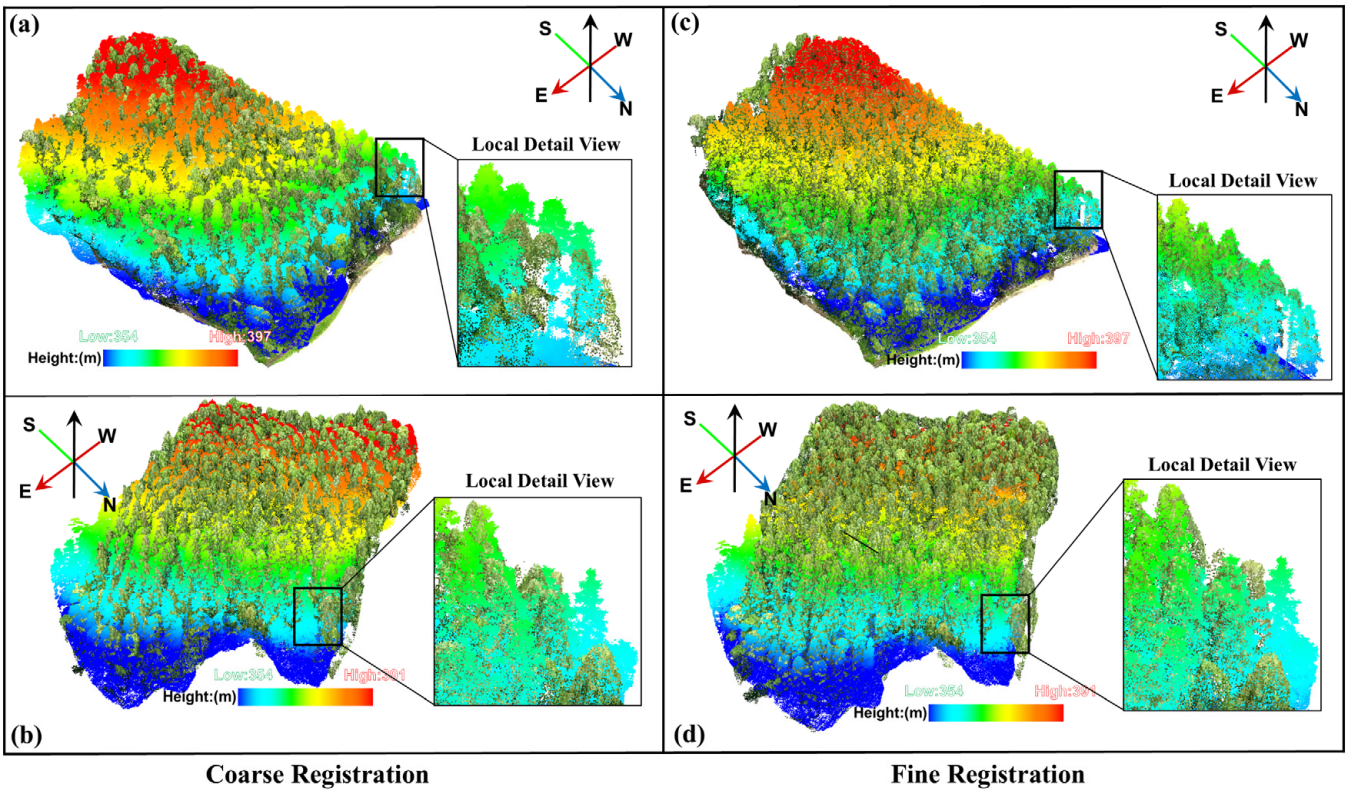


Figure 6. Results of coarse and fine registration of point clouds. (**a**,**b**) represent the coarsely registered point clouds of the Forest 1 and Forest 2 regions, respectively, while (**c**,**d**) depict the finely registered point clouds for the same regions.

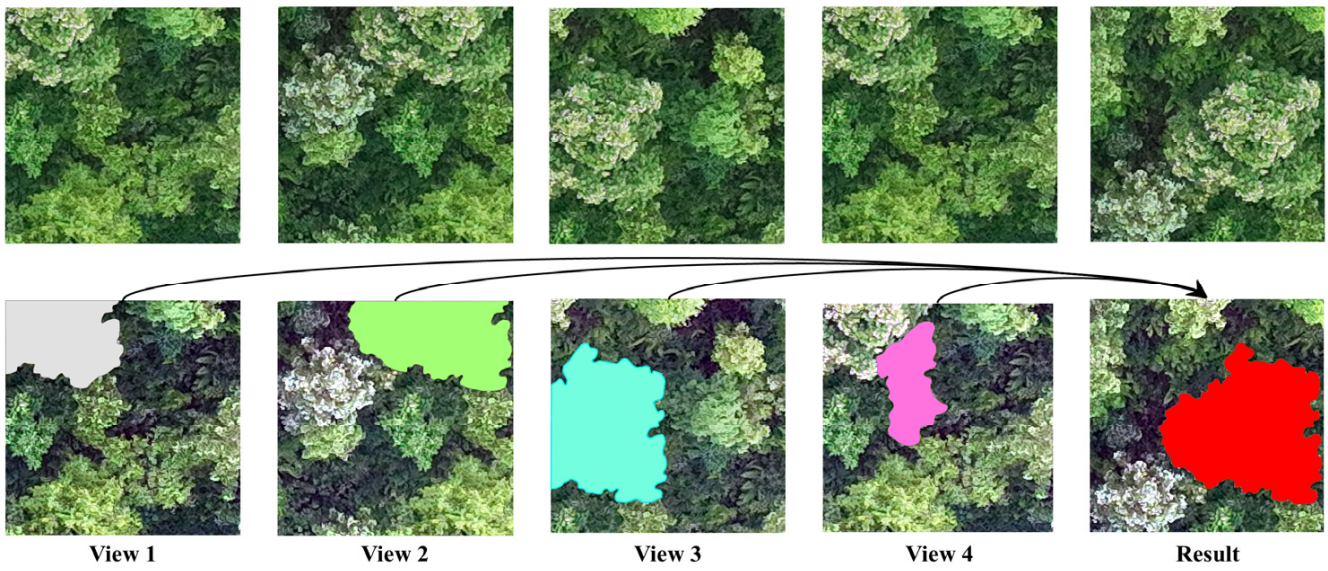


Figure 7. Illustrations of incorrect or erroneous instance segmentation.

To address this challenge, this study combines multi-view geometry with an expectation–maximization (EM) algorithm based on a semantic probability model to construct a semantic probability field for multi-view instance masks and LiDAR point clouds, thereby resolving inconsistencies in instance masks for the same tree across different viewpoints. The algorithm iteratively optimizes the association probabilities between masks and 3D point clouds within the EM framework, explicitly modeling the uncertainty in segmentation results. Let $\left\{\mathrm{LIM}_k^I\right\}$ denote the kth local instance mask in image $\mathcal{I}$. We

iterate over all masks LIM_k^I in all images and create an independent instance container $\mathcal{C} = \{\mathcal{C}_k\}$ for each mask, where $\mathcal{C}_k = \{(\mathcal{I}, k)\}$ denotes the container for the kth mask instance in image. The algorithm primarily consists of five steps (the workflow is shown in Figure 8): initial semantic probability association, expectation step, maximization step, iteration control, and instance generation.

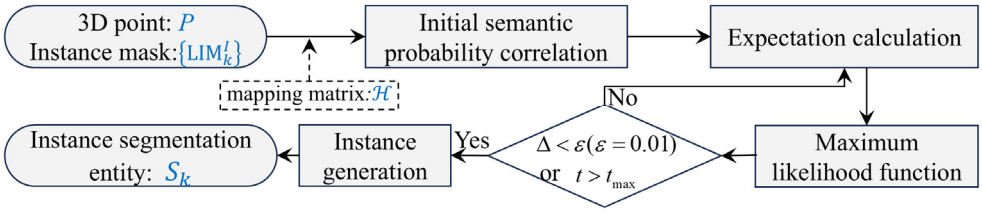


Figure 8. Workflow of cross-modal semantic transfer.

(1) Initial semantic probability association

First, we establish an initial semantic probability association between 3D points and 2D masks. For each 3D point p_i, we use the mapping relationship $\mathcal{H}$ provided by the SfM system to determine all mask instances $\mathcal{V}_i = (I,k) \left| \mathrm{proj}_I(p_i) \in \mathrm{LIM}_k^I \right.$ that cover p_i, calculate the number of coverage viewpoints $n_i = |\mathcal{V}_i|$, and establish the association between points and masks based on n_i:

$$P^{(0)}(p_i \in \mathcal{C}_k) = \begin{cases} \frac{1}{|\mathcal{V}_i|} & if \ \ (\mathcal{I},k) \in \mathcal{V}_i \\ 0 & otherwise \end{cases} \tag{3}$$

(2) Expectation step

In the expectation step of iterative optimization, we introduce multi-view spatial consistency constraints for probability re-estimation based on the semantic probability distribution of the 3D point cloud–mask association in the current round t. The core objective of this step is to solve the misassociation problem caused by the initial uniform distribution by integrating the spatial compatibility of three-dimensional space and two-dimensional images. Specifically, we design two key geometric factors (visibility verification factor $w_{ik}^{(I)}$ and spatial proximity weight $cons_{ik}^{(I)}$) to jointly construct the semantic probability update weights:

$$P^{(t)}(p_i \in \mathcal{C}_k) = \frac{\sum\limits_{(I,k)\in\mathcal{C}_k} w_{ik}^{(I)} \cdot cons_{ik}^{(I)}}{\sum\limits_{m} \sum\limits_{(J,m)\in\mathcal{C}_m} w_{im}^{(J)} \cdot cons_{im}^{(J)}} \tag{4}$$

$$w_{ik}^{(I)} = \begin{cases} 1 & if \quad pi \quad visible \quad in \quad I \\ 0 & otherwise \end{cases} \tag{5}$$

$$cons_{ik}^{(I)} = \exp\left(- \frac{\left\| proj_I(p_i) - centroid\left(\mathrm{LIM}_k^I\right) \right\|^2}{\left(0.05 \times \sqrt{W_I^2 + H_I^2}\right)^2} \right) \tag{6}$$

where $w_{ik}^{(I)}$ represents the visibility judgment of a 3D point, which is mainly determined through the mapping relationship $\mathcal{T}$ of SfM; $cons_{ik}^{(I)}$ represents the spatial consistency between the 3D point and the 2D mask; W and H represent the width and height of the image, respectively; $proj_I(p_i)$ represents the projection point p_i of the point in the image; and $centroid\left(\mathrm{LIM}_k^I\right)$ represents the centroid of the kth mask in the image $\mathcal{I}$. The physical

meaning of Formula (6) is that the closer the 3D point is to the center of the mask, the higher the spatial consistency.

(3) Maximization steps

Based on the geometrically optimized semantic probability, in order to eliminate redundant canopy masks from different perspectives (i.e., the same canopy is divided into multiple fragmented instances in different images), we perform hierarchical clustering of mask instances using semantic probability distribution similarity metrics.

For each pair of mask instances $(\mathcal{C}_a, \mathcal{C}_b)$, calculate the distribution vectors $\vec{P_a} = [P(p_1 \in \mathcal{C}_a), P(p_2 \in \mathcal{C}_a), \ldots]$ and $\vec{P_b} = [P(p_1 \in \mathcal{C}_b), P(p_2 \in \mathcal{C}_b), \ldots]$ of the corresponding 3D points. The formula for calculating the cosine similarity is:

$$S_{ab} = \frac{\sum_i P(p_i \in \mathcal{C}_a) P(p_i \in \mathcal{C}_b)}{\sqrt{\sum_i P^2(p_i \in \mathcal{C}_a)} \cdot \sqrt{\sum_i P^2(p_i \in \mathcal{C}_b)}} \tag{7}$$

The principle of the hierarchical clustering algorithm is presented in Algorithm 1. The algorithm first constructs a similarity matrix $[S_{ab}]_{K \times K}$ and finds the pair $(\mathcal{C}_{a^*}, \mathcal{C}_{b^*})$ with the highest similarity. It then uses a threshold τ to determine whether to merge two mask instances. Compared with fixed threshold segmentation, hierarchical clustering can adaptively handle differences in tree crown size. For large tree crowns, it allows merging with lower similarity (to accommodate distribution diversity), while for small tree crowns, strict matching is required.

Algorithm 1 Hierarchical clustering algorithm

Input: Similarity matrix $[S_{ab}]_{K \times K}$, mask instance $\mathcal{C}$
Initialize: $\tau = 0.9$, $\mathrm{S_{init}} = 1$

While $\mathrm{S_{init}} = 1$ **do**
  Find the maximum similarity mask instance pair
  $(a^*, b^*) = \underset{a \neq b}{\mathrm{argmax}} S_{ab}$
  if $S_{a^* b^*} > \tau$ **then**
    $\mathcal{C}_{new} = \mathcal{C}_{a^*} \cup \mathcal{C}_{b^*}$
    $P_{temp}(p_i \in \mathcal{C}_{new}) = \left(P^{(t)}(p_i \in \mathcal{C}_{a^*}) + P^{(t)}(p_i \in \mathcal{C}_{b^*}) \right) / 2$
  elif $\max S_{ab} < \tau$ **do**
    $\mathrm{S_{init}} = 0$
  end if
end while

Output: Mask instance merge results $\mathcal{C}^*$

(4) Iterative control

After hierarchical clustering merging, since each instance merging changes the association structure between 3D points and masks, it is necessary to dynamically evaluate the optimization process. We use Formula (8) to reallocate the association semantic probability between 3D points and 2D masks, and use Formula (9) to calculate the probability change Δ. If $\Delta < \varepsilon (\varepsilon = 0.01)$ or $t > t_{\max}$, stop iteration; if $\Delta > \varepsilon (\varepsilon = 0.01)$, return to step (2) and continue iteration optimization.

$$P^{(t)}(p_i \in \mathcal{C}_{new}) = \frac{\sum_{k \in \mathcal{C}_{new}} P^{(t)}(p_i \in \mathcal{C}_k)}{|\mathcal{C}_{new}|} \tag{8}$$

$$\Delta = \frac{1}{|\mathcal{P}|}\sum_{i}\sum_{k}\left|P^{(t)}(p_i \in \mathcal{C}_k) - P^{(t-1)}(p_i \in \mathcal{C}_k)\right| \tag{9}$$

(5) Instance generation

Finally, after completing multiple rounds of expectation–maximization iterative optimization, we decoupled the soft probabilistic associations into hard spatial divisions, achieving semantic transfer from a continuous probability field to a discrete instance of a 3D point cloud, and ultimately outputting a 3D tree canopy instance segmentation entity with spatial continuity and semantic consistency.

$$S_k = \left\{ p_i | k = \underset{m}{\operatorname{argmax}} \left(\sum_{(I,m)\in\mathcal{C}_m} P(p_i \in \mathcal{C}_m) \right) \right\} \tag{10}$$

This study integrates multi-view geometry with an EM algorithm based on semantic probability modeling to construct a semantic probability field that aligns multi-view instance masks with LiDAR point clouds. Through multi-stage iterative optimization, the proposed method achieves robust semantic transfer from 2D tree crown masks to 3D point clouds, while addressing inconsistencies in instance segmentation across different viewpoints. The algorithm first establishes initial probabilistic correspondences using SfM projections. It then introduces multi-view spatial consistency constraints—such as visibility validation and spatial proximity weighting—to dynamically refine the probability distribution. A hierarchical clustering approach based on distributional similarity is employed to adaptively merge fragmented instance masks belonging to the same tree, with convergence controlled by monitoring the evolution of the probability field. Finally, a maximum a posteriori decision rule is applied to convert the continuous semantic probability field into discrete 3D tree crown instances, thereby enabling structured semantic-to-point-cloud migration and providing a quantifiable 3D topological foundation for individual tree segmentation.

3.4. Implementation and Evaluation Metrics

All experiments in this study were conducted on a computer equipped with an NVIDIA GeForce RTX 3060 GPU (Nvidia Corporation, Santa Clara, CA, USA) and an Intel i5-12600kf CPU (Intel Corporation, Santa Clara, CA, USA). The training of the 2D single-tree instance segmentation network was performed using the Python 3.7 programming environment and the PyTorch (Version 1.12.0) deep learning framework, with the AdamW optimizer (initial learning rate of 0.001). The batch size and number of epochs were set to 8 and 60, respectively, with a Poly learning rate decay strategy. To mitigate the impact of limited training data and accelerate convergence, the model was pre-trained on the ImageNet dataset [39]. The pre-trained weights were then used to initialize the MSFFNet network. For the single-tree instance segmentation task in forest scenes, the number of classification categories is set to 1, and the confidence threshold is set to 0.7. Running the complete individual tree segmentation pipeline on our 40 m $\times$ 107 m and 30 m $\times$ 70 m benchmark dataset (roughly 3.2 million points) took about approximately 190 min. This includes MSFFNet model training (180 min), MSFFNet inference (5 min), point cloud registration (1 min), and semantic mapping (4 min). The runtime scales approximately linearly with the area of the study region. As research code, our implementation is not optimized for minimal runtime or resource usage.

To validate the effectiveness of the MSFFNet network proposed in this paper for 2D single-tree instance segmentation in complex mountainous scenes, we use overall

accuracy, precision, recall, F1-score, and intersection-over-union (IoU) as evaluation metrics. The formulas are as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{11}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{14}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{15}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

In addition, to validate the effectiveness of the cross-modal label transfer method proposed in this paper, we use segmentation accuracy (SAC), omission error (OME), and commission error (COE) as evaluation metrics for 3D single tree segmentation. These three metrics are widely used in the evaluation of 3D point cloud single tree segmentation [40]. Their calculation formulas are as follows:

$$\mathrm{SAC} = \frac{v_{cs}}{v_{gt}} \tag{16}$$

$$\mathrm{OME} = \frac{v_{us}}{v_{gt}} \tag{17}$$

$$\mathrm{COE} = \frac{v_{fs}}{v_{gt}} \tag{18}$$

where v_{cs} represents the value of correctly segmented tree point clouds, v_{us} represents the value of unsegmented tree point clouds, v_{fs} represents the value of incorrectly segmented tree point clouds, and v_{gt} represents the value of ground truth tree point clouds.

4. Experimental Results and Analysis

4.1. Accuracy of Individual Tree Crown Delineation

In this study, we evaluated the instance segmentation accuracy of the MSFFNet model using 260 tree crown samples distributed across two regions. As shown in Table 2, the MSFFNet model achieved consistently high accuracy across various backbone configurations in both regions, with all metrics—including overall accuracy, precision, recall, F1-score, and IoU—exceeding 85%. Among them, the MSFFNet with the MiT-B4 backbone achieved the best performance, attaining an IoU of 87.60%. In contrast, the model using the MiT-B0 backbone yielded the lowest accuracy, primarily due to its limited model depth, which resulted in suboptimal feature extraction from the CHM and RGB imagery. Consequently, it failed to effectively enhance the deep feature interaction and fusion in the MLFF module. These experimental results demonstrate that increasing model depth significantly improves crown instance segmentation accuracy and further highlight the critical role of the MLFF module in the overall architecture.

Table 2. Tree crown instance segmentation results of MSFFNet using different backbone networks and datasets.

Dataset	Model	OA	Precision	Recall	F1-Score	IoU
Forest 1	MiT-B1	0.9073	0.8725	0.8838	0.8797	0.8651
	MiT-B2	0.9129	0.8746	0.8863	0.8821	0.8689
	MiT-B3	0.9154	0.8770	0.8894	0.8856	0.8714
	MiT-B4	0.9195	0.8817	0.8953	0.8907	0.8760
Forest 2	MiT-B1	0.9055	0.8684	0.8801	0.8724	0.8586
	MiT-B2	0.9081	0.8712	0.8837	0.8751	0.8622
	MiT-B3	0.9092	0.8722	0.8851	0.8785	0.8659
	MiT-B4	0.9110	0.8757	0.8896	0.8836	0.8714

The qualitative results of the MSFFNet model are shown in Figure 9. The segmentation performance is evaluated using the IoU metric, where tree crown instances with IoU > 0.5 are considered correctly segmented (highlighted with blue bounding boxes), while undetected crowns are defined as missed segments (highlighted with red bounding boxes). The model based on the MiT-B3 backbone produced fewer mis-segmented and missed instances, demonstrating superior segmentation performance compared to models with other backbones. In contrast, the MiT-B0-based model exhibited the poorest performance, with a notably higher number of missed detections, resulting in reduced accuracy. It is worth noting that segmentation errors and omissions were more pronounced in the Forest 2 region, where tree crowns are denser, more closely spaced, and exhibit greater overlap. In comparison, the Forest 1 region achieved higher segmentation accuracy due to its relatively sparse and distinct tree crown distribution.

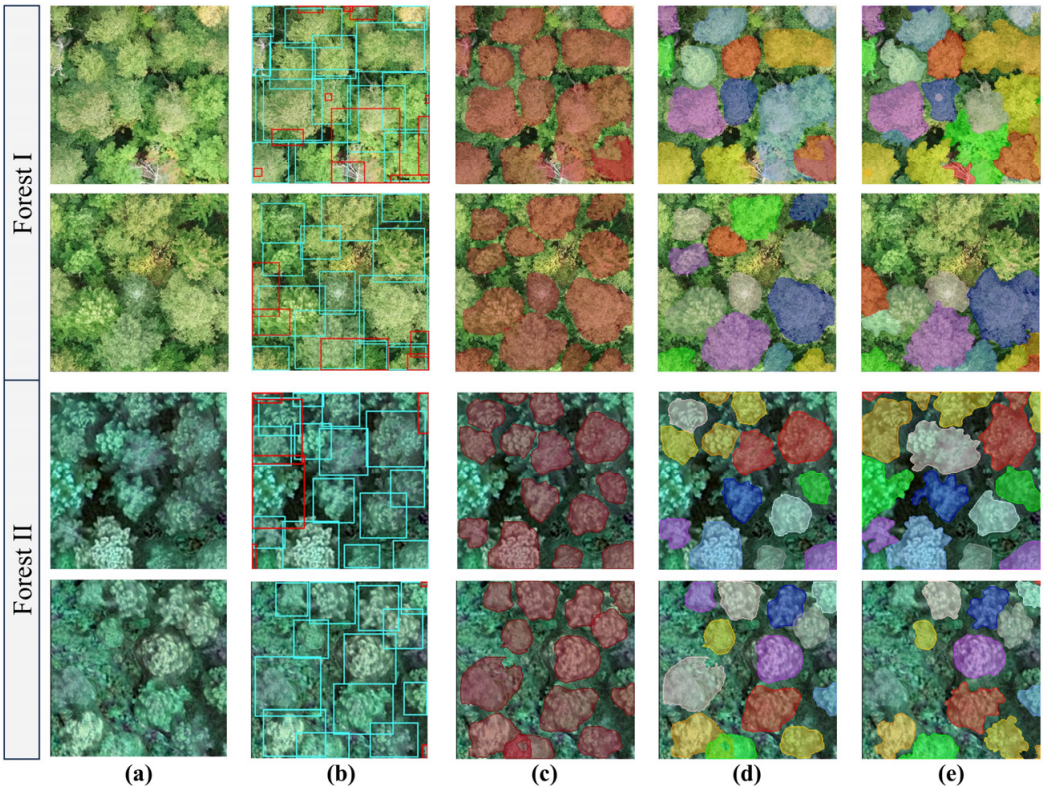


Figure 9. Individual tree instance segmentation results generated by the proposed MSFFNet model. (**a**–**e**) represent the original RGB image, object detection results, semantic segmentation results, instance segmentation results, and instance segmentation labels, respectively.

To further evaluate the performance of the proposed instance segmentation model, we conducted comparative experiments with three state-of-the-art instance segmentation methods: Mask R-CNN, SOLOV2, and Mask2Former. As shown in Table 3, the proposed MSFFNet achieved the highest instance segmentation accuracy, with an IoU of 87.60%. In contrast, Mask R-CNN yielded the lowest segmentation accuracy for tree crowns. This can be attributed to its RoI Align operation, which, while preserving spatial information, still relies heavily on local features for pixel assignment within overlapping regions, making it difficult to accurately distinguish adjacent crown boundaries. MSFFNet addresses this limitation through the MLFF module, which enables deep interaction between the CHM and visible imagery. By effectively extracting complementary features and performing iterative cross-modality feature refinement across layers, MSFFNet significantly enhances tree crown detection and boundary delineation capabilities. Notably, MSFFNet outperformed Mask R-CNN, SOLOV2, and Mask2Former in terms of recall, F1-score, and IoU, with the maximum IoU improvements reaching 3.32%, 2.59%, and 1.60%, respectively.

Table 3. Comparison between MSFFNet and state of the art instance segmentation methods.

Dataset	Model	Precision	Recall	F1-Score	IoU
Forest 1	Mask R-CNN	0.8650	0.8657	0.8596	0.8428
	SOLOV2	0.8719	0.8755	0.8657	0.8501
	Mask2Former	0.8784	0.8859	0.8781	0.8611
	MSFFNet (ours)	0.8817	0.8953	0.8907	0.8760
Forest 2	Mask R-CNN	0.8618	0.8591	0.8554	0.8393
	SOLOV2	0.8652	0.8732	0.8628	0.8480
	Mask2Former	0.8690	0.8787	0.8713	0.8554
	MSFFNet (ours)	0.8757	0.8896	0.8836	0.8714

Figure 10 presents the qualitative results of tree crown instance segmentation obtained by different models. The results indicate that the proposed MSFFNet model yields fewer mis-segmented and missed instances and demonstrates superior capability in distinguishing tree crown boundaries. In the third and fourth rows, the tree crowns exhibit similar spectral characteristics, posing challenges for boundary discrimination. MSFFNet achieves precise boundary segmentation by incorporating the spatial structural information from the CHM, whereas Mask R-CNN, SOLOV2, and Mask2Former fail to effectively separate such spectrally similar yet distinct objects. In contrast, the tree crowns in the first and second rows can be accurately segmented using the texture and spectral features from the visible imagery, highlighting the importance of complementary information from multi-source data. Overall, Mask R-CNN is limited by its convolutional architecture's ability to capture long-range dependencies, resulting in noticeable mis-segmentation and missed detections. In comparison, both Mask2Former and the proposed method exhibit strong global contextual understanding. However, the proposed MSFFNet shows a clear advantage in detecting smaller tree crowns, as illustrated in the second row of Figure 10.

4.2. Ablation Studies

To evaluate the effectiveness of the MSFFNet architecture, we conducted ablation experiments on the model. As shown in Table 4, when both the FAM and FFM modules are used together, MSFFNet achieves the highest accuracy in instance canopy segmentation (F1-score = 89.07%, IoU = 87.60%). In contrast, when the FAM and FFM modules were removed separately, the model's IoU accuracy decreased by 0.48 and 0.74, respectively, indicating that both modules significantly contributed to improving the model's segmentation performance. Additionally, when the MLFF module (FAM + FFM) was not used, the model

achieved the lowest canopy segmentation accuracy, further validating the effectiveness of the MLFF module.

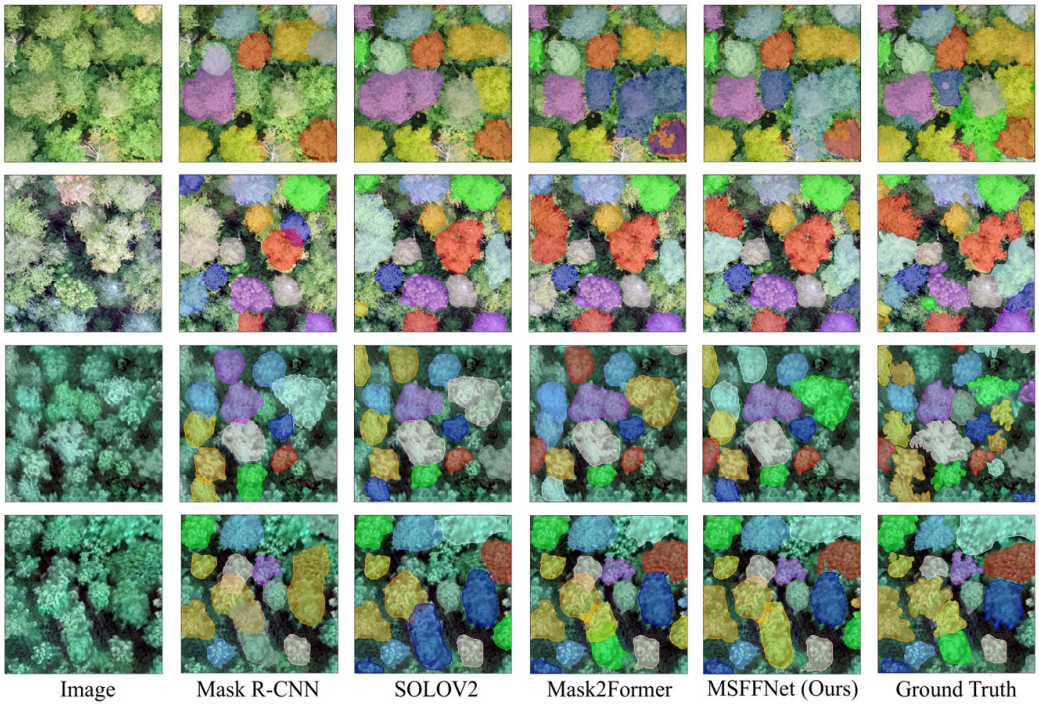


Figure 10. Comparison of canopy segmentation results from different models.

Table 4. MSFFNet model ablation experiment results.

Model	FAM	FFM (CFI + PFI)	F1-Score	IoU
MSFFNet	✗	✗	0.8754	0.8629
	✗	✗	0.8843	0.8686
	✗	✓	0.8851	0.8712
	✓	✓	0.8907	0.8760

4.3. 3D Point Cloud Single-Tree Instance Segmentation Results

Canopy instance segmentation and point cloud registration serve as critical prerequisites for the proposed cross-modal semantic label transfer method, whose ultimate goal is the individual tree segmentation of 3D point clouds based on cross-modal semantic information. To evaluate the effectiveness of the proposed cross-modal semantic transfer framework in complex forest environments, experiments were conducted on two regions, Forest 1 and Forest 2 (as shown in Figure 11). The point clouds of both forest scenes are color-coded based on elevation values (Figure 11(a1,a2)). After separating ground points, non-ground points are colored according to their semantic labels (Figure 11(b1,b2)), where tree points are marked in green and non-tree points in gray. The results of individual tree point cloud segmentation are presented in Figure 12. The proposed method achieved accurate segmentation for trees with varying morphologies and complex spatial distributions, indicating its effectiveness for individual tree segmentation tasks in forest environments. It is worth noting that the segmentation accuracy in Forest 2 is relatively lower than in Forest 1 (see Table 5), primarily due to the higher degree of crown overlap in Forest 2, which increases the difficulty of accurate boundary delineation.

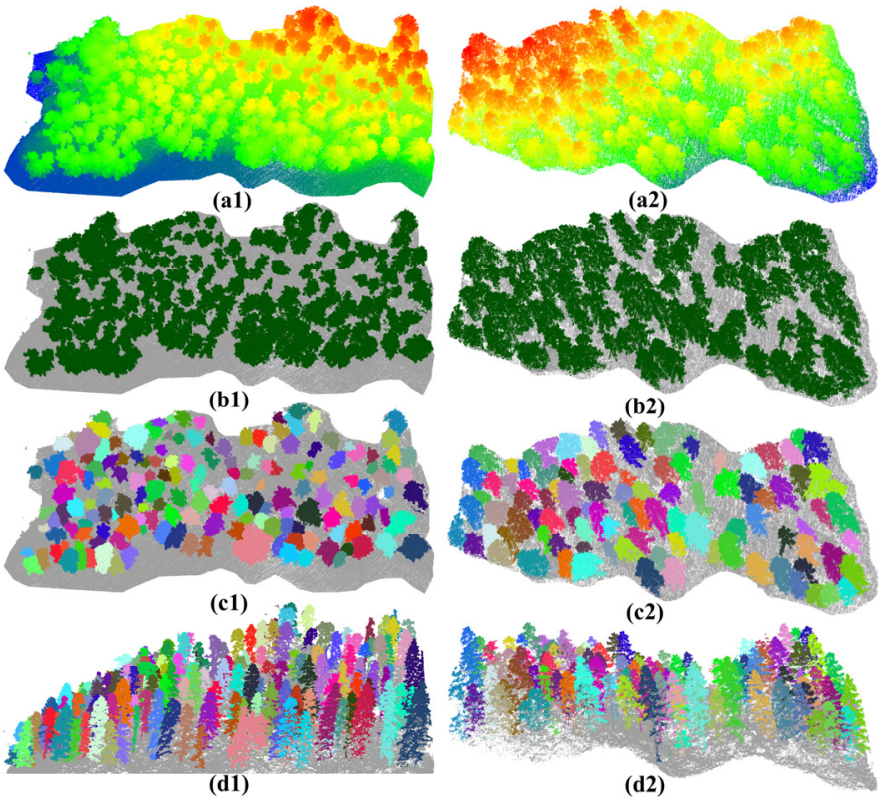


Figure 11. Individual tree segmentation of Forest 1 and Forest 2. (**a1**–**d1**) and (**a2**–**d2**) represent the elevation-rendered point cloud results, binary classified point cloud results, top-down views of individual tree point cloud segmentation results, and side views of individual tree point cloud segmentation results for Forest 1 and Forest 2, respectively.

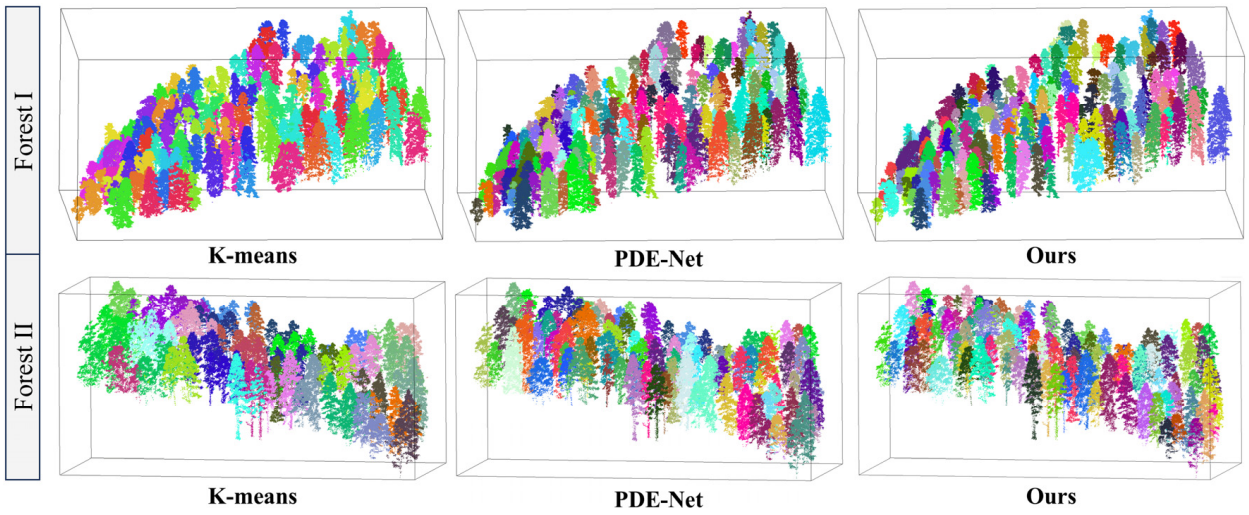


Figure 12. Semantic segmentation results of individual tree point clouds on the Forest 1 and Forest 2 datasets.

To quantitatively evaluate the performance of individual tree point cloud semantic segmentation, all tree point clouds were manually and meticulously segmented using CloudCompare, serving as the ground truth reference. In addition, to assess the effectiveness of the proposed method, we compared it with two widely adopted baseline approaches: K-means [41] and PDE-Net [42]. As shown in Table 5, the proposed method outperformed all comparison methods across three evaluation metrics: SAC, OME, and COE. It is noteworthy that the SAC value for Forest 1 was higher than that for Forest

2, which can be attributed to the relatively simpler spatial distribution of trees in Forest 1. Overall, the proposed method demonstrated superior performance in individual tree semantic segmentation tasks, confirming its strong robustness in handling complex mountainous forest environments.

Table 5. Comparison of the performance of our method and other methods in 3D point cloud individual tree instance segmentation.

Method	Metric	Forest 1	Forest 2	Average SAC
K-means	SAC	0.85	0.84	0.84
	OME	0.32	0.36	
	COE	0.33	0.38	
PDE-Net	SAC	0.88	0.86	0.87
	OME	0.25	0.29	
	COE	0.29	0.33	
Ours	SAC	0.92	0.90	0.91
	OME	0.18	0.20	
	COE	0.15	0.17	

The individual tree point cloud semantic segmentation results obtained by different methods are illustrated in Figures 13 and 14. Multi-view qualitative evaluations demonstrate that the proposed method is capable of effectively segmenting most trees across various forest scenarios. In contrast, the K-means algorithm exhibited the weakest performance due to its sensitivity to noise and outliers, as well as its reliance on a predefined number of clusters. PDE-Net showed comparable performance to the proposed method in segmenting tree crown boundary points; however, in regions with high point cloud overlap, the proposed method outperformed PDE-Net, owing to its superior ability to model topological connectivity within the point cloud. Furthermore, as shown in the enlarged views of Figure 13a,b, for small-sized tree crowns, both K-means and PDE-Net often produce incorrect segmentation results, whereas the proposed method achieved precise segmentation. Overall, the proposed approach demonstrated consistently strong performance across all aspects of individual tree point cloud semantic segmentation.

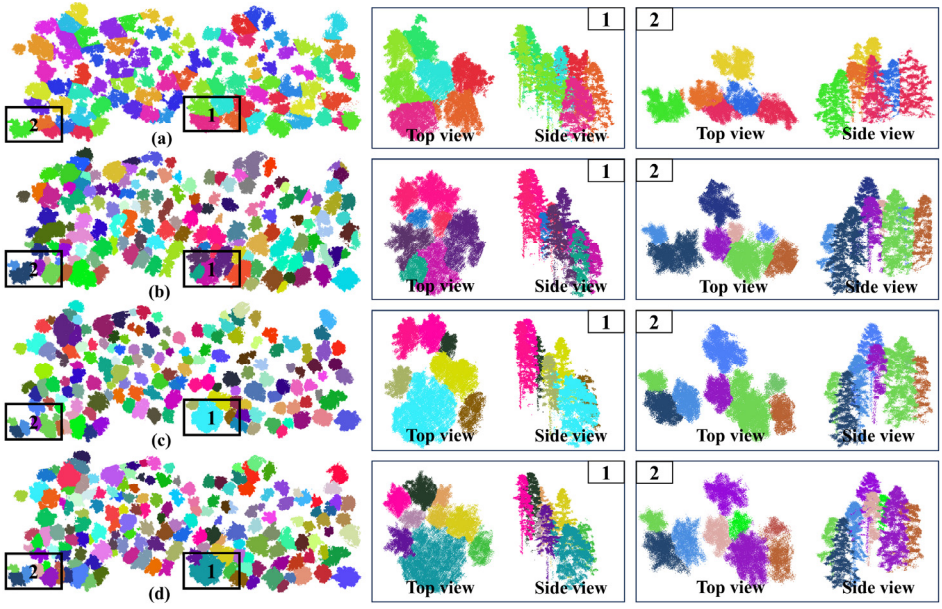


Figure 13. Semantic segmentation results of individual tree point clouds on the Forest 1 dataset. (**a**–**d**) represent the results obtained by the K-means method, the PDE-Net method, the proposed method, and the ground truth, respectively.

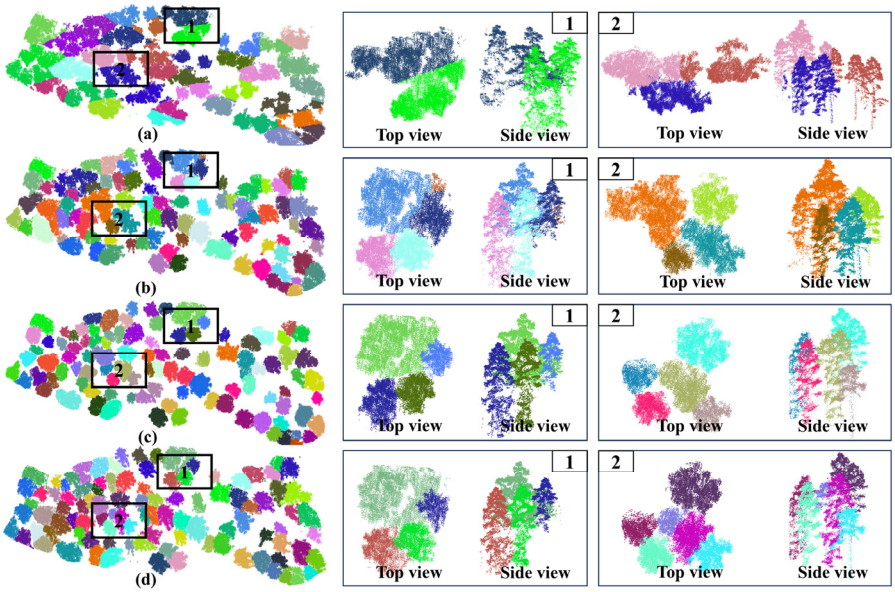


Figure 14. Semantic segmentation results of individual tree point clouds on the Forest 2 dataset. (**a**–**d**) represent the results obtained by the K-means method, the PDE-Net method, the proposed method, and the ground truth, respectively.

5. Discussion

This study proposes a cross-modal semantic transfer framework for individual tree segmentation in complex forest environments, combining UAV imagery and LiDAR data to achieve 3D point cloud parsing without manual annotation. The effectiveness of the method was validated on two subtropical forest datasets, where the MSFFNet model achieved the best performance in 2D instance segmentation (IoU: 87.60%), and the semantic transfer framework significantly outperformed existing methods in 3D point cloud segmentation (SAC: 0.92, OME: 0.18, COE: 0.15). These results demonstrate the framework's ability to address key challenges in forestry remote sensing, such as high annotation costs and cross-domain generalization limitations. Its outstanding performance stems from the synergy of three core components: (1) MSFFNet integrates hierarchical features from RGB and CHM data through deformable alignment and cross-modal interaction (MLFF module), enhancing boundary discrimination in dense canopies. An ablation study confirmed its critical role, as removing the MLFF module caused a notable IoU decrease; (2) Two-stage registration (TCF+ICP) resolves spatial inconsistencies between MPC and LiDAR point clouds, reducing alignment errors from meter-level to sub-pixel accuracy. This ensures geometric consistency in cross-modal mapping, which is particularly important for understory trees where LiDAR returns are sparse; (3) Semantic probability association combines multi-view geometry with EM optimization to address segmentation inconsistencies across viewpoints. By modeling spatial proximity and visibility constraints, the framework achieves robust label fusion, reducing omission errors by 31% compared to PDE-Net.The proposed framework has practical value for forest management. By transferring 2D semantic information into 3D point clouds, it enables efficient extraction of tree structural parameters (e.g., crown volume, tree height) for biodiversity monitoring, timber yield estimation, carbon stock assessment, and ecological interaction analysis. Its runtime (roughly 190 min for 3.2 million points) scales linearly with area, indicating potential for large-scale monitoring. However, several limitations remain. In crown instance segmentation, challenges such as crown overlap, texture similarity among tree species, forest type variations, and different climatic conditions still prevent MSFFNet from extracting some smaller crowns. The method also does not handle understory tree segmentation—a persistent challenge for 2D image-based

forest surveys. In point cloud registration, segmentation accuracy depends on registration precision; while end-to-end deep learning approaches could avoid this dependency, they incur much higher computational costs and require large annotated forest point cloud datasets, which are currently scarce. Furthermore, in scenarios involving diverse sensor types and point cloud generation methods, unifying coordinate systems remains an open research topic.

The proposed 2D–3D semantic mapping strategy efficiently transfers per-crown instance semantics into 3D space, avoiding the heavy annotation burden required by deep learning-based approaches. Nevertheless, the method has not yet been tested in deciduous or tropical broadleaf forests with seasonal foliage variation. Future work will focus on two directions: (1) Optimizing runtime through lightweight network design and parallel computing; (2) Integrating topological constraints to improve segmentation of overlapping crowns.

6. Conclusions

In this study, we propose a cross-modal semantic transfer framework for individual tree segmentation from complex mountainous forest point clouds, enabling large-scale instance-level tree segmentation without the need for extensive manual annotations. The framework integrates 2D semantic extraction with a 3D semantic mapping strategy to reduce dependency on large amounts of labeled data. A multi-source feature fusion model is developed for tree crown instance segmentation, which effectively captures both spectral and spatial features to extract accurate semantic information for each crown. A cross-modal semantic transfer function based on a semantic probability field is constructed to project 2D crown semantics onto 3D point clouds, addressing inconsistencies in tree instances caused by varying viewpoints. The proposed framework is validated in subtropical mountainous forests, achieving a high individual tree segmentation accuracy (SAC) of 0.92, with omission and commission errors as low as 0.18 and 0.15, respectively, demonstrating its effectiveness and robustness in highly undulating terrain. The obtained single-tree semantic segmentation results can provide critical data support for forest resource surveys, timber volume estimation, and logging monitoring (e.g., calculating forest volume). In summary, this study provides an effective solution for single-tree instance segmentation in forests. However, the current method has not fully considered the topological connectivity of tree point clouds. Future work will involve a deeper analysis of point cloud topological structures, integrating 2D and 3D information, and extending the method to broader forest areas.

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

Funding: This research is supported in part by the National Natural Science Foundation of China (No. 42261075, No. 41861062), in part by the Jiangxi Provincial Natural Science Foundation (No. 20224ACB212003), in part by the Jiangxi Provincial Training Project of Disciplinary, Academic, and Technical Leader (No. 20232BCJ22002), in part by the Jiangxi Gan-Po Elite Talents-Innovative High-End Talents Program (No. gpyc20240071), and in part by the Science and Technology Innovation Project Supported by Department of Natural Resources of Jiangxi Province (No. ZRKJ20252604).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: Author Ye Yuan was employed by the company Shenzhen DJI Innovations Technology Co., Ltd. Author Jiahao Liu was employed by the company Jiangxi Helicopter 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. Dai, W.; Yang, B.; Dong, Z.; Shaker, A. A new method for 3D individual tree extraction using multispectral airborne LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **2018**, *144*, 400–411. [CrossRef]
2. Diez, Y.; Kentsch, S.; Fukuda, M.; Caceres, M.L.L.; Moritake, K.; Cabezas, M. Deep learning in forestry using UAV-acquired RGB data: A practical review. *Remote Sens.* **2021**, *13*, 2837. [CrossRef]
3. Burt, A.; Disney, M.; Calders, K. Extracting individual trees from lidar point clouds using treeseg. *Methods Ecol. Evol.* **2019**, *10*, 438–445. [CrossRef]
4. Itakura, K.; Miyatani, S.; Hosoi, F. Estimating tree structural parameters via automatic tree segmentation from LiDAR point cloud data. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *15*, 555–564. [CrossRef]
5. Moorthy, S.M.K.; Calders, K.; Vicari, M.B.; Verbeeck, H. Improved supervised learning-based approach for leaf and wood classification from LiDAR point clouds of forests. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3057–3070. [CrossRef]
6. Dersch, S.; Schoettl, A.; Krzystek, P.; Heurich, M. Towards complete tree crown delineation by instance segmentation with Mask R–CNN and DETR using UAV-based multispectral imagery and lidar data. *ISPRS Open J. Photogramm. Remote Sens.* **2023**, *8*, 100037. [CrossRef]
7. Wu, J.; Yang, G.; Yang, H.; Zhu, Y.; Li, Z.; Lei, L.; Zhao, C. Extracting apple tree crown information from remote imagery using deep learning. *Comput. Electron. Agric.* **2020**, *174*, 105504. [CrossRef]
8. He, H.; Zhou, F.; Xia, Y.; Chen, M.; Chen, T. Parallel fusion neural network considering local and global semantic information for citrus tree canopy segmentation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *17*, 1535–1549. [CrossRef]
9. Liu, Y.; Chen, D.; Fu, S.; Mathiopoulos, P.T.; Sui, M.; Na, J.; Peethambaran, J. Segmentation of individual tree points by combining marker-controlled watershed segmentation and spectral clustering optimization. *Remote Sens.* **2024**, *16*, 610. [CrossRef]
10. Yang, J.; Kang, Z.; Cheng, S.; Yang, Z.; Akwensi, P.H. An individual tree segmentation method based on watershed algorithm and three-dimensional spatial distribution analysis from airborne LiDAR point clouds. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 1055–1067. [CrossRef]
11. Gu, J.; Congalton, R.G. Individual tree crown delineation from UAS imagery based on region growing by over-segments with a competitive mechanism. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4402411. [CrossRef]
12. Ma, Z.; Pang, Y.; Wang, D.; Liang, X.; Chen, B.; Lu, H.; Weinacker, H.; Koch, B. Individual tree crown segmentation of a larch plantation using airborne laser scanning data based on region growing and canopy morphology features. *Remote Sens.* **2020**, *12*, 1078. [CrossRef]
13. Qi, C.; Su, H.; Kaichun, M.; Guibas, L. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
14. Qi, C.; Yi, L.; Su, H.; Guibas, L. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5105–5114.
15. Hu, Q.; Yang, B.; Xie, L.; Rosa, L.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
16. Krisanski, S.; Taskhiri, M.S.; Gonzalez Aracil, S.; Herries, D.; Turner, P. Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning. *Remote Sens.* **2021**, *13*, 1413. [CrossRef]
17. Krispel, G.; Opitz, M.; Waltner, G.; Possegger, H.; Bischof, H. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 1874–1883.
18. Li, S.; Tang, H. Multimodal Alignment and Fusion: A Survey. *arXiv* **2024**, arXiv:2411.17040. [CrossRef]
19. Liu, Q.; Ma, W.; Zhang, J.; Liu, Y.; Xu, D.; Wang, J. Point-cloud segmentation of individual trees in complex natural forest scenes based on a trunk-growth method. *J. For. Res.* **2021**, *32*, 2403–2414. [CrossRef]
20. Zhong, Y.; Qin, J.; Liu, S.; Ma, Z.; Liu, E.; Fan, H. An unsupervised semantic segmentation network for wood-leaf separation from 3D point clouds. *Plant Phenomics* **2025**, *7*, 100064. [CrossRef]
21. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.

22. Wang, D. Unsupervised semantic and instance segmentation of forest point clouds. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 86–97. [CrossRef]
23. Li, X.; Zhang, Z.; Li, Y.; Huang, M.; Zhang, J. SFL-Net: Slight filter learning network for point cloud semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5703914. [CrossRef]
24. Fricker, G.A.; Ventura, J.D.; Wolf, J.A.; North, M.P.; Davis, F.W.; Franklin, J. A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. *Remote Sens.* **2019**, *11*, 2326. [CrossRef]
25. Zhong, H.; Lin, W.; Liu, H.; Ma, N.; Liu, K.; Cao, R.; Wang, T.; Ren, Z. Identification of tree species based on the fusion of UAV hyperspectral image and LiDAR data in a coniferous and broad-leaved mixed forest in Northeast China. *Front. Plant Sci.* **2022**, *13*, 964769. [CrossRef]
26. DJI Terra. 2025. Available online: https://enterprise.dji.com/cn/dji-terra/ (accessed on 16 March 2025).
27. Cai, S.; Zhang, W.; Liang, X.; Wan, P.; Qi, J.; Yu, S.; Yan, G.; Shao, J. Filtering airborne LiDAR data through complementary cloth simulation and progressive TIN densification filters. *Remote Sens.* **2019**, *11*, 1037. [CrossRef]
28. CloudCompare. 2025. Available online: https://www.danielgm.net/cc/ (accessed on 3 July 2024).
29. Zeng, J.; Shen, X.; Zhou, K.; Cao, L. FO-Net: An advanced deep learning network for individual tree identification using UAV high-resolution images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 323–338. [CrossRef]
30. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412 2017.
31. Sun, C.; Huang, C.; Zhang, H.; Chen, B.; An, F.; Wang, L.; Yun, T. Individual tree crown segmentation and crown width extraction from a heightmap derived from aerial laser scanning data using a deep learning framework. *Front. Plant Sci.* **2022**, *13*, 914974. [CrossRef]
32. Jin, D.; Qi, J.; Gonçalves, N.B.; Wei, J.; Huang, H.; Pan, Y. Automated tree crown labeling with 3D radiative transfer modelling achieves human comparable performances for tree segmentation in semi-arid landscapes. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *134*, 104235. [CrossRef]
33. Ulku, I.; Akagündüz, E.; Ghamisi, P. Deep semantic segmentation of trees using multispectral images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2022**, *15*, 7589–7604. [CrossRef]
34. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
35. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis. (IJCV)* **2010**, *88*, 303–338. [CrossRef]
36. Guangshuai, W.; Yi, W.; Yongjun, Z. Registration of airborne LiDAR data and multi-view aerial images constrained by junction structure features. *Int. J. Geogr. Inf. Sci.* **2020**, *22*, 1868–1877.
37. Dai, W.; Kan, H.; Tan, R.; Yang, B.; Guan, Q.; Zhu, N.; Xiao, W.; Dong, Z. Multisource forest point cloud registration with semantic-guided keypoints and robust RANSAC mechanisms. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103105. [CrossRef]
38. Shi, P.; Yan, S.; Xiao, Y.; Liu, X.; Zhang, Y.; Li, J. RANSAC back to SOTA: A two-stage consensus filtering for real-time 3D registration. *IEEE Rob. Autom. Lett.* **2024**, *12*, 11881–11888. [CrossRef]
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; IEEE: New York, NY, USA, 2009; pp. 248–255.
40. Jiang, T.; Liu, S.; Zhang, Q.; Xu, X.; Sun, J.; Wang, Y. Segmentation of individual trees in urban MLS point clouds using a deep learning framework based on cylindrical convolution network. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *123*, 103473. [CrossRef]
41. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]
42. Luo, H.; Khoshelham, K.; Chen, C.; He, H. Individual tree extraction from urban mobile laser scanning point clouds using deep pointwise direction embedding. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 326–339. [CrossRef]

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.