A Comparative Study of Deep Learning and Unsupervised Segmentation Methods for Individual Tree Delineation from LiDAR Point Clouds

Jinhong Wang[1], Wei Yao[2,3*], Tiangang Yin[1]

[1] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[2] Spatial Intelligence and Urban Computing, Institute of Urban Environment,
Chinese Academy of Sciences, Xiamen, China, – wei.yao@ieee.org
[3] State Key Lab for Ecological Security of Regions and Cities, Institute of Urban Environment,
Chinese Academy of Sciences, Xiamen, China

Keywords: Comparative study, Individual tree segmentation, Deep learning, Unsupervised, Forest LiDAR data.

Abstract

Individual tree segmentation from LiDAR is central to automated inventory, yet many evaluations cover only one forest type or one methodological family. It also remains unclear whether point-level and instance-level metrics rank methods in the same order, which complicates operational choices. We therefore harmonise preprocessing, a binary tree versus non-tree semantic task, and a joint evaluation protocol. Representative unsupervised (graph-based TreeIso; region-growing TreeX) and deep learning (vote-based TreeLearn; mask-based ForestFormer3D) pipelines are compared on both public benchmarks and a fused unmanned aerial and mobile laser scanning benchmark from rugged subtropical Hong Kong woodland called HKTrees. Performance is strongest on regular coniferous plots. Broadleaved heterogeneity and crown overlap lower instance recall relative to semantic scores. HKTrees is the hardest regime and shows pronounced domain shift. Point-level and instance-level rankings are not always aligned. We discuss trade-offs in annotation cost, generalisation, and interpretability and outline expansion of HKTrees dataset for larger-scale training.

1. Introduction

The early forest inventory relied heavily on costly and time-consuming field surveys, where individual trees within the sample plots were measured one by one for attributes such as location, species, diameter at breast height, tree height and crown width, and these measurements were then aggregated to derive key properties at the stand-level (Zhen et al., 2016). With the development of LiDAR technology, laser scanning has become a primary sensing approach for digital 3D scene acquisition, largely because it can provide dense three-dimensional point observations with high reliability. The use of LiDAR-acquired 3D point clouds for individual-tree inventory can substantially reduce manual labor and improve the efficiency of forest resource monitoring.

Existing individual tree segmentation (ITS) methods can be broadly categorized into unsupervised approaches and deep learning–based techniques. Traditional ITS are predominantly based on unsupervised segmentation algorithms using distinct geometric features. In recent years, research has progressively shifted toward supervised deep learning (DL) techniques. However, the performance of existing ITS methods (Xi and Hopkinson, 2022, Xiang et al., 2025a, Yao et al., 2013) across diverse forest types has not yet been systematically assessed. They often evaluate within a single forest type, leading to insufficient comparisons of cross-forest-type generalization between deep learning and unsupervised approaches.

Therefore, this study aims to conduct a comparative analysis of ITS methods for forest LiDAR point clouds. For example, the ForAINet framework has demonstrated robust semantic and instance segmentation performance across various forest types and geographic regions, achieving an F1 score exceeding 85% for individual trees in airborne LiDAR point clouds (Xiang et al., 2024). However, such deep learning solutions typically require large, high-quality annotated training sets, whereas unsupervised methods (e.g., graph –based segmentation, region -growing, or clustering algorithms) operate without manual labels but may exhibit lower accuracy in structurally complex stands. A benchmarking study on terrestrial laser scanning point clouds reported that a graph-based approach currently outperforms data-driven models on both plot-level F1-score and tree-level mean F1 score, while also revealing that segmentation performance varies greatly across forest types (Cherlet et al., 2026). This finding underscores that instance segmentation remains difficult to fully automate and highlights the need for diverse training and evaluation data. Furthermore, a recent evaluation of YOLOv5 (Jocher et al., 2022) in the FOR-Instance (Puliti et al., 2023) dataset (which covers a wide range of forest types in multiple countries) reported a detection F1 score of 0.74 and a tree detection rate of 64%, outperforming conventional parameter-dependent methods by a substantial margin (Straker et al., 2023). Nevertheless, such studies are typically confined to a single algorithmic family and do not offer a direct side-by-side comparison - versus unsupervised paradigms across multiple forest types. Thus, we find that three specific research gaps motivate the present study.

(1) It remains unclear whether the superior performance of deep learning models in homogeneous forest settings generalizes reliably to structurally heterogeneous tropical or subtropical forests, and whether unsupervised methods, which require no training data, can offer comparable stability across forest types.

(2) Many studies report only detection-level scores (e.g. F1-score, precision, recall), while others rely on point - accuracy. It is not known whether these two types of metrics yield consistent rankings of method performance. A discrepancy between point-level and instance-level evaluations would have profound implications for method selection in operational forest invent-

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-1125-2026

ory, where both the correct delineation of each individual tree and the precise assignment of every point to its corresponding tree are often required.

(3) Most of the benchmark datasets are collected in well-managed natural or plantation forests under relatively controlled conditions. In contrast, Hong Kong’s urban forests have a completely different environment, characterized by fragmented canopy cover, high species diversity, complex understory vegetation, uneven terrain and pervasive anthropogenic interference (e.g. man-made objects). In particular, these forests are representative of the types of forests commonly found in mountainous regions of South China.

To address these gaps, this study systematically evaluates a representative set of deep learning and unsupervised ITS methods across multiple types of forests (broadleaved and coniferous), including both forests from open-source datasets and the urban-forest interface in Hong Kong. A unified evaluation framework is adopted that incorporates both the point-level and the instance-level metrics to ensure fair and consistent comparison. The findings are expected to provide practical guidance for the selection of methods in various forest inventory applications and to illuminate the relative strengths and limitations of data-driven versus unsupervised paradigms under varying forest conditions.

2. Related Work

2.1 Unsupervised methods

Early individual-tree segmentation relied on unsupervised algorithms based primarily on the geometric features of point clouds and complex heuristic rules. These methods were predominantly designed for sparse airborne laser scan data. Top-down segmentation strategies refer to approaches that first process the canopy layer before delineating individual tree stems. Typically, the canopy height model (CHM) is obtained first by subtracting a terrain model from a LiDAR-derived digital surface model. The potential tree positions are then detected using local maximum algorithms, which serve as a priori knowledge to guide individual tree segmentation through image processing techniques such as the watershed algorithm (Reitberger et al., 2009, Yao et al., 2012, Liu et al., 2015). In this context, they have been widely adopted due to their computational efficiency. However, such strategies often perform poorly in understory vegetation, making lower-canopy trees difficult to detect. To address this limitation, (Duncanson et al., 2014) proposed a multi-level canopy segmentation method that effectively improves the identification of understory vegetation by optimizing segmentation hierarchies and strategies.

As the density of point clouds increases (particularly for terrestrial LiDAR data), the structural details captured of the forest stands become richer, causing greater demands on the robustness of the segmentation algorithms. In this context, compared to traditional top-down strategies, bottom-up approaches demonstrate greater adaptability and are more effective in addressing ITS tasks in a complex forest area. These approaches first identify candidate stem points and then use them to delineate individual tree crowns (Shendryk et al., 2016, Polewski et al., 2015, Polewski et al., 2017, Xi and Hopkinson, 2022). However, reliance on handcrafted rules raises questions about the generalizability of these unsupervised methods in forest scenes with diverse tree species and complex structures, and whether such methods might suffer from compromised segmentation accuracy under varying conditions (Ruoppa et al., 2025, Yao et al., 2014).

2.2 Supervised methods

In recent years, with the rapid advancement of deep learning in computer vision and remote sensing, a series of deep learning-based individual tree segmentation models have been proposed (Chen et al., 2021, Chang et al., 2022, Xiang et al., 2024, Xiang et al., 2025a). These approaches use data-driven strategies to automatically extract features, providing a new solution to overcome generalization limitations and achieve more accurate segmentation performance in complex scenes. Unsupervised methods typically require input of pre-classified tree points rather than raw scan data. Such reliance introduces two key limitations: it depends on a reliable pre-processing step for tree point extraction, and it restricts the degree of automation. Early supervised learning-based methods addressed this by pursuing fully automated segmentation that directly takes raw scan point clouds as input. Typically, a semantic segmentation network is applied to extract tree points, while, the process of individual tree segmentation is then still performed by unsupervised approaches. (Chen et al., 2021, Chang et al., 2022).

This semantic segmentation based framework, while improving automation compared to unsupervised methods, does not fully address the issue of error accumulation. To overcome this limitation, deep learning-based instance segmentation methods have been progressively introduced into tree segmentation tasks. Early attempts relied mainly on object detection techniques, localizing individual trees by predicting bounding boxes (Wang et al., 2023). However, such methods perform poorly in forest environments with dense tree distributions and complex canopy intersection, where substantial overlap and occlusion among bounding boxes compromise detection accuracy. In contrast, strategies based on point offset voting have demonstrated greater adaptability. These approaches predict how far each point is from the corresponding tree’s representative point (such as the trunk center, treetop, or stem base), and then cluster the shifted points to separate individual trees. (Xiang et al., 2024, Henrich et al., 2024). This method reduces sensitivity to predefined parameters and enables more robust individual tree segmentation in complex stand structures.

However, these methods still rely on manually designed components that require extensive tuning, such as the voting mechanism for predicting geometric attributes and the heuristic rules (e.g., proposal aggregation and point set aggregation) for clustering the voting results. The root of this dependence lies in the fact that existing models are not designed with the direct prediction of instance masks as their objective (Schult et al., 2022). To address this issue, tree instance mask prediction methods based on the Transformer architecture have emerged (Xiang et al., 2025a). Unlike the conventional paradigm that relies on offset vector prediction followed by post-processing clustering, this approach directly learns the mapping from raw point clouds to tree instance masks, achieving truly end-to-end segmentation. Its core advantages are twofold: on the one hand, the attention mechanism effectively models the spatial relationships both between trees and among points within a single tree, enhancing the ability to distinguish dense canopies; on the other hand, the design of directly outputting instance masks eliminates the dependence on manual clustering rules, greatly improving segmentation accuracy in complex forest scenarios.

3. Materials and Study Areas

In this study, the experimental dataset comprises multiple open-source ITS datasets (FOR-InstanceV2 (Xiang et al., 2025b), TreeLearn (Henrich et al., 2023), Paracou tropical ULS (Bai et al., 2023) and WythamWoodsTLS (Calders et al., 2022)) as well as manually collected forest data from suburban areas of Hong Kong. Based on the structure of the forest stand, the plots were systematically categorized into two types: broadleaved and coniferous. In contrast, the Hong Kong dataset shows substantial differences in terms of tree structural complexity, canopy overlap, and terrain ruggedness. Given these conditions, we designated our own collected data as a separate forest type, referred to as HKTrees. The overview of the data can be seen in Table 3

3.1 FOR-InstanceV2

FOR-instanceV2 is a curated extension and re-partitioning of the FOR-instance (Puliti et al., 2023) UAV laser scanning benchmark (Puliti et al., 2023), released together with the ForestFormer3D framework (Xiang et al., 2025a). It provides high-density ULS and terrestrial laser scanning (TLS) data from multiple sites across Norway, the Czech Republic, Austria, Australia, and New Zealand. In this study, based on forest stand characteristics, we classify the individual scenes into three types (see in Table 1).

Table 1. Forest type of FOR-InstanceV2.

Dataset	Forest type
CULS	Coniferous
NIBIO	Coniferous
NIBIO2	Coniferous
SCION	Coniferous
TUWIEN	Broadleaved
BlueCat	Broadleaved
RMIT	Broadleaved
Yuchen	Broadleaved

3.2 TreeLearn

The TreeLearn dataset (Henrich et al., 2023) provides mobile laser scanning (MLS) data acquired in temperate European forest canopies. In this study, we only used its manually refined benchmark subset (156 complete trees) to ensure reliable reference annotations. Based on forest stand characteristics, we classify TreeLearn as a broadleave forest type.

3.3 Paracou tropical ULS

The Paracou benchmark (Bai et al., 2023) provides both UAV and TLS data acquired over a lowland tropical moist forest in French Guiana (2021–2022, RIEGL miniVUX-1UAV). The forest has a canopy height of approximately 27m. In this study, we used the ULS data with individual tree labels. Based on forest characteristics, we classify it as broadleaved forest.

3.4 WythamWoodsTLS

The Wytham Woods TLS inventory (Calders et al., 2022) provides leaf-off terrestrial laser scanning (TLS) data for a $\sim$ 1.4ha deciduous temperate broadleaved plot in Oxfordshire, UK (late November 2015 – January 2016, RIEGL VZ-400). The release contains per-tree segmented point clouds for 876 trees. In this study, we classify this dataset as broadleaved forest.

3.5 HKTrees

HKTrees is an internal ITS benchmark acquired on suburban hillside sites in the New Territories, Hong Kong. The working point cloud combines unmanned aerial laser scanning (ULS) from a RIEGL VUX-120 with handheld mobile laser scanning from a Leica BLK2GO, thereby combining a top-down view of the canopy with pedestrian-scale coverage along trails and slopes. The dominant vegetation corresponds to an evergreen broadleaved forest. The fused product exhibits pronounced variability in local sampling density driven by range, incidence angle, occlusion, and the differing scan geometries of the two platforms. Moreover, it together with rugged topography, heterogeneous tree-size distributions, and frequent crown overlap. Non-tree returns and scene clutter from built structures, infrastructure, and understory elements further increase the difficulty of fully automated semantic and instance segmentation relative to the comparatively controlled reference plots in open benchmarks. For these reasons, HKTrees is analyzed throughout the paper as a distinct forest type when assessing cross-site robustness.

Reference annotations were produced manually for three non-overlapping plots (Table 2). Each plot was labeled with *both* point-wise semantic classes and per-tree instance identifiers. The semantic taxonomy follows a binary scheme distinguishing *tree* from *non-tree* points, where the latter aggregates terrain, built objects, understory, and any other returns not attributed to woody vegetation. Instance labels assign a unique tree ID to every echo classified as a tree, allowing joint assessment of semantic filtering and individual-tree segmentation within the same evaluation masks (see Figure 1).

Table 2. Manually annotated HKTrees plots: ground dimensions and inventoried tree counts.

Plot Name	Extent (m)	Area (m^2)	# Trees
AKK	40×59	2360	41
NTM	74×113	8362	49
SMN	21×27	567	18
Total	—	11 289	108

4. Methodology

4.1 Data preprocessing

Let $\mathcal{P} = \{(\mathbf{p}_i, \ell_i, c_i)\}_{i=1}^{N}$ denote a forest point cloud in a plot coordinate frame, where $\mathbf{p}_i = (x_i, y_i, z_i)$ are Cartesian coordinates, $\ell_i \in \{\text{tree}, \text{non-tree}\}$ is a *binary* semantic label after harmonization (Section 3), and $c_i \in \{0, 1, \ldots, K\}$ is the instance (tree) identifier with $c_i = 0$ reserved for non-tree points. Individual-tree segmentation aims to recover $\hat{c}_i$ that agrees with reference c_i on tree points while correctly assigning non-tree returns to the background class.

Semantic harmonisation. Public ITS datasets (FOR-instance V2, TreeLearn, Paracou, WythamWoodsTLS) and internal HKTrees plots use heterogeneous original class schemes. All points are remapped to a single taxonomy with two semantic classes: *tree* (stem, branches, foliage attributed to woody vegetation, as defined by each benchmark's instance mask) and *non-tree* (terrain, low vegetation not assigned a tree ID and man-made objects). The mappings are fixed in a lookup table so that every algorithm sees identical binary masks for training, validation, and evaluation.

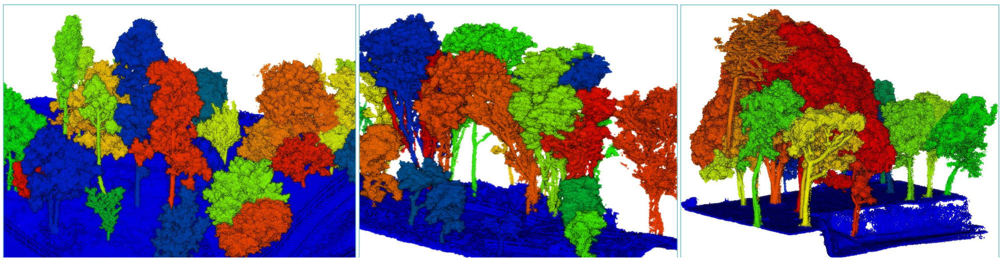

Figure 1. A visual demonstration of the labeled HKTrees dataset across three sites, where different tree instances are shown with randomly assigned colors. The data we collected were obtained from three slope sites, namely AKK, NTM, and SMN.

Table 3. Overview of open datasets used in this study. ULS: UAV-borne (unmanned aerial) laser scanning; MLS: mobile laser scanning; TLS: terrestrial laser scanning. Forest-type labels follow the original data descriptions.

Dataset	Year	Scanner type	Forest type
FOR-instanceV2 (Xiang et al., 2025b)	∼2017–2022 (V2 release 2025)	ULS (RIEGL VUX-1UAV / MiniVUX-1 UAV)	Broadleaved & Coniferous
TreeLearn (Henrich et al., 2023)	2023	MLS	Broadleaved
Paracou tropical ULS (Bai et al., 2023)	2021–2022	ULS (RIEGL miniVUX-1UAV); TLS (RIEGL VZ-400) for label transfer	Broadleaved
WythamWoodsTLS (Calders et al., 2022)	2015–2016	TLS (RIEGL VZ-400)	Broadleaved
HKTrees (ours)	2026	ULS (RIEGL VUX-120); MLS (Leica BLK2GO)	HKTrees

Plot cropping and data split. After harmonization, each study site is clipped to a nearly square window of approximately $40\,\mathrm{m} \times 40\,\mathrm{m}$. This yields a uniform spatial support across sensors and biomes. The resulting plot inventory is divided into training and testing subsets by plot count in a ratio 3:1.

4.2 Compared methods

Deep-learning instance segmentation approaches are grouped into *vote-based* models, which predict geometric offsets or centroids and then group points into trees (Henrich et al., 2024), and *mask-based* models, which regress instance masks or queries directly without a separate hand-crafted clustering stage (Schult et al., 2022, Xiang et al., 2025a). Unsupervised competitors are separated into *graph-based* bottom-up linkage methods that connect stem cues to crowns and clustering / region-growing schemes that expand seeds under local density or shape constraints. These families embody the trade-off highlighted in Section 1: supervised models leverage annotated structure but depend on compute and labels, while unsupervised methods remain lightweight and label-free, but are sensitive to hyperparameters and scene complexity (Cherlet et al., 2026, Ruoppa et al., 2025). Table 4 summarizes ITS methods in comparative experiments.

TreeLearn (Henrich et al., 2024) and **ForestFormer3D** (Xiang et al., 2025a) are each studied under two learning configurations to separate data effects from architectural priors. First, we train models from scratch (following the official architectures, loss functions, and optimization schedules) on the harmonized $40\,\mathrm{m} \times 40\,\mathrm{m}$ training plots with the 3:1 split described above, using only binary tree/non-tree supervision and instance IDs from the remapped labels. Second, we evaluated the publicly released checkpoints distributed by the original authors, without additional fine-tuning, to quantify out-of-the-box generalization.

TreeIso (Xi and Hopkinson, 2022) links candidate stem points along estimated trunk axes and propagates graph connectivity to separate crowns, but the published pipeline does not include a semantic stage that removes non-tree returns. Operating it on raw scanned point clouds would therefore entangle segmentation errors with foreground extraction. Consequently, TreeIso and TreeX experiments use only ground-truth tree points (points with $c_i > 0$ in the reference) as input, which isolates instance grouping under controlled semantics yet does not constitute a fully automatic "scan-to-trees" workflow.

Table 4. Methods in comparative experiments

Category	Representative Methods
Deep Learning (DL)	-
Vote-based	TreeLearn (Henrich et al., 2024)
Mask-based	ForestFormer3D (Xiang et al., 2025a)
Unsupervised	-
Graph-based	TreeIso (Xi and Hopkinson, 2022)
Clustering & Region Growing	TreeX (Burmeister et al., 2025)

4.3 Evaluation metrics

In current evaluation methodologies, the mainstream approach is point-based assessment, which evaluates tree detection performance by comparing matched individual tree points against ground truth data to compute precision, recall, and F1-score. Additionally, the coverage metric (Cov) is used to quantify the alignment between the predicted and ground truth tree boundaries (Wang et al., 2019).

However, this point-based evaluation paradigm can introduce a fundamental bias when comparing unsupervised and DL methods, leading to unfair comparison. The key to this inequity lies in their inherent design principles. Unsupervised methods typically rely on parameters tuned for the general forest environment without any exposure to manual annotations. In contrast, DL models are explicitly trained on labeled datasets, allowing them to learn and replicate the specific segmentation style of the annotators. Thus, we make the assumption that standard point-based metrics, which may always reward precise point-level alignment with this human-generated ground truth, are inherently biased towards supervised DL approaches.

Therefore, to comprehensively evaluate the fairness of point-based methods, we supplement the point-level validation with an additional evaluation framework. This framework shifts the focus from point-level accuracy to object-level agreement. It operates by first establishing a valid correspondence between a predicted and a ground-truth tree when their bounding boxes or segmented regions exhibit an Intersection-over-Union (IoU) exceeding a defined threshold (e.g., 0.5). For all successfully matched pairs, the framework then assesses the accuracy of key structural attributes, such as tree location and height (Reitberger et al., 2009).

4.3.1 Semantic metrics Semantic segmentation is evaluated on a point-wise basis. After mapping to $\{\text{tree}, \text{non-tree}\}$, let TP_c, FP_c, and FN_c denote the true-positive, false-positive, and false-negative point counts for class $c \in \{\mathrm{tr}, \mathrm{nt}\}$. The class-wise intersection-over-union is

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}\,, \tag{1}$$

and we report the mean intersection-over-union (mIoU) as the primary semantic score,

$$\mathrm{mIoU} = \frac{1}{2}\big(\mathrm{IoU}_{\mathrm{tr}} + \mathrm{IoU}_{\mathrm{nt}}\big)\,. \tag{2}$$

4.3.2 Instance metrics For individual-tree segmentation, each reference tree i and each predicted instance j are represented by the 3D point sets $\mathcal{G}_i$ and $\mathcal{P}_j$ assigned to the corresponding IDs. Their volumetric overlap is

$$\mathrm{IoU}(\mathcal{P}_j, \mathcal{G}_i) = \frac{|\mathcal{P}_j \cap \mathcal{G}_i|}{|\mathcal{P}_j \cup \mathcal{G}_i|}\,. \tag{3}$$

Coverage. Following (Xiang et al., 2024), we quantify the quality of the delineation by assigning, to every tree of ground-truth i, the prediction that maximizes overlap,

$$\mathcal{P}_i^\star = \arg\max_j \mathrm{IoU}(\mathcal{P}_j, \mathcal{G}_i)\,, \tag{4}$$

and averaging those best overlaps across the plot,

$$\mathrm{Cov} = \frac{1}{N_{\mathrm{gt}}} \sum_{i=1}^{N_{\mathrm{gt}}} \mathrm{IoU}\big(\mathcal{P}_i^\star, \mathcal{G}_i\big)\,, \tag{5}$$

where N_{gt} is the number of reference trees. High Cov indicates that each GT crown has a prediction that spatially aligns with it, even before hard match decisions (Wang et al., 2019). For detection-style instance scoring, we form one-to-one pairs between $\{\mathcal{P}_j\}$ and $\{\mathcal{G}_i\}$ (Hungarian assignment maximizing IoU). A match counts as a true positive only if $\mathrm{IoU} \geq \tau$ with $\tau = 0.5$; unmatched predictions are false positives and unmatched references are false negatives. With N_{TP}, N_{FP}, and N_{FN} the resulting counts,

$$\begin{aligned} \mathrm{Prec}_{\mathrm{ins}} &= \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}\,, \\ \mathrm{Rec}_{\mathrm{ins}} &= \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}\,, \\ F_{1,\mathrm{ins}} &= \frac{2\,\mathrm{Prec}_{\mathrm{ins}}\,\mathrm{Rec}_{\mathrm{ins}}}{\mathrm{Prec}_{\mathrm{ins}} + \mathrm{Rec}_{\mathrm{ins}}}\,. \end{aligned} \tag{6}$$

4.3.3 Plot-level metrics In order to comprehensively assess the performance of ITS methods. We also add a tree-instance-wise evaluation to each plot (Reitberger et al., 2009). For every tree of ground-truth we take a reference planimetric location $\mathbf{x}_i^{\mathrm{gt}} = (x_i, y_i)$ (e.g. position of the stem map) and height h_i^{gt}; for each predicted tree we derive $(\mathbf{x}_j^{\mathrm{pred}}, h_j^{\mathrm{pred}})$ with the same rule for all methods (e.g. projection of the lowest stem cluster or the highest point in the canopy). A pair (i, j) is *admissible* if

$$\|\mathbf{x}_i^{\mathrm{gt}} - \mathbf{x}_j^{\mathrm{pred}}\|_2 \leq d_{\max} \quad \text{and} \quad |h_i^{\mathrm{gt}} - h_j^{\mathrm{pred}}| \leq h_{\max}\,, \tag{7}$$

with thresholds $d_{\max}$ and $h_{\max}$ shared across competitors. Admissible pairs are then linked with a one-to-one matcher (the Hungary algorithm maximizing a joint score or, equivalently, minimizing distance). Let $N_{\mathrm{TP}}^{\mathrm{loc}}$, $N_{\mathrm{FP}}^{\mathrm{loc}}$, and $N_{\mathrm{FN}}^{\mathrm{loc}}$ denote true positives, false positives, and false negatives on the plot scale. The plot precision, recall, and F_1 are then

$$\begin{aligned} \mathrm{Prec}_{\mathrm{plot}} &= \frac{N_{\mathrm{TP}}^{\mathrm{loc}}}{N_{\mathrm{TP}}^{\mathrm{loc}} + N_{\mathrm{FP}}^{\mathrm{loc}}}\,, \\ \mathrm{Rec}_{\mathrm{plot}} &= \frac{N_{\mathrm{TP}}^{\mathrm{loc}}}{N_{\mathrm{TP}}^{\mathrm{loc}} + N_{\mathrm{FN}}^{\mathrm{loc}}}\,, \\ F_{1,\mathrm{plot}} &= \frac{2\,\mathrm{Prec}_{\mathrm{plot}}\,\mathrm{Rec}_{\mathrm{plot}}}{\mathrm{Prec}_{\mathrm{plot}} + \mathrm{Rec}_{\mathrm{plot}}}\,. \end{aligned} \tag{8}$$

4.4 Implementation details

All unsupervised baselines (TreeIso, TreeX) were run with the default parameters distributed by the respective authors, using the TLS-oriented preset where the software distinguishes acquisition modes and the ULS-oriented preset for airborne-style point densities.

For supervised models, training and inference recipes followed the original configures. In particular, for our self-trained ForestFormer3D checkpoint, the multi-scale grouping sampling radius was set to 8 m (instead of the default 16 m). The publicly released ForestFormer3D and TreeLearn weights were evaluated under the same preprocessing pipeline. All neural-network training and all inferences reported in this study were executed on a single NVIDIA RTX A6000 GPU with 48 GB of device memory.

5. Results

Tables 5-7 report held-out test scores for each forest site (broadleaved, coniferous, HKTrees) under the harmonised protocol (Section 4.3). Semantic columns list binary mean IoU (mIoU) and overall accuracy (OA) after tree/non-tree remapping. Instance scores are split into point-wise panoptic and instance terms. PQ, and precision, recall, F_1, and coverage (Cov) at instance IoU$\geq$ 0.5 (Eqs. (6)–(5)). A plot-level stem-centre F_1 (macro mean over plots, Eqs. (7)–(8)). TreeIso and TreeX runs on ground-truth tree points use reference labels only to isolate tree echoes. They do not perform semantic segmentation of the full scene, so semantic mIoU/OA are not applicable.

Several patterns align with broader benchmarking experience on forest point clouds: coniferous plots with more regular stem structure yield the highest instance and plot-level scores for the best deep model. Broadleaved canopies show larger gaps between high precision and lower recall at the instance level, consistent with crown overlap and fragmented predictions. The

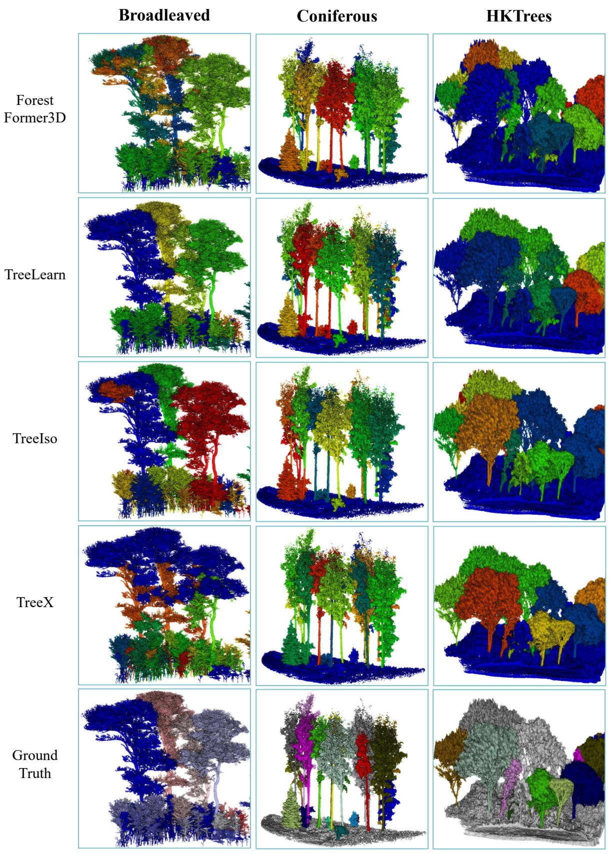

Figure 2. Qualitative individual-tree segmentation with random instance colours. Broadleaved plots show a large height range, where unsupervised methods often over-segment tall crowns vertically yet merge neighbouring shorter stems. Coniferous stands are more regular and yield stronger agreement. HKTrees is affected by rugged terrain, shrub clutter and heavy crown overlap, which lead unsupervised pipelines fail to separate touching canopies, often collapsing multiple trees into one prediction.

HKTrees site remains the most challenging, where even strong models exhibit reduced PQ and plot F_1.

6. Discussion

The experiment results in Tables 5–7 reveal how the forest type shapes algorithmic behavior. The coniferous plots (Table 6) exhibit comparatively regular stem geometry, more separable crowns, and fewer extreme height contrasts within a single neighborhood. Under such conditions, both vote-based and mask-based deep models reach high instance F_1 and plot-level stem agreement, and even unsupervised pipelines benefit because trunk detection, and region-growing seeds align more consistently with physical trees. Broadleave forests (Table 5) embed greater crown heterogeneity and lateral overlap. Thus, several methods retain strong semantic scores, yet show a pronounced drop in instance recall and coverage, which is consistent with merged instances, fragmented predictions, and ambiguous boundaries between touching canopies. Failure modes already visible in the qualitative comparison (Figure 2). HKTrees (Table 7) compound these difficulties with rugged terrain, shrub clutter, and heavy crown overlap.

The high mean IoU for the tree class does not guaranty strong

Table 5. Test result of broadleaved forest.

Method	Semantic		Instance					
	mIoU	OA	Point-wise					Plot
			PQ	P	R	F_1	Cov	$F_{1,\mathrm{plot}}$
ForestFormer3D,	0.983	0.999	0.746	0.854	0.467	0.604	0.467	0.710
ForestFormer3D, pretrain	0.994	0.999	0.414	0.835	0.585	0.688	0.585	0.667
TreeLearn,	0.984	0.999	0.522	0.555	0.304	0.393	0.304	0.380
TreeLearn, pretrain	0.727	0.981	0.354	0.440	0.134	0.205	0.134	0.389
TreeIso	—	—	0.589	0.436	0.153	0.227	0.153	0.474
TreeX	—	—	0.632	0.692	0.216	0.329	0.216	0.465

Table 6. Test result of coniferous forest.

Method	Semantic		Instance					
	mIoU	OA	Point-wise					Plot
			PQ	P	R	F_1	Cov	$F_{1,\mathrm{plot}}$
ForestFormer3D,	0.995	0.999	0.897	0.950	0.849	0.897	0.849	0.938
ForestFormer3D, pretrain	0.994	0.999	0.898	0.958	0.868	0.911	0.868	0.944
TreeLearn,	0.996	0.999	0.626	0.464	0.788	0.584	0.788	0.820
TreeLearn, pretrain	0.868	0.972	0.345	0.225	0.498	0.310	0.498	0.789
TreeIso	—	—	0.671	0.749	0.306	0.435	0.306	0.593
TreeX	—	—	0.644	0.486	0.280	0.356	0.280	0.657

panoptic quality or instance F_1. Coverage can remain informative because it measures best-overlap assignment per reference tree before hard decisions, but large gaps between coverage and F_1 signal systematic under-segmentation or over-segmentation rather than minor boundary noise.

Unsupervised graph and region-growing methods have no annotation cost and offer transparent, parameterised rules that can be adjusted when domain changed. In this study, TreeIso and TreeX were evaluated on ground-truth tree points to isolate the grouping of instances from semantic extraction (Section 4.2). Consequently, their scores should be interpreted as an upper bound on instance performance given perfect foreground filtering, not as end-to-end scan-to-tree accuracy. Deep models instead learn point feature representations from data, which can better disambiguate dense overlap when training domains match test conditions, but they demand labeled plots, GPU resources, and careful handling of class imbalance between tree and non-tree returns. Pretrained checkpoints illustrate that strong out-of-the-box weights can help instance-level behavior on challenging domains, yet semantic heads may remain miscalibrated under new sensors or labeling styles, producing misleadingly low mIoU even when instance overlap is still usable.

6.1 Limitations

The present comparison focuses on in-domain test splits per forest type. Systematic cross-forest training and testing (e.g., coniferous-trained models on broadleaved or HKTrees plots) is deferred to future work, but is needed to quantify transferability of both handcrafted rules and learned representations. Further methodological extensions are summarized in Section 7.

7. Outlook

Expansion of HKTrees. HKTrees currently comprises three manually annotated plots and 108 reference trees. This footprint is adequate for external validation, qualitative failure analysis, and illustrating domain shift relative to open benchmarks, but it is not yet large enough to support high-capacity supervised training comparable to FOR-InstanceV2 (Xiang et al., 2025b) or TreeLearn (Henrich et al., 2023) dataset. We therefore treat HKTrees as a pilot benchmark: ongoing and planned

Table 7. Test result of HKTrees (Hong Kong complex woodland).

Method	Semantic		Instance					
	mIoU	OA	Point-wise					Plot
			PQ	P	R	F_1	Cov	$F_{1,\text{plot}}$
ForestFormer3D,	0.970	0.986	0.641	0.455	0.294	0.357	0.294	0.600
ForestFormer3D, pretrain	0.253	0.301	0.613	0.818	0.529	0.643	0.529	0.666
TreeLearn,	0.961	0.981	0.641	0.367	0.647	0.468	0.647	0.627
TreeLearn, pretrain	0.930	0.965	0.538	0.158	0.353	0.218	0.353	0.545
TreeIso	—	—	0.657	0.500	0.294	0.370	0.294	0.442
TreeX	—	—	0.723	0.667	0.353	0.462	0.353	0.615

campaigns will expand spatial coverage and quantity of independent plots with consistent labeling.

Methods and evaluation. Beyond data expansion, we also see value in (i) semi-supervised and self-supervised pretraining on large unlabeled MLS/ULS blocks and (ii) active learning to target overlapping crowns and steep-slope stems, (iii) cross-sensor domain adaptation between ULS, MLS, and TLS.

8. Conclusion

This paper presented a harmonized comparison of representative unsupervised and deep learning approaches to individual-tree segmentation from forest LiDAR point clouds under complex conditions of broadleaved, coniferous, and Hong Kong forests, using semantic and instance-level metrics. Coniferous stands with regular structure yielded the strongest and most consistent scores, whereas broadleaved heterogeneity and especially the HKTrees domain exposed larger instance-level degradation despite often high point-wise semantic accuracy, highlighting crown overlap, clutter, and sensor fusion as dominant error sources. Point-level and instance-level indicators did not always rank methods identically, underscoring the need for dual reporting when operational goals include stem mapping and tree counts.

When to favor unsupervised methods, Choose unsupervised instance segmentation when labeled training data are unavailable or difficult to obtain, when compute budgets are limited, or when a reliable tree mask is already supplied by another stage and only instance grouping remains. Unsupervised tools are also appropriate for sanity checks and baselines before investing in annotation.

When to favor deep learning; Choose supervised instance segmentation when sufficiently large, representative labeled data exist for the target sensor and forest type, when dense overlap and complex understory make the segmentation difficult.

References

Bai, Y., Vincent, G., Barbier, N., Martin-Ducup, O., 2023. Uva laser scanning labelled las data over tropical moist forest classified as leaf or wood points.

Burmeister, J.-M., Tockner, A., Reder, S., Engel, M., Richter, R., Mund, J.-P., Döllner, J., 2025. treeX: Unsupervised Tree Instance Segmentation in Dense Forest Point Clouds. *arXiv preprint arXiv:2509.03633*.

Calders, K., Verbeeck, H., Burt, A., Origo, N., Nightingale, J., Malhi, Y., Wilkes, P., Raumonen, P., Bunce, R., Disney, M., 2022. Terrestrial laser scanning data Wytham Woods: individual trees and quantitative structure models (QSMs). *Zenodo*.

Chang, L., Fan, H., Zhu, N., Dong, Z., 2022. A two-stage approach for individual tree segmentation from TLS point clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 8682–8693.

Chen, X., Jiang, K., Zhu, Y., Wang, X., Yun, T., 2021. Individual tree crown segmentation directly from UAV-borne LiDAR data using the PointNet of deep learning. *Forests*, 12(2), 131.

Cherlet, W., Dayal, K., Chen, S., Cooper, Z., Disney, M., Hanzl, A., Levick, S., Nightingale, J., Origo, N., Senf, C., Soenens, L., Terryn, L., Van den Broeck, W. A., Calders, K., 2026. Benchmarking tree instance segmentation of terrestrial laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 231, 230-247.

Duncanson, L., Cook, B., Hurtt, G., Dubayah, R., 2014. An efficient, multi-layered crown delineation algorithm for mapping individual tree structure across multiple ecosystems. *Remote Sensing of Environment*, 154, 378–386.

Henrich, J., van Delden, J., Kneib, T., Seidel, D., Ecker, A., 2023. TreeLearn Dataset.

Henrich, J., van Delden, J., Seidel, D., Kneib, T., Ecker, A. S., 2024. TreeLearn: A deep learning method for segmenting individual trees from ground-based LiDAR forest point clouds. *Ecological Informatics*, 84, 102888.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Wong, C., Yifu, Z., Montes, D. et al., 2022. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo*.

Liu, T., Im, J., Quackenbush, L. J., 2015. A novel transferable individual tree crown delineation model based on Fishing Net Dragging and boundary classification. *ISPRS journal of photogrammetry and remote sensing*, 110, 34–47.

Polewski, P., Yao, W., Heurich, M., Krzystek, P., Stilla, U., 2015. Detection of fallen trees in ALS point clouds using a Normalized Cut approach trained by simulation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 252-271.

Polewski, P., Yao, W., Heurich, M., Krzystek, P., Stilla, U., 2017. A voting-based statistical cylinder detection framework applied to fallen tree mapping in terrestrial laser scanning point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 129, 118-130.

Puliti, S., Pearse, G., Surovỳ, P., Wallace, L., Hollaus, M., Wielgosz, M., Astrup, R., 2023. For-instance: a uav laser scanning benchmark dataset for semantic and instance segmentation of individual trees. *arXiv preprint arXiv:2309.01279*.

Reitberger, J., Schnörr, C., Krzystek, P., Stilla, U., 2009. 3D segmentation of single trees exploiting full waveform LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6), 561–574.

Ruoppa, L., Hietala, T., Seppänen, V., Taher, J., Hakala, T., Yu, X., Kukko, A., Kaartinen, H., Hyyppä, J., 2025. Benchmarking individual tree segmentation using multispectral airborne laser scanning data: the FGI-EMIT dataset. *arXiv preprint arXiv:2511.00653*.

Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B., 2022. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*.

Shendryk, I., Broich, M., Tulbure, M. G., Alexandrov, S. V., 2016. Bottom-up delineation of individual trees from full-waveform airborne laser scans in a structurally complex eucalypt forest. *Remote Sensing of Environment*, 173, 69–83.

Straker, A., Puliti, S., Breidenbach, J., Kleinn, C., Pearse, G., Astrup, R., Magdon, P., 2023. Instance segmentation of individual tree crowns with YOLOv5: A comparison of approaches using the ForInstance benchmark LiDAR dataset. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 9, 100045.

Wang, X., Liu, S., Shen, X., Shen, C., Jia, J., 2019. Associatively segmenting instances and semantics in point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4096–4105.

Wang, Z., Li, P., Cui, Y., Lei, S., Kang, Z., 2023. Automatic detection of individual trees in forests based on airborne LiDAR data with a tree region-based convolutional neural network (RCNN). *Remote Sensing*, 15(4), 1024.

Xi, Z., Hopkinson, C., 2022. 3D graph-based individual-tree isolation (Treeiso) from terrestrial laser scanning point clouds. *Remote Sensing*, 14(23), 6116.

Xiang, B., Wielgosz, M., Kontogianni, T., Peters, T., Puliti, S., Astrup, R., Schindler, K., 2024. Automated forest inventory: Analysis of high-density airborne LiDAR point clouds with 3D deep learning. *Remote Sensing of Environment*, 305, 114078.

Xiang, B., Wielgosz, M., Puliti, S., Král, K., Krček, M., Missarov, A., Astrup, R., 2025a. ForestFormer3D: A Unified Framework for End-to-End Segmentation of Forest LiDAR 3D Point Clouds. *arXiv preprint arXiv:2506.16991*.

Xiang, B., Wielgosz, M., PULITI, S., Král, K., Krucek, M., Missarov, A., Astrup, R., 2025b. For-instancev2 dataset and pre-trained model in paper entitled ”forestformer3d: A unified framework for end-to-end segmentation of forest lidar 3d point clouds”.

Yao, W., Krull, J., Krzystek, P., Heurich, M., 2014. Sensitivity Analysis of 3D Individual Tree Detection from LiDAR Point Clouds of Temperate Forests. *Forests*, 5(6), 1122–1142.

Yao, W., Krzystek, P., Heurich, M., 2012. Tree species classification and estimation of stem volume and DBH based on single tree extraction by exploiting airborne full-waveform LiDAR data. *Remote Sensing of Environment*, 123, 368-380.

Yao, W., Krzystek, P., Heurich, M., 2013. Enhanced detection of 3D individual trees in forested areas using airborne full-waveform LiDAR data by combining normalized cuts with spatial density clustering. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5/W2, 349–354.

Zhen, Z., Quackenbush, L. J., Zhang, L., 2016. Trends in automatic individual tree crown detection and delineation—Evolution of LiDAR data. *Remote Sensing*, 8(4), 333.