QC-SF: Improving Computer Vision for Airborne LiDAR Point Clouds of Boreal Forests with Quebec Simulated Forest Dataset

Olivier Stocker, Reza Mahmoudi Kouhi, Omid Reisi Gahrouei, Thierry Badard and Eric Guilbert
CRDIG, Université Laval, Québec, Canada
olivier.stocker.1@ulaval.ca, eric.guilbert@scg.ulaval.ca

Abstract

Boreal forest ecosystems are under immense pressure, and while airborne LiDAR has emerged as a powerful monitoring tool, leveraging its large data volumes requires automated analysis. Deep learning methods offer a solution but are hindered by the scarcity of large-scale, labeled datasets in forestry, contrasting to the data-rich urban environments. To address this gap, we introduce the Quebec Simulated Forest (QC-SF), a large-scale, synthetic airborne LiDAR dataset fully labeled for semantic segmentation. QC-SF contains 60,000 forest plots, each composed of a controlled mixture of the two dominant species in Quebec's boreal forest: Black Spruce and Balsam Fir. Using this benchmark, we evaluate the performance of five renowned and state-of-the-art point cloud networks: KPConv, MinkUNet, DGCNN, Point Mamba and Point Transformer V3. Our results identify Point Transformer V3 as the most effective architecture, achieving 91.66% mIoU. Furthermore, we validate the sim-to-real transferability of our dataset, demonstrating that augmenting a small number of real-world scans with our synthetic data improves segmentation performance by 6 points in mIoU score. QC-SF is available at : qc-sf.org.

1. Introduction

Boreal forest ecosystems are being heavily altered due to escalating stressors such as climate change, wildfires, epidemics, and anthropogenic pressures [27]. Monitoring these vast and vulnerable environments is critical for understanding these changes and formulating robust preservation policies [42]. Given the immense scale of the boreal biome, Airborne Laser Scanning (ALS) has emerged as an indispensable tool. The ability of LiDAR to penetrate vegetation allows for the collection of detailed structural information across the canopy, understory, and ground [23]. However, raw ALS data, which consists of geometric and limited spectral information, cannot be directly utilized by ecologists and forest managers. To be useful, these massive point clouds must be converted into meaningful ecological metrics, such as biomass maps, species distribution, and canopy height models.

The sheer volume of data generated by ALS surveys necessitates automated processing methods. Deep learning has revolutionized 3D point cloud processing, beginning with PointNet [48] and evolving through various strategies, including graph-based [34, 57], voxel-based [14, 39], and point convolution methods [6, 54]. Current state-of-the-art either leverage transformer architectures [65, 68], which utilize self-attention mechanisms [55] or mamba architectures that rely on space-state models [35, 37]. These modern architectures offer superior performance but have substantially more parameters, demanding massive amounts of labeled training data. This requirement for labeled data presents a significant bottleneck in forestry applications. Manually labeling forest point clouds is exceptionally challenging compared to urban environments. Forest structures are complex and irregular; trees are often entwined, their boundaries are fuzzy, and dense understories can obscure the ground. Consequently, manual annotation is a slow, cumbersome process, resulting in the limited quantity and quality of ground truth data. Most recent studies have nonetheless relied on this process [7, 8, 32, 33]. This scarcity hinders the development of dedicated vision tools and makes comparative studies difficult.

Recent efforts have aimed to mitigate this data gap by creating larger datasets, such as PureForest [28] and FORspecies20K [47]. However, these datasets have limitations. PureForest, while large, contains only monospecific plots with plot-wise classification labels, and plot-based information is not enough when facing mixed forests. FORspecies20K, an aggregation of 25 smaller datasets, focuses only on the classification of already individualized trees. Crucially, both datasets are heavily biased towards European forests. This leaves a significant scarcity of data representing the specific characteristics of the North American Eastern Boreal Shield. In domains where manual labeling has proven insufficient for deep learning, simula-

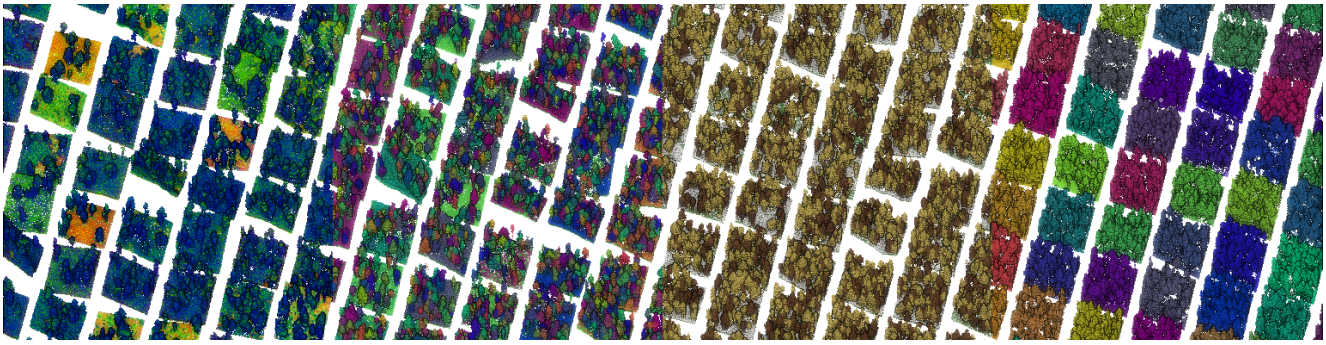


Figure 1. Few plots of the Quebec Simulated Forest (QC-SF) dataset, with colors representing respectively from left to right, intensity, individual trees, semantic classes, and plot index.

tion has emerged as a powerful alternative. Early benchmarks like ModelNet40 [70] used synthetic 3D objects to drive foundational research. In LiDAR simulation, Wu *et al*. [63] demonstrated that synthetic data generated via game engines could significantly aid model training for urban scenes. Following this precedent, we introduce Quebec Simulated Forest (QC-SF), a novel, large-scale synthetic LiDAR dataset designed specifically for the boreal forests of Quebec, focusing on the dominant species: Black Spruce (*Picea Mariana*) and Balsam Fir (*Abies Balsamea*). This benchmark facilitates the training of semantic segmentation models in data-scarce Quebec boreal forest environments. Our primary contributions are:

- A framework for generating realistic, large-scale synthetic LiDAR data representative of the Eastern Boreal Shield Forest.
- The creation and release of QC-SF, a multi-task synthetic LiDAR dataset comprising 60,000 labeled forest plots, equivalent to 2,400 ha.
- A comprehensive benchmark evaluating state-of-the-art deep learning networks for semantic segmentation on QC-SF.
- A validation of sim-to-real transfer learning, demonstrating the efficacy of using simulated data to improve segmentation performance on real-world data, particularly in few-shot contexts.

2. Related Works

2.1. Point Cloud Simulation for Forest Scenes

Lidar simulation has long been integral to sensor development and algorithm evaluation. The available software reflects a fundamental trade-off between physical realism and computational efficiency. On one hand, physically-based radiative transfer models like DART [26, 67], DIRSIG [30], and LESS [41] offer high-fidelity simulations by modeling complex beam-environment interactions, but at a significant computational cost. On the other hand, more efficient geometric-optical models like HELIOS [5] and its successor HELIOS++ [62] use a simplified beam ray tracing representation. This efficiency is critical when the goal is to generate massive volumes of diverse data required for training deep learning models.

The realism of any simulation is fundamentally dependent on the underlying 3D scene modelization. For forests, high-fidelity approaches often reconstruct individual trees from Terrestrial Laser Scanning (TLS) data [11, 20], using Quantitative Structure Models (QSMs) for woody components and voxelized representations for foliage [51, 59]. While initiatives like the RAMI IV experiment [61] produce meticulous digital twins, these reconstruction methods are labor-intensive and difficult to scale. An alternative is procedural generation with software like Xfrog [36] or Arbraro [58], which can generate large libraries of tree models efficiently but often at the cost of ecological authenticity and realism. Overall, if some studies [16, 18, 59] have explored modelization for lidar simulation, no consensus exists on the optimal methodology for producing learning-friendly simulations.

2.2. Deep Learning for Semantic Segmentation in Forestry

The application of deep learning to forest point clouds has primarily focused on tasks like individual tree segmentation [66], species classification [8, 47], and element segmentation (wood/leaf) [33]. However, progress in this field has historically lagged behind urban computer vision due to a critical bottleneck: the scarcity of large, labeled datasets. Early works in forestry understandably relied on less data-hungry architectures like PointNet and PointNet++ [7, 8, 33], with training data often limited to a few hundred manually annotated plots [32].

More recently, efforts to aggregate multiple smaller datasets have enabled the benchmarking of more powerful architectures. For instance, Puliti *et al*. [47] combined 25 datasets to evaluate models like PointNet++[49] (as Ensem-

ble), MinkUNet[14], PointMixer[13], and DGCNN[57]. Xiang *et al*. [66] utilized the FOR-instance dataset [46] to train an adapted MinkUNet[14] model. Despite these advances, even these aggregated datasets are small and heterogeneous compared to urban benchmarks like SemanticKITTI [29]. The intensive manual labor required for annotation remains the primary barrier to progress, highlighting the need for a scalable alternative to labeled data generation.

2.3. Simulation for Machine Learning in Forest Applications

Simulation is emerging as a viable solution to the data bottleneck in forestry. Recent studies have demonstrated its utility for a range of applications, from wildfire fuel identification [15] to validating the impact of point cloud density on segmentation tasks [53]. Specifically for training deep learning models, Bryson *et al*. [9] and Capua *et al*. [12] used simulated point clouds to train networks for leaf/wood separation. Crucially, the work of Schafer *et al*. [51, 52] showed that models trained on simulated data for biomass estimation perform comparably to those trained on real data, and that synthetic data significantly improves results in limited-data training scenarios. Lu *et al*. [40] aimed at simulating a lidar dataset for wood/leaf segmentation, yet only focused on backpack LiDAR and did not evaluate the dataset via deep learning solutions. These studies validate the promise of sim-to-real transfer in forestry. However, they have primarily focused on element-level segmentation (leaf/wood) and individual tree species classification. Our work addresses a more challenging task of coniferous species semantic segmentation.

First, we present a methodology to generate a massive-scale dataset for species-level semantic segmentation in a complex boreal forest environment. Second, we release this dataset as a public benchmark to promote research. Finally, we move beyond by rigorously evaluating the performance of PTv3 to a real world application and provided a quantitative validation of sim-to-real transfer, demonstrating that our simulated data can significantly improve model performance in data-scarce, real-world scenarios.

3. Methods

3.1. Case Study

Our study focuses on the Quebec portion of the Boreal Shield ecozone, specifically the closed-canopy forest subzone between latitudes 48°N and 52°N. This region is dominated by black spruce (*Picea mariana*) and balsam fir (*Abies balsamea*), which form dense stands with characteristically shady undergrowth. Together, these two species represent over 60% of the forest composition in Quebec [17, 21] and are the most prevalent coniferous species in the Boreal Shield East ecozone [31].

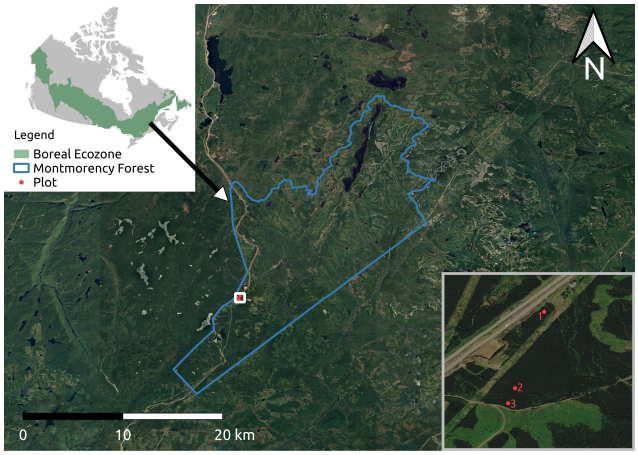


Figure 2. Positions of pure Balsam Fir plot (1), pure Black Spruce (2) and mixed plot (3), on bottom right, relative to Montmorency forest (blue) and the Canadian boreal shield ecozone (top left). Google Earth background.

Our study site is located within the Montmorency Forest (48.28°N, 71.19°W), a research forest representative of the southern edge of the Eastern Boreal Shield (Figure 2). We established three 100m^2 (10 m × 10 m) plots in mature stands, situated away from forest edges to avoid border effects. The plots were selected to represent distinct compositions: one dominated by black spruce (51 trees), a second by balsam fir (31 trees), and a third with a mix of both species (43 trees).

3.2. Real Data Acquisition and Annotation

Field Inventory and ULS Data Collection. The publicly available ALS data for the Montmorency Forest, acquired in 2016 by Quebec's ministry of forests, has a point density of approximately 2.5 points/m^2. This density is too sparse for the detailed structural and species-level analysis required by our study. To create a reference dataset, we conducted a field inventory and a high-density Unmanned Aerial Vehicle (UAV) Laser Scanning (ULS) survey. First, a complete forest inventory was performed for the three plots, where the species and stem position (at breast height) of every tree were precisely recorded using a total station and GNSS. Subsequently, the ULS acquisition was carried out over the plots. The acquisition occurred under sub-optimal meteorological conditions, with winds of approximately 15 km/h and a 5-10 cm layer of snow on the ground. The raw ULS data, which achieved an average density of ∼2900 points/m^2, was post-processed using DJI Terra software. Finally, we completed data collection by acquiring spectral measurements of the needles from both species to obtain reflectance values.

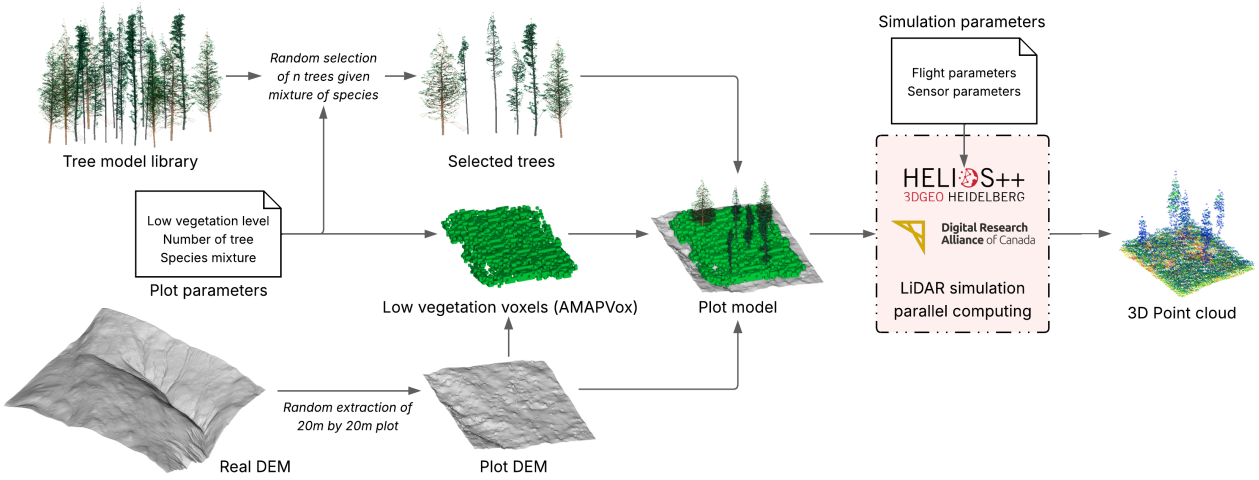


Figure 3. Workflow for generating our virtual LiDAR dataset Quebec Simulated Forest (QC-SF).

Ground Truth Annotation Workflow. To generate the ground truth labels for semantic segmentation, we followed a multi-step workflow. First, the ground was automatically classified using the Simple Morphological Filter (SMRF) algorithm [45] implemented in the PDAL library [10]. Following this, each tree crown was manually segmented from the non-ground points using CloudCompare software. These individualized tree point clouds were then cross-referenced with the field inventory data to assign the correct species label (*Picea mariana* or *Abies balsamea*). Finally, remaining unclassified points located near the ground were assigned to low vegetation class, while all other points were left unlabeled.

Dataset Generation and Augmentation. The final step was to process the dense, annotated ULS point clouds into a dataset suitable for our experiments. To simulate a contemporary ALS acquisition and create our target data, the ULS point clouds were randomly subsampled to a density of 50 points/m^2. To augment the number of available real-world samples, we employed an iterative subsampling strategy. For each of the three high-density plot clouds, we repeatedly generated distinct 50 points/m^2 samples without replacement until the original point cloud was exhausted. This process yielded a final dataset of 158 real-world point cloud samples, each containing approximately 20,000 points. The samples were distributed as follows: 58 from the pure black spruce plot, 58 from the pure balsam fir plot, and 42 from the mixed-species plot.

3.3. Simulated Dataset

We introduce the **Quebec Simulated Forest (QC-SF)**, a synthetic LiDAR dataset composed of 50k procedurally generated virtual forest plots for training and 10k for testing. The design of QC-SF was guided by two primary principles: 1) achieving a high level of realism to minimize the simulation-to-real domain shift, and 2) ensuring maximum representativeness by including a wide range of real-world forest conditions. The general workflow for generating plots is presented in Figure 3.

3.3.1. Forest Plot Modeling

Each virtual forest plot is composed of three primary components: trees, low vegetation (understory), and ground. In contrast to using purely procedural 3D models, like Xfrog or Arbraro for trees, which can accelerate the modeling process at the cost of realism, we prioritized authenticity. For each component, we used models and data that were either validated in previous scientific studies or derived from our acquisitions.

To ensure plot diversity, we designed a three-dimensional parameter space. Each generated plot occupies a unique position in this space, defined by its specific combination of: 1) tree density, 2) species mixture, and 3) understory vegetation level. A representation of that vector space is given in Figure 4.

Ground. We sourced our digital terrain models from two distinct 1 km^2 ALS tiles of forested areas. For the training set, we used a tile from the LidarBC open access dataset [2], selected for its wide variety of slope and terrain types. For the test set, we used a geographically separate tile from Quebec's open forest dataset [1] to ensure model generalization. Ground surfaces for both were extracted using Simple Morphological Filter (SMRF) algorithm. For each plot, a 20 m $\times$ 20 m patch was randomly sampled from the corresponding terrain tile set.

Tree Models. Our tree library is a modified version of the 3D models from Côté *et al*. [16], containing 229 unique tree

instances (145 balsam fir, 84 black spruce). Each tree consists of two separate 3D meshes: one for the wood structure (stem and branches) and another for the foliage (needles). As the original models were highly detailed, we simplified the meshes to a vertex spacing of approximately 5 cm (effective precision of typical ALS systems). The resulting library contains a wide distribution of tree sizes, from young saplings to fully mature trees (Figure 5). To prevent data leakage, the tree model library was partitioned into training (80%) and testing (20%) sets.

Forest Spatial Structure. The spatial structure of each plot was procedurally generated based on its assigned density and species mixture. For each tree to be added, a model was randomly selected from our library (respecting the target species mix) and placed at random coordinates with a random planar rotation. Following the methodology of Schafer *et al*. [52], we did not enforce crown separation, as overlapping crowns are a common and realistic feature of boreal closed-canopy forests. Based on field studies [4], the density of Black Spruce and Balsam Fir stands can range from 1000 to 3000 stems/ha. Our dataset covers and extends this range, with plot densities varying from 5 trees (125 stems/ha) to 128 trees (3200 stems/ha). The species mixture is a continuous value from 0 (100% Balsam Fir) to 1 (100% Black Spruce).

Low Vegetation. We modeled 20 distinct levels of understory vegetation, corresponding to a Leaf Area Index (LAI) ranging from 0.5 to 1.9 m^2/m^2, consistent with values reported by Liu *et al*. [38]. These levels were achieved by procedurally placing on the ground up to four layers of 50 cm turbid voxels (as AMAPvox [56]). The density of these voxels was varied to meet the target LAI for each plot.

Optical Properties. To ensure accurate simulation of intensity signal, we assigned spectral reflectance values to every material at the 1064 nm wavelength. Where direct measurements were unavailable, we referred to values from previous studies:

- **Leaves:** 0.48 for Balsam Fir, 0.32 for Black Spruce (from field acquisition).
- **Bark:** 0.42 for Balsam Fir, 0.21 for Black Spruce (fig. 5 in [50]).
- **Low Vegetation:** 0.56 (avg. reflectance of plants at 1064 nm [43]).
- **Ground:** 0.44 (avg. reflectance at 1064 nm [24, 25]).

3.3.2. LiDAR Simulation

We performed the LiDAR simulations using HELIOS++ [62], a simulator designed for high-performance computing (HPC) environments [22]. We selected this software as it offers an effective balance between simulation speed and physical realism, which was essential for generating our large-scale dataset.

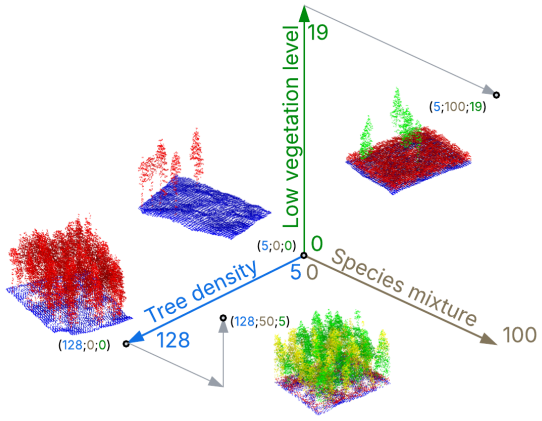


Figure 4. Representation of the vector space of plot parameters of QC-SF dataset. Each plot has a unique coordinate in the 3D space defined by the 3 axis, namely, tree density, species mixture and low vegetation level.

The simulation was parameterized to emulate a typical aerial survey. We defined a virtual platform equipped with a RIEGL VQ-780II sensor, with flight and sensor parameters configured according to the manufacturer's technical specifications [3] and Natural Resources Canada's guidelines for ALS acquisitions [60]. Each flight line was simulated from a 1 km above-ground altitude to achieve an average density of 16.5 points/m^2. To replicate realistic coverage with swath overlaps and reach our final target density, each plot was surveyed with three simulated flight lines from varying angles, resulting in a final density of 50 points/m^2 per plot.

HELIOS++ estimated the LiDAR intensity using a Lambertian reflectance model, as we provided reflectance values at 1064 nm for the scene materials. We chose not to simulate platform or sensor noise, instead treating the idealistic simulated data as a baseline and included acquisition noise at the data augmentation step. This strategy allows for a wider representation of acquisition noises. The simulations were processed on the Digital Research Alliance of Canada's HPC servers. The computation required around 2000h split onto nodes of 128 GB of RAM and 32 cores.

4. Experiments

Using our newly created QC-SF dataset and the corresponding real-world data, we designed two experiments: 1) a semantic segmentation benchmark on the synthetic data, and 2) an evaluation of simulation-to-real transfer learning.

4.1. Semantic Segmentation Benchmarking

Setup. The first experiment uses the semantic segmentation labels of the QC-SF dataset to benchmark the perfor-



Figure 5. Examples of (left) balsam fir (*Abies balsamea*) and (right) black spruce (*Picea mariana*) models from [19]. Colors for visualization only.

mance of five deep learning architectures. These networks were selected to represent the primary strategies in 3D point cloud processing:

- **Voxel-based Convolution:** MinkUNet [14]
- **Point-based Convolution:** KPConv [54]
- **Graph-based Convolution:** DGCNN [57]
- **Space-State-based:** Point Mamba [37]
- **Transformer-based:** Point Transformer v3 [65]

All models were trained using a consistent data augmentation pipeline as applied in the following order: random point dropout, random cuboid dropout [69], random translation, random flips along the X and Y axes, random rotation around the Z axis, random point jittering, and random intensity jittering. Notably, the random point jittering was designed to simulate sensor and platform noise, intentionally omitted during simulation (see 3.3.2), by adding Gaussian noise with a standard deviation of 5 cm. All five of our implementations are available publicly [link to GitHub repository anonymized].

Result discussion. The results of the semantic segmentation benchmark on the QC-SF test set are presented in Table 1. Confirming its status as a state-of-the-art architecture, Point Transformer v3 (PTv3) achieved the highest performance across all metrics, closely followed by Point Mamba.

A detailed analysis reveals that the primary advantage of PTv3 lies in its superior ability to segment ground and low vegetation. Figure 6k shows the low number of segmentation errors on the ground. This suggests its self-attention mechanism is highly effective at capturing the structural patterns and spectral differences that distinguish these classes from the trees. While PTv3 performed well on tree species segmentation (Figure 6), it was outscored by Point Mamba on species segmentation. Especially, Point Mamba proved to be better at identifying black spruce (91.39% IoU).

In contrast, DGCNN's performance lags significantly behind the other models, as illustrated by Figure 6i, where segmentation errors are significantly more visible than the three other methods. This outcome is likely attributable to its limited receptive field; the network was trained on patches of only 4096 points (the original implementation of [57] used 1024 points), representing a small fraction of a full forest plot. As context is crucial for semantic segmentation [44], this constraint likely hindered the model's effectiveness.

Interestingly, as shown by the second row of Figure 6, the majority of segmentation errors are concentrated at the intersection of sub-canopy trees. This finding highlights that networks struggle with the same complex scenarios that make manual labeling difficult, namely, segmenting small, overlapping tree canopies.

These findings align with related work in forestry, such as the benchmarks by Puliti *et al*. [47], which also reported lower performance for DGCNN compared to MinkUNet on tree classification tasks. This underscores a critical point: model performance is highly domain-specific. While DGCNN may be on par with other methods on object-centric benchmarks like ModelNet40 [54], our results show that these capabilities do not necessarily translate to the challenges of semantic segmentation in complex forest scenes.

We provide in the Supplementary Material an in-depth analysis of model performance regarding the simulation parameters: tree density, species mixture, and understory vegetation level.

4.2. Sim-to-Real Transfer Learning

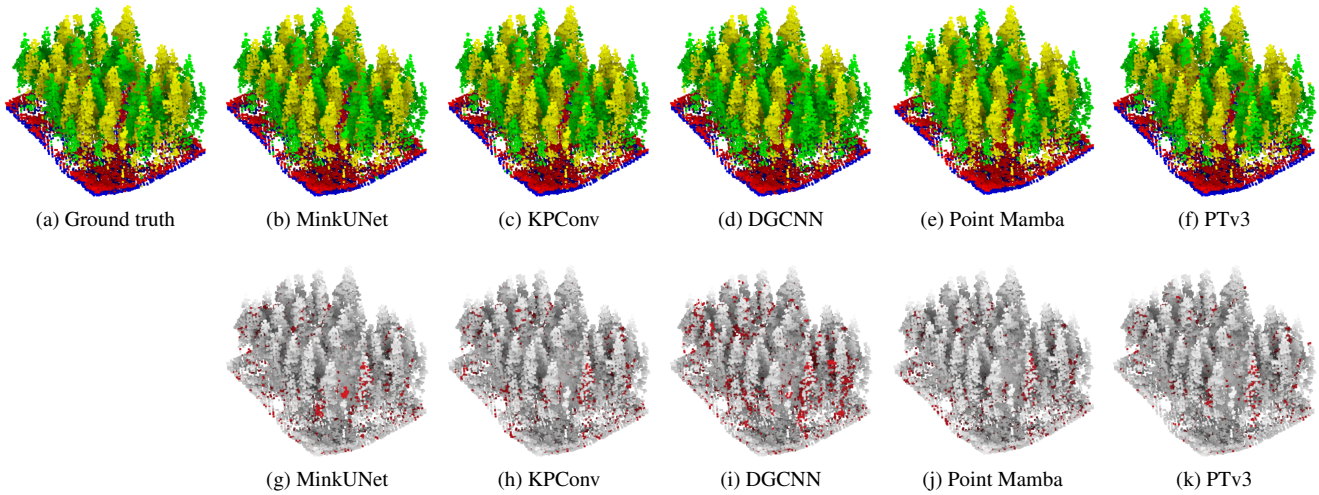
Setup. Our second experiment aimed to evaluate the practical utility of our synthetic dataset. By assessing sim-to-real transfer performance, our goal is to show QC-SF usage outside of the benchmarking. We selected Point Transformer v3 for this task, as it was the top-performing model in our previous experiment. We compared three training strategies :

- **Sim-only**: Trained on QC-SF dataset.
- **Real-only**: Trained on our real-world dataset of 116 samples (baseline).
- **Mixed**: Trained on a combination of the 116 real samples augmented with 500 randomly selected virtual plots from QC-SF. This ratio was choosen to avoid real data dilution.

Models were trained following our first experiment data augmentation procedure.

Table 1. Semantic segmentation results on the QC-SF test set.

Method	OA	mIoU	IoU_{Ground}	IoU_{AB}	IoU_{PM}	IoU_{Low_veg}
MinkUnet [14]	93.47%	87.26%	84.12%	89.40%	90.80%	84.70%
KPConv [54]	93.09%	86.51%	83.79%	89.65%	89.39%	83.21%
DGCNN [57]	91.15%	83.20%	80.96%	87.04%	85.12%	79.67%
Point Mamba [37]	94.48%	89.23%	85.74%	**91.47%**	**91.39%**	88.35%
Point Transformer v3 [65]	**95.61%**	**91.66%**	**93.23%**	**91.40%**	90.81%	**91.20%**



(a) Ground truth (b) MinkUNet (c) KPConv (d) DGCNN (e) Point Mamba (f) PTv3

(g) MinkUNet (h) KPConv (i) DGCNN (j) Point Mamba (k) PTv3

Figure 6. Semantic segmentation results of the benchmarked architectures on a plot from the test set of Quebec Simulated Forest (QC-SF) Dataset. **Top Row:** (a) Ground Truth, and predictions of (b) MinkUNet, (c) KPConv, (d) DGCNN, (e) Point Mamba, and (f) Point Transformer v3. **Bottom Row:** Visualization of segmentation errors for each respective model. Misclassified points are colored in red, while correctly classified points are gray. **Legend:** Ground (blue), Balsam Fir (green), Black Spruce (yellow), Low Vegetation (red).

Ablation study on sim vs real samples. To determine the optimal ratio of synthetic to real data, we conducted an ablation study by training multiple instances of Point Transformer V3. Figure 8 shows the resulting mIoU as a function of the number of synthetic samples added to our 116 real training samples. We found that the ideal range is between 50 and 5,000 synthetic samples. Fewer than 50 samples were insufficient for the model to generalize effectively from the simulation, while adding more than 5,000 samples skewed the training distribution too heavily towards synthetic data, causing the model to disregard the features learned from the real-world examples.

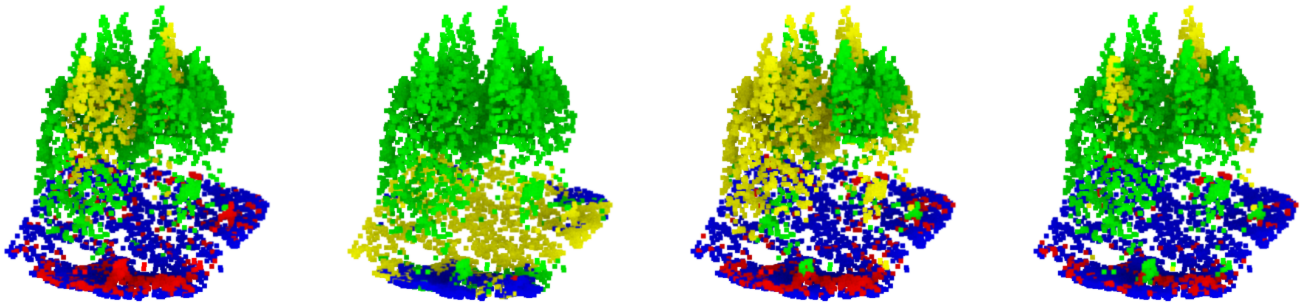
Result discussion. The results are detailed in Table 2. The model trained exclusively on synthetic data (Sim-only) performed poorly when tested on real scans, reaching only 26.33% mIoU. This model completely missed black spruce (0.55% IoU) and low vegetation (0.36% IoU) classes, and Figure 7c highlights the strong confusion between them. This result highlights the existence of a significant domain gap between QC-SF and our real-world dataset. Similar domain gap has been observed in previous studies [63, 64].

Table 2. Performance comparison of Point Transformer v3 trained on simulated, real, and mixed datasets. Tested on the real-world mixed plot (n°3)

Metric	Sim-only	Real-only	Mix (Sim+Real)
OA	59.59%	62.57%	**78.98%**
mIoU	26.33%	46.39%	**52.12%**
IoU_{Ground}	26.64%	79.60%	**81.23%**
IoU_{AB}	77.77%	44.10%	**75.19%**
IoU_{PM}	0.55%	19.79%	**20.62%**
IoU_{Low_V}	0.36%	**42.05%**	31.44%

In contrast, the baseline model trained on the small real-world dataset (Real-only) scored better, reaching 46.39% mIoU. As seen in Figure 7c, this model achieves strong ground detection (79.60% IoU). If the global understanding of the forest structure is present, the real-only model struggles with species segmentation, explaining the low 19.79% IoU score on black spruce class.

Our most significant finding is that training with a mixed



(a) Ground truth (b) Sim-only (c) Real-only (d) Mix (Sim+Real)

Figure 7. Semantic segmentation results of sim-to-real training strategies on real-world plot n°3. (a) Ground Truth, (b) Sim-only model, (c) Real-only model, (d) Mixed (Sim+Real) model. **Observation:** Note the significant reduction of errors in the mixed model (d) compared to the real-only model (c). **Legend:** Ground (blue), Balsam Fir (green), Black Spruce (yellow), Low Vegetation (red).

dataset, even in a few-shot context, gives a substantial performance boost. As illustrated by Figure 7d, the mixed model achieves an mIoU of 52.12%, an improvement of nearly 6 points over the real-only model baseline. This confirms our hypothesis that synthetic data can effectively compensate for real-data scarcity and validates that our modeling and simulation procedures produce meaningful virtual LiDAR data.

The performance gain is most significant for the balsam fir class, where the IoU increased from 44.10% to 75.19%, an increment clearly displayed by Figure 7d. This indicates that the simulated data provides robust examples that help the model generalize better from limited real examples. Furthermore, while the IoU for black spruce only rose by 1% from the baseline, we observed a 16% jump in its contribution to Overall Accuracy, an improvement made clear by comparing Figure 7a and Figure 7d.

However, we observed a trade-off in the low vegetation class, where the mixed model's performance (31.44% IoU) dropped below that of the real-only model (42.05% IoU). This is likely due to a specific domain discrepancy, as the simulated understory does not account for the 5-10 cm snow layer present in the real ULS data. Despite this, the overall gain in mIoU underlines the strong contribution of the simulated data.

5. Conclusion

In this work, we addressed the critical bottleneck of labeled data in forestry by introducing the Quebec Simulated Forest (QC-SF), a large-scale, multi-task, synthetic airborne LiDAR dataset. Representing the Eastern Canadian boreal forest, QC-SF provides 60,000 fully annotated plots. Using QC-SF, we evaluated 3D deep learning architectures on semantic segmentation in a forestry context. Our benchmark identified Point Transformer v3 as the most effective architecture for this environment.

We demonstrated the practical value of our synthetic data through a sim-to-real transfer learning experiment. By augmenting a small number of real-world scans with data from QC-SF, we achieved an improvement of nearly 6% in mIoU. This result validates our data generation framework and confirms that high-quality synthetic data is a viable solution for training powerful models, even when real-world annotated samples are scarce. While this study focused on semantic segmentation, the QC-SF dataset also provides instance and wood/leaf separation labels. In future work, we will provide evaluations for other vision tasks like instance and panoptic segmentation. The QC-SF dataset will be continuously improved, and we plan to expand beyond our two current species (Balsam Fir and Black Spruce) to better represent the boreal forest ecosystems.

QC-SF is a dataset designed to help develop computer vision tools dedicated to forestry. Better computer vision will ultimately allow more accurate ecological models, enabling land managers and foresters to implement more impactful climate change mitigation techniques and better preserve our boreal forests.

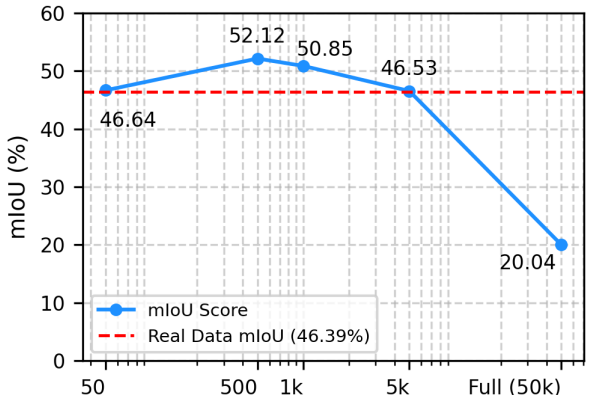


Figure 8. Sim-to-real transfer learning performance (% mIoU) as a function of the number of synthetic training samples on a logarithmic scale. The blue line shows the model's score.

References

[1] Forêt ouverte. 4

[2] LidarBC. 4

[3] RIEGL VQ-780II Product Detail. 5

[4] Dominique Arseneault and Luc Sirois. The millennial dynamics of a boreal forest stand from buried trees. *Journal of Ecology*, 92(3):490–504, 2004. 5

[5] S. Bechtold and B. Höfle. Helios: A multi-purpose lidar simulation framework for research, planning and training of laser scanning operations with airborne, ground-based mobile and stationary platforms. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-3:161–168, 2016. 2

[6] Alexandre Boulch, Gilles Puy, and Renaud Marlet. Fkaconv: Feature-kernel alignment for point cloud convolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[7] S. Briechle, P. Krzystek, and G. Vosselman. Semantic labeling of als point clouds for tree species mapping using the deep neural network pointnet++. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 951–955, 2019. 1, 2

[8] S. Briechle, P. Krzystek, and G. Vosselman. Classification of tree species and standing dead trees by fusing uav-based lidar data and multispectral imagery in the 3d deep neural network pointnet++. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020: 203–210, 2020. 1, 2

[9] Mitch Bryson, Feiyu Wang, and James Allworth. Using Synthetic Tree Data in Deep Learning-Based Tree Segmentation Using LiDAR Point Clouds. *Remote Sensing*, 15(9):2380, 2023. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute. 3

[10] Howard Butler, Andrew Bell, Michael P. Gerlek, chambbj, Pete Gadomski, Connor Manning, Mateusz Łoskot, Bas Couwenberg, Norman Barker, Paul Ramsey, Even Rouault, Julia Dark, Kyle Mann, Guilhem Villemin, Nicolas Chaulet, Claire Foster, Michael Rosen, Ogi Moore, Scott Lewis, Grigory, Kirk McKelvey, Martin Dobias, Isaac Bell, Michael D. Smith, Bram, xantares, Vicky Vergara, Robert Coup, and Kristian Evers. PDAL/PDAL: 2.8.3, 2024. 4

[11] Kim Calders, Glenn Newnham, Andrew Burt, Simon Murphy, Pasi Raumonen, Martin Herold, Darius Culvenor, Valerio Avitabile, Mathias Disney, John Armston, and Mikko Kaasalainen. Nondestructive estimates of above-ground biomass using terrestrial laser scanning. *Methods in Ecology and Evolution*, 6(2):198–208, 2015. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12301. 2

[12] Francisco Raverta Capua, Juan Schandin, and Pablo De Cristóforis. Training point-based deep learning networks for forest segmentation with synthetic data, 2024. arXiv:2403.14115 [cs]. 3

[13] Jaesung Choe, Chunghyun Park, Francois Rameau, Jaesik Park, and In So Kweon. PointMixer: MLP-Mixer for Point Cloud Understanding. In *Computer Vision – ECCV 2022*, pages 620–640. Springer, Cham, 2022. ISSN: 1611-3349. 3

[14] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks, 2019. arXiv:1904.08755 [cs]. 1, 3, 6, 7

[15] Lino Comesaña-Cebral, Joaquín Martínez-Sánchez, Gabriel Suárez-Fernández, and Pedro Arias. Wildfire response of forest species from multispectral LiDAR data. A deep learning approach with synthetic data. *Ecological Informatics*, 81:102612, 2024. 3

[16] Jean-François Côté, Richard A Fournier, and Joan E Luther. Validation of l-architect model for balsam fir and black spruce trees with structural measurements. *Canadian Journal of Remote Sensing*, 39(sup1):S41–S59, 2013. 2, 4

[17] Pierre-Luc Couillard, Serge Payette, Martin Lavoie, and Jason Laflamme. La forêt boréale du québec : influence du gradient longitudinal. *Le Naturaliste canadien*, 143(2):18–32, 2019. 3

[18] Jean-François Côté, Jean-Luc Widlowski, Richard A. Fournier, and Michel M. Verstraete. The structural and radiative consistency of three-dimensional tree reconstructions from terrestrial lidar. *Remote Sensing of Environment*, 113 (5):1067–1081, 2009. 2

[19] Jean-François Côté, Richard A. Fournier, Gordon W. Frazer, and K. Olaf Niemann. A fine-scale architectural model of trees to enhance lidar-derived measurements of forest canopy structure. *Agricultural and Forest Meteorology*, 166-167:72–85, 2012. 6

[20] M. I. Disney, M. Boni Vicari, A. Burt, K. Calders, S. L. Lewis, P. Raumonen, and P. Wilkes. Weighing trees with lasers: advances, challenges and opportunities. *Interface Focus*, 8(2):20170048, 2018. Publisher: Royal Society. 2

[21] Martin Dupuis, Antoine Leboeuf, Éric Vaillancourt, Valérie Roy, Jonathan Arsenault, Arnold Bennett, and Mélanie Major. Photographic Interpretation of Forest Species in Southern Québec. Technical report, 2021. 3

[22] Alberto M. Esmorís, Miguel Yermo, Hannah Weiser, Lukas Winiwarter, Bernhard Höfle, and Francisco F. Rivera. Virtual LiDAR Simulation as a High Performance Computing Challenge: Toward HPC HELIOS++. *IEEE Access*, 10:105052–105073, 2022. 5

[23] Fabian Fassnacht, Hooman Latifi, Krzysztof Stereńczak, Michael Lefsky, Christoph Straub, Lars Waser, Aniruddha Ghosh, and Aneta Modzelewska. Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, 184, 2016. Review Lidar for forestery and species classification. 1

[24] Petri R. Forsström, Aarne Hovi, Jussi Juola, and Miina Rautiainen. Links between light availability and spectral properties of forest floor in European forests. *Agricultural and Forest Meteorology*, 336:109481, 2023. 5

[25] Petri R. Forsström, Jussi Juola, Aarne Hovi, and Miina Rautiainen. Dataset of understory reflectance spectra and fractional cover in a boreal forest area in Finland. 3, 2024. Publisher: Mendeley Data. 5

[26] Jean-Philippe Gastellu-Etchegorry, Tiangang Yin, Nicolas Lauret, Thomas Cajgfinger, Tristan Gregoire, Eloi Grau, Jean-Baptiste Feret, Maïlys Lopes, Jordan Guilleux, Gérard Dedieu, Zbyněk Malenovský, Bruce Douglas Cook, Douglas

Morton, Jeremy Rubio, Sylvie Durrieu, Gregory Cazanave, Emmanuel Martin, and Thomas Ristorcelli. Discrete anisotropic radiative transfer (dart 5) for modeling airborne and satellite spectroradiometer and lidar acquisitions of natural and urban landscapes. *Remote Sensing*, 7(2):1667–1701, 2015. 2

[27] Sylvie Gauthier, Pierre Bernier, Timo Kuuluvainen, Anatoly Z Shvidenko, and Dmitry G Schepaschenko. Boreal forest health and global change. *Science*, 349(6250):819–822, 2015. 1

[28] Charles Gaydon and Floryne Roche. PureForest: A Large-Scale Aerial Lidar and Aerial Imagery Dataset for Tree Species Classification in Monospecific Forests, 2024. arXiv:2404.12064 [cs]. 1

[29] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[30] Adam A. Goodenough and Scott D. Brown. DIRSIG5: Next-Generation Remote Sensing Data and Image Simulation Framework. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(11):4818–4833, 2017. 2

[31] Txomin Hermosilla, Alex Bastyr, Nicholas C. Coops, Joanne C. White, and Michael A. Wulder. Mapping the presence and distribution of tree species in Canada's forested ecosystems. *Remote Sensing of Environment*, 282:113276, 2022. 3

[32] Ekaterina Kalinicheva, Loic Landrieu, Clément Mallet, and Nesrine Chehata. Predicting vegetation stratum occupancy from airborne lidar data with deep learning. *arXiv preprint arXiv:2201.08051*, 2022. 1, 2

[33] Sean Krisanski, Mohammad Sadegh Taskhiri, Susana Gonzalez Aracil, David Herries, and Paul Turner. Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning. *Remote Sensing*, 13(8), 2021. 1, 2

[34] Loïc Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. *CoRR*, abs/1711.09869, 2017. 1

[35] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. PointMamba: A Simple State Space Model for Point Cloud Analysis, 2024. arXiv:2402.10739 [cs]. 1

[36] Bernd Lintermann and Oliver Deussen. Interactive modeling of plants. *IEEE Computer Graphics and Applications*, 19(1): 56–65, 1999. 2

[37] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point Mamba: A Novel Point Cloud Backbone Based on State Space Model with Octree-Based Ordering Strategy, 2024. arXiv:2403.06467 [cs]. 1, 6, 7

[38] Yang Liu, Ronggao Liu, Jan Pisek, and Jing M. Chen. Separating overstory and understory leaf area indices for global needleleaf and deciduous broadleaf forests by fusion of MODIS and MISR data. *Biogeosciences*, 14(5):1093–1110, 2017. Publisher: Copernicus GmbH. 5

[39] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-Voxel CNN for Efficient 3D Deep Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1

[40] Yawen Lu, Zhuoyang Sun, Jinyuan Shao, Qianyu Guo, Yunhan Huang, Songlin Fei, and Yingjie Chen. LiDAR-Forest Dataset: LiDAR Point Cloud Simulation Dataset for Forestry Application, 2024. arXiv:2402.04546 [cs]. 3

[41] Yaotao Luo, Donghui Xie, Jianbo Qi, Kun Zhou, Guangjian Yan, and Xihan Mu. LESS LiDAR: A full-waveform and discrete-return multispectral LiDAR simulator based on ray tracing algorithm. *Remote Sensing*, 15(18):4529, 2023. Publisher: MDPI. 2

[42] JH McCaughey, A Barr, TA Black, B Goodison, EH Hogg, B Amiro, DT Price, RB Stewart, N Stolle, J Chen, et al. The boreal ecosystem research and monitoring sites (berms) initiative: scientific accomplishments. In *The International Boreal Forest Research Association's 2000 International Science Conference*, pages 129–135, 2002. 1

[43] Audrey Mercier, Susanna Karlqvist, Aarne Hovi, and Miina Rautiainen. Reflectance spectra of boreal forest floor elements: ferns, herbaceous plants and decaying wood. 1, 2023. Publisher: Mendeley Data. 5

[44] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-Context Data Augmentation for 3D Scenes, 2021. arXiv:2110.02210 [cs]. 6

[45] Thomas J. Pingel, Keith C. Clarke, and William A. McBride. An improved simple morphological filter for the terrain classification of airborne LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 77:21–30, 2013. 4

[46] Stefano Puliti, Grant Pearse, Peter Surový, Luke Wallace, Markus Hollaus, Maciej Wielgosz, and Rasmus Astrup. FOR-instance: a UAV laser scanning benchmark dataset for semantic and instance segmentation of individual trees, 2023. arXiv:2309.01279 [cs]. 3

[47] Stefano Puliti, Emily R. Lines, Jana Müllerová, Julian Frey, Zoe Schindler, Adrian Straker, Matthew J. Allen, Lukas Winiwarter, Nataliia Rehush, Hristina Hristova, Brent Murray, Kim Calders, Louise Terryn, Nicholas Coops, Bernhard Höfle, Samuli Junttila, Martin Krůček, Grzegorz Krok, Kamil Král, Shaun R. Levick, Linda Luck, Azim Missarov, Martin Mokroš, Harry J. F. Owen, Krzysztof Stereńczak, Timo P. Pitkänen, Nicola Puletti, Ninni Saarinen, Chris Hopkinson, Chiara Torresan, Enrico Tomelleri, Hannah Weiser, and Rasmus Astrup. Benchmarking tree species classification from proximally-sensed laser scanning data: introducing the FOR-species20K dataset, 2024. arXiv:2408.06507 [cs]. 1, 2, 6

[48] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 1

[49] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 2

[50] Miina Rautiainen, Petr Lukeš, Lucie Homolová, Aarne Hovi, Jan Pisek, and Matti Mõttus. Spectral Properties of Conifer-

ous Forests: A Review of In Situ and Laboratory Measurements. *Remote Sensing*, 10(2):207, 2018. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 5

[51] Jannika Schäfer, Hannah Weiser, Lukas Winiwarter, Bernhard Höfle, Sebastian Schmidtlein, and Fabian Ewald Fassnacht. Generating synthetic laser scanning data of forests by combining forest inventory information, a tree point cloud database and an open-source laser scanning simulator. *Forestry: An International Journal of Forest Research*, 96(5):653–671, 2023. 2, 3

[52] Jannika Schäfer, Lukas Winiwarter, Hannah Weiser, Bernhard Höfle, Sebastian Schmidtlein, Jan Novotnỳ, Grzegorz Krok, Krzysztof Stereńczak, Markus Hollaus, and Fabian Ewald Fassnacht. CNN-based transfer learning for forest aboveground biomass prediction from ALS point cloud tomography. *European Journal of Remote Sensing*, 57 (1):2396932, 2024. Publisher: Taylor & Francis. 3, 5

[53] Olivier Stocker, Reza Mahmoudi Kouhi, Eric Guilbert, Antonio Ferraz, and Thierry Badard. Investigating the Impact of Point Cloud Density on Semantic Segmentation Performance Using Virtual Lidar in Boreal Forest. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 978–981, 2023. ISSN: 2153-7003. 3

[54] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *CoRR*, abs/1904.08889, 2019. 1, 6, 7

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[56] Grégoire Vincent, Cécile Antin, Marilyne Laurans, Julien Heurtebize, Sylvie Durrieu, Claudia Lavalley, and Jean Dauzat. Mapping plant area index of tropical evergreen forest by airborne laser scanning. A cross-validation study using LAI2200 optical sensor. *Remote Sensing of Environment*, 198:254–266, 2017. 5

[57] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019. 1, 3, 6, 7

[58] Jason Weber and Joseph Penn. Creation and rendering of realistic trees. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 119–128, 1995. 2

[59] Hannah Weiser, Lukas Winiwarter, Katharina Anders, Fabian Ewald Fassnacht, and Bernhard Höfle. Opaque voxel-based tree models for virtual laser scanning in forestry applications. *Remote Sensing of Environment*, 265:112641, 2021. 2

[60] Joanne C. White, Michael A. Wulder, Andrés Varhola, Mikko Vastaranta, Nicholas C. Coops, Bruce D. Cook, Doug Pitt, and Murray Woods. A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. *The Forestry Chronicle*, 89(06):722–723, 2013. Publisher: Canadian Institute of Forestry. 5

[61] J.-L. Widlowski, B. Pinty, M. Lopatka, C. Atzberger, D. Buzica, M. Chelle, M. Disney, J-P. Gastellu-Etchegorry, M. Gerboles, N. Gobron, E. Grau, H. Huang, A. Kallel, H. Kobayashi, P. E. Lewis, W. Qin, M. Schlerf, J. Stuckens, and D. Xie. The fourth radiation transfer model intercomparison (rami-iv): Proficiency testing of canopy reflectance models with iso-13528. *Journal of Geophysical Research: Atmospheres*, 118(13):6869–6890, 2013. 2

[62] Lukas Winiwarter, Alberto Manuel Esmorís Pena, Hannah Weiser, Katharina Anders, Jorge Martínez Sánchez, Mark Searle, and Bernhard Höfle. Virtual laser scanning with helios++: A novel take on ray tracing-based simulation of topographic full-waveform 3d laser scanning. *Remote Sensing of Environment*, page 112772, 2021. 2, 5

[63] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893, 2018. ISSN: 2577-087X. 2, 7

[64] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. SqueezeSegV2: Improved Model Structure and Unsupervised Domain Adaptation for Road-Object Segmentation from a LiDAR Point Cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4376–4382, 2019. ISSN: 2577-087X. 7

[65] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point Transformer V3: Simpler, Faster, Stronger, 2024. arXiv:2312.10035 [cs]. 1, 6, 7

[66] Binbin Xiang, Maciej Wielgosz, Theodora Kontogianni, Torben Peters, Stefano Puliti, Rasmus Astrup, and Konrad Schindler. Automated forest inventory: Analysis of high-density airborne LiDAR point clouds with 3D deep learning. *Remote Sensing of Environment*, 305:114078, 2024. 2, 3

[67] Tiangang Yin, Nicolas Lauret, and Jean-Philippe Gastellu-Etchegorry. Simulation of satellite, airborne and terrestrial lidar with dart (ii): Als and tls multi-pulse acquisitions, photon counting, and solar noise. *Remote Sensing of Environment*, 184:454–468, 2016. 2

[68] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. 1

[69] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-Supervised Pretraining of 3D Features on Any Point-Cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 6

[70] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015. IEEE. 2