Scaling Up 3D Forest Vision with Synthetic LiDAR

Anonymous CVPR submission

Paper ID *****

Abstract

Accurate tree segmentation from forest laser scans is essential to understanding ecosystem functions in carbon cycling and beyond. We have developed a new synthetic data generation pipeline to do this for forest vision tasks, integrating advances in game-engines with physics-based LiDAR simulation. As a result, we have produced a comprehensive, diverse, annotated 3D forest dataset on an unprecedented scale. Extensive experiments with a state-of-the-art tree segmentation algorithm and a popular real dataset show that our synthetic data can substantially reduce the need for labelled real data. After fine-tuning on just a single, real, forest plot of ***less than 0.1 hectare****, the pretrained model achieves segmentations that are competitive with a model trained on the full scale real data. We have also identified critical factors for successful use of synthetic data: physics, diversity, and scale, paving the way for more robust 3D forest vision systems in the future.*

1. Introduction

Understanding forest dynamics and monitoring their status have long been a priority in the Sustainable Development Goals of the United Nations [17]. For close-range forest sensing to obtain metrics for individual trees, the typical workflow involves obtaining point clouds of forests through laser scanning and segmenting individual trees. The next step is to extract the properties of interest, such as above-ground biomass [4], species [13], tree health [3], crown volume, canopy height and diameter at breast height (DBH) [8], among others. Accurate measurement of these properties is essential for understanding key ecosystem functions related to carbon cycling and beyond (Fig. 1). The accurate segmentation of individual trees is a key bottleneck in the entire pipeline.

The increasing complexity of tree segmentation algorithms requires more data and computation for a stronger and more generalisable model performance [16]. Despite extensive data collection efforts [2, 12, 21], the available 3D forest dataset for the development of tree segmentation algorithms still falls short of the scale typically required for robust machine learning [19]. This is not surprising given that forest scanning requires field work that is limited by multiple factors, such as site accessibility, weather conditions, and seasons. In addition, the intricate canopy structures of forests make it challenging and time consuming to delineate the semantic labels and instance labels. For example, Puliti et al. [12] report that just annotating 2.79 hectares of forest plots takes two skilled researchers six months.

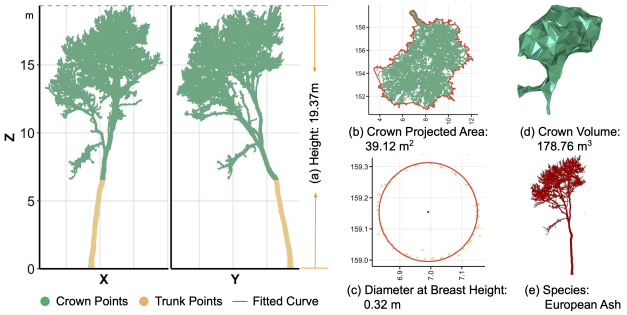

Figure 1. **Individual tree segmentation is essential for deriving key forest properties in ecology.** *This figure shows examples of metrics derived from individual tree segmentation: canopy height for biomass and carbon estimation, and crown area, crown volume, and species composition for understanding forest structure, biodiversity, and ecosystem functions.*

Synthetic data can significantly reduce the effort to collect data in the field and then to annotate the data, as many existing synthetic forest datasets have demonstrated [1, 6, 9, 20]. Nevertheless, its effectiveness depends on whether an algorithm trained on the synthetic data can generalise well to the real world. A key factor is the diversity of synthetic data on large scales: with enough diversity, synthetic data could allow the algorithm trained on it to generalize to real-world forests [11, 18], and with large-scale data, neural networks can benefit from adequate computations for effective weight and bias learning [16]. Under this assumption, the learning paradigm for a forest vision algorithm would involve pretraining on synthetic data, followed by fine-tuning on real data from the region of interest.

Our approach sources forest scenes from the gaming in-

dustry and regenerates them using procedural foliage generation, originally developed in Unreal Engine but, as we find, grounded in ecological forest-growth principles. Experiments show that the algorithm pretrained with our synthetic data gets significant improvement than training from scratch on a few real plots. Fine-tuning the pretrained algorithm on just one forest plot achieves segmentation accuracy comparable to full-scale real data, although this plot can represent as little as 2.2% of the labelled dataset. In summary, our contributions include:

- a synthetic data generation pipeline for forest vision tasks that combine advances in game engines and physics-based simulation;
- a comprehensive, diverse, and large-scale synthetic dataset of 3D point-clouds for tree segmentation tasks;
- experiments showing the effectiveness of the synthetic dataset, and revealing that physics-based simulation, scene diversity, and large dataset scale are all crucial for synthetic data to be effective for training.

2. Methods

2.1. Synthetic data generation pipeline

Procedural foliage generation. We generate forest scenes for simulation using Unreal Engine's procedural foliage generation algorithm, which enables large-scale 3D forest generation from a limited set of foliage models. This procedural foliage generation algorithm, despite having originated from the game industry, is broadly consistent with ecological theory of forests originating and forming canopy structure through light and resource competition [7, 10, 14]: given a bare plot of land, starting from seeds, trees grow and expand to fill all available space, but subject to constraints on light, represented by collision and shade radius in this algorithm. After enough years, or simulation steps, the scene reaches a state of equilibrium — a closed canopy — while still respecting individual space requirements.

With our generation method established, we then select and adapt forest scenes from available datasets and asset packs to build a diverse, simulation-ready collection. We first adapt five natural forest scenes from the SPREAD dataset [5], originally developed for generating 2D images for trunk segmentation. The dataset contains six natural forests, four urban tree scenes, and one plantation; one natural forest scene ("Burned Wood") is excluded, leaving three deciduous forests, one redwood forest, and one rainforest.

Because the SPREAD collection lacks European coniferous forests and has limited European deciduous coverage, we supplement it with seven additional forest scenes originally from the Unreal Engine Marketplace. These scenes are selected for their ecological relevance, specifically European coniferous and deciduous forests, and for the availability of high-quality tree assets at the time of collection. No methodological limit is imposed, so more scenes can be incorporated as suitable assets become available. In total, the dataset comprises twelve adapted forest scenes, covering 75 ha.

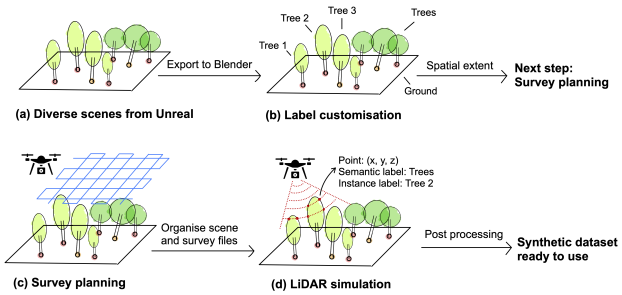

Figure 2. **Synthetic data generation pipeline, written in Python, and fully automated to execute each step in this figure.** *The pipeline processes a forest scene from Unreal Engine by (a) adding files to Blender, (b) customizing instance and semantic labels, (c) planning the UAV flight path, setting up a virtual laser scanner, and (d) structuring files for LiDAR survey. It completes post-processing to produce a machine learning dataset. This process can be executed with a single command line, allowing parameter adjustments through arguments.*

Virtual LiDAR survey. Once we have obtained the forest scenes, we use the HELIOS simulator [23] to simulate UAV laser scans and generate point clouds with the desired annotations. To this end, we have developed a pipeline written in Python that will convert UE forest scenes, customise the labels, plan the virtual UAV flight for the survey and conduct the LiDAR simulation (Fig. 2).

Table 1. **Our virtual survey resembles a real world UAV mission.** *The RIEGL VUX-1 UAV scanner is widely used.*

Parameter	Value
Scanner Model	RIEGL VUX-1 UAV
Pulse Frequency	100 kHz
Max Returns (100 kHz)	15
Flight Pattern	Criss-cross
Flight Spacing	20.0 m
Flight Speed	5.0 m/s
Relative Altitude	60.0 m

An attractive aspect of synthetic data is that it can provide instance-level labels along with the data. Based on the customised forest, our pipeline will generate a survey file for each scene that will be used by HELIOS to conduct the survey. Our virtual survey resembles how UAV-laser scans work in the real world. Tab. 1 gives an overview of the key parameters of the survey. With the forest scenes and survey files ready, the pipeline will proceed with the LiDAR simulation to generate UAV-laser scans with the desired annotations.

2.2. Generated synthetic dataset

The synthetic data generation pipeline simulates UAV laser scans for 12 forest scenes — 5 adapted from the original SPREAD collection, and 7 new scenes. The dataset includes 4 coniferous and 6 deciduous scenes, focusing on European forests (Fig. 3). Two additional scenes, Rainforest and Redwood, are retained from the original SPREAD collection for completeness, but are not our main regional focus. The data set is prepared for machine learning by merging point clouds from each scene, tiling them into 50m $\times$ 50m plots, and splitting them into training (70%), validation (15%) and test sets (15%). All plots have a point density greater than 1000 pts/m^2, adequate for effective tree segmentation learning [22].

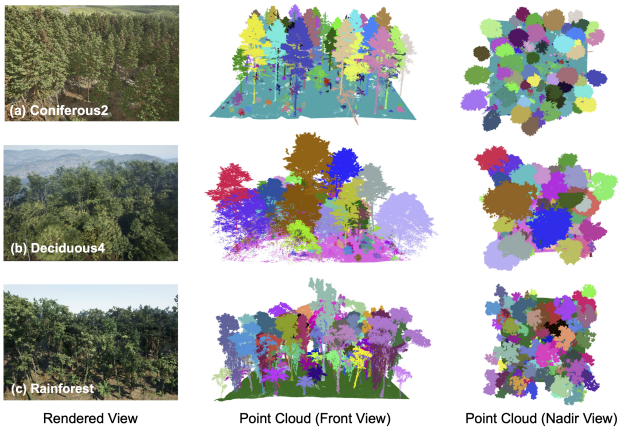

Figure 3. **The synthetic data is diverse, physics-based, and available at large scale.** *Using Unreal Engine, large-scale, photorealistic forest scenes are created using gaming assets and procedural foliage. Then the pipeline simulates laser scans to produce synthetic point clouds with instance labels (shown here in colours). The 50m $\times$ 50m tiled plots are ready for machine learning.*

2.3. Experiments

Benchmark dataset from real world FOR-Instance [12] is a recently published UAV laser scan dataset that has previously been used for tree segmentation development and evaluation [15, 22, 24], making it a suitable choice for this study. The FOR-Instance data set covers 5 forest regions, including 3 coniferous regions (Coniferous-N, Coniferous-S, Coniferous-S) and 2 deciduous regions (Deciduous-R, Deciduous-T), totalling 2.79 hectares of forest plots.

Tree segmentation algorithm For the tree segmentation algorithm, we use ForAINet from Xiang et al. [24] (denoted by $\mathcal{A}$), which was developed using the FOR-Instance dataset and considered the state-of-the-art algorithm for instance and semantic segmentation of forest point clouds.

Few-shot learning We first pre-train $\mathcal{A}$ on the full-scale synthetic $\text{Sim}_{11,187}$ data for 150 epochs, where 'Sim' denotes synthetic data, 'Real' denotes real FOR-Instance data, the first subscript indicates the number of scenes, and the second the number of plots. $\text{Sim}_{11,187}$ consists of 11 scenes, totalling 13635 cylinder samples. This pre-training should provide a strong feature extraction network benefiting from the diversity of the synthetic data that includes coniferous, deciduous, and rainforest scenes. The pretrained model is then fine-tuned on the corresponding coniferous and deciduous forest plots from the real FOR-Instance dataset. Unless otherwise specified, we report metrics as the mean over all cylinder samples obtained during inference, each with its own accuracy, recall, and F-score.

We have built two few shot datasets, subsets of FOR-Instance with few labelled plots to assess model performance with limited data:

- Real-$C_{3,3}$, contains the 3 coniferous scenes from the original FOR-Instance dataset, one plot from each coniferous region (Real-C-$N_{1,1}$, Real-C-$C_{1,1}$, and Real-C-$S_{1,1}$).
- Real-$D_{2,2}$, contains the 2 deciduous scenes from FOR-Instance dataset, one plot from each deciduous region (Real-D-$R_{1,1}$ and Real-D-$T_{1,1}$). (These are the only deciduous plots available in the FOR-Instance dataset due to the difficulty in scanning and annotating deciduous forests.)

We compare our few-shot learning results with the results obtained from training with $\mathcal{A}$ on the full-scale FOR-Instance dataset $\text{Real}_{6,42}$, which contains 6 scenes, totalling 42 plots and 1051 samples (Tab. 2), to see if we can obtain competitive results with only a few forest plots.

Table 2. **Real datasets used in few shot learning experiments.** *The subscripts denote the number of scenes and the number of plots. Besides the full dataset $\text{Real}_{6,42}$, we organised few-shot subsets for coniferous (Real-C) and deciduous forests (Real-D). We report cylinder sample counts from each subset and their percentages of the full dataset.*

Dataset	Scenes	Plots	Samples	Percentage (%)
$\text{Real}_{6,42}$	Coniferous-N, Coniferous-C, Coniferous-S, Coniferous-N2 Deciduous-R, Deciduous-T	42	1051	100
Real-$C_{3,3}$	Coniferous-N, Coniferous-C, Coniferous-S	3	76	7.2
Real-C-$N_{1,1}$	Coniferous-N	1	32	3.1
Real-C-$C_{1,1}$	Coniferous-C	1	31	2.9
Real-C-$S_{1,1}$	Coniferous-S	1	23	2.2
Real-$D_{2,2}$	Deciduous-R, Deciduous-T	2	160	15.2
Real-D-$R_{1,1}$	Deciduous-R	1	72	6.8
Real-D-$T_{1,1}$	Deciduous-T	1	88	8.4

3. Results

Synthetic data can enhance instance segmentation with minimal use of real plots In Tab. 3, we compare the performance of $\mathcal{A}$ trained in different settings in coniferous and deciduous plots.

Training from scratch with the reduced coniferous data set Real-$C_{3,3}$ leads to a significant drop in F1 score of more than 29%. Pretraining with synthetic data and then

Table 3. **Pretraining on synthetic data, and using minimal real data fine-tuning, performs competitively with a much larger, labelled, real dataset.** *F1 scores improve significantly for small, labelled coniferous regions when synthetic (mixed deciduous/coniferous) data is available for pre-training. Deciduous plots show a similar pattern, with the same pretraining data.*

Test	Pretrain	Train/Tune	Experiment	Real Data Used (%)	F1-score (%)
Coniferous-N	-	$\text{Real}_{6,42}$	Training on all real plots from scratch	100	79.8
	-	$\text{Real-C}_{3,3}$	Training on 3 coniferous plots from scratch	7.2	50.5
	$\text{Sim}_{11,187}$	$\text{Real-C}_{3,3}$	Fine-tuning on 3 coniferous plots	7.2	79.4
	$\text{Sim}_{11,187}$	$\text{Real-C-N}_{1,1}$	Fine-tuning on 1 plot from Coniferous-N	3.1	77.1
Deciduous-R	-	$\text{Real}_{6,42}$	Training on all real plots from scratch	100	69.9
	-	$\text{Real-D}_{2,2}$	Training on 2 deciduous plots from scratch	15.2	31.6
	$\text{Sim}_{11,187}$	$\text{Real-D}_{2,2}$	Fine-tuning on 2 deciduous plots	15.2	70.0
	$\text{Sim}_{11,187}$	$\text{Real-D-R}_{1,1}$	Fine-tuning on 1 plot from Deciduous-R	6.8	69.3

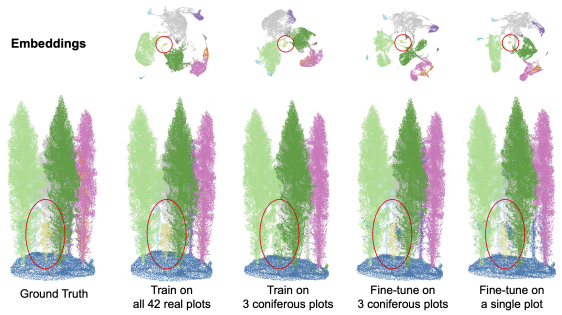

Figure 4. **Pretraining on synthetic data enables the algorithm to learn representations like those from real data, with fine-tuning on just a single forest plot.** *We present a sample from Coniferous-N, where embeddings and points shifted by offsets serve as the foundation, for instance segmentation. Training only on three coniferous plots, the algorithm struggles to learn disentangled representations, as seen in segmentation results where small trees in the understory are not segmented, noted within the red circle.*

fine-tuning with $\text{Real-C}_{3,3}$ fully restores the drop in performance. Even extreme few-shot training, using only 3.1% of the samples, is competitive, dropping just a couple of percentage points on F1. Fig. 4 shows learnt representations and segmentation results for a Coniferous-N sample in various settings; synthetic data pretraining facilitates the learning of comparable representations, even with minimal fine-tuning. A similar pattern of improvement can be observed in Deciduous-R (Tab. 3) and other deciduous plots.

Key elements contributing to the effectiveness of synthetic data In Sec. 1, we point out three key strengths of our synthetic data generation pipeline: diverse forest scenes, large-scale forest generation, and physics-based laser scans for point generation. We perform ablation studies to study the role of each of the three. We pre-train $\mathcal{A}$ on each of these synthetic datasets in the same setting of pre-training on $\text{Sim}_{11,187}$, the full scale synthetic data (Tab. 4). Then the zero-shot learning experiments directly evaluate the model obtained in each forest region from the real datasets.

We use four forest scenes from the SPREAD dataset [5], each with 5 plots, totalling 20 plots. The data set incorporates tree mesh points that arguably approximate point clouds. We refer to the point clouds from tree mesh points as $\text{Nodal}_{4,20}$, and those from UAV-LiDAR simulation with HELIOS as $\text{Sim}_{4,20}$.

Tab. 4 shows the results for Coniferous-C. Training with UAV-LiDAR simulation ($\text{Sim}_{4,20}$) significantly enhances performance, whereas nodal point clouds from tree mesh ($\text{Nodal}_{4,20}$) performs poorly: F1 score in Tab. 4 improved from 5.4% to 42.6%.

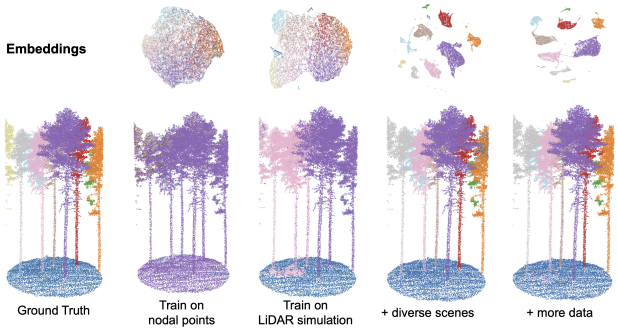

Figure 5. **Synthetic data helps the algorithm learn disentangled representations.**

We extend $\text{Sim}_{4,20}$ to $\text{Sim}_{11,55}$ to study scene diversity. $\text{Sim}_{11,20}$ adds three deciduous and four coniferous scenes not in $\text{Sim}_{4,20}$ and $\text{Sim}_{4,20}$, with each scene having 5 plots like $\text{Sim}_{4,20}$. There is consistent improvement across all regions, with Coniferous-C's F-1 score increasing by 38.9% from 42.6% to 81.5%.

We use the full-scale synthetic data set $\text{Sim}_{11,187}$ to study the effects of the size of the data set. Compared with $\text{Sim}_{11,55}$, it includes all available plots (17 per forest scene) for the 11 scenes. More training data helps the model learn scene variations and enhances F1 performance from 81.5% to 87%.

Table 4. **Three factors enhance the value of synthetic data for training: physics-based laser scans, scene diversity, and increased data.**

Test	Train	Ablation Experiment	F1-score (%)
Coniferous-C	$\text{Nodal}_{4,20}$	Nodal points from 4 scenes (5 plots per scene)	5.4
	$\text{Sim}_{4,20}$	Simulated LiDAR from the same 4 scenes (5 plots per scene)	42.6
	$\text{Sim}_{11,55}$	+7 additional scenes (total 11 scenes, 5 plots each)	81.5
	$\text{Sim}_{11,187}$	+More plots per scene (total 11 scenes, 17 plots each)	87.0

4. Conclusion

We have introduced a novel synthetic data generation pipeline for 3D forest vision to address a key challenge in this domain— the scarcity of labelled data for supervised machine learning. We used the pipeline to create an extensive 3D forest dataset that covers 12 scenes and 75 hectares, exceeding the size of any real dataset in this field that we have encountered so far. Comprehensive experiments confirmed the efficacy of the synthetic dataset for learning representations for instance segmentation.

References

[1] Mitch Bryson, Ahalya Ravendran, Celine Mercier, Tancred Frickey, Sadeepa Jayathunga, Grant Pearse, and Robin JL Hartley. Domain adaptation of deep neural networks for tree part segmentation using synthetic forest trees. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 14: 100078, 2024. 1

[2] Kim Calders, Hans Verbeeck, Andrew Burt, Niall Origo, Joanne Nightingale, Yadvinder Malhi, Phil Wilkes, Pasi Raumonen, Robert GH Bunce, and Mathias Disney. Laser scanning reveals potential underestimation of biomass carbon in temperate forest. *Ecological Solutions and Evidence*, 3(4): e12197, 2022. 1

[3] Aland HY Chan, Chloe Barnes, Tom Swinfield, and David A Coomes. Monitoring ash dieback (hymenoscyphus fraxineus) in british forests using hyperspectral remote sensing. *Remote Sensing in Ecology and Conservation*, 7(2): 306–320, 2021. 1

[4] David A Coomes, Michele Dalponte, Tommaso Jucker, Gregory P Asner, Lindsay F Banin, David FRP Burslem, Simon L Lewis, Reuben Nilus, Oliver L Phillips, Mui-How Phua, et al. Area-based vs tree-centric approaches to mapping forest carbon in southeast asian forests from airborne laser scanning data. *Remote Sensing of Environment*, 194: 77–88, 2017. 1

[5] Zhengpeng Feng, Yihang She, and Srinivasan Keshav. Spread: A large-scale, high-fidelity synthetic dataset for multiple forest vision tasks. *Ecological Informatics*, 87: 103085, 2025. 2, 4

[6] Vincent Grondin, Jean-Michel Fortin, François Pomerleau, and Philippe Giguère. Tree detection and diameter estimation based on deep learning. *Forestry: An International Journal of Forest Research*, 96(2):264–276, 2022. 1

[7] Francis Hallé and Roelof AA Oldeman. Essay on the architecture and dynamics of growth of tropical trees. 1970. 2

[8] Xinlian Liang, Yunsheng Wang, Jiri Pyörälä, Matti Lehtomäki, Xiaowei Yu, Harri Kaartinen, Antero Kukko, Eija Honkavaara, Aimad EI Issaoui, Olli Nevalainen, et al. Forest in situ observations using unmanned aerial vehicle as an alternative of terrestrial measurements. *Forest ecosystems*, 6(1):1–16, 2019. 1

[9] Yawen Lu, Yunhan Huang, Su Sun, Tansi Zhang, Xuewen Zhang, Songlin Fei, and Victor Chen. M2fnet: Multi-modal forest monitoring network on large-scale virtual dataset. *arXiv preprint arXiv:2402.04534*, 2024. 1

[10] Stephen W Pacala, Charles D Canham, John Saponara, John A Silander Jr, Richard K Kobe, and Eric Ribbens. Forest models defined by field measurements: estimation, error analysis and dynamics. *Ecological monographs*, 66(1):1–43, 1996. 2

[11] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE, 2019. 1

[12] Stefano Puliti, Grant Pearse, Peter Surovỳ, 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. *arXiv preprint arXiv:2309.01279*, 2023. 1, 3

[13] Stefano Puliti, Emily R Lines, Jana Müllerová, Julian Frey, Zoe Schindler, Adrian Straker, Matthew J Allen, Lukas Winiwarter, Nataliia Rehush, Hristina Hristova, et al. Benchmarking tree species classification from proximally-sensed laser scanning data: introducing the for-species20k dataset. *arXiv preprint arXiv:2408.06507*, 2024. 1

[14] Drew W Purves, Jeremy W Lichstein, and Stephen W Pacala. Crown plasticity and competition for canopy space: a new spatially implicit model parameterized for 250 north american tree species. *PloS one*, 2(9):e870, 2007. 2

[15] Adrian Straker, Stefano Puliti, Johannes Breidenbach, Christoph Kleinn, Grant Pearse, Rasmus Astrup, and Paul Magdon. 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, 2023. 3

[16] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019. 1

[17] Latha Swamy, Erika Drazen, Whitney R Johnson, and Jacob J Bukoski. The future of tropical forests under the united nations sustainable development goals. *Journal of sustainable forestry*, 37(2):221–256, 2018. 1

[18] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 1

[19] Shahadat Uddin and Haohui Lu. Dataset meta-level and statistical features affect machine learning performance. *Scientific Reports*, 14(1):1670, 2024. 1

[20] Di Wang. Unsupervised semantic and instance segmentation of forest point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:86–97, 2020. 1

[21] Hannah Weiser, Jannika Schäfer, Lukas Winiwarter, Nina Krašovec, Fabian E Fassnacht, and Bernhard Höfle. Individual tree point clouds and tree measurements from multi-platform laser scanning in german forests. *Earth System Science Data*, 14(7):2989–3012, 2022. 1

[22] Maciej Wielgosz, Stefano Puliti, Binbin Xiang, Konrad Schindler, and Rasmus Astrup. Segmentanytree: A sensor and platform agnostic deep learning model for tree segmentation using laser scanning data. *arXiv preprint arXiv:2401.15739*, 2024. 3

[23] 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*, 269:112772, 2022. 2

[24] 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. 3