


## CNN-based transfer learning for forest aboveground biomass prediction from ALS point cloud tomography

Jannika Schäfer<sup>a</sup>, Lukas Winiwarter<sup>b,c</sup>, Hannah Weiser<sup>d</sup>, Bernhard Höfle<sup>d,e</sup>, Sebastian Schmidlein<sup>a</sup>, Jan Novotný<sup>f</sup>, Grzegorz Krok<sup>g</sup>, Krzysztof Stereńczak<sup>g</sup>, Markus Hollaus<sup>b</sup> and Fabian Ewald Fassnacht<sup>h</sup> 

<sup>a</sup>Institute of Geography and Geoecology (IFGG), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany; <sup>b</sup>Photogrammetry Research Area, Department of Geodesy and Geoinformation, Wien, Austria; <sup>c</sup>Department of Basic Sciences in Engineering Sciences, University of Innsbruck, Innsbruck, Austria; <sup>d</sup>3DGeo Research Group, Institute of Geography, Heidelberg University, Heidelberg, Germany; <sup>e</sup>Interdisciplinary Centre for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany; <sup>f</sup>Global Change Research Institute of the Czech Academy of Sciences, Brno, Czech Republic; <sup>g</sup>Department of Geomatics, Forest Research Institute, Sekocin Stary, Raszyn, Poland; <sup>h</sup>Remote Sensing and Geoinformatics, Freie Universität Berlin, Berlin, Germany

### ABSTRACT

This study presents a new approach for predicting forest aboveground biomass (AGB) from airborne laser scanning (ALS) data: AGB is predicted from sequences of images depicting vertical cross-sections through the ALS point clouds. A 3D version of the VGG16 convolutional neural network (CNN) with initial weights transferred from pre-training on the ImageNet dataset was used. The approach was tested on datasets from Canada, Poland, and the Czech Republic. To analyse the effect of training sample size on model performance, different-sized samples ranging from 10 to 375 ground plots were used. The CNNs were compared with random forest models (RFs) trained on point cloud metrics. At the maximum number of training samples, the difference in RMSE between observed and predicted AGB of CNNs and RFs ranged from  $-2$  t/ha to 5 t/ha, and the difference in squared Pearson correlation coefficient ranged from  $-0.05$  to 0.06. Additional pre-training on synthetic data derived from virtual laser scanning of simulated forest stands could only improve the prediction performance of the CNNs when only a few real training samples (10–40) were available. While 3D CNNs trained on cross-section images derived from real data showed promising results, RFs remain a competitive alternative.

### ARTICLE HISTORY

Received 7 April 2024  
Revised 30 July 2024  
Accepted 22 August 2024

### KEYWORDS



Forest; airborne laser scanning (ALS); deep learning; random forest; virtual laser scanning; synthetic data

## Introduction

Forests play an important role in the global carbon cycle, being the main terrestrial carbon sink (Pan et al., 2011). Deforestation and forest degradation contribute to anthropogenic carbon emissions, while carbon is sequestered through forest growth and the expansion of forest areas (Dixon et al., 1994). To effectively monitor forests and to investigate the effects of anthropogenic and non-anthropogenic influences on forest status, large-scale data on forest structure is required. Airborne laser scanning (ALS) allows to non-destructively obtain information on the three-dimensional structure of forests over large areas. It is therefore increasingly used for estimating stocks and changes of forest aboveground biomass (AGB) (Strimbu et al., 2023).

In the most commonly applied area-based approach (Næsset, 2002), ground measurements of AGB are linked to metrics describing the distribution of the spatially co-located laser scanning returns. Regression models can subsequently be employed for wall-to-wall AGB prediction from the ALS data. Both non-linear and linear models as well as non-

parametric machine learning methods, such as nearest neighbour interpolation, support vector machines, and random forest (RF), are frequently applied for relating the point cloud metrics to the AGB observations (Fassnacht et al., 2014). While many studies have shown that AGB can be predicted from point cloud metrics (e.g. Bouvier et al., 2015; Sheridan et al., 2015; Zhao et al., 2009), it remains an open question whether AGB estimates could be further improved by using more information from individual returns, i.e. the xyz-coordinates of the returns inherent in discrete return ALS data, rather than aggregated point cloud metrics. Depending on the point density, information about individual trees (e.g. their location, height, and volume) may be present in the raw point cloud data that is lost when using aggregated metrics at the plot level. Deep learning algorithms offer the potential to test this hypothesis without the need to specifically detect individual trees as done in earlier studies (e.g. Dalponte et al., 2018; Jucker et al., 2017), as they do not require handcrafted and pre-extracted

**CONTACT** Jannika Schäfer  [jannika.schaefer@kit.edu](mailto:jannika.schaefer@kit.edu)  Institute of Geography and Geoecology (IFGG), Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, Karlsruhe 76131, Germany

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

features. They can process raw data such as images or point clouds, thereby learning a latent representation of the data during the model optimization (LeCun et al., 2015).

Previous studies demonstrated that deep learning on ALS data can be used to classify tree species, coniferous and deciduous trees, as well as dead trees and snags (e.g. Briechle et al., 2021; Hamraz et al., 2019; Hell et al., 2022), and to estimate forest attributes such as growing stock volume and AGB (e.g. Ayrey & Hayes, 2018; Ayrey et al., 2021; Balazs et al., 2022; Oehmcke et al., 2022; Seely et al., 2023). The applied methods include 2D convolutional neural networks (CNNs) that are applied on 2D projections of the point clouds (e.g. Balazs et al., 2022; Briechle et al., 2021; Hamraz et al., 2019), 3D CNNs for which the point clouds are binned into a voxel space (e.g. Ayrey & Hayes, 2018; Ayrey et al., 2021; Balazs et al., 2022; Oehmcke et al., 2022), and deep learning algorithms that take the raw point clouds as input, such as PointNet, KPConv, 3DmFV-Net, or PointCNN (e.g. Hell et al., 2022; Oehmcke et al., 2022; Seely et al., 2023).

The limiting factor for the development of deep learning applications for inferring forest attributes from ALS data is the large demand for labelled training data (Hamedianfar et al., 2022). In the aforementioned studies that predicted forest attributes on a plot level, 1044–17432 plots were used for model training, an additional 225–1000 plots for model validation and 225–3000 plots for model testing. Such large sample sizes are rarely available in forestry applications. In a review on remote sensing-based forest AGB estimations, Fassnacht et al. (2014) reported that only 9 of 90 reviewed studies had a sample size between 200 and 500 plots, and 66 of 90 studies had a sample size smaller than 100 plots.

Common techniques for dealing with limited data availability are data augmentation and transfer learning (Hamedianfar et al., 2022). Data augmentation can increase the number of training data, for example, by flipping, rotating, and cropping images (Shorten & Khoshgoftaar, 2019), or rotating, scaling, jittering, and point-wise displacement of point clouds (R. Li et al., 2020). As the neural network learns how to represent the data in the optimization process, these examples show transformations of the data that are invariate with respect to the output. However, many of these methods were developed for classification or object detection but not for regression tasks and may not be directly applicable for some datasets, for example, if the scales in the image are related to the response variable (Hwang & Whang, 2022). In transfer learning, models pre-trained on other data are further trained on a small sample of the target data, reducing the amount of labelled target data required (Hamedianfar et al., 2022). A lot of training efforts

are required to recognize basic shapes like edges and corners, and these efforts can be transferred across domains. Transfer is generally more successful the closer the target and the source domain are, but it has been shown to work even across quite contrasting domain pairs (Niu et al., 2021). There are many pre-defined CNN architectures available with weights derived from training on large image datasets, such as those from the ImageNet database (Deng et al., 2009; Kattenborn et al., 2021). In contrast, pre-trained models do not yet exist for point-based deep learning methods for vegetation analysis (Winiwarter et al., 2022).

Pre-trained CNNs have been successfully applied for forestry applications, e.g. Briechle et al. (2021) used ResNet-18 models with pre-trained ImageNet weights for the classification of tree species and standing dead trees. However, there is a domain gap between the ImageNet images and those images that can be derived from laser scanning or other remote sensing techniques of forest. Accordingly, Fuller et al. (2022) found that pre-training on satellite images instead of ImageNet images improved the performance of land-cover classification from satellite images when using a vision transformer architecture. Another possibility for model pre-training is the use of simulated data (Winiwarter et al., 2022). Data simulation is a cost- and time-efficient way to generate large amounts of labelled training data. Luo et al. (2022) used synthetic forest scenes composed of randomly placed individual tree point clouds to train a deep learning model for tree detection in forest laser scanning point clouds acquired by an unoccupied aerial vehicle (UAV) and Sun et al. (2022) trained a deep learning model on synthetic images generated by Generative Adversarial Networks to segment individual tree crowns from ALS canopy height models.

In this study, we investigated whether deep learning can be applied to predict AGB from point clouds when the size of the training dataset is rather small. To compensate for the limited amount of training data, we used a transfer learning approach. We employed a CNN architecture developed by Solovyev et al. (2022) that is fed with sequences of 2D frames. Solovyev et al. (2022) created 3D versions of popular 2D CNNs that have been pre-trained on ImageNET images. These 3D CNNs have been successfully used to detect stalled brain capillaries in stacks of mouse brain images. Here, we apply this method to predict AGB from ALS tomography, i.e. the sequences of images derived from vertical cross-sections through ALS point clouds. Since the cross-section images look quite different from the ImageNet images, we tested whether implementing an additional pre-training step with cross-section images derived from simulated laser scanning point clouds can help in the domain transfer and increase prediction performance.

These synthetic data were generated by combining virtual forest stands and a laser scanning simulator. The performance of the 3D CNNs was evaluated using datasets obtained from four study sites. The number of ground plots used for training and validation was varied, starting from 10 and ranging up to 35, 97, 167, and 375, depending on the study site. As a benchmark of model performance, AGB was also predicted from point cloud metrics using a random forest model.

The research questions we addressed in this study were:

- (1) To what extent can AGB of forest stands be estimated from stacked cross-section images derived from the ALS point clouds using a 3D version of the VGG16 CNN pre-trained on the ImageNet dataset?
- (2) How does the training sample size influence the prediction performance of the CNN?
- (3) How does additional pre-training on synthetic data influence the prediction performance of the CNN?

## Material and methods

### Study sites

We used ALS point clouds and corresponding forest inventory data from four sites that were collected 1) in the Petawawa Research Forest (PRF) in Ontario, Canada, 2) in the Milicz Forest (MF) district in the south-west of Poland, 3) in the Silesian Beskids (SB) in the east of the Czech Republic, and 4) in the DendroNET sites (DN) that are spread across the Czech Republic. Table 1 provides an overview on

ground data including main tree species and ALS data acquisitions at the four sites. For MF, individual tree information (species, diameter at breast height ( $D_{1.3}$ ), and tree height) was available and AGB was estimated using the same models as for the synthetic data (see next section). For the other sites, we used the AGB reference values that were provided by the original data owners.

PRF is a remote sensing supersite that covers approximately 10000 ha of mixed-wood forests. The open-access data of PRF have been described in detail by White et al. (2019). The stand density in the 223 circular ground plots ranged from 32 to 13024 trees/ha, with a mean value of 2500 trees/ha (Figure 1). The AGB ranged from 1 to 529 t/ha, with a mean value of 158 t/ha.

In MF, 70% of the 500 circular ground plots were located in pure stands of *Pinus sylvestris* L (Stereńczak et al., 2018). The stand density ranged from 20 to 3957 trees/ha and AGB ranged from 15 to 368 t/ha, with mean values of 952 trees/ha and 160 t/ha, respectively.

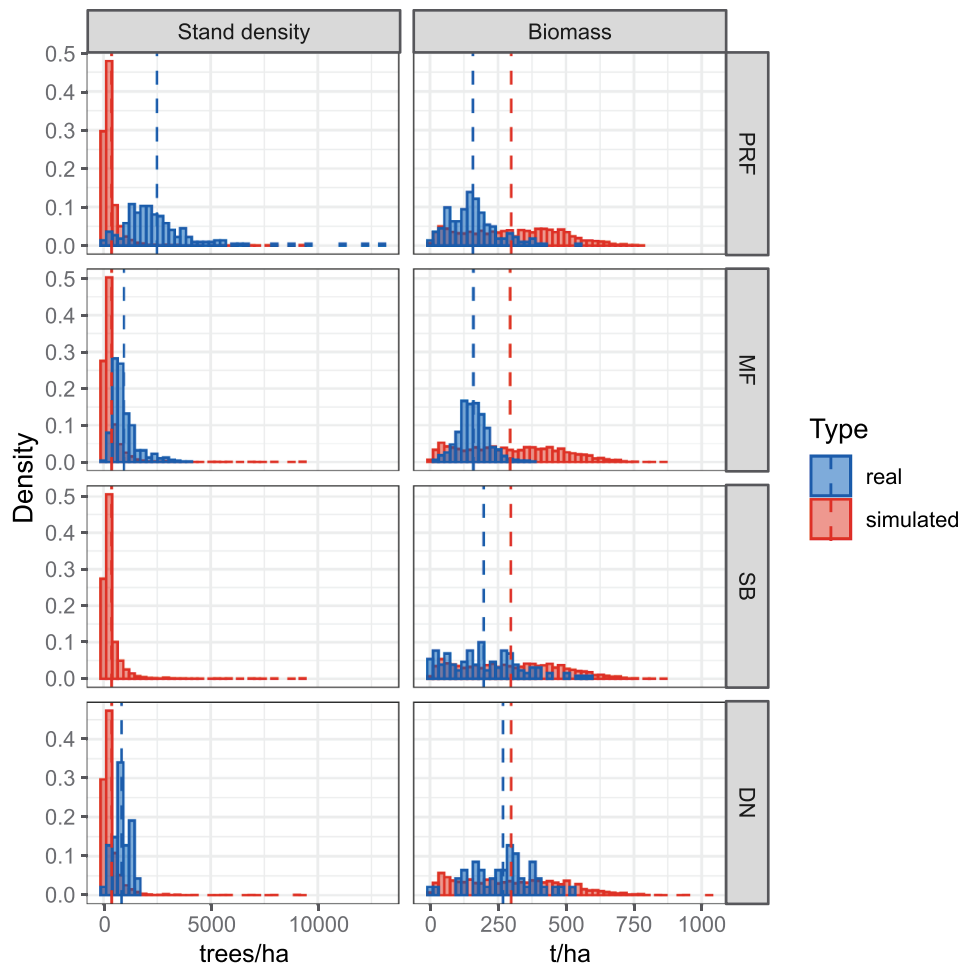
For SB, ground data of 130 plots were available. Tree data were collected from nested circular plots. Trees with a  $D_{1.3} > 7$  cm were sampled within a radius of 3 m, whereas trees with a  $D_{1.3} > 12$  cm were sampled within a radius of 12.62 m. Information on stand density was not available. The average AGB was 198 t/ha, ranging from 2 to 583 t/ha. More information on the SB data can be found in Brovkina et al. (2022).

The 47 DN plots were mostly located in pure forest stands. Ground data were collected from square plots of 30 m × 30 m. ALS data were available for 20 m × 20 m plots, therefore plot AGB values have been calculated from trees within these smaller plots based on individual tree positions. The stand density (measured in 30 m × 30 m) ranged from 89 to 1600 trees/ha, with

**Table 1.** Ground data collection and laser scanning acquisition settings, and the resulting mean pulse density and mean planar point density of the four study sites. Numbers in square brackets indicate settings/values of the laser scanning simulations differing from the reported settings/values of the real acquisitions.

	Petawawa Research Forest (PRF)	Milicz Forest (MF)	Silesian Beskids (SB)	DendroNET (DN)
Forest inventory time	2014	Summer 2015	July 2019	October 2021
Ground plot design	circular plots, radius: 14.1 m	circular plots, radius: 12.62 m	nested circular plots, max. radius: 12.62 m	square plots, side length: 30 m
Number of ground plots	223	500	130	47
Main tree species	<i>Pinus strobus</i> L., <i>Populus tremuloides</i> Michx., <i>Quercus rubra</i> L., <i>Pinus resinosa</i> Ait., <i>Betula papyrifera</i> , <i>Picea glauca</i> Moench	<i>Pinus sylvestris</i> L., <i>Fagus sylvatica</i> L., <i>Quercus</i> spp. L.	<i>Picea abies</i> (L.) H. Karst, <i>Fagus sylvatica</i> L.	<i>Picea abies</i> (L.) H. Karst, <i>Pinus sylvestris</i> L., <i>Fagus sylvatica</i> L.
ALS acquisition time	2012	Summer 2015	July 2019	October 2021
Sensor	RIEGL LMS-Q680i	RIEGL LMS-Q680i	RIEGL LMS-Q780	RIEGL LMS-Q780
Altitude above ground	750 m	480–620 m [480 m]	819 m	515 m
Flight speed	? [54 m/s]	54 m/s	56 m/s	56 m/s
Flight line distance	250 m [≈ 242 m]	? [≈ 296 m]	440 m	–
Flight pattern	parallel <sup>a</sup>	parallel <sup>a</sup>	parallel <sup>a</sup>	perpendicular <sup>a</sup>
Mean pulse density	5.4 pulses/m <sup>2</sup> [4.3 pulses/m <sup>2</sup> ]	9.4 pulses/m <sup>2</sup> [12.4 pulses/m <sup>2</sup> ]	7.0 pulses/m <sup>2</sup> [9.7 pulses/m <sup>2</sup> ]	13.2 pulses/m <sup>2</sup> [18.0 pulses/m <sup>2</sup> ]
Mean planar point density	11.7 points/m <sup>2</sup> [8.3 points/m <sup>2</sup> ]	19.9 points/m <sup>2</sup> [24.4 points/m <sup>2</sup> ]	12.8 points/m <sup>2</sup> [17.8 points/m <sup>2</sup> ]	23.9 points/m <sup>2</sup> [31.7 points/m <sup>2</sup> ]

<sup>a</sup>The laser scanning simulations were performed with flight strips that were not perfectly parallel/perpendicular to reflect deviations from the flight pattern in the real data.



**Figure 1.** Stand density and aboveground biomass (AGB) of the real and simulated forest plots for the four study areas of Petawawa Research Forest (PRF), Milicz Forest (MF), Silesian Beskids (SB), and DendroNET sites (DN). Mean values are indicated by the dashed vertical lines. Information on stand density was not available for the Silesian Beskids (SB).

a mean value of 846 trees/ha. The AGB ranged from 1 to 528 t/ha, with a mean value of 268 t/ha.

### Synthetic data

Four datasets of synthetic forest inventory information and corresponding simulated ALS point clouds were generated, one for each study site. Forest stand compositions were simulated using Forest Factory 2.0 (Henniger et al., 2023), a software that makes use of the individual-based forest gap model FORMIND (Köhler & Huth, 1998). The original Forest Factory 2.0 version is calibrated to generate square forest plots of 20 m × 20 m. Because the ground plots of the real data exceeded this size, we used a modified version of Forest Factory 2.0 that enables to generate forest plots of 30 m × 30 m. The virtual forest stands were composed of *Pinus sylvestris* L., *Picea abies* (L.) H. Karst, *Fagus sylvatica* L., and *Quercus* spp. L. Tree biomass was calculated using species-specific allometric models of the German National Forest Inventory that are implemented in the R package “rBDAT” (Vonderach et al., 2021). By default, Forest Factory 2.0 generates

many more small AGB plots than large AGB plots (Schäfer, Winiwarter, et al., 2023). To avoid the effects of unbalanced training data, we sought to simulate data with AGB values equally distributed over a range of the real datasets. We therefore simulated a large number of forest stands (4565200). Of these virtual stands, all stands with an AGB of 0–600 t/ha were grouped into 12 bins, each bin with an AGB range of 50 t/ha. We then randomly sampled 200 stands per bin, resulting in a selection of 2400 forest stands in total. Forest Factory 2.0 randomly assigns tree locations within a forest stand, which means that trees can be located unrealistically close together. We therefore implemented a workflow to generate new tree locations: For each forest stand, we created a grid of possible tree locations with an Euclidean distance of 1 cm between the locations and a minimum distance of 20 cm to the plot borders. We randomly selected one of these possible locations and assigned it to the first tree in the plot. All locations that were within the crown radius of that tree were then excluded from the remaining possible locations. This allowed for partial overlaps of trees, as the stem of

the second tree could be placed directly at the edge of the first tree's crown. This procedure was continued for all trees within the plot.

Laser scanning of the virtual stands was simulated with the open-source laser scanning simulation framework HELIOS++ version 1.1.2 (Winiwarter et al., 2022). 3D representations of the virtual forest stands were created using tree point clouds that were extracted from real laser scanning data of temperate forests in the south-west of Germany. These data were acquired by a RIEGL miniVUX-1UAV mounted on a UAV during leaf-on conditions. The dataset of tree point clouds and the corresponding tree metrics has been published by Weiser, Schäfer, Winiwarter, Krašovec, Seitz, et al. (2022) and described in detail by Weiser, Schäfer, Winiwarter, Krašovec, Fassnacht, et al. (2022). For each tree in the virtual forest stands, a tree point cloud of matching tree species and a tree height within  $\pm 4$  m of the virtual tree's height were randomly selected and placed at the location of the virtual tree. If no tree point cloud was available in the specified height range, the one with the smallest difference in height was selected. The tree point clouds were randomly rotated around the z-axis and uniformly scaled in all three dimensions such that the height of the point cloud matched the height of the virtual tree. All points outside of the  $30 \text{ m} \times 30 \text{ m}$  stands were removed. Laser scanning was simulated for scenes of  $90 \text{ m} \times 90 \text{ m}$  composed of nine virtual forest stands. The forest point clouds were voxelized with a voxel size of 3 cm. Filled voxels (with at least one point) were set to be opaque, and empty voxels were transparent for the simulation (following Weiser et al., 2021). The ground was represented by a horizontal plane. Airborne laser scanning of the virtual forest scenes was simulated according to the acquisition settings of the real campaigns (Table 1), i.e. four different ALS simulations were conducted over the same scenes.

The virtual forest stands and simulated ALS point clouds were cropped to match the real ground plots (circular plots with radii of 12.62 m or 14.1 m, and square plots with side lengths of 20 m, depending on the dataset used, see Table 1). When there was no tree located within the cropped plot area, the synthetic forest stand was removed from the dataset. This resulted in 2379 synthetic plots for PRF, 2362 for SB, and 2340 for DN. Due to an error in the sampling of the virtual forest stands, the dataset for MF consisted of 2426 synthetic plots (of which 2362 resulted from the sampling with regard to a uniform AGB distribution). Since the distribution was barely influenced by the additional 64 plots, the sampling was not repeated. Plot AGB was calculated as the sum of AGB of all trees within a plot divided by the plot area. The characteristics of the synthetic forest stands differed slightly depending on plot size and shape. The stand density

ranged from 16 to 9414 trees/ha and the AGB ranged from 1 to 1028 t/ha, with mean values of 377 trees/ha and 298 t/ha, respectively (Figure 1).

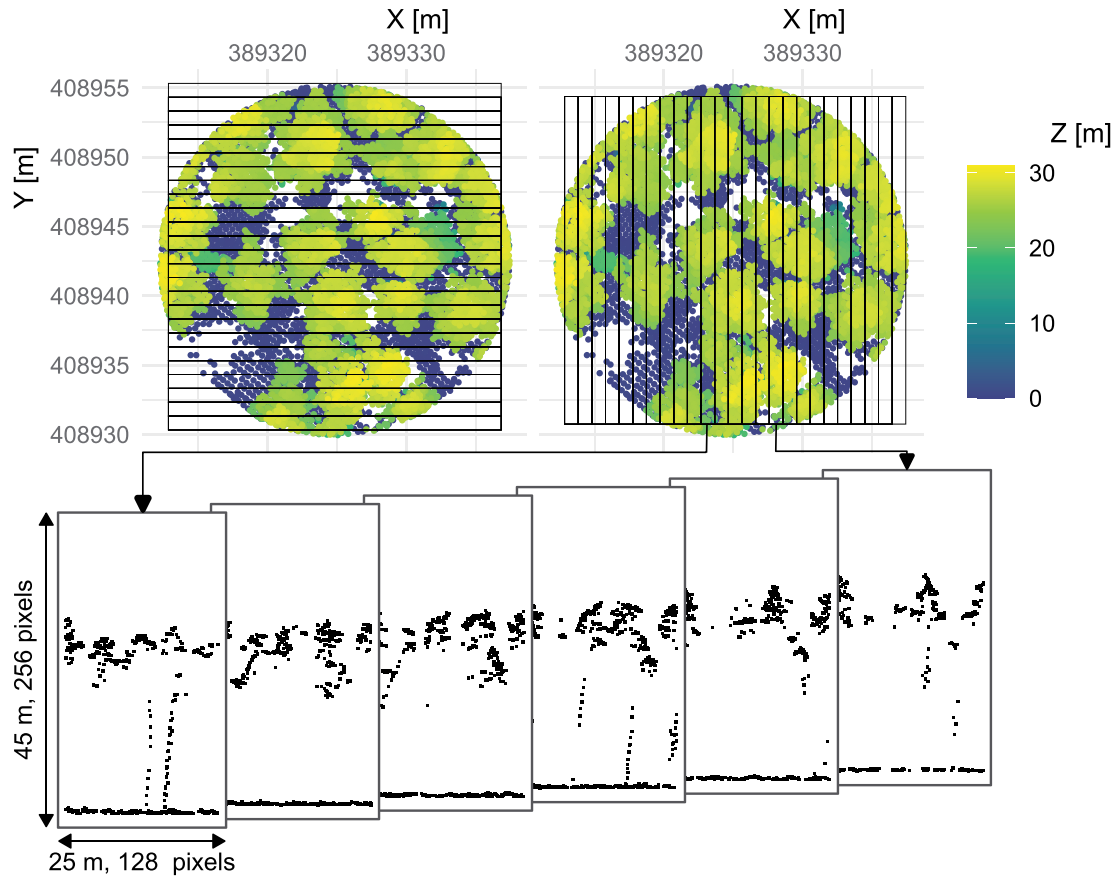
### Cross-section images

In order to feed point cloud data into a CNN, a rasterization is required. Therefore, vertical cross-section images were extracted from real and simulated ALS point clouds. To that end, the point clouds were height normalized using the “normalize height” function of the R package “lidR” (Roussel et al., 2020), and cut into 1 m thick vertical non-overlapping slices, both in the x-direction and in the y-direction. The slices were 45 m high (corresponding to the tallest trees) and the width corresponded to the respective plot extent (20–28.2 m). A binary image of  $128$  (width)  $\times$   $256$  (height) pixels was generated from each slice. Pixel values were set to black (0) if the volume represented by the pixel contained at least one ALS return and to white (1) otherwise. Preliminary experiments on the MF dataset revealed that using RGB-images with a height-related colour map (viridis) did not improve the results. Figure 2 shows exemplary cross-section images for one MF ground plot.

### Experimental set-up

All experiments were conducted individually for each of the four study sites. The real data were split into 25% test and 75% training data using a stratified sampling approach, so that both test and training datasets represented the full range of AGB values of the respective study sites. This resulted in 56, 125, 33, and 12 test plots for PRF, MF, SB, and DN, respectively. The remaining plots were used for model training and validation.

For model training, we utilized CNNs that had been pre-trained on ImageNet data (Section 2.4.1). Then, we compared the performance of models trained on cross-section images obtained from real data with models that were first additionally pre-trained on cross-section images derived from synthetic data before being trained on cross-section images derived from real data. As a benchmark, we employed random forest models (RFs) (Breiman, 2001) with point cloud metrics as AGB predictors. Both CNNs and RFs were trained on randomly selected subsamples of varying sizes, as well as on the complete training datasets of each study site (167 plots for PRF, 375 plots for MF, 97 plots for SB, and 35 plots for DN). The size of the subsamples ranged from 10 to 100 (in increments of 10) training plots for PRF and MF, 10–90 plots for SB, and 10–30 plots for DN (due to the different numbers of available plots for each study site). For the CNNs, the training datasets were further randomly split into 80% actual training data and 20% validation data, used



**Figure 2.** Generation of cross-section images from an ALS point cloud of a ground plot located in the Milicz Forest. For better visualization, the black pixels in the cross-section images have been enlarged compared to the actual images and only a selection of images is shown. Each pixel represents a volume of approximately  $19.5 \text{ cm} \times 17.6 \text{ cm} \times 100 \text{ cm}$ . The vertical and horizontal black lines indicate the borders of the transects extracted from each plot point cloud.

for optimizing the model hyperparameters. Both CNNs and RFs were also trained using synthetic data only (no real forest plots).

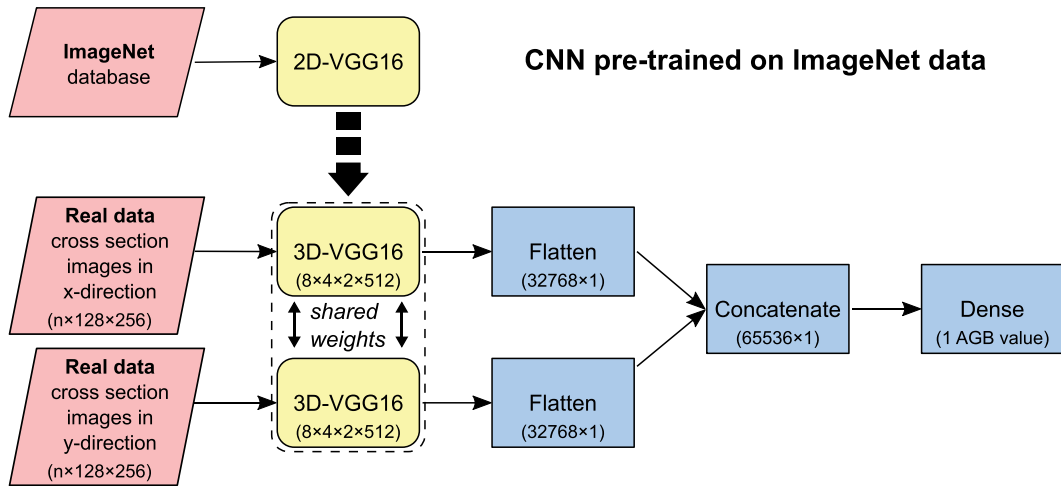
For each sample size, the random sampling of the training data and subsequent model training were repeated 10 times to enhance the informative value of the performance metrics to account for the influence of the randomly selected subsamples on the performance metrics. Due to the extensive time required for training, conducting additional repetitions was not feasible. Model performance was assessed using the median values of the root mean squared error (RMSE), the squared Pearson correlation coefficient ( $r^2$ ) and, as a measure of systematic error, the mean error (ME) of observed and predicted AGB values for the test datasets.

The CNNs were conducted on a system with an NVIDIA RTX A4000 GPU, 16 GB VRAM, 256 GB RAM, and an Intel® Xeon(R) Silver 4210 R CPU @ 2.40 GHz  $\times$  40. The RFs were conducted on a system with 256 GB RAM, an Intel® Xeon(R) CPU E5-2630 v3 @ 2.40 GHz  $\times$  16, and no dedicated GPU. We did not systematically assess the run time of the models. CNN training took between 4 minutes and 6 hours, depending on the training sample size and the number of epochs before early stopping. Because of the large

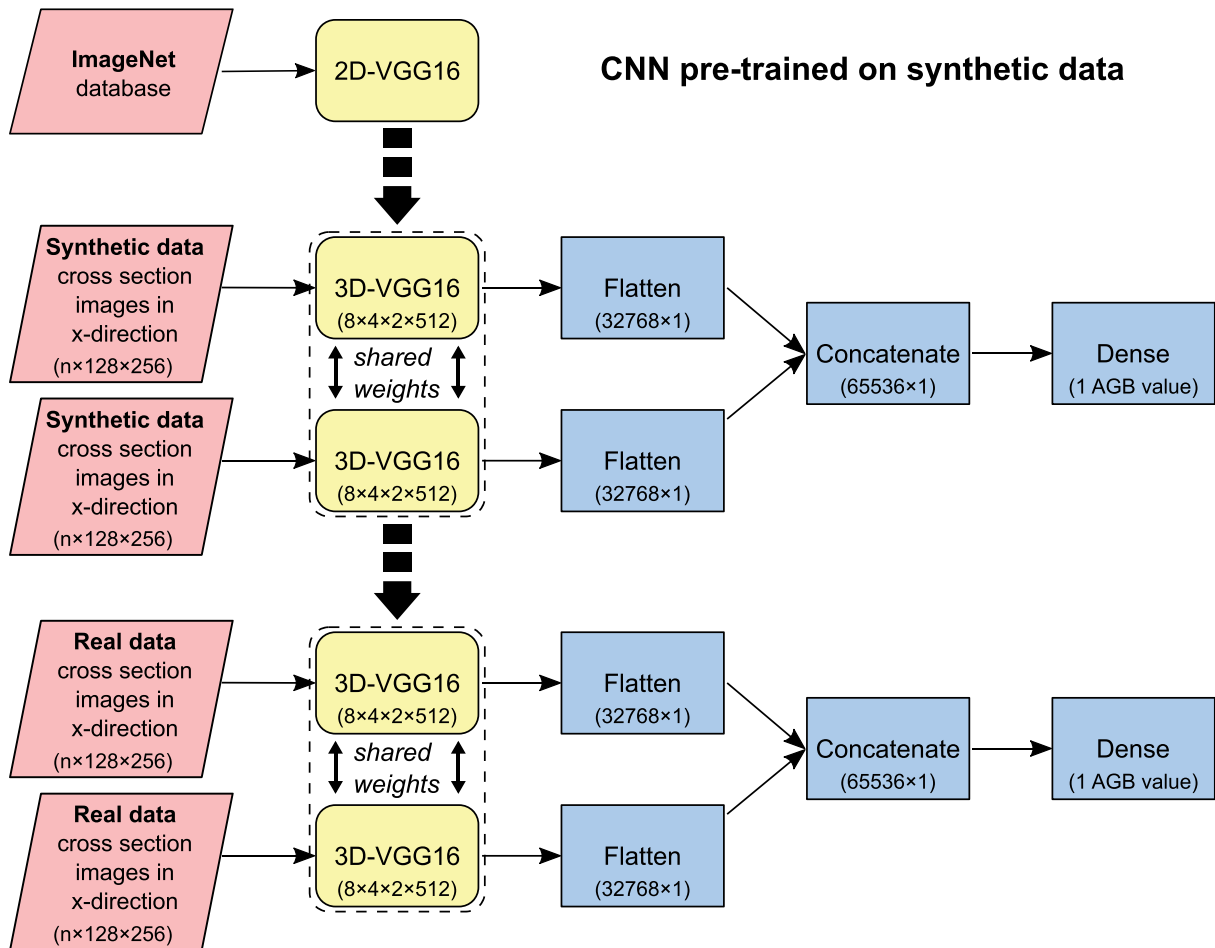
training data size, pre-training on the synthetic data took more than a day. In contrast, the RF training took only a few seconds for each run.

### Neural network architecture

The backbone of our method is a VGG16 2D CNN (Simonyan & Zisserman, 2015) pre-trained on the ImageNet dataset (Deng et al., 2009). In order to apply it on 3D data, we used a re-designed architecture by Solovyev et al. (2022) in which all 2D convolutions are replaced by 3D convolutions. Thereby, the convolutions can be moved in x-, y- and z-direction through a stack of images. The initial weights of the 3D kernel were transferred from the weights of the pre-trained VGG16. We fed the cross-sections, ordered by depth, in both x- and y-directions separately through the CNN and concatenated the two feature vectors resulting from the two directions. We then passed the concatenated vector through a dense layer to obtain a scalar AGB value. The network architecture is shown in Figures 3 and 4. For optimization of the weights, we used the Adam algorithm (Kingma & Ba, 2017), employing an exponentially decaying learning rate. The following hyperparameter values were selected based on the results of parameter tuning in preliminary experiments: initial learning rate =  $10^{-7}$ ,



**Figure 3.** Neural network architecture of the CNN trained on real data only. For a single data point, a block of images in x- and y-directions are fed through the 3D-VGG16 separately, but with shared weights between the networks. The output is flattened and concatenated, before being fed through a dense layer, further reducing the dimensionality to 1, i.e. the scalar AGB value as the regression target. Initial weights were derived from pre-training on the ImageNet database. Numbers in brackets indicate the output shape of each layer.



**Figure 4.** Neural network architecture of the CNN pre-trained on synthetic data. For a single data point, a block of images in x- and y-directions are fed through the 3D-VGG16 separately, but with shared weights between the networks. The output is flattened and concatenated, before being fed through a dense layer, further reducing the dimensionality to 1, i.e. the scalar AGB value as the regression target. Initial weights were derived from pre-training on the ImageNet database. An additional pre-training step was performed on synthetic forest data. Numbers in brackets indicate the output shape of each layer.

decay steps = 100 000, decay rate = 0.96, staircase = TRUE. Because of the limited GPU memory, the batch size was set to 1. Training was carried out for up to 600 epochs using the mean squared error as loss metric. An epoch refers to the complete iteration of the entire training dataset through the CNN during which the model weights are updated. After each epoch, the validation RMSE was calculated from the withheld validation data which was never used for training. Early stopping of model training was applied if the RMSE did not decline over 20 consecutive epochs. For the pre-training on the synthetic data, the same neural network architecture was first applied on cross-section images derived from the simulated point clouds. The weights of the best model according to the validation on the synthetic dataset were then used as initial weights in the further training on the real data cross-section images.

### Random forest models

For the benchmark models, point cloud metrics were derived from all returns, first returns, and all returns with a normalized height > 2 m. We used a subset of the pre-defined standard metrics from the “cloud\_metrics” function implemented in the R package “lidR” (Roussel et al., 2020), precisely: the mean and the maximum of return heights, the standard deviation, the entropy, the kurtosis, and the skewness of the height distribution, the percentage of returns with a height > 2 m, the percentage of returns above the mean height, the percentage of the 1st–5th returns, the percentage of ground returns, the 5th to 95th height percentiles in increments of 5%, the cumulative percentage of returns in each of nine equally spaced height layers, and the total number of returns.

The point cloud metrics served as predictors in a random forest regression. The function “tuneRF” of the “randomForest” package (Liaw & Wiener, 2002) in R was employed to search for the optimal number of predictors to sample at each split, starting at 14 predictors and inflating or deflating the number of predictors by 2 in each iteration. The number of trees was set to 500.

## Results

### Model performance of CNNs (without pre-training on synthetic data) compared to random forest models

The difference in CNN and RF performance varied between the study sites. Performance metrics were aggregated by taking the median over 10 repetitions, for each of which the training/test split was randomized. Figure 5 shows the performance metrics for the different models and training sample sizes (the accuracy metrics are also

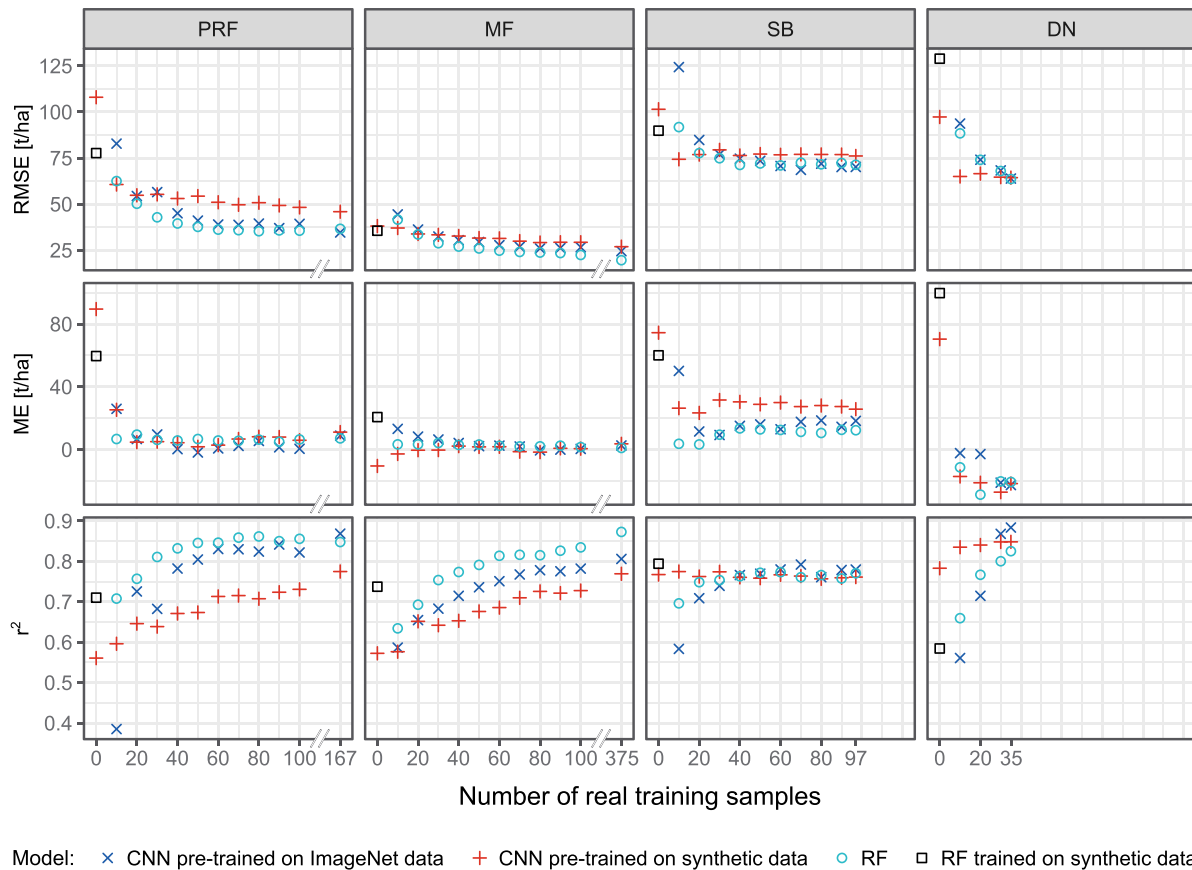
provided in Tables A1–A4 in the Appendix). For all study sites, using the maximum number of samples for model training resulted in the highest model accuracies as expressed by RMSE and  $r^2$ . For PRF and SB, the best results were achieved by CNNs, with a lowest median RMSE of 34 t/ha and 69 t/ha and a highest median  $r^2$  of 0.87 and 0.79 per respective study site. For MF, the lowest median RMSE was 20 t/ha and the highest median  $r^2$  was 0.87, both achieved by RFs. For DN, using RFs for model training resulted in the lowest median RMSE of 63 t/ha, while using CNNs resulted in the highest median  $r^2$  of 0.88. However, when using the maximum number of training samples, differences in prediction performance and systematic error between CNNs and RFs were small. The difference in RMSE between observed and predicted AGB of CNNs and RFs ranged from –2 t/ha to 5 t/ha (20–71 t/ha for RFs and 24–70 t/ha for CNNs), and the difference in  $r^2$  ranged from –0.05 to 0.06 (0.77–0.87 for RFs and 0.78–0.88 for CNNs), depending on the study site. The ME of the CNNs ranged from –23 t/ha for DN (indicating an overprediction of AGB) to +18 t/ha for SB (indicating an underprediction of AGB), and from –21 t/ha to +12 t/ha when using RFs. The absolute ME of the RFs was 2–6 t/ha lower than that of the CNNs. For PRF, SB, and DN, the CNNs performed similarly or slightly better than the RFs with regard to RMSE and  $r^2$ , whereas they performed slightly worse for MF.

The prediction performance of both CNNs and RFs depended on the number of samples that were used for model training. When the training sample sizes were small, the CNNs had – in most cases – a slightly lower prediction performance than RFs, but the predictive performance of CNNs and RFs tended to converge as the training sample size increased. For PRF and MF, the RFs performed better than the CNNs when trained on 10–100 samples. The difference in model performance was more pronounced in  $r^2$  than in RMSE. For both model types, the systematic error was rather small, especially for MF. Only at a training sample size of 10, the CNNs showed a much higher mean error than the RFs did.

For SB, the RFs outperformed the CNNs when trained on 10–20 samples. At higher sample sizes, the difference in prediction performance between both model types was small. At some sample sizes, the CNNs resulted in a higher prediction performance, while at other sample sizes, the RF predictions were better. The systematic error was in most cases slightly lower for the RFs than for the CNNs.

The results for the DN dataset differed the most from the other study sites. Because of the small number of ground plots, models could only be trained on 10–35 samples (including validation





**Figure 5.** Median root mean squared error (RMSE), mean error (ME) and squared Pearson correlation coefficient ( $r^2$ ) of the AGB predictions for different model types and sample sizes. Training sample count included 20% validation data for the CNNs. Model training and testing was repeated 10 times for each training sample size, except for the CNNs that were only trained on synthetic data. For these models, model training and testing was not repeated due to the long computing times.

data). The RMSE of the CNNs and the RFs were similar for sample sizes  $> 10$ , while it was slightly higher for the CNNs at a sample size of 10. The systematic error of the CNNs was much lower than that of the RFs for sample sizes of 10–20 (ME of  $-2$  to  $3$  t/ha for the CNNs,  $-21$  to  $23$  t/ha for the RFs), and similar for sample sizes of 30–35. With regard to the  $r^2$ , CNNs did not reach the prediction performance of RFs when trained on 10–20 samples, but outperformed RFs when trained on 30 and 35 samples.

### Pre-training on synthetic data

Using CNNs that were pre-trained on synthetic data only improved the prediction performance when the number of real training samples was very small (Figure 5, blue and red cross marks). For PRF and SB, the performance of the CNNs in terms of RMSE and  $r^2$  only improved by pre-training when no more than 10 and 20 samples were used for model training, respectively. A positive effect of pre-training on the systematic error could be observed when using 10–30 training samples for PRF, and 10–40 samples for MF. For SB,

RMSE and  $r^2$  of the predictions could be improved by pre-training on synthetic data for 10–20 and 10–30 training samples, respectively. Pre-training could only reduce the systematic error at a sample size of 10.

The results for MF, PRF, and SB datasets showed that pre-training often not only failed to improve the models but rather substantially worsened them. At larger sample sizes, pre-training strongly increased the RMSE for PRF, decreased  $r^2$  for PRF and MF, and increased the underprediction of AGB for SB. The most positive effect of pre-training on synthetic data was observed for the DN dataset. Here, the pre-training decreased the RMSE by 4–29 t/ha and increased  $r^2$  by 0.13–0.28 for training sample sizes of 10–30 and 10–20, respectively. For higher training sample sizes, the pre-trained CNNs performed slightly worse than the non-pre-trained models, but the differences were small (0.5 t/ha for RMSE, 0.02–0.03 for  $r^2$ ). In contrast, pre-training strongly increased the overprediction of the CNNs for 10–30 training samples.

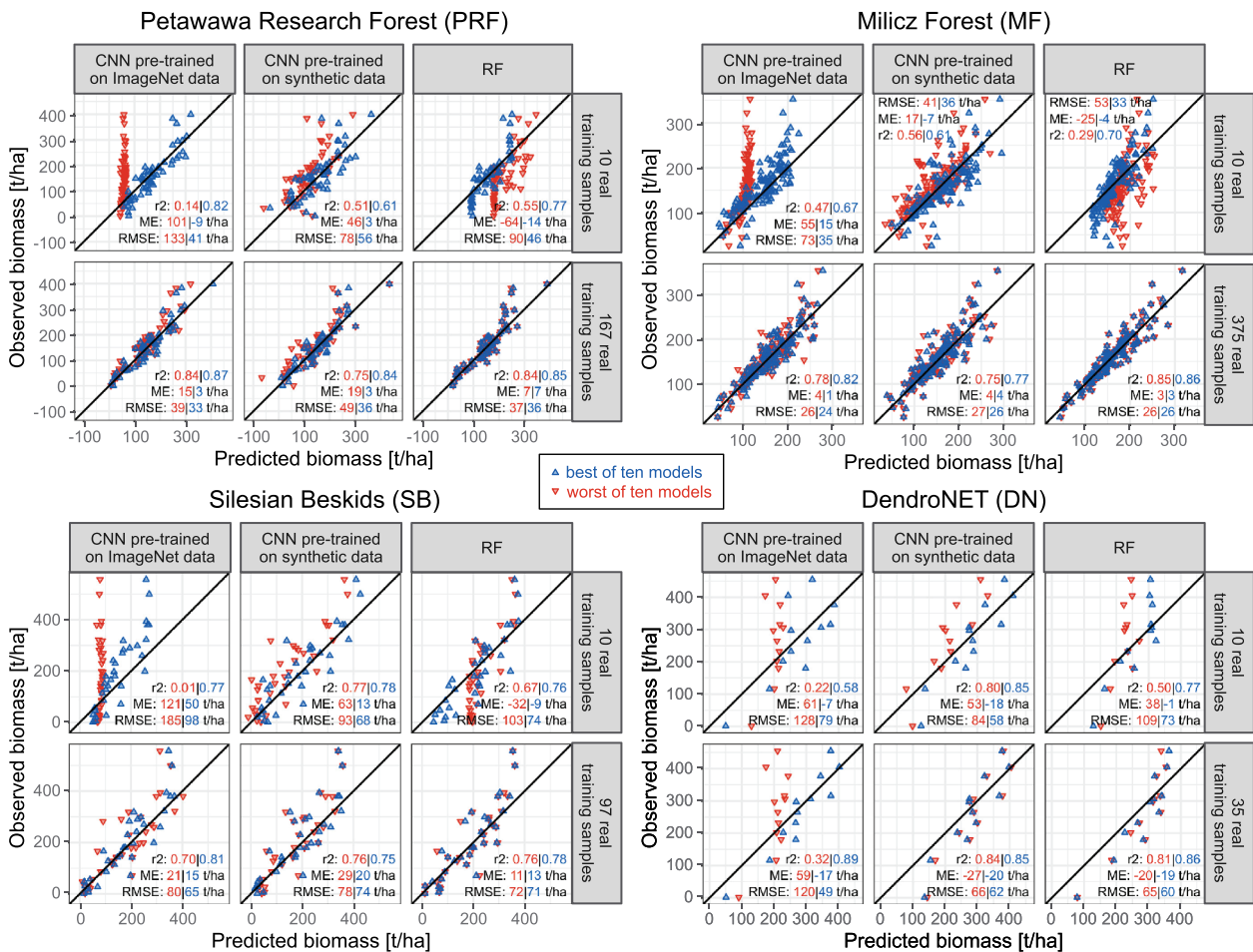
When the models were trained only on synthetic data, the RFs performed better than the CNNs for PRF

and SB (Figure 5, values for 0 real training samples). For MF, CNNs and RFs performed similarly in terms of RMSE, while  $r^2$  was better for the RFs and the ME was better for the CNNs. For DN, the CNNs outperformed the RFs when no real data were used for model training.

### Model stability

Figure 6 shows scatter plots of predicted and observed AGB of the four study sites resulting from the best and worst models for each model type (CNN pre-trained on ImageNet data, CNN pre-trained on synthetic forest data, and RF) and the minimum (10) and maximum (dependent on the study site) number of training samples. Differences in the prediction performance between the best and the worst of the randomized repetitions of models indicate how the models were influenced by the selection of training data and random processes within the models. The samples that were used for model training were randomly sampled from the training datasets. In case of the CNNs (both ImageNet pre-trained and synthetic forest pre-

trained), 20% of these data were further removed to serve as validation data. Hence, when 10 training samples were available, eight were used as actual training samples and two as validation samples. When comparing models trained and validated on 10 samples, it is striking that in the case of the CNNs that were not pre-trained on synthetic data, the worst models resulted in a very narrow range of predicted AGB values, albeit not the mean value of the training data AGB (Figure 6, top left panel of each study site: red markers appearing in an approximate vertical line). For example, for PRF, the worst CNN trained on 10 samples predicted AGB values of 43–66 t/ha while the reference was in the range of 1–399 t/ha. This effect did not occur for the largest sample size nor for CNNs that were pre-trained on synthetic data. We could not find a direct relation between the prediction range and the training data samples, as the AGB range of the training data of these models was much wider than the predicted AGB. However, using only two samples for validation could also have negatively affected model performance if validation accuracy leads to stopping model training too early.



**Figure 6.** Predicted and observed AGB of the ground plots in the test datasets of each study site. Results are shown for models trained on 10 real training samples and on the maximum number of available training samples. Model training and prediction were repeated 10 times for each model type and training sample size, resulting in 10 predictions per ground observation. The training datasets consisting of 10 samples were randomly selected from the total training data in each run. Only the best and the worst predictions (in terms of RMSE) of the 10 model iterations are depicted.

In most cases, the difference in performance between best and worst prediction was highest for the ImageNet pre-trained CNNs and lowest for the synthetic forest pre-trained CNNs. Accordingly, pre-training on synthetic data had a stabilizing effect on the CNNs. The positive effect of the pre-training diminishes with increasing training sample size.

## Discussion

In our study, CNNs using images of ALS point cloud cross-sections performed similar to RFs using traditional point cloud metrics in the prediction of AGB on plot level, but there were differences in model performance between study sites and depending on how many samples were used for model training. In an attempt to identify patterns of when CNNs performed substantially better or worse than RFs, we visually examined point clouds of some of the test plots. However, there were no obvious connections between performance differences and forest structure types, e.g. clearings, stand density, and subcanopy layers.

### *Differences between real and synthetic data*

Regarding the use of pre-trained CNNs, our results showed that the use of simulated data for additional pre-training (on top of the ImageNet weights) did not improve model performance with the exception of cases of extremely limited training data availability. In all other cases, performance was actually decreased by the addition of the pre-training step with simulated data, suggesting that the CNN finds patterns in the synthetic data that do not exist in a similar way in the real data. This effect may be caused by either the synthetic forest stands or by the simulation of laser scanning. The simulated forest stands that were selected for the synthetic training datasets differ in their composition from the real forest stands at our study sites: They have a different species composition and they have on average higher AGB values but a lower number of trees per hectare than the real stands (cf. [Figure 1](#)). In addition, they lack understorey elements, and the crowns of neighbouring trees may overlap unrealistically due to our simple approach of assigning tree positions. It should be tested whether synthetic datasets generated using alternative forest growth simulators, such as SILVA (Pretzsch et al., 2002), which incorporate competition between neighbouring trees at the individual tree level, output actual realistic tree positions, and allow the simulation of different forest management strategies, would improve the performance of the pre-trained CNNs. The differences between simulated and real pulse density and planar point density (cf. [Table 1](#)) indicate that our simulations could not exactly replicate the real laser scanning of the four study sites, which can in

part be explained by the missing understorey. In a previous study, we also observed that the height distribution of the simulated laser scanning returns can differ substantially from the real one, depending on the stand characteristics (Schäfer, Weiser, et al., 2023). Other studies using HELIOS++ for virtual laser scanning have found that the quality of the generated point cloud is also subject to the representation of the 3D scene and can be improved either by precise fine-tuning of the voxelization model (Weiser et al., 2021) or by the use of procedurally generated, highly detailed mesh models of trees (Esmorís et al., 2024). In the latter study, a successful transfer of a deep-learning model trained on purely synthetic data to a real dataset was shown for the case of leaf-wood separation. We conclude that more effort is needed to fine-tune the scene model (e.g. use a different representation, such as high-detail mesh models of trees) and the parameters for the HELIOS++ simulations to make the resulting point clouds more realistic. To investigate whether the poor results for the pre-training on synthetic data are more affected by the forest stand simulations or the laser scanning simulations, two potential pathways exist: 1) the use of virtual laser scanning based on real forest inventory data, thereby excluding effects from the forest stand synthetization, and 2) the use of real laser scanning point clouds of trees stitched together based on the compositions given from the synthetic forest stands, excluding the laser scanning simulation. For the latter case, the flight- and sensor-parameters of the available tree point cloud database would have to match the ones of the real training and test data, which was not the case in our study.

### *Uncertainty of AGB reference data*

To precisely evaluate model performance, accurate reference values (i.e. AGB) are necessary. AGB values in this study were calculated using allometric models, which have been shown to contain significant uncertainty (Vorster et al., 2020). Additionally, tree crowns reaching out of or into the ground plots contribute to errors in the reference AGB (Knapp et al., 2021), as the AGB plot is calculated as the total AGB of all trees with a stem position within the plot. While these sources of uncertainty can only be removed by extensive and potentially destructive fieldwork with real data, the use of 3D mesh models for synthetic data allows the accurate quantification of wood volume and thus the derivation of AGB estimates that are not affected by allometric errors. Synthetic data also include information on tree parts exceeding or reaching into the plot, making it easy to precisely quantify the amount of AGB that is within the plot area. Accounting for these boundary effects in real data is much more difficult as it requires detailed tree information and is subject to uncertainties due to assumptions about

tree shape and crown projection area that need to be made (Kleinn et al., 2020).

### Differences between study sites

In a previous study on ALS-based AGB predictions that solely investigated RFs and their response to simulated training data using the same real-world ALS and ground data, we also observed that results differed substantially between the four study sites (Schäfer, Winiwarter, et al., 2023). We note that these differences partially result from different AGB distributions, which are not represented equally well by the simulated data. Most substantially, in the current study, the mean AGB value of the simulated data of DN was much closer to the one of the real data than for the other study sites (cf. Figure 1). This was also the dataset for which pre-training of the CNNs on synthetic data was most successful.

### Augmentation of training data

While the metrics used in the RF models usually describe the vertical distribution of the returns, and the horizontal distribution is less frequently taken into account (Bouvier et al., 2015), the CNNs are able to consider both vertical and horizontal distributions in the convolutions. Limited data augmentation was carried out to achieve larger training sets but could be exploited more in the future, e.g. by mirroring images or by extracting cross-sections in other directions. As this would not solve any issues related to the domain of AGB values present in the training data, the effect of such efforts may, however, be limited.

### Comparison between CNNs and RF

Although the predictive performance of CNNs and RFs was similar, there are several reasons to use RFs rather than CNNs, both in terms of data pre-processing and the modeling itself. Generating images of point cloud cross-sections is much more complicated, takes more time and needs more disk space than extracting point cloud metrics. In addition, CNNs require a high-performance GPU to satisfy the computational demands and still take much longer to train than RFs. A drawback of using CNNs is also the “black box” characteristic of the approach that makes it more difficult to interpret the results, e.g. to explain why the pre-trained CNNs sometimes predicted negative AGB values for one plot in PRF (cf. Figure 6).

Other studies comparing deep-learning methods to traditional machine learning models for AGB predictions from ALS data often found that deep learning results in higher prediction performance. Ayrey and Hayes (2018) adapted several 2D CNNs (LeNet, AlexNet, GoogLeNet, Inception-V3, and ResNet-50) to run on 3D voxel representations of the ALS point

clouds and compared the model performance to RF and linear mixed models trained on point cloud metrics. In their study, all deep learning models except for AlexNet resulted in lower RMSE but higher or similar systematic error compared to RF and linear mixed models. Oehmcke et al. (2022) applied PointNet, the kernel point convolution (KPConv) approach, and the Minkowski CNN in comparison to linear regression and power regression models for ALS-based AGB predictions and found that their adaptations of the Minkowski CNN and KPConv clearly surpassed the linear regression and power regression, while PointNet performed worse. When comparing RF to an Octree CNN-HRNet and a Dynamic Graph CNN, Seely et al. (2023) showed that AGB predictions with RF had a slightly lower  $R^2$  and a slightly higher RMSE than the deep learning predictions. In contrast to our study, Ayrey and Hayes (2018) used 15373 samples for model training and 1000 samples for validation, Oehmcke et al. (2022) used 4271 and 919 samples, and Seely et al. (2023) used 1635 and 350, respectively. Compared to our sample sizes of up to 35–375 plots (including 20% validation samples), these datasets are much better suited for deep learning approaches. We expect that the performance of the AGB prediction on cross-section images with CNNs could be improved when using more training samples. While we hypothesized that synthetic data could be used to extend training datasets when limited data are available, our results did not support that claim.

### Outlook

We see potential for future research in multiple directions:

- Improvement of a) the synthetic forest stand composition and the positions of the individual trees within a plot, b) the 3D models of individual trees, and c) the laser scanning simulation parameters.
- Systematic investigations on how the CNN performance is affected by forest structure as well as by ground plot and point cloud characteristics (e.g. stand density, tree species, plot size and shape, point density, and penetration into subcanopy layers).
- Experiments with hyperparameter tuning and different deep learning network architectures for which pre-trained ImageNet weights are available (e.g. ResNet, EfficientNet, and DenseNet), as well as investigation of the effect of these pre-training efforts by running models on randomly initialized weights using the same model architecture for comparison.
- Additional data augmentation by rotation, mirroring, and random jittering of points, as shown in previous studies using CNNs for point cloud

tasks (Briechle et al., 2021; H. Li et al., 2020; Oehmcke et al., 2022).

## Conclusion

This study demonstrated that CNNs can predict AGB from cross-section images and achieve similar accuracies as RFs trained on traditional point cloud metrics. When the maximum number of available training samples was used, the CNN performance slightly surpassed the performance of the RFs for three of the four study sites, indicating that the CNN performance could be further improved by increasing the training sample size. We investigated whether the need of deep learning models for large amounts of training data could be satisfied by data simulations but found that pre-training on synthetic data did only improve model performance when very little training data were available. Notably, pre-training on synthetic data even decreased model performance at larger training sample sizes. Since the use of simulated data has been shown to provide benefits in other applications, even in the domain of forestry remote sensing, there is reason to believe that a gap between real and simulated data needs to be closed before such transfer can be successful for our use-case. For the time being, RF remains a competitive alternative to data-hungry deep learning models.

## Acknowledgments

We are grateful to Joanne White and the Canadian Forest Service for providing data of the Petawawa Research Forest, and to technicians from the Institute of Forest Ecosystem Research and the Global Change Research Institute for data collection at the DendroNetwork and Silesian Beskids sites. We thank Roman Solovjev for sharing the pre-trained 3D CNNs and Hans Henniger for providing the modified version of Forest Factory 2.0. The first author is extremely grateful for her research stay in the Photogrammetry Research Unit at TU Wien, funded by the Graduate School for Climate and Environment of KIT.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the frame of the project SYSSIFOSS - 411263134/2019–2022; by the Polish State Forests National Forest Holding in the frame of the project ‘Development of the method of forest inventory using the results of the REMBIOFOR project’ [Project No. 500463, agreement No. EO.271.3.12.2019, signed on 14.10.2019]; and by the National Centre for Research and Development (Poland)

in the frame of the REMBIOFOR project ‘Remote sensing-based assessment of woody biomass and carbon storage in forests’ as part of the BIOSTRATEG programme [Agreement No. BIOSTRATEG1/267755/4/NCBR/2015].

## ORCID

Fabian Ewald Fassnacht  <http://orcid.org/0000-0003-1284-9573>

## Data availability statement

The individual tree point clouds used for generating the synthetic scenes are published at PANGAEA, at <https://doi.pangaea.de/10.1594/PANGAEA.942856>. R code for the creation synthetic forest stands is available on GitHub, at <https://github.com/JannikaSchaefer/Syssifoss>.

## CRedit author statement

**Jannika Schäfer:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Lukas Winiwarter:** Investigation, Methodology, Writing – Original Draft, Writing – Review & Editing. **Hannah Weiser:** Investigation, Data curation, Writing – Review & Editing. **Bernhard Höfle:** Conceptualization, Data curation, Resources, Writing – Review & Editing, Supervision, Funding acquisition. **Sebastian Schmidlein:** Resources, Writing – Review & Editing, Supervision. **Jan Novotný:** Data curation, Writing – Review & Editing. **Grzegorz Krok:** Data curation, Writing – Review & Editing. **Krzysztof Stereńczak:** Investigation, Data curation, Writing – Review & Editing. **Markus Hollaus:** Resources, Writing – Review & Editing. **Fabian Ewald Fassnacht:** Conceptualization, Methodology, Software, Resources, Writing – Review & Editing, Supervision, Funding acquisition.

## References

- Ayrey, E., & Hayes, D. J. (2018, April). The use of three-dimensional convolutional neural networks to interpret LiDAR for forest inventory. *Remote Sensing*, 10(4), 649. <https://doi.org/10.3390/rs10040649>
- Ayrey, E., Hayes, D. J., Kilbride, J. B., Fraver, S., Kershaw, J. A., Cook, B. D., & Weiskittel, A. R. (2021, December). Synthesizing disparate LiDAR and satellite datasets through deep learning to generate wall-to-wall regional forest inventories for the complex, mixed-species forests of the eastern United States. *Remote Sensing*, 13(24), 5113. <https://doi.org/10.3390/rs13245113>
- Balazs, A., Liski, E., Tuominen, S., & Kangas, A. (2022, April). Comparison of neural networks and k-nearest neighbors methods in forest stand variable estimation using airborne laser data. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 4, 100012. <https://doi.org/10.1016/j.ophoto.2022.100012>
- Bouvier, M., Durrieu, S., Fournier, R. A., & Renaud, J.-P. (2015, January). Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. *Remote Sensing of Environment*, 156, 322–334. <https://doi.org/10.1016/j.rse.2014.10.004>

- Breiman, L. (2001, October). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Briechele, S., Krzystek, P., & Vosselman, G. (2021, June). Silvi-net – a dual-cnn approach for combined classification of tree species and standing dead trees from remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 98, 102292. <https://doi.org/10.1016/j.jag.2020.102292>
- Brovkina, O., Navrátilová, B., Novotný, J., Albert, J., Slezák, L., & Cienciala, E. (2022, September). Influences of vegetation, model, and data parameters on forest aboveground biomass assessment using an area-based approach. *Ecological Informatics*, 70, 101754. <https://doi.org/10.1016/j.ecoinf.2022.101754>
- Dalponte, M., Frizzera, L., Ørka, H. O., Gobakken, T., Næsset, E., & Gianelle, D. (2018, February). Predicting stem diameters and aboveground biomass of individual trees using remote sensing data. *Ecological Indicators*, 85, 367–376. <https://doi.org/10.1016/j.ecolind.2017.10.066>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009, June). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. (pp. 248–255). Miami, FL, USA. <https://doi.org/10.1109/CVPR.2009.5206848>
- Dixon, R. K., Solomon, A. M., Brown, S., Houghton, R. A., Trexler, M. C., & Wisniewski, J. (1994, January). Carbon pools and flux of global forest ecosystems. *Science*, 263(5144), 185–190. <https://doi.org/10.1126/science.263.5144.185>
- Esmoris, A. M., Weiser, H., Winiwarter, L., Cabaleiro, J. C., & Höfle, B. (2024). Deep learning with simulated laser scanning data for 3d point cloud classification. (Preprint available on Earth ArXiv). *Isprs Journal of Photogrammetry & Remote Sensing*, 215, 192–213. <https://doi.org/10.31223/X53Q3Q>
- Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014, November). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, 154, 102–114. <https://doi.org/10.1016/j.rse.2014.07.028>
- Fuller, A., Millard, K., & Green, J. R. (2022). SatViT: Pretraining transformers for earth observation. *IEEE Geoscience & Remote Sensing Letters*, 19, 1–5. <https://doi.org/10.1109/LGRS.2022.3201489>
- Hamedianfar, A., Mohamedou, C., Kangas, A., & Vauhkonen, J. (2022, February). Deep learning for forest inventory and planning: A critical review on the remote sensing approaches so far and prospects for further applications. *Forestry: An International Journal of Forest Research*, 95(4), 451–465. <https://doi.org/10.1093/forestry/cpac002>
- Hamraz, H., Jacobs, N. B., Contreras, M. A., & Clark, C. H. (2019, December). Deep learning for conifer/deciduous classification of airborne LiDAR 3D point clouds representing individual trees. *Isprs Journal of Photogrammetry & Remote Sensing*, 158, 219–230. <https://doi.org/10.1016/j.isprs.2019.10.011>
- Hell, M., Brandmeier, M., Briechele, S., & Krzystek, P. (2022, April). Classification of tree species and standing dead trees with lidar point clouds using two deep neural networks: PointCNN and 3DmFV-net. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 90(2), 103–121. <https://doi.org/10.1007/s41064-022-00200-4>
- Henniger, H., Huth, A., Frank, K., & Bohn, F. J. (2023, September). Creating virtual forests around the globe and analysing their state space. *Ecological Modelling*, 483, 110404. <https://doi.org/10.1016/j.ecolmodel.2023.110404>
- Hwang, S.-H., & Whang, S. E. (2022, August). RegMix: Data mixing augmentation for regression (No. arXiv:2106.03374). arXiv.<https://doi.org/10.48550/arXiv.2106.03374>
- Jucker, T., Caspersen, J., Chave, J., Antin, C., Barbier, N., Bongers, F. & Coomes, D. A. (2017). Allometric equations for integrating remote sensing imagery into forest monitoring programmes. *Global Change Biology*, 23(1), 177–190. <https://doi.org/10.1111/gcb.13388>
- Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021, March). Review on convolutional neural networks (CNN) in vegetation remote sensing. *Isprs Journal of Photogrammetry & Remote Sensing*, 173, 24–49. <https://doi.org/10.1016/j.isprs.2020.12.010>
- Kingma, D. P., & Ba, J. (2017, January). Adam: A method for stochastic optimization. (Eprint available on arXiv), <https://doi.org/10.48550/arXiv.1412.6980>
- Kleinn, C., Magnussen, S., Nölke, N., Magdon, P., Álvarez-González, J. G., Fehrmann, L., & Pérez-Cruzado, C. (2020, October). Improving precision of field inventory estimation of aboveground biomass through an alternative view on plot biomass. *Forest Ecosystems*, 7(1), 57. <https://doi.org/10.1186/s40663-020-00268-7>
- Knapp, N., Huth, A., & Fischer, R. (2021, January). Tree crowns cause border effects in area-based biomass estimations from remote sensing. *Remote Sensing*, 13(8), 1592. <https://doi.org/10.3390/rs13081592>
- Köhler, P., & Huth, A. (1998, June). The effects of tree species grouping in tropical rainforest modelling: Simulations with the individual-based model Formind. *Ecological Modelling*, 109(3), 301–321. [https://doi.org/10.1016/S0304-3800\(98\)00066-0](https://doi.org/10.1016/S0304-3800(98)00066-0)
- LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Li, H., Hu, B., Li, Q., & Jing, L. (2020, September). CNN-Based tree species classification using airborne lidar data and high-resolution satellite image. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2679–2682). Waikoloa, HI, USA.
- Li, R., Li, X., Heng, P.-A., & Fu, C.-W. (2020, June). PointAugment: An auto-augmentation framework for point cloud classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6377–6386). IEEE, Seattle, WA, USA.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22. <https://cran.r-project.org/web/packages/randomForest>
- Luo, Z., Zhang, Z., Li, W., Chen, Y., Wang, C., Nurunnabi, A. A. M., & Li, J. (2022). Detection of individual trees in UAV LiDAR point clouds using a deep learning framework based on multichannel representation. *IEEE Transactions on Geoscience & Remote Sensing*, 60, 1–15. <https://doi.org/10.1109/TGRS.2022.3230051>
- Næsset, E. (2002, April). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, 80(1), 88–99. [https://doi.org/10.1016/S0034-4257\(01\)00290-5](https://doi.org/10.1016/S0034-4257(01)00290-5)

- Niu, S., Liu, M., Liu, Y., Wang, J., & Song, H. (2021, October). Distant domain transfer learning for medical imaging. *IEEE Journal of Biomedical and Health Informatics*, 25(10), 3784–3793. <https://doi.org/10.1109/JBHI.2021.3051470>
- Oehmcke, S., Li, L., Revenga, J. C., Nord-Larsen, T., Trepekli, K., Gieseke, F., & Igel, C. (2022, November). Deep learning based 3D point cloud regression for estimating forest biomass. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems* (pp. 1–4). ACM, Seattle Washington.
- Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A. & Hayes, D. (2011, August). A large and persistent carbon sink in the world's forests. *Science*, 333(6045), 988–993. <https://doi.org/10.1126/science.1201609>
- Pretzsch, H., Biber, P., & Durský, J. (2002, June). The single tree-based stand simulator SILVA: Construction, application and evaluation. *Forest Ecology & Management*, 162(1), 3–21. [https://doi.org/10.1016/S0378-1127\(02\)00047-6](https://doi.org/10.1016/S0378-1127(02)00047-6)
- Roussel, J.-R., Auty, D., Coops, N. C., Tompalski, P., Goodbody, T. R., Meador, A. S. & Achim, A. (2020). lidar: An R package for analysis of airborne laser scanning (ALS) data. *Remote Sensing of Environment*, 251, 112061. <https://doi.org/10.1016/j.rse.2020.112061>
- Schäfer, J., Weiser, H., Winiwarter, L., Höfle, B., Schmidlein, S., & Fassnacht, F. E. (2023, April). 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. <https://doi.org/10.1093/forestry/cpad006>
- Schäfer, J., Winiwarter, L., Weiser, H., Novotný, J., Höfle, B., Schmidlein, S. & Fassnacht, F. E. (2023, December). Assessing the potential of synthetic and ex situ airborne laser scanning and ground plot data to train forest biomass models. *Forestry: An International Journal of Forest Research*, 97(4), 512–530. <https://doi.org/10.1093/forestry/cpad061>
- Seely, H., Coops, N. C., White, J. C., Montwé, D., Winiwarter, L., & Ragab, A. (2023, December). Modelling tree biomass using direct and additive methods with point cloud deep learning in a temperate mixed forest. *Science of Remote Sensing*, 8, 100110. <https://doi.org/10.1016/j.srs.2023.100110>
- Sheridan, R. D., Popescu, S. C., Gatzliolis, D., Morgan, C. L. S., & Ku, N.-W. (2015, January). Modeling forest aboveground biomass and volume using airborne LiDAR metrics and forest inventory and analysis data in the Pacific Northwest. *Remote Sensing*, 7(1), 229–255. <https://doi.org/10.3390/rs70100229>
- Shorten, C., & Khoshgoftaar, T. M. (2019, December). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. <https://doi.org/10.1186/s40537-019-0197-0>
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. (Eprint available on arXiv, <https://doi.org/10.48550/arXiv.1409.1556>)
- Solovyev, R., Kalinin, A. A., & Gabruseva, T. (2022, February). 3D convolutional neural networks for stalled brain capillary detection. *Computers in Biology & Medicine*, 141, 105089. <https://doi.org/10.1016/j.compbiomed.2021.105089>
- Stereńczak, K., Lisańczuk, M., Parkitna, K., Mitelsztedt, K., Mroczek, P., & Miścicki, S. (2018). The influence of number and size of sample plots on modelling growing stock volume based on airborne laser scanning. *Drewno Prace Naukowe Doniesienia Komunikaty*, 61(201), 5–22. <https://doi.org/10.12841/wood.1644-3985.D11.04>
- Strimbu, V. F., Næsset, E., Ørka, H. O., Liski, J., Petersson, H., & Gobakken, T. (2023, May). Estimating biomass and soil carbon change at the level of forest stands using repeated forest surveys assisted by airborne laser scanner data. *Carbon Balance and Management*, 18(1), 10. <https://doi.org/10.1186/s13021-023-00222-4>
- Sun, C., Huang, C., Zhang, H., Chen, B., An, F., Wang, L., & Yun, T. (2022). Individual tree crown segmentation and crown width extraction from a heightmap derived from aerial laser scanning data using a deep learning framework. *Frontiers in Plant Science*, 13. <https://doi.org/10.3389/fpls.2022.9149740>
- Vonderach, C., Kublin, E., Bösch, B., & Kändler, G. (2021). *rBDAT: Implementation of BDAT tree taper fortran functions [computer software manual]*. <https://CRAN.R-project.org/package=rBDAT> (R package version 0.9.8).
- Vorster, A. G., Evangelista, P. H., Stovall, A. E. L., & Ex, S. (2020, May). Variability and uncertainty in forest biomass estimates from the tree to landscape scale: The role of allometric equations. *Carbon Balance and Management*, 15(1), 8. <https://doi.org/10.1186/s13021-020-00143-6>
- Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Fassnacht, F. E., & Höfle, B. (2022, July). Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests. *Earth System Science Data*, 14(7), 2989–3012. <https://doi.org/10.5194/essd-14-2989-2022>
- Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Seitz, C., Schimka, M. & Höfle, B. (2022, March). Terrestrial, uav-borne, and airborne laser scanning point clouds of central European forest plots, Germany, with extracted individual trees and manual forest inventory measurements [data set].
- Weiser, H., Winiwarter, L., Anders, K., Fassnacht, F. E., & Höfle, B. (2021, November). Opaque voxel-based tree models for virtual laser scanning in forestry applications. *Remote Sensing of Environment*, 265, 112641. <https://doi.org/10.1016/j.rse.2021.112641>
- White, J. C., Chen, H., Woods, M. E., Low, B., & Nasonova, S. (2019, December). The petawawa research forest: Establishment of a remote sensing supersite. *The Forestry Chronicle*, 95(3), 149–156. <https://doi.org/10.5558/tfc2019-024>
- Winiwarter, L., Esmoris Pena, A. M., Weiser, H., Anders, K., Martínez Sánchez, J., Searle, M., & Höfle, B. (2022, February). 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. <https://doi.org/10.1016/j.rse.2021.112772>
- Zhao, K., Popescu, S., & Nelson, R. (2009, January). Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. *Remote Sensing of Environment*, 113(1), 182–196. <https://doi.org/10.1016/j.rse.2008.09.009>

## Appendix A. Model accuracies of CNNs and RFs for all training sample sizes and all study sites

**Table A1.** Model accuracies for Petawawa Research Forest (PRF). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and  $r^2$  is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	$r^2$
CNN pre-trained on ImageNet data	10	82.82	25.97	0.39
CNN pre-trained on ImageNet data	20	54.41	6.76	0.73
CNN pre-trained on ImageNet data	30	56.53	9.64	0.68
CNN pre-trained on ImageNet data	40	45.04	0.30	0.78
CNN pre-trained on ImageNet data	50	41.06	-1.84	0.80
CNN pre-trained on ImageNet data	60	38.94	0.75	0.83
CNN pre-trained on ImageNet data	70	38.76	2.34	0.83
CNN pre-trained on ImageNet data	80	39.46	5.52	0.82
CNN pre-trained on ImageNet data	90	37.00	1.46	0.84
CNN pre-trained on ImageNet data	100	39.30	0.52	0.82
CNN pre-trained on ImageNet data	167	34.62	8.90	0.87
CNN pre-trained on synthetic data	0	108.00	89.53	0.56
CNN pre-trained on synthetic data	10	60.65	25.35	0.60
CNN pre-trained on synthetic data	20	54.80	4.75	0.65
CNN pre-trained on synthetic data	30	55.30	5.06	0.64
CNN pre-trained on synthetic data	40	53.05	4.38	0.67
CNN pre-trained on synthetic data	50	54.34	1.80	0.67
CNN pre-trained on synthetic data	60	51.02	2.85	0.71
CNN pre-trained on synthetic data	70	49.65	6.74	0.72
CNN pre-trained on synthetic data	80	50.78	8.18	0.71
CNN pre-trained on synthetic data	90	49.30	8.00	0.72
CNN pre-trained on synthetic data	100	48.25	5.95	0.73
CNN pre-trained on synthetic data	167	45.90	11.15	0.77
RF trained on real data	10	62.50	6.74	0.71
RF trained on real data	20	50.20	9.54	0.76
RF trained on real data	30	42.84	6.19	0.81
RF trained on real data	40	39.50	5.80	0.83
RF trained on real data	50	37.65	6.76	0.85
RF trained on real data	60	36.17	5.79	0.85
RF trained on real data	70	35.86	5.60	0.86
RF trained on real data	80	35.28	6.39	0.86
RF trained on real data	90	35.74	5.01	0.85
RF trained on real data	100	35.59	6.47	0.86
RF trained on real data	167	36.61	7.13	0.85
RF trained on synthetic data	0	77.70	59.60	0.71



**Table A2.** Model accuracies for Milicz Forest (MF). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and  $r^2$  is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	$r^2$
CNN pre-trained on ImageNet data	10	44.42	13.22	0.59
CNN pre-trained on ImageNet data	20	36.21	8.28	0.65
CNN pre-trained on ImageNet data	30	32.42	6.43	0.68
CNN pre-trained on ImageNet data	40	30.44	4.21	0.71
CNN pre-trained on ImageNet data	50	29.44	2.20	0.74
CNN pre-trained on ImageNet data	60	27.74	2.52	0.75
CNN pre-trained on ImageNet data	70	26.87	1.99	0.77
CNN pre-trained on ImageNet data	80	25.94	-0.48	0.78
CNN pre-trained on ImageNet data	90	26.19	-0.02	0.78
CNN pre-trained on ImageNet data	100	26.61	0.42	0.78
CNN pre-trained on ImageNet data	375	24.48	2.63	0.81
CNN pre-trained on synthetic data	0	38.10	-10.41	0.57
CNN pre-trained on synthetic data	10	37.04	-2.76	0.58
CNN pre-trained on synthetic data	20	33.82	-0.37	0.65
CNN pre-trained on synthetic data	30	33.35	-0.29	0.64
CNN pre-trained on synthetic data	40	32.67	2.23	0.65
CNN pre-trained on synthetic data	50	31.59	1.74	0.68
CNN pre-trained on synthetic data	60	31.40	1.83	0.69
CNN pre-trained on synthetic data	70	29.91	-1.32	0.71
CNN pre-trained on synthetic data	80	29.15	-1.60	0.73
CNN pre-trained on synthetic data	90	29.31	0.82	0.72
CNN pre-trained on synthetic data	100	29.31	0.61	0.73
CNN pre-trained on synthetic data	375	26.91	3.67	0.77
RF trained on real data	10	41.59	3.33	0.63
RF trained on real data	20	33.54	3.35	0.69
RF trained on real data	30	28.79	4.21	0.75
RF trained on real data	40	26.96	2.95	0.77
RF trained on real data	50	25.94	3.22	0.79
RF trained on real data	60	24.75	2.43	0.81
RF trained on real data	70	24.05	2.03	0.82
RF trained on real data	80	23.71	2.07	0.82
RF trained on real data	90	23.45	2.50	0.83
RF trained on real data	100	22.40	1.61	0.83
RF trained on real data	375	19.57	1.05	0.87
RF trained on synthetic data	0	35.57	20.72	0.74

**Table A3.** Model accuracies for Silesian Beskids (SB). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and  $r^2$  is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	$r^2$
CNN pre-trained on ImageNet data	10	124.36	50.13	0.58
CNN pre-trained on ImageNet data	20	84.80	11.55	0.71
CNN pre-trained on ImageNet data	30	77.05	9.30	0.74
CNN pre-trained on ImageNet data	40	74.68	15.13	0.77
CNN pre-trained on ImageNet data	50	73.50	16.06	0.77
CNN pre-trained on ImageNet data	60	70.69	12.97	0.78
CNN pre-trained on ImageNet data	70	68.58	17.70	0.79
CNN pre-trained on ImageNet data	80	71.93	18.57	0.76
CNN pre-trained on ImageNet data	90	70.12	14.52	0.78
CNN pre-trained on ImageNet data	97	70.16	18.29	0.78
CNN pre-trained on synthetic data	0	101.48	74.49	0.77
CNN pre-trained on synthetic data	10	74.34	26.40	0.77
CNN pre-trained on synthetic data	20	76.89	23.39	0.76
CNN pre-trained on synthetic data	30	79.44	31.60	0.77
CNN pre-trained on synthetic data	40	76.38	30.41	0.76
CNN pre-trained on synthetic data	50	77.17	28.84	0.76
CNN pre-trained on synthetic data	60	76.76	29.97	0.77
CNN pre-trained on synthetic data	70	77.01	27.41	0.76
CNN pre-trained on synthetic data	80	76.95	28.08	0.76
CNN pre-trained on synthetic data	90	76.89	27.44	0.76
CNN pre-trained on synthetic data	97	76.13	25.76	0.76
RF trained on real data	10	91.81	3.70	0.70
RF trained on real data	20	77.67	3.30	0.75
RF trained on real data	30	74.89	9.49	0.75
RF trained on real data	40	71.25	13.49	0.76
RF trained on real data	50	72.06	12.87	0.77
RF trained on real data	60	71.00	12.52	0.77
RF trained on real data	70	72.52	11.30	0.76
RF trained on real data	80	71.60	10.57	0.77
RF trained on real data	90	72.31	12.58	0.76
RF trained on real data	97	71.34	12.34	0.77
RF trained on synthetic data	0	89.86	60.07	0.79

**Table A4.** Model accuracies for DendroNET sites (DN). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and  $r^2$  is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	$r^2$
CNN pre-trained on ImageNet data	10	93.72	-2.24	0.56
CNN pre-trained on ImageNet data	20	74.08	-2.86	0.71
CNN pre-trained on ImageNet data	30	68.31	-21.08	0.87
CNN pre-trained on ImageNet data	35	63.87	-22.58	0.88
CNN pre-trained on synthetic data	0	97.3	70.4	0.78
CNN pre-trained on synthetic data	10	65.01	-17.12	0.84
CNN pre-trained on synthetic data	20	66.54	-21.13	0.84
CNN pre-trained on synthetic data	30	64.65	-27.11	0.85
CNN pre-trained on synthetic data	35	64.37	-21.65	0.85
RF trained on real data	10	88.4	-11.32	0.66
RF trained on real data	20	74.05	-28.69	0.77
RF trained on real data	30	67.96	-20.27	0.8
RF trained on real data	35	63.64	-20.64	0.83
RF trained on synthetic data	0	128.92	99.71	0.58