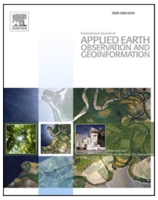
International Journal of Applied Earth Observations and Geoinformation 98 (2021) 102292

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Silvi-Net – A dual-CNN approach for combined classification of tree species and standing dead trees from remote sensing data

S. Briechle [a], P. Krzystek [a], G. Vosselman [b,*]

[a] Munich University of Applied Sciences, Munich, Germany
[b] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
ALS
Multispectral imagery
3D vegetation mapping
Dead trees
CNN
Transfer learning

ABSTRACT

Forest managers and nature conservationists rely on precise mapping of single trees from remote sensing data for efficient estimation of forest attributes. In recent years, additional quantification of dead wood in particular has garnered interest. However, tree-level approaches utilizing segmented single trees are still limited in accuracy and their application is therefore mostly restricted to research studies. Furthermore, the combined classification of presegmented single trees with respect to tree species and health status is important for practical use but has been insufficiently investigated so far. Therefore, we introduce Silvi-Net, an approach based on convolutional neural networks (CNNs) fusing airborne lidar data and multispectral (MS) images for 3D object classification. First, we segment single 3D trees from the lidar point cloud, render multiple silhouette-like side-view images, and enrich them with calibrated laser echo characteristics. Second, projected outlines of the segmented trees are used to crop and mask the MS orthomosaic and to generate MS image patches for each tree. Third, we independently train two ResNet-18 networks to learn meaningful features from both datasets. This optimization process is based on pretrained CNN weights and recursive retraining of model parameters. Finally, the extracted features are fused for a final classification step based on a standard multi-layer perceptron and majority voting. We analyzed the network's performance on data captured in two study areas, the Chernobyl Exclusion Zone (ChEZ) and the Bavarian Forest National Park (BFNP). For both study areas, the lidar point density was approximately 55 points/m^2 and the ground sampling distance values of the true orthophotos were 10 cm (ChEZ) and 20 cm (BFNP). In general, the trained models showed high generalization capacity on independent test data, achieving an overall accuracy (OA) of 96.1% for the classification of pines, birches, alders, and dead trees (ChEZ) - and 91.5% for coniferous, deciduous, snags, and dead trees (BFNP). Interestingly, lidar-based imagery increased the OA by 2.5% (ChEZ) and 5.9% (BFNP) compared to experiments only utilizing MS imagery. Moreover, Silvi-Net also demonstrated superior OA compared to the baseline method PointNet++ by 11.3% (ChEZ) and 2.2% (BFNP). Overall, the effectiveness of our approach was proven using 2D and 3D datasets from two natural forest areas (400–530 trees/ha), acquired with different sensor models, and varying geometric and spectral resolution. Using the technique of transfer learning, Silvi-Net facilitates fast model convergence, even for datasets with a reduced number of samples. Consequently, operators can generate reliable maps that are of major importance in applications such as automated inventory and monitoring projects.

1. Introduction

In forestry, precise and reliable mapping of tree species is a fundamental concern. The classification of dead wood in particular is of increasing importance because forests are suffering from changing climatic conditions. Furthermore, tree-level approaches are increasingly of interest in area-wide forest inventory. For instance, forest attributes such as above-ground biomass and growing stock can be estimated based on tree-specific allometric models (Chave et al., 2014). Moreover, forest managers and nature conservationists require quantitative mapping results to investigate the robustness and sustainability of various forest compositions (Overbeck and Schmidt, 2012). Besides these conventional applications of vegetation mapping, tree species information can also be advantageous in more unusual cases. For example, Briechle et al. (2020b) showed that observed vegetation anomalies are helpful for the detection of unknown radioactive waste sites in the Chernobyl

* Corresponding author.
E-mail addresses: sebastian.briechle@hm.edu (S. Briechle), peter.krzystek@hm.edu (P. Krzystek), george.vosselman@utwente.nl (G. Vosselman).

https://doi.org/10.1016/j.jag.2020.102292
Received 3 November 2020; Received in revised form 22 December 2020; Accepted 23 December 2020
Available online 9 February 2021

Exclusion Zone (ChEZ).

1.1. Conventional approaches

Traditionally, forest inventory has been based on manual field measurements. Forest managers have typically relied on sample-based procedures followed by area-wide extrapolation (McRoberts and Tomppo, 2007). Nevertheless, in situ inventory is labor intensive and, therefore, both time-consuming and expensive. For a temperate forest area of around 300 km^2, Latifi et al. (2015) demonstrated that lidar-based data collection is 90% less expensive compared to a conventional forest inventory. Fassnacht et al. (2016) reviewed the work of various researchers who have investigated forest parameter estimation at the single tree level using remote sensing data. Airplanes, helicopters, and innovative platforms such as unmanned aerial vehicles (UAVs) equipped with lidar sensors and multispectral (MS) or hyperspectral cameras enable acquisition of high-resolution data from a bird's eye view. In particular, the fusion of lidar point clouds and optical multichannel imagery is the most prominent option for the inventory of forest structural variables (Latifi and Heurich, 2019). In a preprocessing step, single trees are typically delineated from airborne laser scanning (ALS) data. This tree segmentation is mostly based on a canopy height model (CHM) (Pyysalo and Hyyppä, 2002; Solberg et al., 2006) or on the original 3D point cloud (Reitberger et al., 2009; Wu et al., 2016). After the segmentation process, extracted single tree objects can be classified according to tree species. Therefore, the majority of previous studies typically relied on a two-step approach. First, handcrafted feature sets describing the geometry and radiometry of single trees were generated from the remote sensing data. Second, appropriate machine learning (ML) classifiers, such as support vector machines (SVMs) or random forests (RFs), were applied for classification. For example, Heinzel and Koch (2012) investigated different feature sets derived from full-waveform lidar data, hyperspectral data, and color infrared (CIR) images in a temperate forest. Their SVM-based method could classify pine (*Pinus sylvestris*), spruce (*Picea abies*), oak (*Quercus petraea*), and beech (*Fagus sylvatica*) with an overall accuracy (OA) of 89.7%, 88.7%, 83.1%, and 90.7%, respectively. Dalponte et al. (2012) used airborne hyperspectral imagery and lidar data from a mountain area in the Southern Alps. They investigated the performance of both RF and SVM classifiers on different feature subsets generated from data with varying spatial resolution. Overall, seven species and a "non-forest" class were classified with an OA of 83.0%. In a mixed temperate forest, Shi et al. (2018) categorized five species by fusing ALS data with hyperspectral imagery (OA = 83.7%). The authors successfully combined plant functional traits (e.g. equivalent water thickness, leaf mass per area and leaf chlorophyll), spectral features, and lidar metrics.

Recently, the classification of dead trees has become increasingly important. Most previous studies regarded this task as a binary problem and classified tree objects into dead or living. For instance, Yao et al. (2012) utilized an SVM classifier and handcrafted features generated from full waveform lidar data (25 points/m^2) captured in a mixed mountain forest in the Bavarian Forest National Park (BFNP). Based on features derived from the 3D point cloud, laser intensity, and laser pulse width, their method classified dead and living trees with an OA of 73% for leaf-on trees and 71% for leaf-off trees. Polewski et al. (2015) presented an active learning-based approach to detect standing dead trees (snags) in the BFNP. Using features from ALS point clouds and CIR imagery, manually labeled single trees could be classified into dead and living with an OA of 89%. Casas et al. (2016) proposed a classification model based on single-tree ALS metrics and separated snags from living trees with an OA of 92%. In a comprehensive study, Kaminska et al. (2018) trained an RF classifier using intensity and structural variables from multi-temporal ALS data (6 points/m^2) and spectral information generated from 20 cm leaf-on CIR images. Their method classified three tree species (spruce, pine, and deciduous), and further categorized them as "dead" or "alive" (OA = 94.3%). More recently, Krzystek et al. (2020) conducted a large-scale experiment in an area of 924 km^2 to classify single trees in the BFNP. Based on ALS data and CIR imagery, their binary classifier separated dead from living trees with an OA of 93%. In summary, the overall performance of approaches for individual tree species classification in dense (and thus complex) temperate forests is still insufficient for practical use, requiring an OA of at least 90% for multi-class tasks.

1.2. Deep learning-based approaches

In recent years, utilizing high-performing deep learning (DL) methods as classification tools has garnered a large amount of interest, outperforming standard ML approaches in various tasks (Voulodimos et al., 2018). Presumably, the biggest advantage of these deep neural networks (DNNs) is their so-called representation learning, which characterizes the automatic extraction of features as part of the training process (LeCun et al., 2004). For scene understanding from irregular and unordered 3D point clouds, Griffiths and Boehm (2019) outlined four general types of DL approaches. On the one hand, the authors reviewed methods that either render multi-view images (Qi et al., 2016) or transform input data into RGB-depth images (Zhao et al., 2018). Thus, proven and efficient 2D convolutional neural networks (CNNs) such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2015), and ResNet (He et al., 2015) can be applied. On the other hand, the authors discussed volumetric approaches that discretize raw 3D data as regular 3D voxel grids and subsequently use 3D convolutions to extract meaningful information (Zhou and Tuzel, 2018). Recently, powerful network architectures such as PointNet++ (Qi et al., 2017) and PointCNN (Li et al., 2018) have been developed. These 3D DNNs enable direct input of raw and unstructured point clouds without the need for prior rasterization or voxelization. Therefore, they allow end-to-end classification of 3D point clouds.

So far, the application of DL methods for the classification of pre-segmented single trees based on lidar data has been rarely investigated. Presumably, one reason for this research gap is the lack of large training datasets. In a natural forest (330 stems/ha), Hamraz et al. (2019) utilized a CNN to classify overstory coniferous and deciduous trees. By generating images from leaf-off and leaf-on ALS point clouds (50 points/m^2), a cross-validated classification accuracy of 92% for coniferous trees and 87% for deciduous trees could be reached. Overall, the CNN was up to 14% more effective than traditional learning methods using handcrafted features. In an urban study area, Hartling et al. (2019) fused data from satellite imagery and lidar data. Using DenseNet (Huang et al., 2017), an overall accuracy (OA) of 83% in classifying eight individual tree species was achieved. Moreover, their approach was clearly superior to both RF (OA = 52%) and SVM (OA = 52%) classifiers, even with restricted training sample quantities. In a tropical wetland located in South China, Sun et al. (2019b) developed a patch-based classification algorithm for seven classes, including six individual tree classes (1388 training samples, 362 test samples). Initially, single trees were segmented by calculating a CHM from the lidar point cloud (5–8 points/m^2). Then, the segment information was utilized to generate 64x64 image patches by cropping 10 cm aerial RGB images. Their most effective model was a modified version of ResNet-50, which classified image patches with an OA of 90%. In the same research area, Sun et al. (2019a) mapped 18 tree species using ALS data and high-resolution RGB images, achieving an OA of 73% at the single-tree level. Their approach involved the application of three well-known CNNs (AlexNet, VGG-16, and ResNet-50). Recently, Briechle et al. (2020a) classified three tree species (pine, birch, and alder) and standing dead pines with crowns using PointNet++ along with UAV-based lidar data and MS imagery. Aside from 3D geometry, laser echo pulse width (EW) values and MS features were also integrated into the classification process. Overall, their DL-based method (OA = 90%) successfully used raw 3D data and was superior to a baseline method using an RF classifier and handcrafted features (OA = 85%).

1.3. Key idea and main issues

In the present paper, the objective was to classify presegmented 3D single tree objects with respect to tree species and dead trees in a combined approach. Therefore, because of its proven outstanding performance, a CNN-based procedure was chosen. Additionally, we applied the technique of transfer learning to tree species classification. Instead of training all model parameters from scratch, this approach is based on pretrained parameters that are fine-tuned on the basis of a task-specific dataset. Especially for relatively small datasets, this procedure allows effective model adaptation, even if there is a considerable domain shift between an existing image collection and a new dataset (Prabha et al., 2020). Essentially, our approach was supported by the idea that a person would likely classify a single tree by looking at its silhouette from different angles. Our approach was further supported by a review of DL methods for 3D data, which found that systems using discrete 2D representations of 3D data typically outperform approaches based on 3D voxel representations (Ioannidou et al., 2017). We therefore wanted to investigate whether multi-view 2D images could also exceed point networks. Thus, the key idea of this study was to train a CNN fusing MS image patches and multiple side-view images generated from UAV-based and helicopter-based lidar data. In addition to the geometric information, we also incorporated calibrated laser echo characteristics (EC) into the classification pipeline. The experiments conducted examined the following research questions:

- Can this new method successfully be applied to data from two regions captured with different lidar sensors and MS cameras?
- Compared to the baseline approach using PointNet++, is there an improvement in classification accuracy when utilizing a CNN approach and multiple 2D representations of single trees?

Furthermore, we investigated some relevant practical issues:

- Is the masking of MS image patches necessary?
- Can the incorporation of laser EC improve performance?
- Which classes can be classified more accurately than others? Why are some classes particularly difficult to distinguish?

The most innovative contribution of our pipeline for single-tree classification is the fusion of MS image patches and multi-view images generated from 3D point clouds in a dual-CNN approach. Furthermore, we initialize the CNN models using pretrained weights and optimize the network parameters by recursive retraining. To visualize the networks' decisions, we use class activation mapping (CAM). In the following sections, we address the study areas, sensors, data preprocessing, and reference data. Subsequently, we present our methodology for tree species classification and the baseline method. Then, we outline the conducted experiments and the main outcomes, including a comparison of both methods. Finally, we discuss the results in relation to previous research and draw conclusions.

2. Materials

2.1. Study areas

In this paper, we present experiments building on datasets from two study areas. The first study area, Chernobyl Exclusion Zone (ChEZ), is densely vegetated with a tree density of approximately 400 trees/ha. The main tree species are Scots pine (*Pinus sylvestris*), silver birch (*Betula pendula*), and black alder (*Alnus glutinosa*), with tree heights up to 30 m (Bonzom et al., 2016). Overall, the forest stand is dominated by Scots pine planted after the nuclear disaster of 1986 (Yoschenko et al., 2011), comprising approximately 50% of all trees. Based on visual interpretation of aerial imagery, we roughly estimated the distribution of pines, birches, and alders to be 50%, 20%, and 30%, respectively. The second study area, Bavarian Forest National Park (BFNP), was established in 1970 and is part of the Natura 2000 network, which was founded to protect the most endangered habitats and species in Europe. The BFNP contains protected flora and fauna of exceptional natural value (Zenáhlíková et al., 2015). The forest area is dominated by Norway spruce (*Picea abies*), European beech (*Fagus sylvatica*), silver fir (*Abies alba*), and larch (*Larix*). Furthermore, other tree species appear less frequently, such as silver birch (*Betula pendula*), sycamore maple (*Acer pseudoplatanus*), and common rowan (*Sorbus aucuparia*) (Cailleret et al., 2014). Due to bark beetle infestation, extensive areas are covered with dead wood – fallen dead trees, standing dead trees, and standing dead trees without crowns (also known as snags).

2.2. Data acquisition and preprocessing

In the ChEZ, we utilized an octocopter developed by a team from the Department of Nuclear Physics Technologies of the Institute of Environment Geochemistry of the National Academy of Sciences of Ukraine. All flights were carried out in fully automatic mode using Global Navigation Satellite System (GNSS) waypoints. Data collection was performed during sunny and partly cloudy weather conditions at a mostly constant wind speed (2–3 m/s). In April 2018, lidar data were collected by a YellowScan Mapper I laser scanner, resulting in a nominal point density of approximately 53 points/m^2. Before the data collection, a calibration flight over a building was conducted to check the boresight angles (BayesMap Solutions LLC, 2018) preset by the manufacturer. Differential GNSS postprocessing (NovAtel Inc., 2017) incorporating GNSS measurements collected by a Trimble R4 base station ensured flight trajectories with centimeter-level precision. Overall, the mean discrepancy between adjacent lidar strips was approximately 5 cm, which is in the range of the measurement accuracy of the instrument. Absolute 3D georeferencing with an accuracy of a few centimeters was achieved by fitting the ALS point cloud to the enclosing polygons of a nearby building. Moreover, the recorded lidar data were radiometrically corrected based on the data-driven method presented in Briechle et al. (2020b). Additionally, we captured MS images using two MicaSense RedEdge cameras that were mounted in a twisted configuration with an angle of approximately 23°. Compared to a field of view (FOV) of 47° for a single camera setup, this setup guaranteed a 50% side overlap of the two camera footprints, thereby increasing the total FOV to approximately 70°. To compensate for changing lighting conditions during and between the flights, we utilized MicaSense's calibrated reflectance panel and downwelling light sensor. These accessories provided useful information for the subsequent reflectance calibration in Agisoft PhotoScan Professional 1.4.1 (Agisoft LLC, 2018). Next, all images were aligned in a bundle adjustment, resulting in a mean reprojection error of 1.3 pixels. Finally, 10 cm MS true orthophotos were generated using the lidar-based surface model as a reference.

In the BFNP, airborne full waveform data were acquired in June 2017 (leaf-on condition) using a Riegl LMS-Q680i instrument carried by a helicopter. The resulting average point density was 55 points/m^2. Additionally, a calibration flight was conducted on a nearby airfield and enabled the correction of the raw amplitude values with regard to traveling distance of the laser beam (Amiri et al., 2019). Next, georeferencing quality was checked based on in-field measurements of vertical and planimetric objects, such as flat areas and enclosed building polygons, respectively. On average, the mean 3D displacements of the lidar data were less than 10 cm. MS aerial imagery in the BFNP was also acquired in June 2017, using a Leica DMC III camera. GNSS data and Inertial Navigation System data provided initial values for the exterior camera orientation. Using the software package Agisoft PhotoScan Professional 1.4.1, the aerotriangulation was performed based on aerial images, a camera calibration model, and ground control points, leading to a sigma naught of 30% of the ground sampling distance (GSD). Next, we generated true orthophotos on the basis of the lidar-based digital surface model. Finally, single trees were delineated from the lidar point

Table 1
Study areas and sensor equipment.

	ChEZ	BFNP
Location	51°23'N, 30°04'E	49°04'N, 13°18'E
Size of study area	37 ha	8.3 km^2
Tree density	400 trees/ha	530 trees/ha
Tree height	15–30 m	15–50 m
Platform	UAV (octocopter)	Helicopter D-HFCE/AS350
Lidar sensor	YellowScan Mapper I	Riegl LMS-Q680i
Laser wavelength	905 nm	1550 nm
Echo characteristics	Pulse width	Intensity
Flight altitude	50 m	550 m
Flight speed	6 m/s	30 m/s
Point density	53 points/m^2 (leaf-off)	55 points/m^2 (leaf on)
MS camera	MicaSense RedEdge	Leica DMC III
Focal length	5.5 mm	92 mm
MS bands	blue (B), green (G), red (R), red edge (RE), near infrared (NIR)	B, G, R, NIR
Flight altitude	130 m	2880 m
Flight speed	9 m/s	30 m/s
End/side lap [%]	79/50	80/60
GSD of orthomosaics	10 cm	20 cm

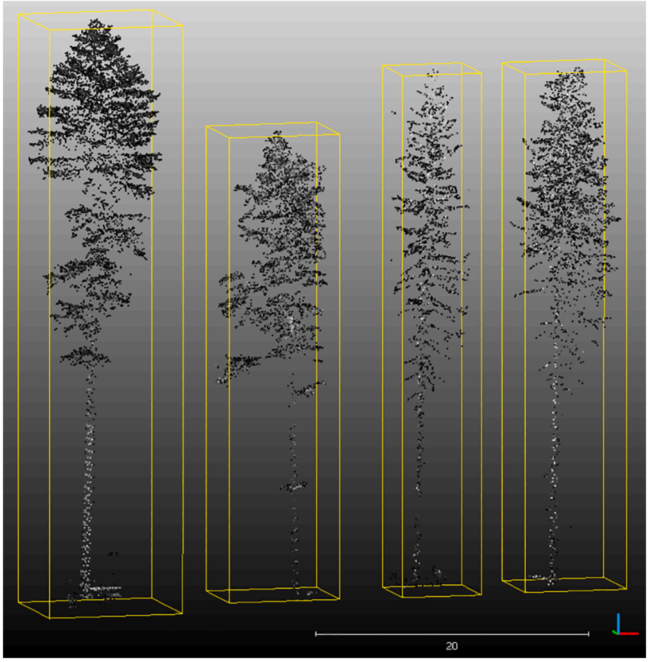


Fig. 1. 3D point clouds for selected samples from BFNP dataset, coloured by normalized intensity from black (0) to white (1). From left to right: coniferous, deciduous, snag, dead tree.

cloud in both study areas utilizing the normalized cut algorithm presented by Reitberger et al. (2009). Following the authors' recommendations, we set the static stopping criterion of the normalized cut segmentation to 0.16. The segmentation quality was not tested quantitatively, however visual inspection helped to verify that no major oversegmentation or undersegmentation occurred. Aside from individual point clouds, the segmentation also provided projected 2D polygons for each tree. Table 1 shows an overview of study areas, sensor platforms, sensor equipment, and data acquisition parameters.

Table 2
Number of samples for study area ChEZ; train/val/test split: 56%/14%/30%.

Tree class	Training samples	Validation samples	Test samples
pine	93	23	51
birch	93	23	51
alder	93	23	51
dead tree	93	23	51
Σ	372	92	204

Table 3
Number of samples for study area BFNP; train/val/test split: 51%/22%/27%.

Tree class	Training samples	Validation samples	Test samples
coniferous	345	149	259
deciduous	345	149	202
snag	345	149	139
dead tree	345	149	145
Σ	1380	596	745

2.3. Reference data

Based on visual interpretation, single tree segments were manually labeled using an interactive tool. Note that incorrect segments were generally not considered in the labeling process to make our classification results independent of the segmentation quality. In the user interface, randomly chosen tree segments are displayed in 3D. The point cloud can be rotated, thereby supporting the annotator in verifying the class label. Furthermore, the corresponding 2D polygon for each segment is superimposed on the aerial image. In detail, the trees in the ChEZ were manually subdivided into the classes "pine", "birch", "alder", and "dead tree". In the BFNP, we labeled the trees with the categories "coniferous" (mostly spruce), "deciduous" (mostly beech and larch), "snag", and "dead tree" (Fig. 1). Here, "snag" refers to a partly or completely dead tree missing a crown or most of the smaller branches (Yao et al., 2012). In contrast, trees labeled "dead tree" are dead trees with crowns. The distinction between "snag" and "dead tree" was based on the subjective perception of three different research assistants. Subsequently, the labeled samples were randomly sorted into training, validation, and test datasets (see Tables 2 and 3). Note that we also included class balancing for both training and validation data.

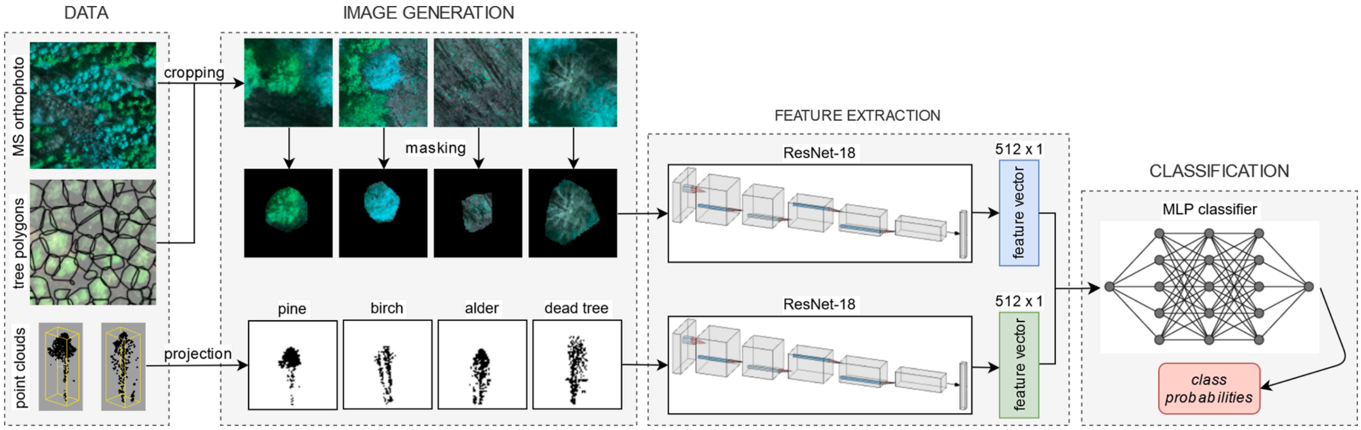


Fig. 2. Outline of the proposed method, Silvi-Net$_{single}$.

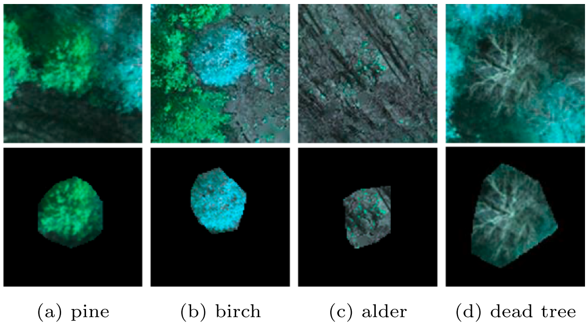


Fig. 3. *MS_unmasked* images (first row) and *MS* images (second row) generated from MS orthomosaics in the ChEZ study area; false-color images (RGB, RE, NIR). Image size corresponds to 10 m × 10 m.

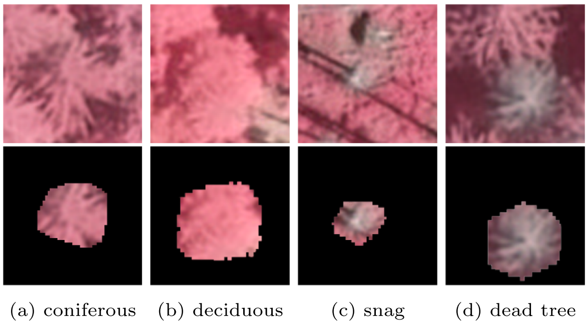


Fig. 4. *MS_unmasked* images (first row) and MS images (second row) generated from MS orthomosaics in the BFNP study area; CIR images (G, R, NIR). Image size corresponds to 12 m × 12 m.

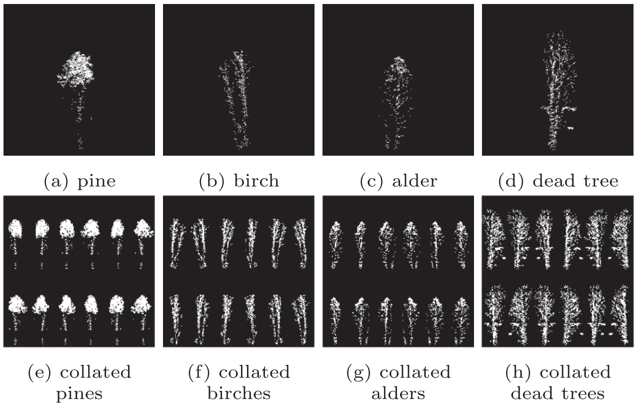


Fig. 5. *GEOM* images generated from UAV-based lidar data in the ChEZ study area. First row (a-d): single images, image size corresponds to 26 m × 26 m. Second row: collated multi-view images. Image size corresponds to 52 m × 52 m.

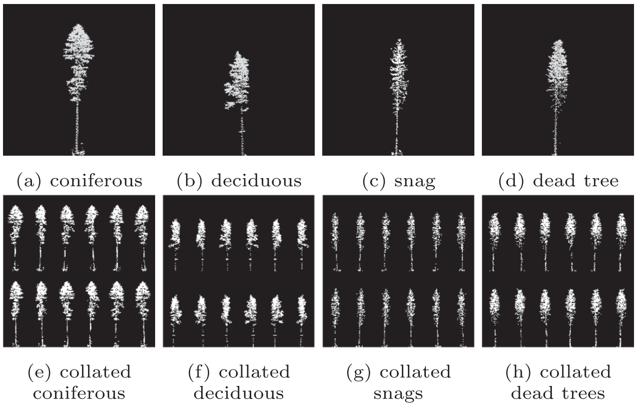


Fig. 6. *GEOM* images generated from ALS data in study area BFNP; First row (a-d): single images, image size is corresponding to 50 m × 50 m; Second row: collated multi-view images, image size is corresponding to 100 m × 100 m.

3. Methodology

3.1. Outline of the proposed method

In general, our network architecture is inspired by DualNet (Hou et al., 2017), a DNN which includes two parallel CNNs and a subsequent aggregation of complementary features in a final classifier. For a better understanding of the overall processing pipeline, important steps of Silvi-Net are illustrated in Fig. 2. Initially, 2D representations of the single trees were created in an image generation process. For each tree, an MS image patch was cropped to place the tree crown in the image center. In this step, we utilized the polygon outlines generated by the lidar-based tree segmentation. To maintain the relative dimensions of the crowns, image patches with the same quadratic size were produced. Thus, we ensured that even the largest tree crowns were included in the images in their entirety. Outlines of the projected 2D tree polygons were used to mask pixels not corresponding to the actual tree. In addition to the MS images, we rendered multiple side-view images from the segmented 3D lidar point clouds of single trees, representing the trees' silhouettes. Optionally, these images were enriched with laser EC values. Basically, we created two types of image sets – one with 12 individual images per tree, and one with an image collage comprised of all 12 side-view images per tree. In the following sections, the approaches utilizing these datasets are referred to as Silvi-Net$_{single}$, and Silvi-Net$_{collages}$ respectively. After the image generation process, features were automatically extracted using two independently trained ResNet-18 models, optimized for both the MS and side-view images. Here, we applied the idea of transfer learning and pretrained weights. To visualize the model's decisions, we produced CAM images and superimposed them on the input images. Overall, we generated 512 features per side-view image or image collage, and additional 512 features from each MS image. Next, the feature vectors were fused and fed into a standard multi-layer perceptron (MLP) that was trained to estimate class probabilities for each sample. For Silvi-Net$_{single}$, the classification led to 12 predicted labels per tree. Therefore, we introduced Silvi-Net$_{majVot}$, an additional evaluation strategy applying majority voting to these 12 predictions. The idea was to outvote individual, falsely classified side-view images and to obtain one label per tree. In the following sections, all of the steps of our approach are described in greater detail.

3.2. Generation of image patches

First, we present the methodology for image generation from MS orthomosaics. Because CNN-based image classification supported by transfer learning typically utilizes three-channel imagery, we reduced the five-channel images in the ChEZ area. More specifically, we transformed the B, G, and R channels into a single gray scale channel by calculating the mean value of these three channels for each pixel. Next, we added the RE and NIR channels, resulting in a three-channel image. Because of the difference in sensors, an alternative procedure was conducted for the BFNP data. In this study area, we removed the blue channel from the raw imagery and utilized the resulting CIR images. For both study areas, we normalized the three image channels independently to values between 0 and 1. For each tree segment, the corresponding polygon was projected onto the orthomosaic. Then, a cropped image patch was produced covering a predefined quadratic region around the polygon center. The image size resulted from the maximum crown dimension (ChEZ: 10 m × 10 m, BFNP: 12 m × 12 m) and the pixel size of the orthomosaics (ChEZ: 10 cm, BFNP: 20 cm). Thus, all tree crowns fit within the image dimensions. Ultimately, this process led to images sized 100 × 100 pixels for ChEZ and 60 × 60 pixels for BFNP. Optionally, pixels outside the tree polygon were masked out and set equal to 0. Thus, *MS_unmasked* and *MS* datasets were prepared on the basis of the ChEZ dataset (Fig. 3) and the BFNP dataset (Fig. 4). In total, 668 MS patches were generated in the ChEZ area, and 2,721 in the BFNP area – each with a masked and an unmasked version.

Additionally, we prepared two different types of images from the 3D lidar point clouds – one with 12 individual images per tree and one with an image collage comprised of all 12 side-view images per tree. In a first step, the point clouds were rotated in constant steps around the z axis to simulate multi-view positions. After visual interpretation, we decided to set the rotation angle to multiples of 30°, leading to multi-view image stacks of 12 images per tree. This was deemed an acceptable balance between information loss and redundancy. Next, we rendered binary silhouette-like images for both study areas (see Fig. 5 a-d and Fig. 6 a-d) by projecting the 3D data onto a virtual vertical raster. The image resolution was set to 10 cm per pixel. Because the images should completely cover even the largest trees, the image size – 260 × 260 px in the ChEZ and 500 × 500 px in the BFNP – was determined by the maximum tree height in the corresponding study area. Note that the image size in the BFNP was much larger than the required input size of ResNet-18. The average tree height was 16.5 m (std = 1.2 m) in the ChEZ and 28.1 m (std = 6.1 m) in the BFNP, respectively. As a consequence, 90% of all trees covered at least half of the image height in the ChEZ. With 70%, this ratio was clearly lower in the BFNP. In case of working with data from forests with an even larger range of tree sizes, we assume that the image size should be calculated in a different way. Otherwise, decreasing results are likely to appear because of significant loss of detail for the average and smaller trees. Furthermore, we wanted to analyze the impact of EC on our classification method. Therefore, we included EW values in the ChEZ dataset and intensity (INT) values in the BFNP dataset. Incorporating the normalized EW and INT values, we also generated 8–bit grayscale images. These two image datasets are referred to as *GEOM* (binary images) and *GEOM_EC* (grayscale images), respectively. Overall, the preprocessing of samples for Silvi-Net$_{single}$ led to 8,016 *GEOM* and *GEOM_EC* images each in the ChEZ area (training: 4,464; validation: 1,104; test: 2,448) and 32,652 *GEOM* and *GEOM_EC* images each in the BFNP (training: 16,560, validation: 7,152, test: 8,940). Subsequently, we rendered image collages utilized in the approach Silvi-Net$_{collages}$, including all 12 views per tree in one image. Therefore, assuming that only the middle third of the quadratic side-view images contains useful information, we cut out these essential image parts. Next, we randomly arranged them as matrices comprising two rows with six images each (see Fig. 5 e-h and Fig. 6 e-h). Thus, unlike the previously created single-view images, the number of samples was equivalent to the number of tree objects (see Tables 2 and 3).

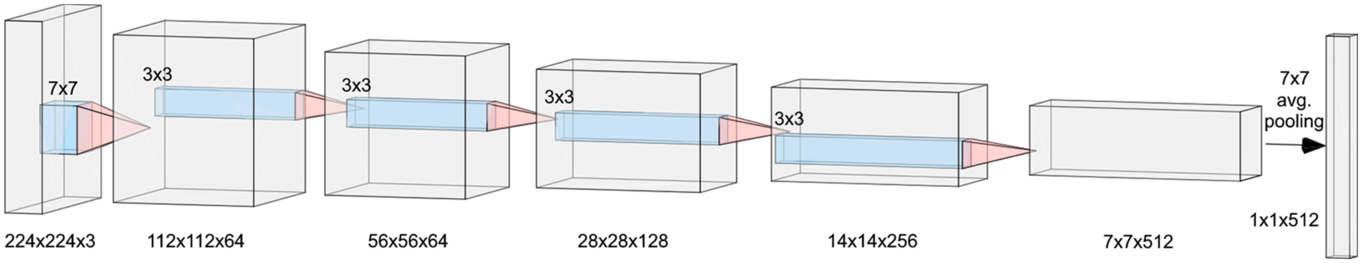


Fig. 7. Simplified ResNet-18 architecture, created with a neural network drawing tool (LeNail, 2019).

3.3. CNN-based feature extraction

In our approach, automatic extraction of features was performed using a standard CNN. Our decision was motivated by the fact that CNNs are well-established neural networks for image-based deep supervised learning that are capable of achieving excellent results in the fields of pattern recognition and machine learning (Schmidhuber, 2015). Moreover, CNNs are especially designed to process sensor data represented as multiple 2D arrays. By considering local and global stationary properties, CNNs have achieved state-of-the-art results in popular image classification tasks, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). In general, CNNs are deep neural networks consisting of numerous stacked layers. The first part, also known as the feature extractor, is mainly comprised of convolutional blocks – sequences of convolutional layers, activation layers, and pooling layers. Thereby, convolutional layers utilize filter kernels to extract low-level image features. Moreover, the kernel depth is equal to the number of image channels. The output of a convolutional layer is a feature map with one channel per filter kernel. Activation layers, such as the rectified linear unit (ReLU), account for non-linear effects. Practically speaking, ReLU sets negative values to the value 0 and reduces the problem of vanishing gradients – an effect that occurs with deep neural networks. Additionally, ReLU layers are computationally inexpensive and enable faster model convergence. Pooling layers essentially subsample the feature maps, using common methods such as average pooling and maximum pooling. The second part of a CNN is the actual classifier. Here, the final output feature maps are flattened into a one-dimensional vector, followed by fully connected classification layers (LeCun et al., 2015). Overall, CNNs include a huge set of model parameters – weights and biases – that need to be estimated. To reduce model overfitting, regularization techniques such as dropout and batch normalization are often included in typical CNN architectures. By adding dropout layers, co-adaptation of neurons can be prevented. Moreover, this technique approximates the idea of ensemble models and allows a higher learning rate (LR). However, it also usually leads to slower model training. Additionally, batch normalization layers can help improve model stability and quality (Ioffe and Szegedy, 2015). Practically speaking, these layers apply channel-wise normalization of the feature maps and result in faster model convergence.

3.3.1. CNN architecture

In our classification pipeline, we utilized two standard ResNet-18 models (He et al., 2015) implemented with the PyTorch framework, version 1.1.0 (Paszke et al., 2019), which is an optimized tensor library for DL using GPUs and CPUs. With their proposed idea of residual blocks, the developers of ResNet successfully minimized the problem of vanishing gradients. At this time, the problem was that the training accuracy of multi-layer CNNs dropped as the number of layers increased. Therefore, the authors proposed to use a reference to the previous layer to compute the output at a given layer. As a result of these so-called skip connections (also termed shortcuts), the training of much deeper CNNs was facilitated. Moreover, these deep residual networks can achieve improved accuracy due to considerably increased depth. In 2015, He et al. (2015) won the ILSVRC using ResNet ensembles with a depth of up to 152 layers. Our decision to use ResNet-18 was motivated by preliminary studies testing different CNN architectures included in the "models" subpackage of the "torchvision 0.3.0" module in PyTorch. Here, both VGG-16 (Simonyan and Zisserman, 2015) and Densenet-121 (Huang et al., 2017) performed significantly worse than ResNet-18 when initialized using weights pretrained on the ImageNet dataset (Deng et al., 2009). Adopting the deeper version, ResNet-50, did not improve the results. Presumably, our dataset was too small to retrain all 23.6 million ResNet-50 parameters in an effective way. In contrast, we were able to robustly retrain all 11.2 million ResNet-18 parameters and optimize the network for our task. In this way, we achieved a suitable trade-off between network depth and dataset size. Fig. 7 shows the architecture of ResNet-18 in a simplified way, including dimensions of tensors and filters and final feature vector. The feature extractor of ResNet-18 consists of four residual blocks. Each block is comprised of a stack of two basic blocks of 2–3 convolutional layers, followed by batch normalization and ReLU layers. Finally, an average pooling layer extracts 512 features per 224 × 224 × 3 input image.

3.3.2. CNN training

We utilized two separate ResNet-18 models optimized for the classification of *GEOM/GEOM_EC* images and *MS* images, respectively. At the beginning, all images were loaded and resampled to the required image size of 224 × 224 pixels. Next, the images in the range [0, 255] were converted to floating tensors in the range [0.0, 1.0]. Then, the three channels of these image tensors were standardized separately using the mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225) values of ImageNet. For each channel, the mean of the data was 0 and the standard deviation was 1. Next, data augmentation was performed on the training data. This method to artificially increase the number of training samples is helpful to avoid overfitting and usually results in better generalization properties of the trained model (Goodfellow et al., 2016). In our approach, we applied a combination of random affine transformation and random horizontal flip to both training and validation data. Moreover, we randomly flipped the *MS* images vertically. Note that this transformation was not performed for the lidar-based side-view images to maintain the vertical orientation of trees. The affine transformation of image coordinates x and y into new image coordinates x′ and y′ can be described as a sequence of rotation, shearing, scaling, and translation (Eq. 1):

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix}\begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}\begin{bmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} \tag{1}$$

where α is the rotation angle, s is the shearing parameter, s_x and s_y are the scale parameters for both coordinate axes, and t_x and t_y are the components of the 2D translation vector. In or approach, we allowed a relative maximum image translation of ±10% in horizontal and vertical directions and scaling parameters s_x and s_y in the interval of [0.80, 1.25]. The shearing parameter s was set to 0. For the rotation angle α, the range defining the maximum random value was set depending on the image type: ±20° for *GEOM* and *GEOM_EC* images and ±180° for *MS* images.

For model training, we utilized the concept of transfer learning. Numerous researchers have shown that DL-based models are able to learn features that – to a certain extent – transfer well across datasets (Hu et al., 2015; Shin et al., 2016). Instead of starting with random parameter values, models can be initialized with weights optimized for extensive and standardized databases like ImageNet. Although the ImageNet dataset includes 1000 object classes and is clearly different from our tree data, we assumed that it would be adaptable for the task of tree species classification. Besides relatively quick convergence, effective fine-tuning of DNNs typically requires much less samples compared to training from scratch (Ng et al., 2015). Therefore, we initialized the ResNet-18 models by utilizing pretrained ImageNet weights. Moreover, we set the maximum number of epochs to 100 and implemented an early stopping criterion defined as 10 epochs with no improvement in validation loss. More precisely, the criterion for model evaluation was based on a cross-entropy loss function. For each image batch (batch_size = 32), we calculated the loss with shared class weights (Eq. 2) and averaged all losses per epoch.

$$loss(x, class) = -log\frac{exp(x[class])}{\Sigma_j exp(x[j])} \tag{2}$$

The model hyperparameters were optimized using an Adam optimizer (Kingma and Ba, 2015). Here, we relied on the default values of PyTorch implementation. Every seven epochs, an exponential learning rate (LR) scheduler decayed the initial LR of 0.001 by a factor of $\gamma = 0.1$ (Eq. 3):

$$LR_{i+1} = LR_i{}^*(1-\gamma). \tag{3}$$

Overall, ResNet-18 is comprised of approximately 11.2 million trainable parameters. In our classification pipeline, the crucial factor for the generalization of well-performing models was a systematic recalculation of the model parameters for each dataset. The procedure was as follows: First, we set all parameters of the feature extractor to be invariable and only retrained the 16,548 parameters of the fully connected layers. Second, we iteratively "unfreezed" the trainable parameters of the four residual blocks. Starting with the deepest block (8.4 million parameters), the number of trainable parameters increased to 10.5 million, 11.0 million, and finally 11.2 million. Third, for each dataset, the model showing the lowest cross-entropy loss on the validation dataset was stored. Finally, this best performing model was set to evaluation mode and was subsequently used to extract 512 features per image. Thus, the optimized ResNet-18 models were practically utilized as automatic feature extractors for the training, validation, and test datasets. In our implementation, we registered a so-called "forward hook" to enable feature extraction from the average pooling layer. In PyTorch, this step was performed utilizing the "register_forward_hook" function in the "nn" (neural network) package.

3.4. MLP-based tree classification

In our classification pipeline, optimized ResNet-18 models were used as automatic feature extractors. To perform tree species classification utilizing 2D representations rendered from both airborne lidar and MS data, we combined the feature sets generated from the side-view images (*GEOM* and *GEOM_EC*) and *MS* images. Next, we inputted the fused feature vectors comprising 1,024 features and the corresponding class labels to a standard MLP classifier. An MLP is a non-parametric neural network classifier with shallow structures containing only a few feature representation levels. Typically, an MLP is composed of interconnected nodes in multiple layers (namely input, hidden, and output layers), with each layer fully connected to both the preceding and succeeding layers (Del Frate et al., 2007). Moreover, the outputs of each node are weighted units followed by a nonlinear activation function (Pacifici et al., 2009). In summary, in a feed-forward manner, an MLP maps a set of input features onto a set of labels (Atkinson and Tatnall, 1997). In our method, we utilized the "MLPClassifier" class from the "sklearn" module "neural_network" (Pedregosa et al., 2011). More specifically, we implemented an MLP with three hidden layers composed of seven neurons each, and set the hyperparameters to the default values. The particular types of feature vectors generated from the *MS* images and side-view images, were weighted 50% each. The MLP classifier was trained using the combined feature set calculated from the training and validation datasets. Finally, we evaluated the MLP on the independent test datasets and derived confusion matrices and standard metrics from the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. We calculated the OA (Eq. 4), precision (Eq. 5), recall (Eq. 6), and F_1 score (Eq. 7).

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

3.5. Baseline method (PointNet++)

For the classification of 3D objects such as trees, it is unknown whether working with rendered multiple 2D images or raw 3D point clouds is favorable. Thus, we compared our new CNN-based method to the approach presented in Briechle et al. (2020a), using a PyTorch implementation of PointNet++ (Wijmans, 2018) for object classification. In the following sections, the most important steps of this baseline method will be explained, including data preparation, network training, and validation.

3.5.1. Preparation of dataset

Before training the network, the 3D dataset had to be prepared appropriately. Typically, PointNet++ can only manage a constant number of 3D points per sample. Therefore, we applied a combined sampling approach to achieve balance between upsampling and downsampling of data, resulting in 1,024 points per tree. Moreover, as generally proposed when working with DNNs, data was standardized. Next, we calculated the surface normals for all 3D points using the "estimate_normals" function from the open source library Open3D (Zhou et al., 2018). Then, handcrafted MS features were generated and integrated into the dataset. Here, we relied on a selection of statistical MS features computed with different vegetation indexes (VI). Depending on the available spectral channels, the number of VIs differed between the study areas. In the BFNP, we derived the normalized difference vegetation index (NDVI; Rouse et al. (1973)) from the CIR images. In the ChEZ area, the five-channel orthomosaics also enabled the calculation of the red edge normalized difference vegetation index (RENDVI; Gitelson and Merzlyak (1994)), the red edge difference vegetation index (REDVI; Briechle et al. (2020a)), the modified red edge simple ratio (MRESR; Datt (1999)), and the modified chlorophyll absorption ratio index (MCARI; Daughtry et al. (2000)). Projections of the tree polygons were utilized to filter VI pixels belonging to a single tree. Subsequently, we computed 12 object-based statistical features from these pixels for each VI – the maximum value (max), minimum value (min), range (max - min), mean value, standard deviation, mode[1], skewness[2], kurtosis[3], as well as the 25th ("1st quartile"), 50th ("median"), 75th ("3rd quartile"), and 90th percentile (perc).

To make the classifier more robust and to avoid overfitting, the feature space was reduced to the five most important MS features. Here, we relied on an RF-based feature selection technique which has been recommended in the literature (Ma et al., 2017; Gregorutti et al., 2017), to generate a ranking of all input features according to their relative importance on the prediction. Then, for each study area, the five most decisive features were selected: *NDVI_skewness*, *MRESR_perc90*, *NDVI_perc90*, *RENDVI_mode*, and *MRESR_mode* in the ChEZ, and *NDVI_perc25*, *NDVI_skewness*, *NDVI_range*, *NDVI_mean*, and *NDVI_min* in the BFNP. Afterwards, the values of these top five MS features were standardized and assigned to each 3D point of each object, resulting in additional point attributes. Overall, the final dataset comprised 12 attributes per 3D point: the 3D coordinates and surface normals, one EC value, and five handcrafted MS features.

Table 4
Hyperparameter settings for PointNet++.

Hyperparameter	Value	Declaration
NUM_CLASSES	4	Number of object categories
NUM_POINT	1024	Number of points per sample
BATCH_SIZE	8	Number of samples per batch
MAX_EPOCH	100	Maximum number of training epochs
MAX_DROPOUT	0.5	Maximum dropout rate
OPTIMIZER	Adam	Optimization algorithm
BASE_LR	1e-3	Initial learning rate
LR_DECAY	0.7	Initial learning decay
BN_MOMENTUM	0.5	Initial momentum for batch normalization
BNM_DECAY	0.5	Decay of batch normalization momentum
WEIGHT_DECAY	1e-4	L2 regularization coefficient

[1] Most frequent value.

[2] Measure of the asymmetry of the probability distribution.

[3] Measure of the tailedness of the probability distribution.

Table 5
Silvi-Net results using only MS image patches (test dataset).

	ChEZ		BFNP	
	OA	F_1 scores per class	OA	F_1 scores per class
MS_unmasked	0.922	0.95/0.89/0.91/0.93	0.830	0.90/0.95/0.65/0.71
MS	**0.936**	0.99/0.92/0.90/0.93	**0.856**	0.90/0.92/0.78/0.76

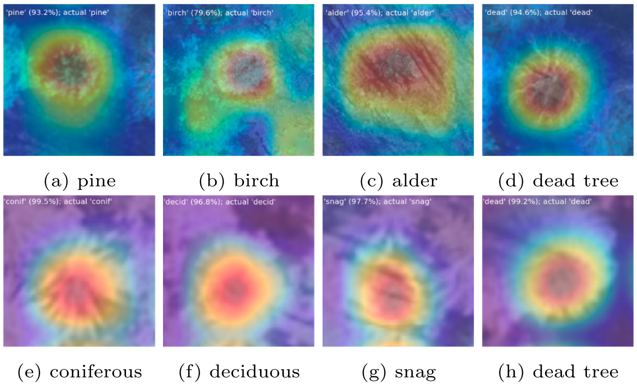

'pine' (93.2%); actual 'pine'
'birch' (79.6%); actual 'birch'
'alder' (95.4%); actual 'alder'
'dead' (94.6%); actual 'dead'
'conif' (99.5%); actual 'conif'
'decid' (96.8%); actual 'decid'
'snag' (97.7%); actual 'snag'
'dead' (99.2%); actual 'dead'
(a) pine (b) birch (c) alder (d) dead tree
(e) coniferous (f) deciduous (g) snag (h) dead tree

Fig. 8. Examples for correct classification of *MS_unmasked* images in the ChEZ (a-d) and BFNP (e-h); CAM overlay.

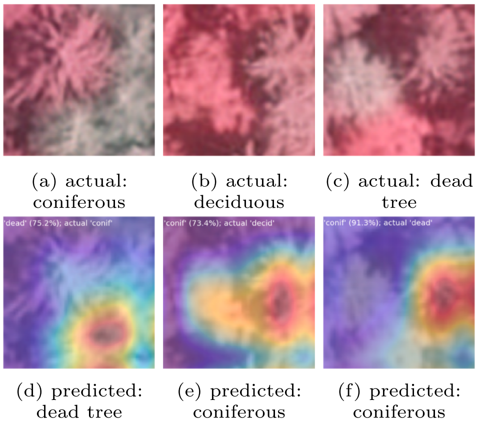

'dead' (75.2%); actual 'conif'
'conif' (73.4%); actual 'decid'
'conif' (91.3%); actual 'dead'
(a) actual: coniferous (b) actual: deciduous (c) actual: dead tree
(d) predicted: dead tree (e) predicted: coniferous (f) predicted: coniferous

Fig. 9. Examples for incorrect classification of *MS_unmasked* images in the BFNP; CAM overlay.

3.5.2. Training and validation

To successfully adapt PointNet++ for the task of tree species classification, we optimized the most decisive hyperparameters of the neural network (Table 4). Therefore, we used a combination of manual search and automated grid search. During model training, we also performed data augmentation to avoid model overfitting and to build generalizable models. Because the trained final model should be robust against object variation, we implemented random transformations of the 3D objects, including scaling in the range [0.80, 1.25], rotation around the vertical axis with an angle of the range [0, 2π], jittering with Gaussian noise (±0.05 m), and 3D translation of the entire point cloud by ±0.1 m. Moreover, setting the random input dropout parameter MAX_DROPOUT to 50% increased the robustness against varying point density and occluded object parts. Finally, we evaluated the model showing the lowest validation loss on the test dataset and generated classification metrics (OA, precision, recall, F_1 score).

4. Experiments

For both study areas, we conducted experiments based on different input datasets. Initially, we utilized sets of binary side-view images only (*GEOM*). To analyze the impact of laser EC on classification results, we trained Silvi-Net with *GEOM_EC* images. Subsequently, we classified single tree objects using only masked (*MS*) and unmasked (*MS_unmasked*) MS images. Finally, we fused automatically extracted features from both lidar-based image sets (*GEOM*, respectively *GEOM_EC*) and *MS* images for classification (*GEOM* + *MS*, respectively *GEOM_EC* + *MS*). In all experiments, we explored three different evaluation strategies – Silvi-Net$_{single}$, Silvi-Net$_{collages}$, and Silvi-Net$_{majVot}$. Furthermore, we integrated CAM technique into our pipeline to better understand the model's decisions on new independent data. Demystifying CNNs' status as "black box" systems, CAM can help to highlight class-specific, distinctive image regions (Zhou et al., 2016). To generate CAM images, the predicted class score in the range [0, 1] was mapped back to the final convolutional layer. In detail, CAM can be described as the dot product of the extracted weights from the final layer and the feature map. In our ResNet-based approach, the resulting CAM images sized 7 × 7 px were bilinearly upsampled and superimposed on the input images sized 224 × 224 px. In the following sections, classification results are presented for Silvi-Net and compared to those of the baseline, PointNet++.

4.1. Masking MS data

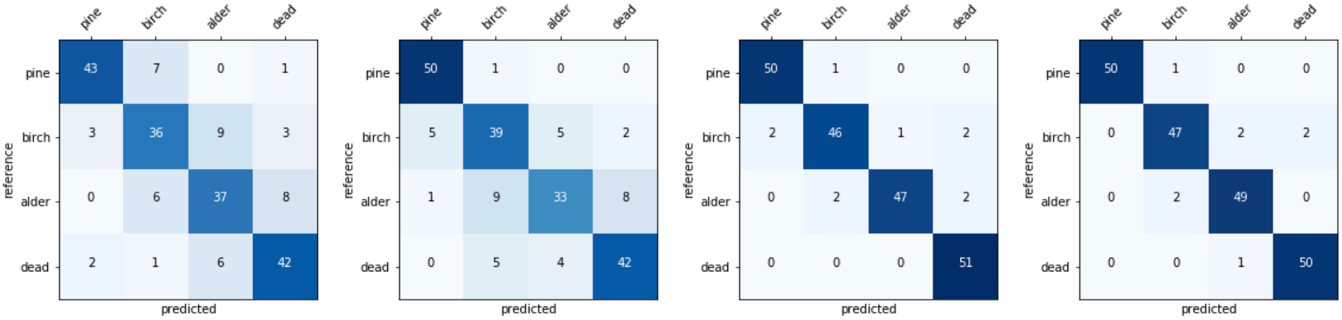
Initially, we investigated whether masking MS image patches (see Section 3.2) would improve classification results. Therefore, classification was performed with both *MS_unmasked* images and *MS* images. In general, we observed a positive impact when masking MS image patches utilizing single tree polygons for both study areas. The gain in OA was 1.4% in the ChEZ and 2.6% in the BFNP (Table 5). These relative values represent 3 of 204 test samples (ChEZ), respectively 19 of 745 test samples (BFNP). Furthermore, *MS* images yielded F_1 scores between 0.90 and 0.99 in the ChEZ. Here, masking improved the results for pine and birch. In particular, pine trees were classified almost perfectly (F_1 score = 0.99). In our second dataset (BFNP), masking pixels located in the surrounding area boosted the classification of snags and dead trees. Nevertheless, the F_1 scores for snags (0.78) and dead trees (0.76) were still relatively low. Remarkably, classification based on CIR imagery led to reasonable accuracy for the coniferous (F_1 score = 0.90) and deciduous (F_1 score = 0.92) classes in this study area. By superimposing *MS_unmasked* images with CAM images, it can be observed that in most cases the neural network automatically identified the crucial tree crowns in the image center (Fig. 8). However, in some cases, neighboring tree pixels in unmasked images affected the results. As a consequence, classification errors were produced because the CNN occasionally focused on nearby trees from different classes (Fig. 9). Thus, we relied on masked MS images in the following experiments.

4.2. Results for ChEZ

The task of classifying pine, birch, alder, and dead trees in the ChEZ was generally performed best by Silvi-Net$_{majVot}$ (Table 6). Moreover, this approach outperformed baseline method PointNet++ (OA = 84.8%) by 11.3%, reaching an OA of 96.1%. Compared to the results based on only *MS* images (OA = 93.6%; Table 5), incorporating geometry information and EC improved the results by 2.5% in this study area. Note that the discrepancy between results for validation data and test data was less than 3% in all experiments. This demonstrates the high generalization capacity of Silvi-Net. Now, we want to focus on more detailed results regarding single data subsets. When using only side-view images of the point clouds (*GEOM* images), the classification results (OA = 77.5%) were 2.0% better than PointNet++ when classifying raw 3D point clouds of the single trees (OA = 75.5%). Although geometric information was partially reduced during image generation, the F_1 scores were higher for all classes except alders. Classification based on *GEOM_EC* images (OA = 80.4%) was superior to *GEOM* images. By incorporating EC, the gain in OA was 2.9%. Specifically, pine as the only coniferous

Table 6
Results for Silvi-Net and PointNet++ on ChEZ test dataset. For each feature subset, the highest OA is displayed in bold letters, and the highest F_1 scores per class are underlined.

	Silvi-Net$_{single}$		Silvi-Net$_{collages}$		Silvi-Net$_{majVot}$		**PointNet++**	
	OA	F_1 scores per class	OA	F_1 scores per class	OA	F_1 scores per class	OA	F_1 scores per class
GEOM	0.732	0.81/0.64/0.71/0.77	0.721	0.82/0.60/0.71/0.74	**0.775**	0.87/0.71/0.72/0.80	0.755	0.83/0.67/0.73/0.79
GEOM_EC	0.765	0.89/0.69/0.67/0.78	0.716	0.82/0.62/0.71/0.72	**0.804**	0.93/0.74/0.71/0.82	0.779	0.92/0.69/0.74/0.78
GEOM + *MS*	0.912	0.94/0.89/0.90/0.92	**0.951**	0.96/0.91/0.97/0.96	**0.951**	0.97/0.92/0.95/0.96	0.821	0.81/0.78/0.81/0.88
GEOM_EC + *MS*	0.937	0.98/0.92/0.92/0.93	0.917	0.98/0.87/0.90/0.92	**0.961**	0.99/0.93/0.95/0.97	0.848	0.89/0.80/0.83/0.87



	pine	birch	alder	dead
pine	43	7	0	1
birch	3	36	9	3
alder	0	6	37	8
dead	2	1	6	42

(a) *GEOM*

	pine	birch	alder	dead
pine	50	1	0	0
birch	5	39	5	2
alder	1	9	33	8
dead	0	5	4	42

(b) *GEOM_EC*

	pine	birch	alder	dead
pine	50	1	0	0
birch	2	46	1	2
alder	0	2	47	2
dead	0	0	0	51

(c) *GEOM+MS*

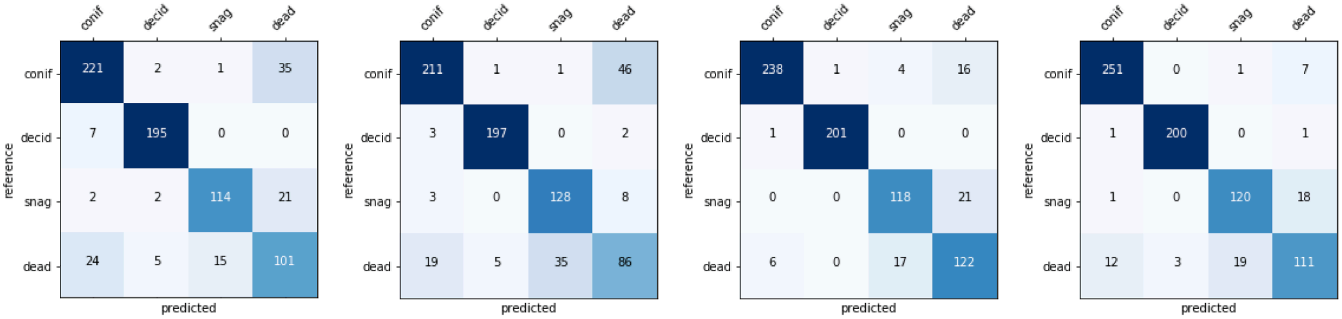
	pine	birch	alder	dead
pine	50	1	0	0
birch	0	47	2	2
alder	0	2	49	0
dead	0	0	1	50

(d) *GEOM_EC+MS*

Fig. 10. Confusion matrices for Silvi-Net$_{majVot}$ (ChEZ test data). Subfigures show results for different feature sets.

Table 7
Results for Silvi-Net and PointNet++ on BFNP test dataset. For each feature subset, the highest OA is displayed in bold letters, and the highest F_1 scores per class are underlined.

	Silvi-Net$_{single}$		Silvi-Net$_{collages}$		Silvi-Net$_{majVot}$		**PointNet++**	
	OA	F_1 scores per class	OA	F_1 scores per class	OA	F_1 scores per class	OA	F_1 scores per class
GEOM	0.811	0.80/0.95/0.86/0.60	0.744	0.68/0.96/0.84/0.51	0.847	0.86/0.96/0.85/0.67	**0.857**	0.85/0.97/0.88/0.72
GEOM_EC	0.808	0.80/0.96/0.84/0.59	0.800	0.81/0.96/0.81/0.59	0.835	0.85/0.97/0.84/0.60	**0.867**	0.88/0.95/0.87/0.76
GEOM + *MS*	**0.911**	0.94/0.99/0.86/0.80	0.909	0.95/0.99/0.84/0.79	**0.911**	0.94/0.99/0.85/0.80	0.882	0.90/0.97/0.89/0.77
GEOM_EC + *MS*	0.899	0.93/0.98/0.86/0.76	0.905	0.95/0.99/0.82/0.79	**0.915**	0.96/0.99/0.86/0.79	0.893	0.92/0.97/0.88/0.80



	conif	decid	snag	dead
conif	221	2	1	35
decid	7	195	0	0
snag	2	2	114	21
dead	24	5	15	101

(a) *GEOM*

	conif	decid	snag	dead
conif	211	1	1	46
decid	3	197	0	2
snag	3	0	128	8
dead	19	5	35	86

(b) *GEOM_EC*

	conif	decid	snag	dead
conif	238	1	4	16
decid	1	201	0	0
snag	0	0	118	21
dead	6	0	17	122

(c) *GEOM+MS*

	conif	decid	snag	dead
conif	251	0	1	7
decid	1	200	0	1
snag	1	0	120	18
dead	12	3	19	111

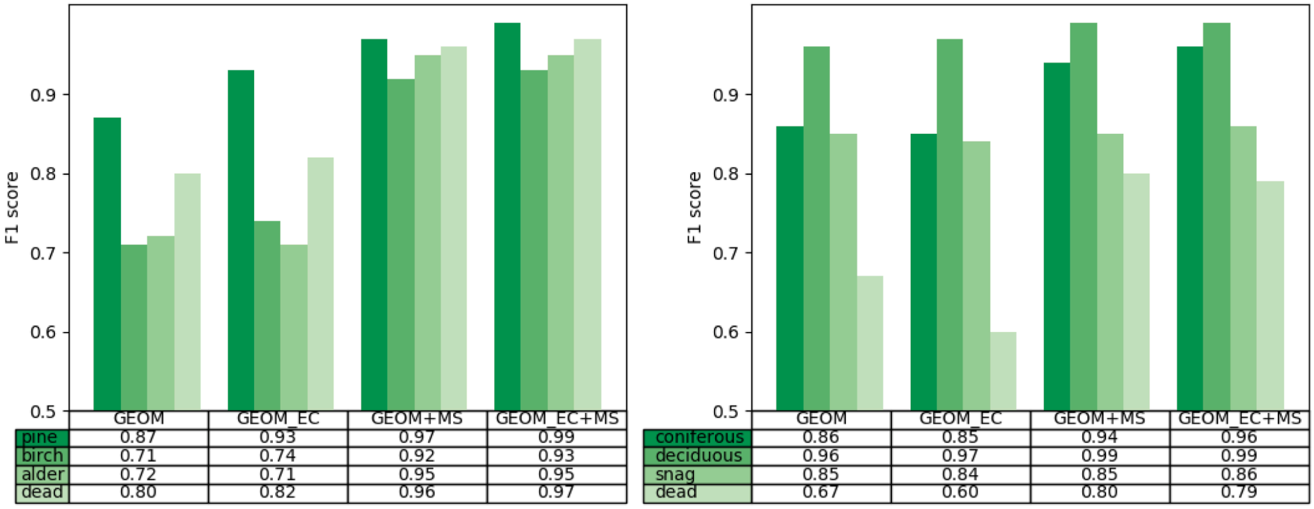
(d) *GEOM_EC+MS*

Fig. 11. Confusion matrices for Silvi-Net$_{majVot}$ (BFNP test data). Subfigures show results for different feature sets.

tree in the dataset benefited most. For both experiments based on side-view imagery generated from 3D point clouds, confusion was biggest between birch and alder (Fig. 10a and 10b). When combining automatically extracted features from *GEOM* images and *MS* images, the classification results clearly increased (OA = 95.1%). By integrating MS information, the F_1 score raised by more than 0.20 for the two deciduous species alder and birch. Consequentially, confusion between these two classes was almost completely resolved (Fig. 10c). Moreover, Silvi-Net$_{majVot}$ performed 13.0% better than PointNet++ based on raw 3D point clouds enriched with the top five hand-crafted MS features (OA = 82.1%). Furthermore, it is noteworthy that for the *GEOM* + *MS* experiment, Silvi-Net$_{collages}$ was equal to Silvi-Net$_{majVot}$. Fusing *GEOM_EC* images and *MS* images yielded the best results. Here, the OA for Silvi-Net$_{majVot}$ reached 96.1%, with F_1 scores ranging between 0.93 (birch) and 0.99 (pine), and a minor remaining confusion between birch and alder (Fig. 10d).

4.3. Results for BFNP

When studying BFNP, general results for the classification of single



	GEOM	GEOM_EC	GEOM+MS	GEOM_EC+MS
pine	0.87	0.93	0.97	0.99
birch	0.71	0.74	0.92	0.93
alder	0.72	0.71	0.95	0.95
dead	0.80	0.82	0.96	0.97

(a)

	GEOM	GEOM_EC	GEOM+MS	GEOM_EC+MS
coniferous	0.86	0.85	0.94	0.96
deciduous	0.96	0.97	0.99	0.99
snag	0.85	0.84	0.85	0.86
dead	0.67	0.60	0.80	0.79

(b)

Fig. 12. F_1 scores per class for Silvi-Net$_{majVot}$ in the ChEZ (a) and in the BFNP (b), using different feature sets.

trees into coniferous, deciduous, snag, and dead tree were partly different from those obtained in the ChEZ. Particularly, in two of four experiments, Silvi-Net$_{majVot}$ was slightly inferior to the baseline PointNet++. However, when fusing side-view images and MS imagery, Silvi-Net$_{majVot}$ was the method of choice (OA = 91.5%; Table 7), exceeding the baseline method (OA = 89.3%) by 2.2%. Furthermore, our experiments show that embedding *GEOM_EC* images helped improve results by 5.9%, in contrast to an OA of 85.6% based on only *MS* images (Table 5). Moreover, the difference between validation data and test data was again in the range of a few percentage points, showing that Silvi-Net generalised well. For a more detailed analysis, we want to draw attention to the experiments examining the impact of single data subsets. Using only geometry data, PointNet++ (OA = 85.7%) performed slightly better than our ResNet-based approach and majority voting (OA = 84.7%). Here, the results generated by Silvi-Net$_{majVot}$ showed a considerable confusion between coniferous and dead trees (Fig. 11a). In detail, the low F_1 score for dead trees (0.67) was mainly due to the fact that 16.6% (24/145) of dead trees were classified as coniferous, and 13.5% (35/259) vice versa. Nevertheless, deciduous trees were classified almost perfectly (F_1 score = 0.96). Moreover, Silvi-Net$_{majVot}$ was clearly superior to the approaches based on single or collated imagery. When utilizing *GEOM_EC* images, we observed that the incorporation of EC was not advantageous for tree classification in this study area, especially since the F_1 score of dead trees dropped by 0.07 (Fig. 11b). All other F_1 scores remained almost unchanged (±0.01). Surprisingly, the baseline method using PointNet++ (OA = 86.7%) benefited from EC by 1.0%. When combining geometry data and masked MS data (*GEOM* + *MS*), Silvi-Net$_{majVot}$ (OA = 91.1%) performed 2.9% better than the baseline method (OA = 88.2%). Here, incorporating *MS* images clearly enhanced the results by 6.4% and especially improved the confusion between coniferous trees and dead trees (Fig. 11c). Note that the classification of snags was not improved by MS data. Using EC (*GEOM_EC* + *MS*) slightly improved Silvi-Net$_{majVot}$ (0.4%), reaching the best result in this study area (OA = 91.5%). However, we observed an unsolved moderate confusion between dead trees and snags (Fig. 11d), with 13.1% (19/145) of dead trees being classified as snags and 12.9% (18/139) vice versa.

5. Discussion

5.1. Main results

Overall, the newly introduced methodology for the classification of single tree species and standing dead trees was successfully applied in two study areas. In general, we achieved an OA of 96.1% using the ChEZ dataset (Fig. 12a) and 91.5% using the BFNP dataset (Fig. 12b). Note that the datasets vary in terms of forest types and sensor models, as well as geometric and spectral resolution. Therefore, the superior results in the ChEZ are mostly due to the fact that both the ground resolution and the number of spectral channels in MS images are much higher. As a result, MS images in this study area contain more extractable information for tree classification. Compared to PointNet++, our approach yielded OA values that were 11.3% (ChEZ) and 2.2% (BFNP) better. Here, the clear lead in the ChEZ was presumably an effect of the relatively small dataset. In this study area, the network parameters of PointNet++ could not be perfectly trained from scratch. In contrast, Silvi-Net was able to deal with a reduced number of samples. Using transfer learning, model parameters were successfully retrained. In summary, the crucial factor for successful performance in our approach was the fusion of lidar data and MS images.

5.2. Detailed results

In total, we pursued three different classification strategies, differing in the way of dealing with the 2D representations of the 3D point clouds. Overall, the conducted experiments revealed that Silvi-Net$_{majVot}$ generally performed better than Silvi-Net$_{single}$, respectively Silvi-Net$_{collages}$. Both Silvi-Net$_{single}$ and Silvi-Net$_{majVot}$ preserved most 3D information contained in the point cloud and information loss was limited to the subsampling process in the dataloader. In contrast, generating collages produced overall poorer image resolution (Silvi-Net$_{collages}$). The final size of a single tree in a collated image was 50% smaller than the tree size in a single view. Nevertheless, CAM overlays of collated *GEOM_EC* images in the ChEZ (Fig. A.1) and in the BFNP (Fig. A.2) demonstrate that Silvi-Net$_{collages}$ still identified most decisive image regions. The advantage of collated images definitely was to process all 12 views of a single tree in one sample and to enable end-to-end classification. Finally, Silvi-Net$_{majVot}$ handled the trade-off between view number and image resolution, and single misclassified samples could be outvoted (see Fig. A.3 and Fig. A.4).

Let us now focus on the different classes. In the ChEZ, classification of pine, birch, alder, and dead tree was already high when using only *MS* images (OA = 93.6%; see Table 5). Here, automatically extracted features from the 10 cm five-channel MS imagery seemed sufficient to classify single trees. Note that the MS-based classification results are still considerably dependent on previously conducted lidar-based tree segmentation. When incorporating geometry information and EC, the gain in OA was 2.5%, resulting in a remarkable OA of 96.1% (Table 6). When only binary side-view images of point clouds were available (*GEOM*

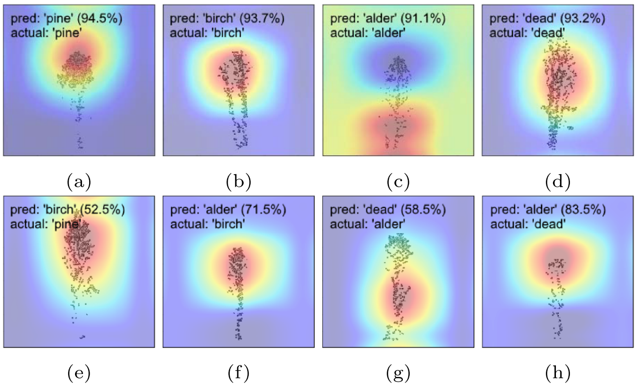


Fig. 13. Examples of correct (a-d) and incorrect (e-h) classification of *GEOM_EC* images in the ChEZ; CAM overlay.

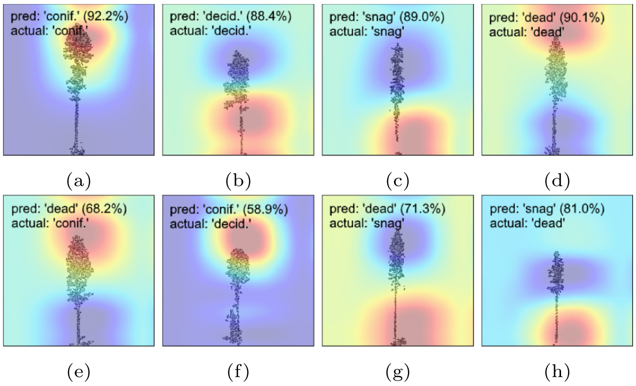


Fig. 14. Examples of correct (a-d) and incorrect (e-h) classification of *GEOM_EC* images in the BFNP; CAM overlay.

images), the OA reached a respectable 77.5%. Moreover, incorporating laser EC improved the results (OA = 80.4%) by 2.9%. Apparently, when using only lidar-based information, pine as the only conifer in the dataset could be classified almost perfectly (F_1 score = 0.93; see Fig. 12a). Note that confusion was biggest between birch and alder without MS imagery (Fig. 10b). However, by integrating MS information, the confusion between these two classes was almost completely resolved (Fig. 10d). Due to their relatively similar shape, these two deciduous species were difficult to differentiate when classification was only based on geometric properties and EC. Finally, it is notable that features automatically extracted from the *MS* images clearly improved the classification of dead trees, increasing the F_1 score from 0.82 to 0.97. Here, the spectral properties included in the infrared channels (NIR, RE) enhanced the separation of dead and living trees.

In the BFNP, the general results for the classification of single trees into coniferous, deciduous, snag, and dead were partly different from those obtained in the ChEZ. Here, when using only lidar-based data (*GEOM_EC*), PointNet++ (OA = 86.7%) performed 3.2% better than Silvi-Net (OA = 83.5%; see Table 7). In this study area, information loss through generating multiple 2D representations of raw 3D point clouds exceeded the advantages of applying pretrained CNNs. The side-view images were still sufficient to classify deciduous trees almost perfectly (F_1 score = 0.97; see Fig. 12b). However, confusion between dead trees and coniferous trees was considerable (Fig. 11b) because most dead trees with crowns are dead coniferous trees. Thus, these two classes do not differ much in their geometric shape and, therefore, could not be separated well. Surprisingly, the incorporation of EC in side-view images negatively affected the Silvi-Net results, whereas the baseline method profited from this information by 1.0%. Specifically, numerous dead trees were classified as snags and, hence, the F_1 score for dead trees dropped by 0.07. When MS information was included in the classification process (*GEOM_EC* + *MS*), Silvi-Net reached an OA of 91.5%, exceeding PointNet++ (OA = 89.3%) by 2.2%. Interestingly, compared to the results based on only *MS* images (OA = 85.6%; see Table 5), the incorporation of lidar-based side-view images improved the OA by 5.9%. Unsurprisingly, MS information reduced the confusion between coniferous and dead trees (Fig. 11d). Presumably, the NIR channel was decisive when differentiating these two classes. However, the impact on the F_1 score for snags was negligible. A plausible reason for that is that snags were insufficiently represented from a bird's eye view. We noticed an unsolved moderate confusion between dead trees and snags induced by the manual labeling process. Since the transition between these two classes representing different stages of a dying tree is fluent, some dead trees were erroneously assigned during visual inspection. Without the subdivision into snag and dead trees with crowns, we assume that our approach would have generated better results for a combined class of dead trees in general. Overall, we want to emphasize that Silvi-Net achieved remarkable results for the classification of coniferous (F_1 score = 0.96) and deciduous trees (F_1 score = 0.99), and is ready for practical use.

5.3. Practical issues

Our experiments clearly demonstrated that masking MS image patches with single tree polygons has a positive impact on network performance. The gain in OA for independent test data was 1.4% in the ChEZ and 2.6% in the BFNP. In particular, the classification of snags and dead trees could be clearly improved in the BFNP. Moreover, CAM images of falsely classified *MS_unmasked* samples revealed that, in some cases, ResNet-18 ignored the crucial tree crowns in the image center, focusing instead on nearby trees from different classes. Note that these misclassifications only occurred in some demanding scenarios with high stand density and, thus, complex tree canopies or even crown overlap (see Fig. 9). Consequently, masking of aerial image patches is even more important in these challenging situations. In summary, we would definitely recommend using masked MS images for classification. Nevertheless, from a practical point of view, we want to point out that an adequate quality of both tree segmentation and data registration is essential for successful lidar-based masking.

We conducted experiments using 2D representations of 3D point clouds and found that embedding EC slightly improved the OA of Silvi-Net (1.0%) in the ChEZ and in the BFNP (0.4%). However, regarding the single tree classes, we did not notice a significant change in results. To visualize the network's decisions, we plotted CAM overlays of exemplary side-view *GEOM_EC* images for both correct classification and misclassification. Generally, ResNet-18 identified tree crowns as the most decisive regions in the ChEZ dataset (Fig. 13), but in some cases (e.g., Fig. 13c), stem information was crucial. Fig. 13g clearly shows that a protruding branch falsely led to the prediction “dead tree”. In the BFNP, Silvi-Net correctly classified 83.5% of the trees, with the CAM images demonstrating that the neural network was attentive to either the crown or stem parts. However, for 123 out of 745 samples (16.5%), the predictions were wrong, such as the coniferous sample in Fig. 14e being classified as a dead tree due to its obvious similarity to one (Fig. 14d). When we look at Fig. 14g and 14h, we can understand the confusion between snags and dead trees, but some incorrect predictions were implausible, such as confusion between coniferous and deciduous samples (e.g., Fig. 14f).

5.4. Evaluation of Silvi-Net

Overall, we can name numerous advantages for our CNN-based approach to tree species classification, but we want to point out that Silvi-Net enables a comfortable fusion of 2D and 3D data captured by different sensor types. We successfully combined information comprising object geometry, laser EC for each 3D point, and reflectance in the visible and NIR spectra. Undeniably, the automatic extraction of meaningful features from previously generated 2D representations is the key factor. The technique of transfer learning using pretrained weights also facilitates fast model convergence, even for relatively small datasets. Despite these clear advantages, we would also like to address the limitations of our approach. When generating multiple side-view images by projecting 3D point clouds, some information is lost, and all images undergo resolution reduction when placed in the dataloader. Compared to PointNet++'s performance with raw 3D point clouds, information loss is considerable but unavoidable. Furthermore, we want to make clear that a well-performing upstream segmentation of single trees is mandatory for Silvi-Net to work well. In our study, we used almost perfectly delineated single trees generated by the normalized cut segmentation algorithm by manually labeling optimal segments, thereby minimizing the effect of undersegmentation or oversegmentation. However, from a practical point of view, many tree segmentation techniques will cause issues in forests with an even higher stand density and more complexity of the canopy.

5.5. *Comparison to related work*

Investigating related work indicates that Silvi-Net achieves promising and competitive results. Yet, it is challenging to provide a comprehensive and fair comparison to other studies that have addressed object-based classification of individual standing dead trees and snags. On the one hand, utilized datasets strongly differ in spatial, spectral, and temporal resolution. On the other hand, the type of study area (urban, natural, managed) and number of samples and classes fluctuate. Using binary classifiers, Krzystek et al. (2020) classified standing dead trees, snags, and living trees in the BFNP. Overall, their approach separated standing dead trees (F_1 score = 0.92; 310 test samples) from living trees (F_1 score = 0.89; 761 test samples) with an OA of 93%. Snags (76 test samples) could be differentiated from living trees (1513 test samples) with an OA of 96%. Interestingly, living trees were classified with an F_1 score of 0.97, whereas the F_1 score for snags was relatively low (0.61). A comparison of multiple classifiers optimized for specific binary tasks to our holistic approach is unfeasible.

To the best of our knowledge, only a few studies have analyzed the combined classification of single tree species and dead wood. On an imbalanced test dataset, Kaminska et al. (2018) reached an OA of 94.3% for the classification of three tree species (spruce, pine, deciduous), each of them further categorized as "dead" or "alive". Instead of F_1 scores, the authors listed producer's accuracy (PA) for the single classes, which is equal to recall. In detail, spruce (146 test samples, PA = 92.4%), pine (148 test samples, PA = 94.1%), deciduous (209 test samples, PA = 99.5%), dead spruce (118 test samples, PA = 90.2%), and dead deciduous (18 test samples, PA = 94%) were classified with high accuracy. However, the PA for dead pine (13 test samples) only reached 69.2%. Recently, Amiri et al. (2019) reported a combined classification of tree species in the BFNP, namely spruce (F_1 score = 0.94), beech (F_1 score = 0.85), fir (F_1 score = 0.59), and dead spruce (F_1 score = 0.74). Based on a huge feature set generated from multi-wavelength ALS data (200 points/m^2), the classifier obtained an OA of 82.1%. In summary, Silvi-Net is clearly better than the RF-based approach presented in Amiri et al. (2019). However, a reasonable comparison to Kaminska et al. (2018) is not possible, because data resolution and classification task differ too much. This denotes an urgent need for objective benchmark forest area datasets comprised of annotated high-resolution lidar data and MS or hyperspectral imagery.

6. Conclusions

In this work, we have presented Silvi-Net, a dual-CNN-based approach for the combined classification of presegmented 3D tree objects with respect to tree species and dead wood in particular. We achieved results superior to those of the baseline method, PointNet++, especially for datasets with a reduced number of samples. Our approach proved to work with data from two natural forests with similar stand density (400–530 trees/ha). Furthermore, lidar data and MS imagery were acquired with different sensor models and, thus, varying geometric and spectral resolution. The trained models showed high generalization capacity on independent test data. The innovative contribution of our study is the fusion of MS image patches and multiple side-view images, rendered from 3D lidar data, in a CNN-based approach. Compared to experiments conducted using only MS images, the fusion of lidar-based side-view images increased the OA by 2.5% in the ChEZ and 5.9% in the BFNP.

We automatically extracted features using two independently trained ResNet-18 networks, and utilized a standard MLP and majority voting for final object classification. Our optimization process is based on the pretrained weights and recursive retraining of CNN model parameters. For practice, we suggest a combination of high-density lidar data and multi-channel high-resolution MS images. Our results proved that lidar data are of special importance for both the tree segmentation and classification. Because snags are insufficiently represented in bird's eye images, their classification benefited most from the lidar data. By contrast, the NIR channels of MS images allow for the enhancement of dead and living tree definitions as well as tree species classification. Because of its positive impact on the network performance, we also recommend masking MS image patches and embedding calibrated laser EC into the classification process.

In future work, the challenge will be to reliably classify ten or more individual tree species and structurally complex forests. This objective can be supported by improved optical sensors providing high-quality lidar point clouds and high-resolution multi-channel images. In addition, off-the-shelf CNNs and transfer learning can be applied to the specific task of tree species classification, even for relatively small datasets. An interesting task for future work would be the application of panoptic segmentation to forest datasets. This fully DL-based method enables combined delineation and classification of single objects, utilizing prominent image-based neural networks, such as Mask R-CNN (He et al., 2017). Consequently, precise and reliable mapping results could contribute to automatic forest inventory, and support monitoring projects investigating the robustness and sustainability of different forest compositions.

CRediT authorship contribution statement

S. Briechle: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **P. Krzystek:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing - review & editing. **G. Vosselman:** Conceptualization, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

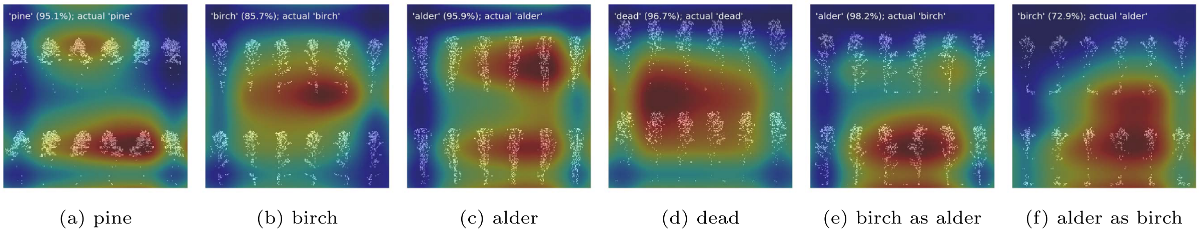
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Prof. Dr. Marco Heurich, Head of the Department of Visitor Management and National Park Monitoring, for providing the remote sensing data in the study area BFNP (lidar data and MS images). We also highly appreciate the support from Dr. Norbert Molitor (Plejades GmbH; Griesheim, Germany) for organizing the field trips in the ChEZ. Moreover, we want to thank our Ukrainian pilots from Flycamstudio, as well as V. Antropov and O. Tretyak from the State Central Enterprise for Radioactive Waste Management, and the collegues from Central Radioactive Waste Management Enterprise for the technical support. This research was funded by Federal Ministry of Education and Research (BMBF), Germany (Grant No. 03FH004IX6).

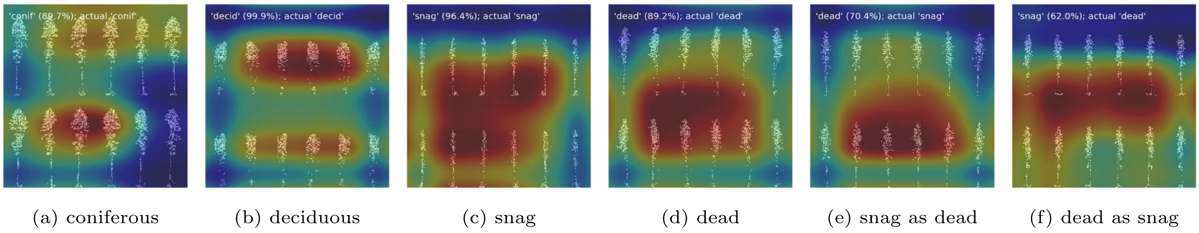
Appendix A

See Figs. A.1, A.2, A.3 and A.4.


'pine' (95.1%); actual 'pine'
'birch' (85.7%); actual 'birch'
'alder' (95.9%); actual 'alder'
'dead' (96.7%); actual 'dead'
'alder' (98.2%); actual 'birch'
'birch' (72.9%); actual 'alder'

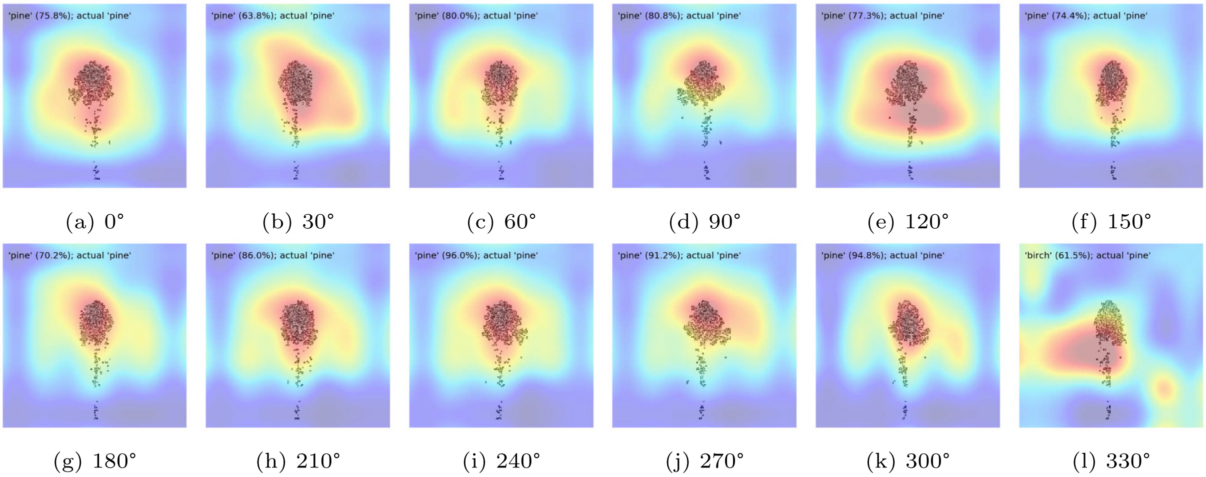
(a) pine (b) birch (c) alder (d) dead (e) birch as alder (f) alder as birch

Fig. A.1. CAM overlays on collated *GEOM_EC* images in the ChEZ; examples for correct classification (a-d) and misclassification (e-f).


'conif' (89.7%); actual 'conif'
'decid' (99.9%); actual 'decid'
'snag' (96.4%); actual 'snag'
'dead' (89.2%); actual 'dead'
'dead' (70.4%); actual 'snag'
'snag' (62.0%); actual 'dead'

(a) coniferous (b) deciduous (c) snag (d) dead (e) snag as dead (f) dead as snag

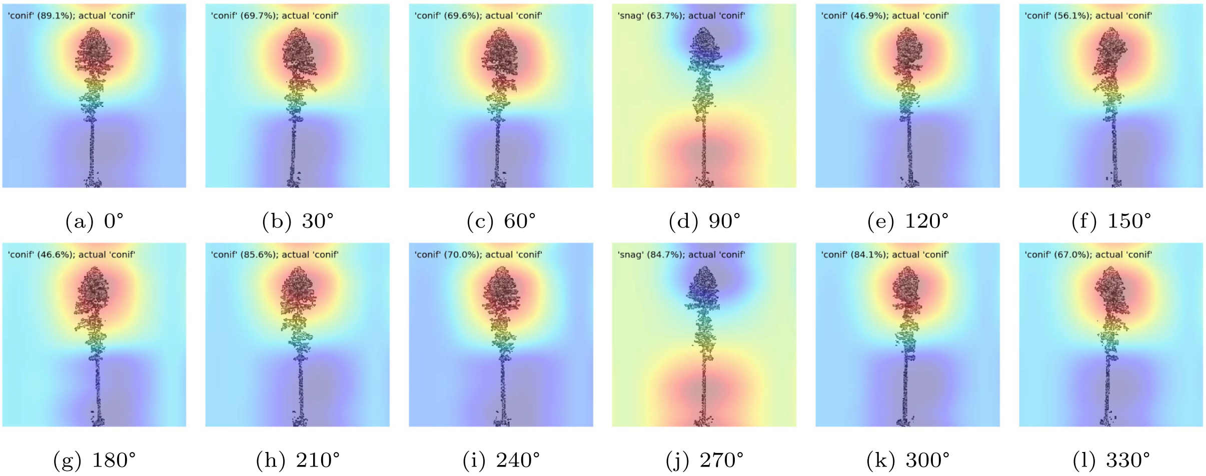
Fig. A.2. CAM overlays on collated *GEOM_EC* images in the BFNP; examples for correct classification (a-d) and misclassification (e-f).


'pine' (75.8%); actual 'pine'
'pine' (63.8%); actual 'pine'
'pine' (80.0%); actual 'pine'
'pine' (80.8%); actual 'pine'
'pine' (77.3%); actual 'pine'
'pine' (74.4%); actual 'pine'
'pine' (70.2%); actual 'pine'
'pine' (86.0%); actual 'pine'
'pine' (96.0%); actual 'pine'
'pine' (91.2%); actual 'pine'
'pine' (94.8%); actual 'pine'
'birch' (61.5%); actual 'pine'

(a) 0° (b) 30° (c) 60° (d) 90° (e) 120° (f) 150°

(g) 180° (h) 210° (i) 240° (j) 270° (k) 300° (l) 330°

Fig. A.3. Majority voting of 11 true (a-k) and 1 false (l) predictions leads to correct final prediction; CAM overlays on *GEOM_EC* images for exemplary pine tree in the ChEZ.


'conif' (89.1%); actual 'conif'
'conif' (69.7%); actual 'conif'
'conif' (69.6%); actual 'conif'
'snag' (63.7%); actual 'conif'
'conif' (46.9%); actual 'conif'
'conif' (56.1%); actual 'conif'
'conif' (46.6%); actual 'conif'
'conif' (85.6%); actual 'conif'
'conif' (70.0%); actual 'conif'
'snag' (84.7%); actual 'conif'
'conif' (84.1%); actual 'conif'
'conif' (67.0%); actual 'conif'

(a) 0° (b) 30° (c) 60° (d) 90° (e) 120° (f) 150°

(g) 180° (h) 210° (i) 240° (j) 270° (k) 300° (l) 330°

Fig. A.4. Majority voting of 10 true (a-c, e-i, k-l) and 2 false (d and j) predictions leads to correct final prediction; CAM overlays on *GEOM_EC* images for exemplary coniferous tree in the BFNP.

References

Agisoft LLC, 2018. Agisoft PhotoScan Professional 1.4.1. https://www.agisoft.com/. Accessed: 2020-09-11.

Amiri, N., Krzystek, P., Heurich, M., Skidmore, A., 2019. Classification of tree species as well as standing dead trees using triple wavelength ALS in a temperate forest. Remote Sensing 11. https://doi.org/10.3390/rs11222614.

Atkinson, P., Tatnall, A., 1997. Introduction neural networks in remote sensing. Int. J. Remote Sens. 18, 699–709. https://doi.org/10.1080/014311697218700.

BayesMap Solutions LLC, 2018BayesStripAlign 2.0. http://bayesmap.com/products/bayesstripalign/. Accessed: 2020-09-11.

Bonzom, J.-M., Hättenschwiler, S., Lecomte-Pradines, C., Chauvet, E., Gaschak, S., Beaugelin-Seiller, K., Della-Vedova, C., Dubourg, N., Maksimenko, A., Garnier-Laplace, J., Adam-Guillermin, C., 2016. Effects of radionuclide contamination on leaf litter decomposition in the Chernobyl Exclusion Zone. Sci. Total Environ. 562, 596–603. https://doi.org/10.1016/j.scitotenv.2016.04.006.

Briechle, S., Krzystek, P., Vosselman, G., 2020a. Classification of tree species and standing dead trees by fusing UAV-based lidar data and multispectral imagery in the 3D deep neural network PointNet++. ISPRS Ann. Photogramm. Remote Sens. Spatial Informat. Sci., V-2-2020, 203–210. doi:10.5194/isprs-annals-V-2-2020-203-2020.

Briechle, S., Molitor, N., Krzystek, P., Vosselman, G., 2020b. Detection of radioactive waste sites in the Chornobyl Exclusion Zone using UAV-based lidar data and multispectral imagery. ISPRS J. Photogramm. Remote Sens. 167, 345–362. https://doi.org/10.1016/j.isprsjprs.2020.06.015.

Cailleret, M., Heurich, M., Bugmann, H., 2014. Reduction in browsing intensity may not compensate climate change effects on tree species composition in the Bavarian Forest National Park. For. Ecol. Manage. 328, 179–192. https://doi.org/10.1016/j.foreco.2014.05.030.

Casas, A., García, M., Siegel, R., Koltunov, A., Ramírez, C., Ustin, S., 2016. Burned forest characterization at single-tree level with airborne laser scanning for assessing wildlife habitat. Remote Sens. Environ. 175, 231–241. https://doi.org/10.1016/j.rse.2015.12.044.

Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M., Delitti, W., Duque, A., Eid, T., Fearnside, P., Goodman, R., Henry, M., Martínez-Yrízar, A., Mugasha, W., Muller-Landau, H., Mencuccini, M., Nelson, B., Ngomanda, A., Nogueira, E., Ortiz-Malavassi, E., Pélissier, R., Ploton, P., Ryan, C., Saldarriaga, J., Vieilledent, G., 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Glob. Change Biol. 20, 3177–3190. https://doi.org/10.1111/gcb.12629.

Dalponte, M., Bruzzone, L., Gianelle, D., 2012. Tree species classification in the southern alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and lidar data. Remote Sens. Environ. 123, 258–270. https://doi.org/10.1016/j.rse.2012.03.013.

Datt, B., 1999. A new reflectance index for remote sensing of chlorophyll content in higher plants: Tests using eucalyptus leaves. J. Plant Physiol. 154, 30–36. https://doi.org/10.1016/S0176-1617(99)80314-9.

Daughtry, C., Walthall, C., Kim, M., De Colstoun, E., McMurtrey III, J., 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sens. Environ. 74, 229–239. https://doi.org/10.1016/S0034-4257(00)00113-9.

Del Frate, F., Pacifici, F., Schiavon, G., Solimini, C., 2007. Use of neural networks for automatic classification from high-resolution images. IEEE Trans. Geosci. Remote Sens. 45, 800–809. https://doi.org/10.1109/TGRS.2007.892009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Fassnacht, F., Latifi, H., Sterenczak, K., Modzelewska, A., Lefsky, M., Waser, L., Straub, C., Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sens. Environ. 186, 64–87. https://doi.org/10.1016/j.rse.2016.08.013.

Gitelson, A., Merzlyak, M.N., 1994. Spectral reflectance changes associated with autumn senescence of Aesculus hippocastanum L. and Acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation. J. Plant Physiol. 143, 286–292. https://doi.org/10.1016/S0176-1617(11)81633-0.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org. Accessed: 2020-09-11.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Stat. Comput. 27, 659–678. https://doi.org/10.1007/s11222-016-9646-1.

Griffiths, D., Boehm, J., 2019. A review on deep learning techniques for 3D sensed data classification. Remote Sensing 11. https://doi.org/10.3390/rs11121499.

Hamraz, H., Jacobs, N., Contreras, M., Clark, C., 2019. Deep learning for conifer/deciduous classification of airborne lidar 3D point clouds representing individual trees. ISPRS J. Photogramm. Remote Sens. 158, 219–230. https://doi.org/10.1016/j.isprsjprs.2019.10.011.

Hartling, S., Sagan, V., Sidike, P., Maimaitijiang, M., Carron, J., 2019. Urban tree species classification using a WorldView-2/3 and lidar data fusion approach and deep learning. Sensors (Switzerland) 19. https://doi.org/10.3390/s19061284.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR, abs/1512.03385. http://arxiv.org/abs/1512.03385. Accessed: 2020-09-11.

Heinzel, J., Koch, B., 2012. Investigating multiple data sources for tree species classification in temperate forest and use for single tree delineation. Int. J. Appl. Earth Obs. Geoinf. 18, 101–110. https://doi.org/10.1016/j.jag.2012.01.025.

Hou, S., Liu, X., Wang, Z., 2017. DualNet: Learn complementary features for image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 502–510. https://doi.org/10.1109/ICCV.2017.62.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing 7, 14680–14707. https://doi.org/10.3390/rs71114680.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Ioannidou, A., Chatzilari, E., Nikolopoulos, S., Kompatsiaris, I., 2017. Deep learning advances in computer vision with 3D data: A survey. ACM Comput. Surv. 50 https://doi.org/10.1145/3042064.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://arxiv.org/pdf/1502.03167.pdf. Accessed: 2020-09-11.

Kaminska, A., Lisiewicz, M., Sterenczak, K., Kraszewski, B., Sadkowski, R., 2018. Species-related single dead tree detection using multi-temporal ALS data and CIR imagery. Remote Sens. Environ. 219, 31–43. https://doi.org/10.1016/j.rse.2018.10.005.

Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://arxiv.org/pdf/1412.6980.pdf. Accessed: 2020-09-11.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Informat. Process. Syst., 2, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed: 2020-09-11.

Krzystek, P., Serebryanyk, A., Schnörr, C., Cervenka, J., Heurich, M., 2020. Large-scale mapping of tree species and dead trees in Šumava National Park and Bavarian Forest National Park using lidar and multispectral imagery. Remote Sensing 12. https://doi.org/10.3390/rs12040661.

Latifi, H., Fassnacht, F., Müller, J., Tharani, A., Dech, S., Heurich, M., 2015. Forest inventories by lidar data: A comparison of single tree segmentation and metric-based methods for inventories of a heterogeneous temperate forest. Int. J. Appl. Earth Obs. Geoinf. 42, 162–174. https://doi.org/10.1016/j.jag.2015.06.008.

Latifi, H., Heurich, M., 2019. Multi-scale remote sensing-assisted forest inventory: A glimpse of the state-of-the-art and future prospects. Remote Sensing 11. https://doi.org/10.3390/rs11111260.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Huang, F., Bottou, L., 2004. Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 97–104. doi:10.1109/CVPR.2004.1315150.

LeNail, A., 2019. NN-SVG: Publication-ready neural network architecture schematics. J. Open Source Softw., 4, 747. http://alexlenail.me/NN-SVG/AlexNet.html. Accessed: 2020-09-11.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. PointCNN: Convolution on X-transformed points. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems vol. 31, Curran Associates, Inc, pp. 820–830, http://papers.nips.cc/paper/7362-pointcnn-convolution-on-x-transformed-points.pdf Accessed: 2020-09-11.

Ma, L., Fu, T., Blaschke, T., Li, M., Tiede, D., Zhou, Z., Ma, X., Chen, D., 2017. Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. ISPRS Int. J. Geo-Informat. 6 https://doi.org/10.3390/ijgi6020051.

McRoberts, R., Tomppo, E., 2007. Remote sensing support for national forest inventories. Remote Sens. Environ. 110, 412–419. https://doi.org/10.1016/j.rse.2006.09.034.

Ng, H.-W., Nguyen, V., Vonikakis, V., Winkler, S., 2015. Deep learning for emotion recognition on small datasets using transfer learning. In: ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction, pp. 443–449. https://doi.org/10.1145/2818346.2830593.

NovAtel Inc., 2017. Inertial Explorer 8.70 - GNSS and inertial post-processing software. https://www.novatel.com/products/software/inertial-explorer/. Accessed: 2020-09-11.

Overbeck, M., Schmidt, M., 2012. Modelling infestation risk of Norway spruce by Ips typographus (L.) in the Lower Saxon Harz Mountains (Germany). For. Ecol. Manage. 266, 115–125. https://doi.org/10.1016/j.foreco.2011.11.011.

Pacifici, F., Chini, M., Emery, W., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113, 1276–1292. https://doi.org/10.1016/j.rse.2009.02.014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc, pp. 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf Accessed: 2020-09-11.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in

Python. J. Machine Learn. Res., 12, 2825–2830. https://arxiv.org/pdf/1201.0490.pdf. Accessed: 2020-09-11.

Polewski, P., Yao, W., Heurich, M., Krzystek, P., Stilla, U., 2015. Active learning approach to detecting standing dead trees from ALS point clouds combined with aerial infrared imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 10–18. doi:10.1109/CVPRW.2015.7301378.

Prabha, R., Tom, M., Rothermel, M., Baltsavias, E., Leal-Taixe, L., Schindler, K., 2020. Lake ice monitoring with webcams and crowd-sourced images. ISPRS Ann. Photogramm. Remote Sens. Spatial Informat. Sci. V-2-2020, 549–556. doi:10.5194/isprs-annals-V-2-2020-549-2020.

Pyysalo, U., Hyyppä, H., 2002. Reconstructing tree crowns from laser scanner data for feature extraction. International Archives of the Photogrammetry. Remote Sensing Spatial Informat. Sci.- ISPRS Archives 34. URL https://pdfs.semanticscholar.org/47b8/6e912bcde3c1c56c1cf71781737e12df90f4.pdf. Accessed: 2020-09-11.

Qi, C., Su, H., Niebner, M., Dai, A., Yan, M., Guibas, L., 2016. Volumetric and multi-view CNNs for object classification on 3D data. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5648–5656. doi:10.1109/CVPR.2016.609.

Qi, C., Yi, L., Su, H., Guibas, L., 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Informat. Process. Syst., 2017, 5100–5109. https://arxiv.org/pdf/1706.02413.pdf. Accessed: 2020-09-11.

Reitberger, J., Schnörr, C., Krzystek, P., Stilla, U., 2009. 3D segmentation of single trees exploiting full waveform lidar data. ISPRS J. Photogramm. Remote Sens. 64, 561–574. https://doi.org/10.1016/j.isprsjprs.2009.04.002.

Rouse Jr, J., Haas, R., Schell, J., Deering, D., 1973. Monitoring vegetation systems in the great plains with ERTS. In: Third ERTS Symposium, NASA, SP-351, pp. 309–317. https://ntrs.nasa.gov/api/citations/19740022614/downloads/19740022614.pdf. Accessed: 2020-09-11.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Shi, Y., Skidmore, A., Wang, T., Holzwarth, S., Heiden, U., Pinnel, N., Zhu, X., Heurich, M., 2018. Tree species classification using plant functional traits from lidar and hyperspectral data. Int. J. Appl. Earth Obs. Geoinf. 73, 207–219. https://doi.org/10.1016/j.jag.2018.06.018.

Shin, H.-C., Roth, H., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35, 1285–1298. https://doi.org/10.1109/TMI.2016.2528162.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR) - Conference Track Proceedings. https://arxiv.org/pdf/1409.1556.pdf. Accessed: 2020-09-11.

Solberg, S., Naesset, E., Bollandsas, O., 2006. Single tree segmentation using airborne laser scanner data in a structurally heterogeneous spruce forest. Photogramm. Eng. Remote Sens. 72, 1369–1378. https://doi.org/10.14358/PERS.72.12.1369.

Sun, Y., Huang, J., Ao, Z., Lao, D., Xin, Q., 2019a. Deep learning approaches for the mapping of tree species diversity in a tropical wetland using airborne lidar and high-spatial-resolution remote sensing images. Forests 10. https://doi.org/10.3390/F10111047.

Sun, Y., Xin, Q., Huang, J., Huang, B., Zhang, H., 2019b. Characterizing tree species of a tropical wetland in southern china at the individual tree level based on convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obser. Remote Sens. 12, 4415–4425. https://doi.org/10.1109/JSTARS.2019.2950721.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Comput. Intell. Neurosci. 2018 https://doi.org/10.1155/2018/7068349.

Wijmans, E.,. 2018. PointNet++ PyTorch. https://github.com/erikwijmans/Pointnet2_PyTorch. Accessed: 2020-09-11.

Wu, B., Yu, B., Wu, Q., Huang, Y., Chen, Z., Wu, J., 2016. Individual tree crown delineation using localized contour tree method and airborne lidar data in coniferous forests. Int. J. Appl. Earth Obs. Geoinf. 52, 82–94. https://doi.org/10.1016/j.jag.2016.06.003.

Yao, W., Krzystek, P., Heurich, M., 2012. Identifying standing dead trees in forest areas based on 3D single tree detection from full waveform lidar data. ISPRS Ann. Photogramm. Remote Sensing Spatial Informat. Sci. 1, 359–364. https://doi.org/10.5194/isprsannals-I-7-359-2012.

Yoschenko, V., Kashparov, V., Melnychuk, M., Levchuk, S., Bondar, Y., Lazarev, M., Yoschenko, M., Farfán, E., Jannik, G., 2011. Chronic irradiation of Scots pine trees (Pinus Sylvestris) in the Chernobyl Exclusion Zone: Dosimetry and radiobiological effects. Health Phys. 101, 393–408. https://doi.org/10.1097/HP.0b013e3182118094.

Zenáhlíková, J., Červenka, J., Čížková, P., Bečka, P., Starỳ, M., Marek, P., Křenová, Z., Svoboda, M., 2015. The Biomonitoring project–monitoring of forest ecosystems in non-intervention areas of the Šumava National Park. Silva Gabreta, 21, 95–104. https://pdfs.semanticscholar.org/8070/5c4e3a6771212622a97c46ba96267246cd55.pdf. Accessed: 2020-09-11.

Zhao, R., Pang, M., Wang, J., 2018. Classifying airborne lidar point clouds via deep features learned by a multi-scale convolutional neural network. Int. J. Geograph. Informat. Sci. 32, 960–979. https://doi.org/10.1080/13658816.2018.1431840.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929. doi:10.1109/CVPR.2016.319.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. https://arxiv.org/pdf/1801.09847.pdf. Accessed: 2020-09-11.

Zhou, Y., Tuzel, O., 2018. VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4490–4499. https://doi.org/10.1109/CVPR.2018.00472.