Defoliation estimation of forest trees from ground-level images

Ursula Kälin[a], Nico Lang[a], Christian Hug[b], Arthur Gessler[b], Jan Dirk Wegner[a]

[a]*EcoVision Lab, ETH Zürich, Switzerland*
[b]*Forest Dynamics, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland*

Abstract

In this paper, we propose to estimate tree defoliation from ground-level RGB photos with convolutional neural networks (CNN). Tree defoliation is usually assessed with field campaigns, where experts estimate multiple tree health indicators per sample site. Campaigns span entire countries to come up with a holistic, nation-wide picture of forest health. Surveys are very laborous, expensive, time-consuming and need a large number of experts. We aim at making the monitoring process more efficient by casting tree defoliation estimation as an image interpretation problem. What makes this task challenging is strong variance in lighting, viewpoint, scale, tree species, and defoliation types. Instead of accounting for each factor separately through explicit modelling, we learn a joint distribution directly from a large set of annotated training images following the end-to-end learning paradigm of deep learning. We evaluate our supervised method on three data sets with different level of difficulty acquired in Swiss forests and compare them to human performance. Results show that tree defoliation estimation from images with CNNs works well and achieves performance very close to human experts.

1. Introduction

Forests are of vital importance to human life and the environment in general. They provide a vast array of ecosystem services, from global to local scales. They store large amounts of carbon, and by continued sequestration they strongly contribute to the terrestrial carbon sink (Pan et al., 2011). They modulate local to global climate via their albedo, surface roughness and their influence on the water cycle (Bonan, 2008). Forests provide timber as construction material and fibers for bioenergy, paper production and the chemical industry (Richardson et al., 2006), and are also biodiversity hotspots (Foley et al., 2005). However, forest functions and their provided services are threatened by various factors such as land use change, air pollutants and climate change (Millar and Stephenson, 2015). Since the 1980s much effort has gone into large-scale forest monitoring in Europe to understand the impact of environmental pollution and a changing climate on tree health (Lorenz, 1995). Defoliation is a main indicator of tree health and is usually assessed by experts that survey tree stands visually in the field at selected sites across the whole country (e.g., a 16×16 km grid in Switzerland Dobbertin and Brang (2001)). A lot of research has been done to automate parts of this tedious, time-consuming, and costly monitoring process. Many works aim

at detecting trees and classifying them into their species (Larsen et al., 2011; Kaartinen et al., 2012). In this regard, the primary data source is remote sensing like multi-spectral aerial (Leckie et al., 2005; Waser et al., 2011) or satellite images (Pu and Landry, 2012), hyper-spectral data (Clark et al., 2005; Roth et al., 2015), dense (full-waveform) LiDAR point clouds (Brandtberg, 2007; Yao et al., 2012b), or a combination of LiDAR and multispectral images (Heikkinen et al., 2011; Korpela et al., 2011; Heinzel and Koch, 2012). Another option, which we follow in our work presented in this paper, is analysis of photos acquired on the ground (Du et al., 2007; Kumar et al., 2012; Mouine et al., 2013; Goëau et al., 2013, 2014; Wegner et al., 2016; Branson et al., 2018). Ground-level images have higher spatial resolution than remote sensing data and provide a horizontal view on the entire tree comparable to the view of the expert assessing the tree visually in the field (Solberg and Strand, 1999). They can be acquired in a more flexible way and do not require expensive, dedicated flight campaigns. This also allows to enthuse citizen scientists for helping with the monitoring effort, for example, through mobile phone apps like *Pl@ntNet* (Goëau et al., 2013, 2014), *Leafsnap* (Kumar et al., 2012), or iNaturalist[1].

A downside of ground-level photos is the absence of an infrared channel, which contributes most evidence for tree health assessment from aerial or satellite imagery (e.g., (Huete et al., 1997)). Here, we investigate whether we can estimate defoliation of forest trees from ground-level images based solely on RGB image texture. We use convolutional neural networks (CNN) as a workhorse, which have seen huge success for a wide range of applications since the comeback of deep learning (Krizhevsky et al., 2012). In contrast to traditional image interpretation that builds on hand-engineered (texture) features, CNNs learn a rich set of discriminative features directly from the data. They have further properties like translation invariance and robustness to (slight) changes of viewpoint and scale that are very useful in our case. We adapt a ResNet (He et al., 2016) CNN architecture to the task and evaluate our method on three different image data sets. All images show forest trees in Switzerland acquired with hand-held photo cameras on the ground. Photos were captured during field surveys and come with expert labels for defoliation. Our results indicate that automated image interpretation with CNNs performs very well for estimating tree defoliation, close to human performance. We view the work presented here as a first step towards an automated tool that can estimate tree health in order to (i) make professional field campaigns less costly, (ii) facilitate quicker collection of much larger data sets, (iii) and enable contributions by citizen scientists via building our system into an app.

2. Related work

Tree stress estimation is usually approached either by expert assessments in situ or by analysis of overhead images that acquire infra-red and near-infrared channels. A good overview of in situ tree monitoring is given in (Eichhorn et al., 2010; Morgenroth and Östberg, 2017) whereas a comprehensive review of forest health assessment methods with remote sensing is given in (Lausch et al., 2016, 2017). In situ tree monitoring in forests

[1] `http://www.inaturalist.org`

often comprises the visual assessment of defoliation and the resulting increase in crown transparency together with estimates of leaf discolouration. In remote sensing, usually multi-spectral imagery, hyperspectral data or airborne laserscanning are used to estimate health indicators at the level of individual trees. For example, Polewski et al. (2015) propose a multi-step approach to detect standing dead trees from color infrared aerial images. They employ a Gaussian mixture model (GMM) that captures dead trees, living trees, and shadows. Tree candidate regions are refined using a level set segmentation method enriched with shape and pixel intensity priors obtained from a set of reference trees. Eitel et al. (2011) use red-edge information of the RapidEye satellite constellation to detect stress of forest trees from space. They compare different indices like the Normalized Difference Vegetation Index (NDVI) and Green NDVI against the Normalized Difference Red-Edge Index (NDRE) and find the latter to perform best. Very high-resolution drone images provide additional means to monitor forest health via spectral indices (Dash et al., 2017). Another widely used method to estimate tree stress is airborne laserscanning in combination with a supervised classification (Yao et al., 2012a; Shendryk et al., 2016). Shendryk et al. (2016) combine full-waveform airborne laser scans with hyperspectral data and train a random forest classifier to estimate dieback and transparency. Mak and Hu (2014) determine the health status of ash trees with ground-based mobile laserscanning as a function of the point density. Lin et al. (2014) apply hyperspectral imagery to the task. They identify if tree leafs change color and humidity depending on their health status and season.

Tree defoliation estimation is a major component of tree stress estimation and especially the laserscanning methods mentioned previously are primarily based on estimating defoliation (Yao et al., 2012a; Lin et al., 2014; Shendryk et al., 2016). A wide variety of research applies remote sensing to defoliation estimation of forest trees (Kantola et al., 2010; Mozgeris and Augustaitis, 2013; Marx and Kleinschmit, 2017; Hawrylo et al., 2018). Kantola et al. (2010) combine airborne laserscanning and aerial images, Mozgeris and Augustaitis (2013) use only aerial images, Marx and Kleinschmit (2017) analyse RapidEye satellite imagery, and Hawrylo et al. (2018) rely on Sentinel-2 satellite images. All works have in common that they propose a traditional supervised classification approach that first extracts a small set of hand-engineered features, usually with spectral indices as their main component, which is then classified with Random Forests, for example.

Estimating tree defoliation from ground-level images has a long history, too, starting with early works like (Lee et al., 1983). Further works like (Mizoue, 2002) and (Dobbertin et al., 2004, 2005) design multi-step image processing workflows that sequentially detect tree regions in the images, extract features, and finally classify often in a semi-supervised way. All methods have in common that they do not apply to the raw image, but need exact delineation of the tree beforehand (Borianne et al., 2017), which is error-prone and often involves user interaction.

In this work, we present a completely automated approach that learns tree defoliation directly from the original images without any previous tree shilhouette extraction. To the best of our knowledge, our work is also the first to propose convolutional neural networks for tree defoliation estimatiom from images. Moreover, most related works test their methods on very small data sets since collection of ground truth stress data is very tedious. For

(a) (b)

Figure 1: Example from the *Urmeter* data set of (a) the original image (675×1013 *pixels*) and (b) image patch (256×256 *pixels*) used as input to the CNN after anisotropic down-scaling.

example, the data set of (Shendryk et al., 2016) consists of only 54 individual trees of a single species (eucalypt) and (Eitel et al., 2011) validate their approach on a single test site with only conifers. In contrast, we train and validate our method on three data sets with over 2000 images of a large variety of different species, shapes, and complex scenarios of different, dense forests. Additionally, we solely rely on standard RGB imagery acquired with hand-held cameras on the ground. We compensate the lack of pixel-wise spectral information or height values with the power of deep CNNs that can learn discriminative texture patterns directly from large amounts of training data. It turns out that CNNs with their very high modelling capacity can learn very complex multi-variate distributions across different tree species, shapes, lighting conditions, varying scale and viewpoint to estimate defoliation with an accuracy similar to human experts.

3. Defoliation estimation

Estimating tree defoliation from ground-level RGB photos is a hard task. Image interpretation has to completely rely on texture (no infrared channel), viewpoint and lighting variations have to be compensated, and it has to work across a wide range of tree species. Traditional methods that rely on hand-engineered features (Lee et al., 1983; Mizoue, 2002;

Dobbertin et al., 2004, 2005; Borianne et al., 2017) have to carefully compensate for all disturbing effects with their model design. In contrast, CNNs learn very high-dimensional multi-variate distributions over species and defoliation directly. Viewpoint changes are compensated via build-in translation invariance. A rich set of learned filters in combination with activation functions (rectified linear units in our case, abbreviated with ReLu) result in discriminative, non-linear texture representations. We therefore choose CNNs for tree defoliation estimation because they learn features and the classification model jointly end-to-end for a specific task and training data set. Circumnavigating manual feature design by learning the most discriminative features directly from the given data is a major reason for success in comparison to more traditional methods.

An important step before training is data pre-processing, which involves cropping or down-sampling, normalization of the radiometric distributions, and data augmentation. In a first step, we transform all images into square patches of size 256×256 *pixels* in order to achieve a single input image size across all samples. Since all our images are larger and of rectangular shape, squeezing them involves anisotropic downsampling (Fig. 1). Relatively small patches save memory and help staying within the memory budget of our GPU[2], but still allow for sufficiently high batch sizes (32 in our case). Higher numbers of images per batch allow for higher learning rates, which in turn accelerates training. In order to ensure sufficient detail in the down-sampled images for training tree defoliation, we ran preliminary tests with higher image resolution of 512×512 *pixels*. However, we did not observe significant performance gains but training time increased significantly. We thus decided to keep 256×256 *pixels* image size for all experiments. A second step for pre-processing is the normalization of the image channels to achieve the same radiometric distribution across all input images. We zero-center (subtraction of the mean) and scale (division by the standard deviation) separately for all three channels. Note that mean and standard deviation are computed only on the training set and applied to both, train and test set. The final pre-processing step is data augmentation, which aims at artificially increasing the (labeled) training data set by applying transformations to the existing images. This simulation of additional training samples often helps generalizing the model because the distribution over the training data set is closer to the true distribution. However, this is only valid if we use transformations for augmentation that would naturally occur in the data (but are under-represented). In our case, we mirror each image along the vertical axis, zoom slightly and add a bit of rotation. This makes the trained model more robust against viewpoint changes, scaling effects, and small tilt variations of either camera or tree.

We train a standard ResNet architecture (He et al., 2016) for regression to estimate tree defoliation[3] for all experiments. ResNet is current state-of-the-art in image classification and in the following we briefly remind the reader of its main technical advantages over alternative architectures like VGG Simonyan and Zisserman (2015), for example. The ResNet architecture manages to generate deeper networks that behave well during the training processes. Usually, the construction of deeper networks results in dramatic increase in number

[2]We run all experiments on a NVIDIA GeForce GTX 1080 Ti GPU with 11 GB memory.

[3]We implement the approach using the deep learning library Keras.

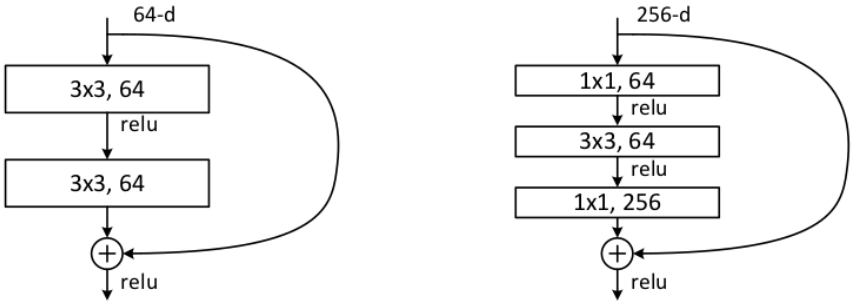


Figure 2: Convolutional ResNet blocks with identity mapping (curved arrows) (He et al., 2016).

of weights as well as exploding or vanishing gradients during the learning process. ResNet reduces this problem significantly with a specific design of convolutional building blocks (Fig. 2), where each block contains two branches. One is propagated through the layers as usual (vertical arrows in Fig. 2), whereas the other branch is used as an identity mapping that is added again at the end of each building block (curved arrow in Fig. 2). Thus, the network learns a residual function, which behaves more stable during the training process. In this work, we use a ResNet50 architecture, which consists of 16 convolutional blocks in total. Our architecture is shown in Tab. 1. Convolutional building blocks always consist of a series of image convolutions with filters of relatively small kernel size (maximum 7×7 *pixels* for *conv1*). Naturally, the filter number per convolutional layer is identical to the number of output activation maps. Each building block can be repeated multiple times in a row (e.g., six times *conv4*). Max pooling spatially down-scales the output of *conv1* by replacing the nine values in the 3×3 kernel with their maximum one. Recall that the average pooling operation before *fc1* works like max pooling, but replaces the maximum value with the average across the kernel. The final, fully-connected layer *fc2* generates a 1-dimensional vector and results in a single output, i.e. the defoliation estimate for the input image patch.

Defoliation is estimated as a continuous value and we thus have to solve a regression task. Two rather simple loss functions are usually applied for regression that either minimize the norm of the difference between prediction and ground truth during training (L1-loss) or the squared norm (L2-loss). Note that L1 bascially translates to minimizing the mean absolute error whereas L2 minimizes the mean squared error. We tried both L1 and L2 but did not find a significant performance gap. We thus opt for the L2-loss for all our experiments, which is the most widely used loss function for regression.

CNNs are trained via backpropagation that updates network weights along high gradients (steepest decrease of the loss) to minimize the descrepancy between predicted defoliation (computed with a forward pass) and the reference label. A quasi-standard today is mini-batch stochastic gradient descent. Its main idea is to randomly sample a small subset from the training data set (32 images per batch in our case) and doing backpropogation repeatedly, alternating with forward passes. A new mini-batch of training images is sampled for each backpropagation. An important hyper-parameter for tuning the learning process is the learning rate. The learning rate defines how much weight we give to the gradients computed

Building blocks	[Kernel size , # Filters]	# Repetitions	# Parameters
input image patch	256×256	-	-
conv1	$\begin{bmatrix} 7 \times 7, & 64 \end{bmatrix}$	1	$9'427$
maxpool	$\begin{bmatrix} 3 \times 3, & - \end{bmatrix}$	1	256
conv2	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix}$	3	$220'032$
conv3	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix}$	4	$1'230'336$
conv4	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix}$	6	$7'129'088$
conv5	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix}$	3	$14'998'528$
averagepool	$\begin{bmatrix} 3 \times 3, & - \end{bmatrix}$	1	-
fc1	-	1	524'544
fc2	-	1	257

Table 1: ResNet architecture (He et al., 2016) as used for all our experiments with $24'112'513$ parameters in total. # Repetitions refers to the number each building block (in square brackets) is executed sequentially (e.g., *conv2* runs three times in a row).

during backpropagation to update the trainable weights. With a reasonably large batch size of 32 images that can be assumed to well represent the distribution of the entire training data set, we can apply higher learning rates resulting in more rapid decrease of the loss. In contrast, smaller batch sizes (e.g., only four images) result in less memory consumption on the GPU. However, this comes at a cost: we cannot trust the subset of images to be a representative sample of the full training data set as much and thus have to set a smaller learning rate. This results in much longer training times. Moreover, a smaller learning rate risks the solver getting stuck in local minima early on during the training process. We thus always aim at fitting a larger batch-size into GPU memory because this allows for smoother optimization and it globally accelerates training.

Often, the learning rate is decayed during the course of training. At the beginning, bigger steps are needed to quickly move towards a favorable solution without getting stuck in local

minima, whereas small steps towards the end prevent the training procedure from fluctuating around minima. Early optimizers use the same learning rate for adjusting all weights and were hard to steer to achieve a stable training process. Newer implementations of optimizers such as Adagrad (Duchi et al., 2011), RMSProp (Hinton et al., 2012), or Adam (Kingma and Ba, 2014) include complexer strategies that can handle the decay and are capable of tuning the learning rate for each learnable parameter individually. We choose to use Adam as optimizer for all experiments because Kingma and Ba (2014) show that it outperforms RMSProp towards the end of the learning process when gradients become sparser.

4. Experiments

Visual inspection of trees is very challenging due to their vast variation in species, shape, height, texture and color. Although species-specific models could potentially predict tree stress more accurately, we lack sufficient training data per species to train a strong model. We thus build a single, species-agnostic model to predict individual tree health from ground-level images. Since deciduous trees lose their leafs in winter, visual health indicators like the dieback, leaf discoloration, and crown transparency are invalid outside the growing season. Therefore, all images were taken between beginning of July and end of August, when also the in situ assessments by the experts are carried out. In the following, we first describe the data sets we use for testing our approach and our evaluation strategy.

4.1. Data and Test Site

We evaluate our method on three different data sets acquired during long-term campaigns of the Swiss Federal Research Institute WSL. All datasets have in common that a trained observer took photos of individual trees during a field survey. Photos are mostly well centered on a single tree and captured with an appropriate zoom level. However, due to dense, complex forest scenarios, the tree of interest can be partially occluded, dense forest may appear in its background, and lighting conditions vary largely. We base our experimental evaluation on data sets *Urmeter*, *Parcours*, and *WSI* that represent different levels of difficulty. A variety of species that are typical for Swiss forests is contained in the data sets (Fig. 3). All three data sets have in common that samples of high defoliation are rare whereas most samples show low defoliation (Fig. 4).

Urmeter consists of 384 images of size 3600×5400 pixels originally recorded as slides from 1987 – 1992 and later scanned. Images contain tree crowns that are standing out of their environment usually with bright sky as background and occluded parts are rare. This set of images was originally constructed as reference image set to teach new observers for field campaigns. Ground-truth defoliation labels were estimated by averaging across image-based assessments of six different experts. Defoliation on the images was assessed according to the ICP Forests manual (Eichhorn et al., 2010). Each photo has a label for the defoliation of the target tree in percent (i.e., the relative leaf or needle loss of the given tree compared to a theoretical reference tree with no defoliation). Note that the expert assessment results in values with 5% increments but due to the averaging of assessments of six persons intermediate values occur. This data set serves two different purposes. First, it

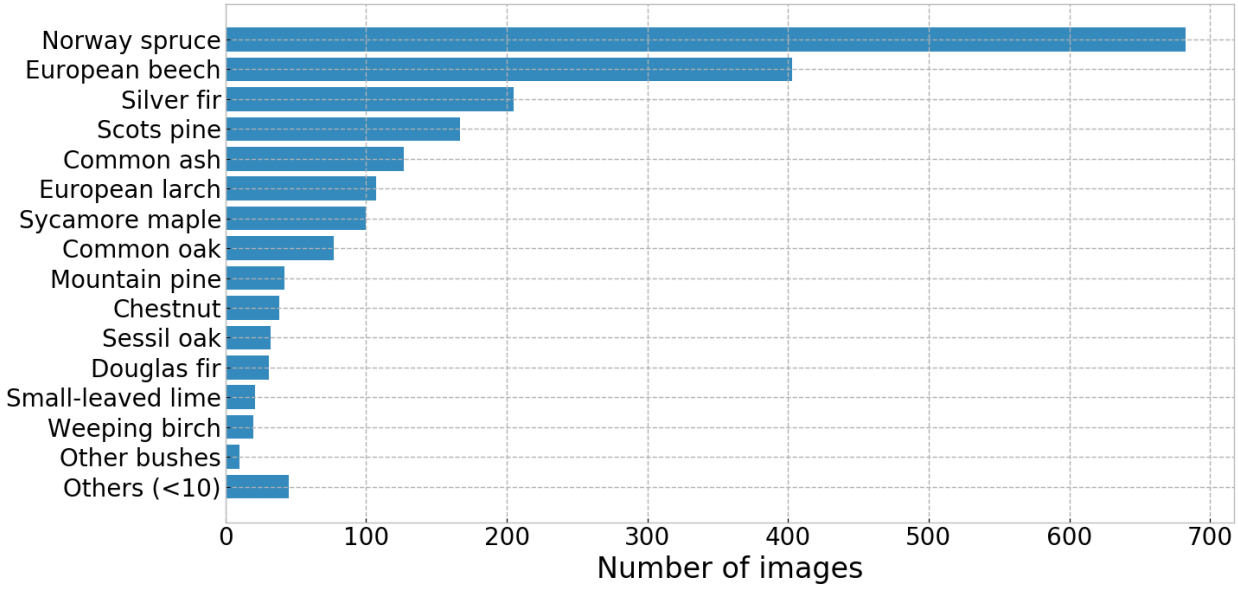


Figure 3: Distribution of tree species over all three data sets *Urmeter*, *Parcours*, and *WSI*.

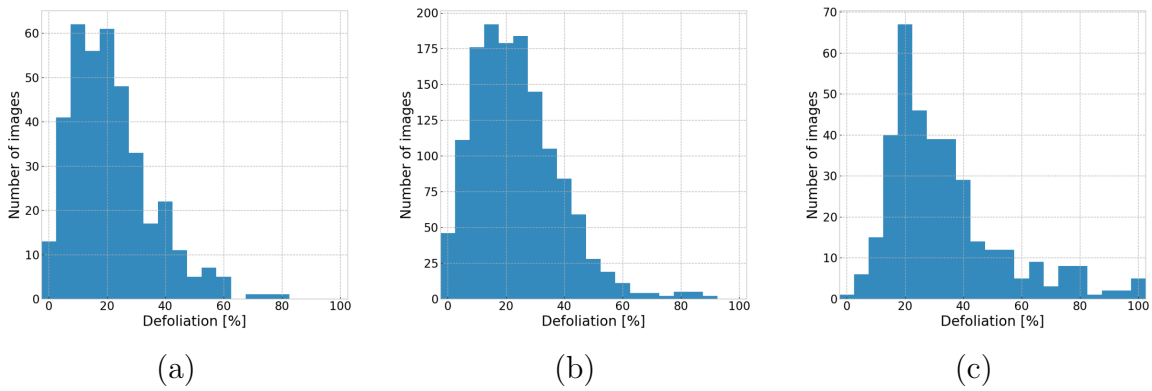


Figure 4: Distribution of samples with respect to defoliation: (a) *Urmeter*, (b) *Parcours*, and (c) *WSI*.

contains images with easily recognizable, individual trees acquired under favorable lighting conditions (Fig. 5(a,b)). Images are relatively easy to interpret and we thus view results with this data set as the best possible or upper bound of what can be expected. Second, the image-based expert assessments serve as a baseline for our automated method. It raises the question of how good a deep learning approach can perform compared to human experts.

Parcours consists of 1361 images of size 900×1370 pixels recorded between 1990 and 1997 originally acquired as slides and later scanned. Images were acquired at ten different locations across Switzerland. Tree defoliation assessments were carried out in the field in contrast to *Urmeter*, where this was done based on the images. Expert assessments were done according to the same protocol as described above. Individual trees are harder to recognize compared to the first data set (*Urmeter*) because they are partially occluded, lighting

(a) (b) (c) (d) (e) (f)

Figure 5: Examples photos of datasets *Urmeter* (a,b), *Parcours* (c,d), and *WSI* (e,f).

conditions are less favorable, and denser forest often covers the background (Fig. 5(c,d)). It is of intermediate difficulty compared to *Urmeter* (easy) and *WSI* (hard).

WSI consists of 363 images of size 3744×5616 pixels and 4159×6239 pixels recorded with digital cameras in 2017 for the purpose of this study. Photos were taken by experts who estimated defoliation values directly in the field with the same protocol as described above. In contrast to both other data sets, this set of images does not consist of a selection of especially typical or well assessable trees. It rather represents realistic conditions in regular forest plots and includes many images where the tree of interest is highly occluded (Fig. 5(e,f)). Additionally, trees are mostly photogaphed with a steeper viewing angle due to less space in front of the tree. Many trees are embedded in dense forest and it is often hard to distinguish individual tree crowns. The radiometric distribution is different (greater variability of brightness and color) with respect to *Urmeter* and *Parcours* because photos were acquired with a modern digital camera instead of scanning analogue slides. Overall, we expect this to be the hardest data set due to steep viewing angle, many occlusions, and dense forest background.

4.2. Evaluation strategy

We perform 5-fold cross-validation for our experimental evaluation in order to avoid any train-test split bias. We ensure a roughly equal distribution of defoliation values across all five folds. Total numbers for train and test split per data set are given in Tab. 2. Figures of results always show average performance across the test splits of all folds and

	Train	Test
Urmeter	307	77
Parcours	1088	273
WSI	290	73

Table 2: Total number of images for train and test sets per data set. Four folds make up the training sample while the fifth is reserved for testing.

plot prediction (vertical axis) versus ground truth values (horizontal axis). Individual results are accumulated into 5% tiles for better visualization of the distribution. We run training per fold for 150 epochs with an initial learning rate of 1^{-4}. Preliminary tests showed that using pre-trained weights (on ImageNet) did neither improve results nor reduce training time nor accelerate convergence. The most likely explanation is that our images differ significantly from standard computer vision images in several ways. Our tree images show mostly high-frequency information (leafs, branches) and the loss function aimed at estimating defoliation is focusing on this detailed evidence. Typical computer vision images are dominated by low-frequency evidence, i.e. few objects cover large portions (> 100 *pixels*) of the image. We thus start training from scratch and initialize network weights randomly.

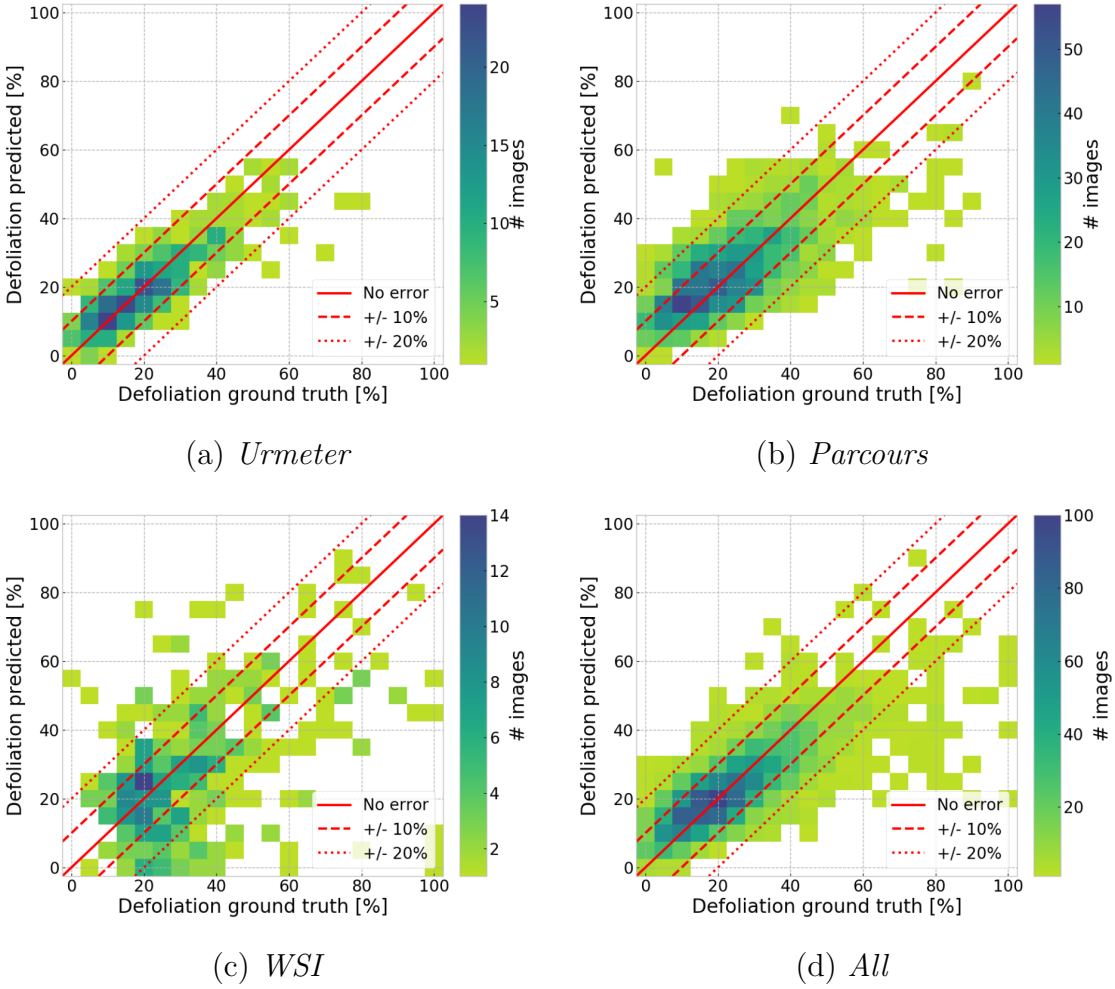
4.3. Tree defoliation estimation results

We present results of our experiments with the three data sets in Tab. 3 and Fig. 6. Predicted defoliation is plotted versus the ground truth defoliation. Dotted red lines indicate error margins of 10% and 20%, while light green (few) to dark blue (many) colors encode the number of images that fall into each cell. Prediction results of *Urmeter* (Fig. 6(a)) and *Parcours* (Fig. 6(b)) show a good correlation to the ground-truth. Only very few samples of *Urmeter* fall outside the 20% error margin and the majority of images is within the 10% margin. The mean absolute error (MAE) of *Urmeter* predictions is 5.5%, which is very small. In all cases of gross errors, the CNN underestimates defoliation which can be explained by lack of training data for trees with high defoliation (cf. Fig. 4(a)). Results

	Separate Training		Joint Training	
	MAE [%]	StdDev [%]	MAE [%]	StdDev [%]
Urmeter	5.5	0.7	5.7	0.2
Parcours	8.3	0.3	7.4	0.4
WSI	14.8	6.9	10.4	0.9
All	7.6	0.3	7.6	0.3

Table 3: Tree stress estimation results after 5-fold cross-validation for (left) training and predicting on all data sets separately (*Separate Training*) and (right) for training across all data jointly but predicting for the individual data sets (*Joint Training*). Mean average errors (MAE) are mean values across the five test splits per experiment with respective standard deviations (StdDev).

for the (harder) *Parcours* data set are slightly worse, but still only few samples fall outside the 20% error margin while most are within 10% deviation from ground truth. The MAE of *Parcours* predictions is 8.3%. Again, there is a tendency to underestimate samples with high defoliation due to too few training samples. In contrast, results for *WSI* (Fig. 6(c)), the hardest data set with the most complex images, are clearly inferior to both other data sets. Although most images fall into the 20% error margin, a trend that would clearly correlate



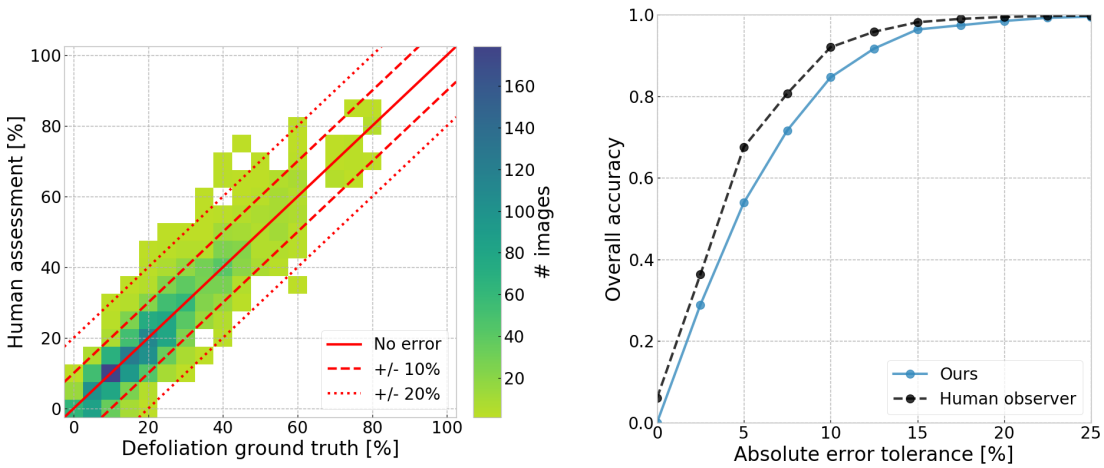
(a) *Urmeter* (b) *Parcours*

(c) *WSI* (d) *All*

Figure 6: Results of 5-fold cross-validation for training and predicting on *Urmeter* (a), *Parcours* (b), *WSI* (c), and all three data sets jointly (d). Red-dotted lines indicate error margins in [%] and light green to dark blue color represents the number of images that fall into an interval (see color bar right of each plot).

predictions with ground truth is much weaker. And once more the model has problems with predictions of high defoliation values, which results in underestimating many samples. It seems that the amount of training data of *WSI* is too low for learning the multi-variate distributions across all complex scenes. This also leads to a higher mean absolute error of 14.8%. To compensate for missing training data we run another experiment where we combine all datasets and train a single model (Fig. 6(d)). Results show that the model can adapt well to the differences across datasets leading to a mean absolute error of 7.6%, which is below the one for *Parcours* despite a remaining tendency to underestimate high defoliation samples.

In order to test whether the larger amount of training data is key to success, we train on the combined data set but predict for each data set separately. It turns out that joint training across all three data sets decreases the mean absolute error from 14.8% to 10.4%



(a) (b)

Figure 7: Comparison of human expert defoliation assessments and our model on the *Urmeter* dataset. (a) Distribution of individual human assessments (vertical axis) with respect to ground truth (horizontal axis, average across all six observers), compare to performance of our CNN model in Fig. 6(a), (b) accuracies of individual human assessments and our CNN model with respect to the error tolerance (dotted red lines in (a)).

for *WSI* and from 8.3% to 7.4% for *Parcours*. The standard deviation across all five folds for *WSI* drops significantly from 6.9% to 0.9%. This finding indicates that the individual data set *WSI* is not big enough to allow the CNN model to capture the highly varying appearance of trees during training. The collection of many more images by WSL in the coming years will most probably improve results. Only *Urmeter* cannot benefit from the joint training data. However, the general systematic error of underestimating samples with high defoliation remains because there are not enough training samples even if combining all three data sets.

We further compare performance of the deep learning approach to human performance. Six different experts judged defoliation of trees in the *Urmeter* dataset by looking at the images. The average defoliation value across those experts serves as ground truth and individual predictions can be plotted like for all previous experiments. Results plotted in Fig. 7(a) do not show a big difference compared to the CNN performance in Fig. 6(a). We further plot overall accuracy versus error tolerance in Fig. 7(b), which indicates that the CNN is not far away from human performance. The mean error of the human experts allows comparing human defoliation estimation based on images to the ones predicted by the CNN. If we allow a 10% error tolerance, our CNN approach achieves only 5 points less on overall accuracy than the human experts.

4.4. Discussion of failure cases

In order to better understand the limitations of the automated CNN approach and to discover promising directions for future improvements, we look at the most dominant failure

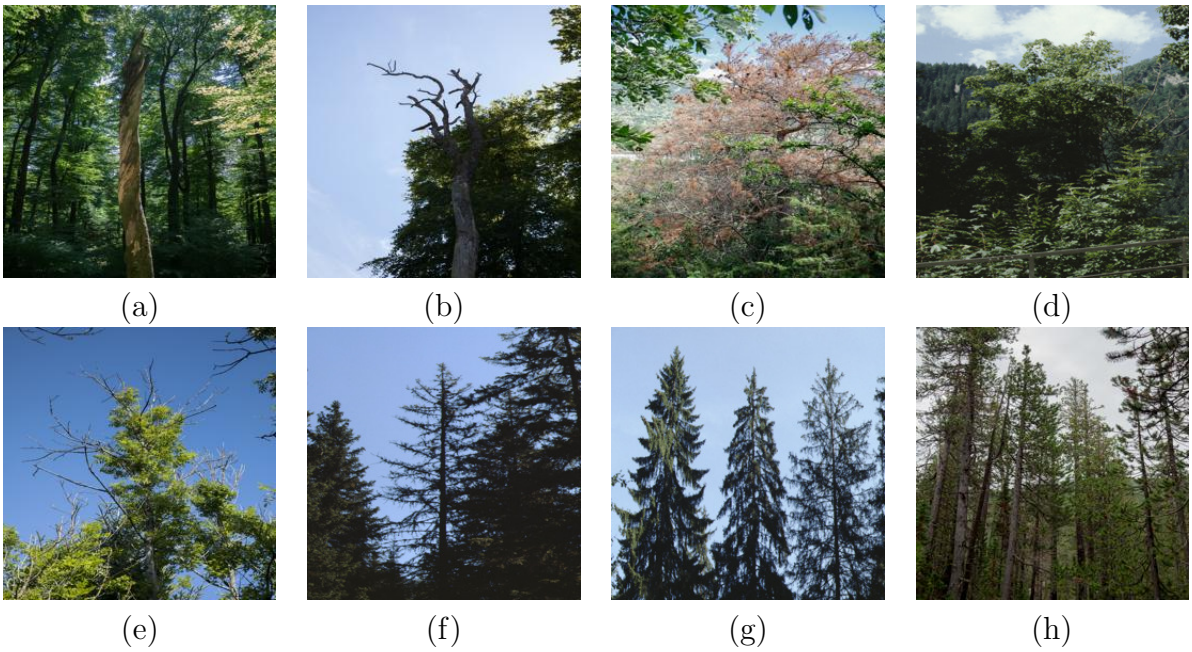


Figure 8: Images with very high prediction errors. The top row shows images where defoliation was predicted too low, whereas the bottom row shows images where defoliation was estimated severely too high (numbers in [%] defoliation, GT = ground truth, PR = predicted): (a) GT: 100, PR: 23.4, (b) GT: 100, PR: 31.0, (c) GT: 100, PR: 45.8, (d) GT: 80, PR: 26.5, (e) GT: 35, PR: 72.9, (f) GT: 30, PR: 65.3, (g) GT: 5, PR: 38.7, (h) GT: 20, PR: 47.6.

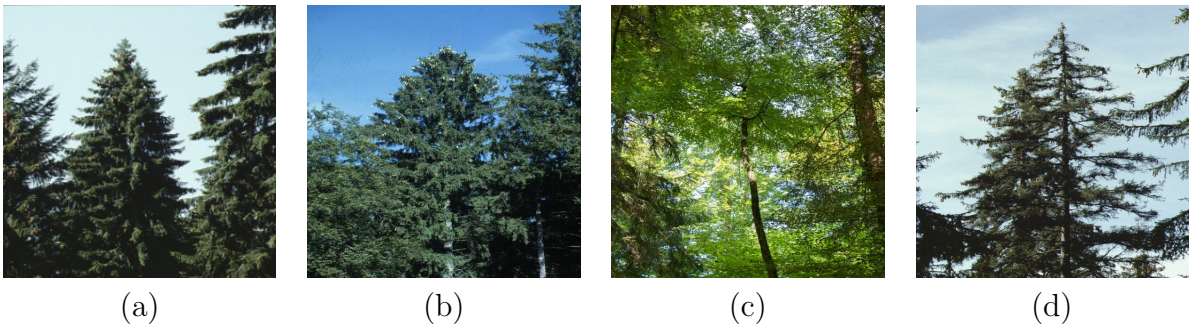


Figure 9: Examples of images where defoliation estimation with our CNN model works very well (prediction errors below 1 percent point): (a) GT: 0, PR: 0, (b) GT: 10, PR: 10.0, (c) GT: 25, PR: 24.5, (d) GT: 55, PR: 55.1.

cases as well as examples of images where our method works well. We show a collection of typical images with very high prediction errors in Fig. 8 and good ones in Fig. 9. The top row in Fig. 8 shows cases where the CNN grossly underestimates defoliation. The first three examples in Fig. 8(a)-(c) show trees without any leafs (100% defoliation). Both trunks in (a, b) are standing in front of dense, green canopy. Although the CNN does predict quite some defoliation (23.4% and 31.0%) in both cases, it still underestimates significantly, most likely because it erroneously assigns background canopy to the foreground trunk. A very hard case is shown in Fig. 8(d), where the defoliated bottom part of the tree crown is mostly

hidden by vegetation in the foreground. While it seems possible to improve predictions for single, completely defoliated trunks (Fig. 8(a-c)) if more training data would be available, occluded cases cannot be resolved by the model. A better strategy would be teaching field crews to take photos from viewpoints that allow observing those parts of the tree that are representative for its overall state.

Gross overestimates of defoliation are shown in the bottom row of Fig. 8. Examples (f)-(h) contain conifers with bright backlighting, interpreted by the CNN model as high defoliation. Cases with a large number of trees and occlusions like in (h), where a human observer cannot tell which tree is the one of interest, are basically unsolvable by the method. Again, teaching field crews to take photos such that the individual tree of interest can be recognized would be a viable option. Example images of trees where the CNN approach achieves very good performance (Fig. 9) indicate that the model can well handle complex cases, too.

5. Conclusions

We have presented a deep machine learning approach for tree defoliation estimation from RGB images. Adaption of a CNN with a ResNet architecture to tree defoliation shows good results for three data sets of different level of difficulty. As with basically all (deep) machine learning methods, the model performs well for all situations with sufficient training data whereas performance suffers in case the number of training samples is very low. Complex scenes where highly defoliated trees are photographed in front of dense, green canopy of very low defoliation are causing failures of the system. This is, on the one hand, due to too few training samples for trees of very high defoliation. On the other hand, telling apart the tree of interest in densely populated scenes poses severe challenges to the method.

However, despite few, extreme cases where the method fails, the overall performance is very promising. Potential future directions could involve combining ground-level photos with aerial or satellite imagery to densely predict defoliation across large regions. Moreover, WSL is continuing to collect photos during their field campaigns, which will lead to several hundred more images each year. We hope that our approach presented in this paper is one step forward towards making monitoring of forest trees more efficient. Our findings can help providing an additional tool for large-scale forest tree monitoring in the long-term. Our approach might allow for assessing many more trees than those few inspected visually by the experts in the field. Taking photos of trees outside the sample plots by the experts can be easily and quickly achieved and multiplies the number of trees with defoliation information in a region. In a next step, crowd sourced images from a citizen science project could be included to increase the spatial coverage of the defoliation assessments.

References

Bonan, G., 2008. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. Science 320, 1444–1449.

Borianne, P., Subsol, G., Caraglio, Y., 2017. Automated efficient computation of crown transparency from tree silhouette images. Computers and Electronics in Agriculture 133, 108–118.

Brandtberg, T., 2007. Classifying individual tree species under leaf-off and leaf-on conditions using airborne lidar. ISPRS Journal of Photogrammetry and Remote Sensing 61, 325–340.

Branson, S., Wegner, J., Hall, D., Lang, N., Schindler, K., Perona, P., 2018. From Google Maps to a Fine-Grained Catalog of Street trees. ISPRS Journal of Photogrammetry and Remote Sensing 135, 13–30.

Clark, M.L., Roberts, D.A., Clark, D.B., 2005. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. Remote Sensing of Environment 96, 375–398.

Dash, J., Watt, M., Pearse, G., Heaphy, M., Dungey, H., 2017. Assessing very high resolution uav imagery for monitoring forest health during a simulated disease outbreak. ISPRS Journal of Photogrammetry and Remote Sensing 131, 1–14.

Dobbertin, M., Brang, P., 2001. Crown defoliation improves tree mortality models. Forest Ecol Manag. 141, 271–284.

Dobbertin, M., Hug, C., Mizoue, N., 2004. Using slides to test for changes in crown defoliation assessment methods. Part I: Visual Assessment of Slides. Environmental Monitoring and Assessment 98, 295–306.

Dobbertin, M., Hug, C., Mizoue, N., 2005. Using slides to test for changes in crown defoliation assessment methods. Part II: Application of the image analysis system CROCO. Environmental Monitoring and Assessment 102, 167–178.

Du, J.X., Wang, X.F., Zhang, G.J., 2007. Leaf shape based plant species recognition. Applied Mathematics and Computation 185, 883–893.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12, 2121–2159.

Eichhorn, J., Roskams, P., Ferretti, M., Mues, V., Szepesi, A., Durrant, D., 2010. Visual assessment of crown condition and damaging agents. manual part iv. Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests 49.

Eitel, J., Vierling, L., Litvak, M., Long, D., Schulthess, U., Ager, A., Krofcheck, D., Stoschek, L., 2011. Broadband, red-edge information from satellites improves early stress detection in a new mexico conifer woodland. Remote Sensing of Environment 115, 3640–3646.

Foley, J., DeFries, R., Asner, G., Barford, C., Bonan, G., Carpenter, S., Chapin, F., Coe, M., Daily, G., Gibbs, H., Helkowski, J., Holloway, T., Howard, E., Kucharik, C., Monfreda, C., Patz, J., Prentice, I., Ramankutty, N., Snyder, P., 2005. Global consequences of land use. Science 309, 570–574.

Goëau, H., Bonnet, P., Joly, A., Affouard, A., Bakić, V., Barbe, J., Dufour, S., Selmi, S., Yahiaoui, I., Vignau, C., Barthélémy, D., Boujemaa, N., 2014. Pl@ntNet mobile 2014: Android port and new features, in: ACM International Conference on Multimedia Retrieval.

Goëau, H., Bonnet, P., Joly, A., Bakić, V., Yahiaoui, J.B.I., Selmi, S., Carré, J., Barthélémy, D., Boujemaa, N., Molino, J.F., Duché, G., Péronnet, A., 2013. Pl@ntNet mobile app, in: ACM International Conference on Multimedia.

Hawrylo, P., Bednarz, B., Wezyk, P., Szostak, M., 2018. Estimating defoliation of Scots pine stands using machine learning methods and vegetation indices of Sentinel-2. European Journal of Remote Sensing 51, 194–205.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heikkinen, V., Korpela, I., Tokola, T., Honkavaara, E., Parkkinen, J., 2011. An SVM Classification of Tree Species Radiometric Signatures Based on the Leica ADS40 Sensor. IEEE Transactions on Geoscience and Remote Sensing 49, 4539–4551.

Heinzel, J., Koch, B., 2012. Investigating multiple data sources for tree species classification in temperate forest and use for single tree delineation. International Journal of Applied Earth Observation and Geoinformation 18, 101–110.

Hinton, G., Srivastava, N., Swersky, K., 2012. Neural Networks for Machine Learning - Lecture 6a: Overview of mini-batch gradient descent. `https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`. [Online; accessed 11-July-2018].

Huete, A., Liu, H., Batchily, K., van Leeuwen, W., 1997. A comparison of vegetation indices over a global set of tm images for eos-modis. Remote Sensing of Environment 59, 440–451.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopainen, M., Heipke, C., Hirschmugl, M., Morsdorf, F., Nsset, E., Pitkänen, J., Popescu, S., Solberg, S., Wolf, B.M., Wu, J.C., 2012. An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning. Remote Sensing 4, 950–974.

Kantola, T., Vastaranta, M., Yu, X., Lyytikainen-Saarenmaa, P., Holopainen, M., Talvitie, M., Kaasalainen, S., Solberg, S., Hyyppa, J., 2010. Classification of defoliated trees using tree-level airborne laser scanning data combined with aerial images. Remote Sensing 2, 2665–2679.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization, in: Proceedings of the 3rd International Conference on Learning Representations.

Korpela, I., Heikkinen, V., Honkavaara, E., Rohrbach, F., Tokola, T., 2011. Variation and directional anisotropy of reflectance at the crown scale – implications for tree species classification in digital aerial images. Remote Sensing of Environment 115, 2062–2074.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks, in: Conference on Neural Information Processing Systems, pp. 1106–1114.

Kumar, N., , Belhumeur, P., Biswas, A., Jacobs, D., Kress, J., Lopez, I., Soares, J., 2012. Leafsnap: A computer vision system for automatic plant species identification, in: European Conference on Computer Vision, pp. 502–516.

Larsen, M., Eriksson, M., Descombes, X., Perrin, G., Brandtberg, T., Gougeon, F.A., 2011. Comparison of six individual tree crown detection algorithms evaluated under varying forest conditions. International Journal of Remote Sensing 32, 5827–5852.

Lausch, A., Erasmi, S., King, D.J., Magdon, P., Heurich, M., 2016. Understanding forest health with remote sensing -part ia review of spectral traits, processes and remote-sensing characteristics. Remote Sensing 8, 1–44.

Lausch, A., Erasmi, S., King, D.J., Magdon, P., Heurich, M., 2017. Understanding forest health with remote sensing-part iia review of approaches and data models. Remote Sensing 9, 1–33.

Leckie, D.G., Gougeon, F.A., Tinis, S., Nelson, T., Burnett, C.N., Paradine, D., 2005. Automated tree recognition in old growth conifer stands with high resolution digital imagery. Remote Sensing of Environment 94, 311–326.

Lee, Y., Alfaro, R., Sickle, G., 1983. Tree-crown defoliation measurement from digitized photographs. Canadian Journal of Forest Research 13, 956–961.

Lin, H., Yan, E., Wang, G., Song, R., 2014. Analysis of Hyperspectral Bands for the Health Diagnosis of Tree Species, in: IEEE Third International Workshop on Earth Observation and Remote Sensing Applications.

Lorenz, M., 1995. International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests-ICP Forests. Water Air Soil Poll. 85, 1221–1226.

Mak, H., Hu, B., 2014. Tree species identification and subsequent health determination from mobile lidar data, in: IEEE Geoscience and Remote Sensing Symposium, pp. 1365–1368.

Marx, A., Kleinschmit, B., 2017. Sensitivity analysis of RapidEye spectral bands and derived vegetation indices for insect defoliation detection in pure Scots pine stands. iForest - Biogeosciences and Forestry 4, 659–668.

Millar, C., Stephenson, N., 2015. Temperate forest health in an era of emerging megadisturbance. Science 349, 823–823.

Mizoue, N., 2002. CROCO: Semi-automatic Image Analysis System for Crown Condition Assessment in Forest Health Monitoring. Journal of Forest Planning 8, 17–24.

Morgenroth, J., Östberg, J., 2017. Measuring and monitoring urban trees and urban forests, in: Ferrini, F., Konijnendijk van den Bosch, C., Fini, A. (Eds.), Routledge Handbook of Urban Forestry. Routledge Handbooks Online, Taylor and Francis Group, Abingdon, New York. chapter 3, pp. 33–48.

Mouine, S., Yahiaoui, I., Verroust-Blondet, A., 2013. Combining Leaf Salient Points and Leaf Contour Descriptions for Plant Species Recognition, in: International Conference on Image Analysis and Recognition, pp. 205–214.

Mozgeris, G., Augustaitis, A., 2013. Estimating crown defoliation of Scots pine (Pinus sylvestris L.) trees using small format digital aerial images. iForest - Biogeosciences and Forestry 6, 15–22.

Pan, Y., Birdsey, R., Fang, J., Houghton, R., Kauppi, P., Kurz, W., Phillips, O., Shvidenko, A., Lewis, S., Canadell, J., 2011. A large and persistent carbon sink in the worlds forests. Science 333, 988–993.

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: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 10–18.

Pu, R., Landry, S., 2012. A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species. Remote Sensing of Environment 124, 516–533.

Richardson, J., Björheden, R., Hakkila, P., Lowe, A., Smith, C., 2006. Bioenergy from sustainable forestry: guiding principles and practice. Springer Science & Business Media 71.

Roth, K.L., Roberts, D.A., Dennison, P.E., Alonzo, M., Peterson, S.H., Bland, M., 2015. Differentiating plant species within and across diverse ecosystems with imaging spectroscopy. Remote Sensing of Environment 167, 135–151.

Shendryk, I., Broich, M., Tulbure, M.G., McGrath, A., Keith, D., Alexandrov, S.V., 2016. Mapping individual tree health using full-waveform airborne laser scans and imaging spectroscopy: A case study for a floodplain eucalypt forest. Remote Sensing of Environment 187, 202–217.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations.

Solberg, S., Strand, L., 1999. Crown density assessments, control surveys and reproducibility. Environ. Monit. Assess. 56, 75–86.

Waser, L., Ginzler, C., Kuechler, M., Baltsavias, E., Hurni, L., 2011. Semi-automatic classification of tree species in different forest ecosystems by spectral and geometric variables derived from Airborne Digital Sensor (ADS40) and RC30 data. Remote Sensing of Environment 115, 76–85.

Wegner, J.D., Branson, S., Hall, D., Schindler, K., Perona, P., 2016. Cataloging Public Objects Using Aerial and Street-Level Images - Urban Trees, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6014–6023.

Yao, W., Krzystek, P., Heurich, M., 2012a. Identifying standing dead trees in forest areas based on 3d single tree detection from full waveform lidar data. ISPRS Annals of the Protogrammetry, Remote Sensing and Spatial Information Sciences I-7, 359–364.

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