Monitoring Vitality of Urban Trees Combining Airborne Imagery and Deep Learning

Master Thesis
Spring Semester 2022
04.07.2022

Luca Gaia
gaial@student.ethz.ch

Dr. Petra D'Odorico (WSL)
Riccardo De Lutio (ETH)
Christian Ginzler (WSL)
Prof. Dr. Jan Dirk Wegner (ETH/UZH)
Prof. Dr. Konrad Schindler (ETH)

EcoVision Lab – ETH-Zurich
Photogrammetry and Remote Sensing Group – ETH-Zurich
Remote Sensing Group – WSL – Birmensdorf

Acknowledgements

I would like to thank some people who made this work possible.

Thanks to Sabine Braun from the IAP for providing the ground observation of the monitored trees in the city of Basel, and to be always available to answer the questions about trees in Basel. Thanks also to Mauro Marty for processing and preparing the aerial images and for the explanations about the processing steps. A particular thank goes to Christian Ginzler and the whole Remote Sensing Group at the WSL for giving me the possibility to spend part of the time at the WSL in Birmensdorf and to be integrated into the group.

A special thanks go to my two supervisors Dr. Petra D'Odorico and Riccardo de Lutio for supervising me with competence and helpfulness.

I would also like to thank my parents and my sisters for always supporting and motivating me during all the work.

Abstract

Urban trees are important because provide important ecosystem services. To measure their vitality defoliation can be used as an indicator. This work investigates the estimation of the defoliation of urban trees at the single tree level in an urban environment using a deep learning approach. Ground observation of the defoliation of the trees and 4-channels aerial images acquired with a manned aircraft are used. The input data are provided for six different years for the City of Basel. The extraction of the area around the single tree is done using a simple approach based on a bounding box of a predefined and fixed tree crown size, usually 10 meters, centred on the tree coordinates. As CNN a ResNet 50 is used, with as input data pairs of label and aerial image of a single tree. Firstly, the best setup for the network is investigated considering only the data from the year 2012. Then following points are investigated: the impact of the temporal mismatch between the ground observations and the aerial image acquisition in single years, the capability of the model to generalize well over time and over new trees when considering observations of more years together, and the impact of the bounding box size for the tree extraction. Considering the results for the best model using only the data for the year 2012 an overall accuracy of 42%, a balanced accuracy of 21% and a MAE of 4.01% were achieved, and it was shown that the model tends to overpredict the defoliation class 20% and that the number of observations pro defoliation class plays an important role for the capability of the model to predict.

Table of contents

1 Introduction

Trees, forests and vegetation, which grow in or very near an urbanized environment, where there is a strong human presence, are defined according to Vogt (2020) as urban forests, typically growing along streets, in parks or gardens (Vogt, 2020).

The presence of trees in urban areas plays an important role for society by providing useful ecosystem services (Velasquez-Camacho et al., 2021). In addition to carbon storage and sequestration, they also help to contain the temperature by reducing heat islands, reduce stormwater runoff, improve air quality by filtering it and in general, have a positive effect on human health (Lüttge and Buckeridge, 2020; Velasquez-Camacho et al., 2021). The positive impact is not only for humans as they can also influence positively the biodiversity (Velasquez-Camacho et al., 2021).

Plants in an urban setting, face different growing conditions compared to plants in a natural environment. The trees along the streets have often limited space for roots development in the underground and for crowns in the aboveground due to man-made constructions and are further exposed to a range of stressors, including construction works, pollution and road salt use in winter (IAPa, 2022; IAPb, 2022; Vogt, 2020).

In addition to these typically urban stress factors, climate change will probably cause intense droughts in the next decades (Dai, 2013).

Therefore, fast and effective urban tree vitality monitoring strategies are now more important than ever.

Tree vitality is defined according to Ognjenović et al. (2022), based on Brang (1998) as "[…] *the ability of a tree to assimilate, to survive stress, to react to changing conditions, and to reproduce*" and different indicators can be used to quantify it, these include radial growth but also crown defoliation and discoloration (Brang, 1998).

To assess the conditions (health and vitality) of forests in Europe, the most used indicator is tree crown defoliation (Gottardini et al., 2020). For example, in the Sanasilva inventory, which has the goal to monitor the forest condition in Switzerland since 1985, the tree crown defoliation is used as a health indicator (Dobbertin et al., 2016).

Also on a European level in the large-scale forest condition monitoring done by the International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (ICP), defoliation is used as an important indicator (ICP, 2022).

The crown defoliation is defined according to the ICP as the loss of leaves (or needles) in a crown when compared to the crown of a tree used as a reference, this value is given as a percentage value in steps of 5% (Eichhorn et al., 2020).

Figure 1 shows an example of different defoliation levels for a Norway Spruce in the frame of the Sanasilva forest health inventory (WSLa, 2022).

Figure 1: Example of different defoliation classes for a Norway Spruce. Low percentage values indicate low defoliation. The images show forest trees, but the principle can be applied also to urban trees (Image source: WSLa (2022)).

A standard method to estimate defoliation is through visual observations (Dobbertin, 2005). An example of this approach is shown in Figure 2. With this kind of observation the defoliation is prone to an intrinsic observer bias, so to try to limit this effect training of the observers should be done (Dobbertin, 2005). Such time series of observations might reach back several decades, representing a valuable historical record.

Figure 2: Example of acquisition of ground observation to monitor the conditions of the trees (for example the defoliation) done by observers in the city of Basel. This is also the same methodology applied to acquire the ground data used in this project. (Image source: (IAPa, 2022).)

However, it is difficult to scale such type of ground assessment over large areas, because of the big effort in terms of monetary and human resources, which is also a disadvantage in urban environments if the trees are not concentrated at the same location but are dispersed over vast areas (Velasquez-Camacho et al., 2021).

In recent years, new technologies have opened the way to new approaches to assess defoliation and health state of the trees relying on the use of digital images (Fraser and Congalton, 2021).

These approaches include the use of images taken from a side view on the ground to estimate the defoliation at the single tree level, which is similar to the view of a ground observer (Kälin et al., 2019), or images taken with a top view using different platforms: unmanned aerial vehicles (UAV) (Fraser and Congalton, 2021; Lehmann et al., 2015; Otsu et al., 2019), manned aircraft (Chiang et al., 2020; Fraser and Congalton, 2021) or satellite (Lottering et al., 2019).
The platforms often determine differences in spatial resolution and in frequency of acquisition. Usually, the areas covered using UAV compared to the manned aircrafts are smaller, but the image resolution is higher (Zhuo et al., 2017), in addition, the costs of the UAV are lower and the flexibility higher (Velasquez-Camacho et al., 2021; Zhuo et al., 2017).
To extract the necessary information to assess defoliation or health state of trees from the input images, the above-mentioned studies use different techniques: for example computing vegetation indices (Otsu et al., 2019), machine learning approaches, using random forest and support vector machine (Fraser and Congalton, 2021) or deep learning, using artificial neural networks and image texture (Lottering et al., 2019), using a Mask RCNN (Chiang et al., 2020) or using CNN (Kälin et al., 2019).
Considering the machine learning approaches, starting from 2015 the importance of deep learning has increased in remote sensing applications (Zhu et al., 2017). Deep learning can be considered a subpart of machine learning and has the ability to map input to labels using a big quantity of data (Alzubaidi et al., 2021), and this approach is based on Neural Networks, with more hidden layers (Zhu et al., 2017). Different types of Neural Networks were developed, and among others also the Convolutional Neural Network (CNN), the most used type (Alzubaidi et al., 2021). CNNs are widely used for computer vision tasks like classification, object detection and segmentation (Patel and Patel, 2020). These types of neural networks have the advantage that no feature selection should be done, but the algorithm can automatically learn the most important features (Alzubaidi et al., 2021).
Given the above-described limitation of the traditional field-based surveys, related to the difficulties to apply them to a large area, and also given the potential of deep learning approaches,
this work tries to deeper investigate the use of aerial images and deep learning methods for the estimation of defoliation.

The goal of this master thesis is

- *to investigate the possibility to estimate the defoliation of trees in an urban environment using airborne imagery and CNN at the single tree level*.

A very similar methodology as in the work by Kälin et al. (2019) is applied, but in this thesis, the focus is more on the use of aerial images to be able to cover a large area without the need to visit each tree individually.

2 Data

The focus of the work is on the region of the City of Basel, in the North-West part of Switzerland. Following data are used: aerial images, ground observations of the state of the trees (defoliation) and coordinates of the trees. There are aerial images and ground observations for six years: 2008, 2009, 2012, 2014, 2018 and 2021.
Figure 3 shows the total project area with the location of the single trees for which there are observations in the year 2012, and additionally two regions more in detail. In the following, the single input data are discussed.

Figure 3: Map of the region analyzed in the work (City of Basel, Switzerland) with the location of the trees (as red points) with ground observations for the year 2012 acquired by the Institut für Angewandte Pflanzenbiologie (IAP). In addition, two areas are shown more in detail. The aerial image shown on the left part of the figure is not used in this project, it is only for visualization purposes. The two detailed areas indeed show the aerial images used for the analysis.

2.1 Aerial Images

The aerial images provided by the Federal Office of Topography (swisstopo) were acquired using manned aircraft in the years 2008, 2009, 2012, 2014, 2018 and 2021, and they have 4 channels: RGB and near-infrared. The resolution of the images changes between 25 and 10 cm and they were taken during routine flights of swisstopo, except for the year 2012 which was a special flight. This explains the higher number of flight lines above the region of Basel for the

year 2012 compared to the other years (see Table 1). The images were not taken at the same times and periods in the different years, in addition, also two types of sensors were used to acquire the images (as a consequence the acquired bands are slightly different (Leica, 2016, 2011)). More details about the image acquisition can be found in the columns on the left part of Table 1.

The raw images provided by swisstopo were not directly used, but they were firstly processed, resulting in 4 channels tif images, with pixel values encoded in uint16. Details about the processing are given in Chapter 3.3. Single strips images (and not an orthomosaic of the complete project area) are used in this work.

Additional information related to the aerial images, which is also provided by swisstopo, are the flight lines. This information shows the track flown during the image acquisition and contains additional information (i.e. acquisition time).

Table 1: Summary of the most important characteristics of the data: aerial images provided by swisstopo and ground observations of the state of the trees provided by the IAP. The values in the brackets refer to all the aerial images which are provided, the values outside the brackets refer only to the images which cover the tree of interest and are the images that are actually used in this work. For the part about the ground observation, the values in the brackets refer to the trees which remain after the processing steps described in Chapter 3.3, the values outside the brackets refer to the trees of interest, which are actually the trees that are used in this work,
The trees of interest are the trees of the year 2012, so for each year, it is worked only with the same trees that can be found also in 2012 after the processing.

	Aerial images					Ground observations	
Year	Data acquisition	Time acquisition	Pixel size [cm]	Number of flight lines	Sensor	Data acquisition	Number of observed trees
2008	12/10	9:36-10:53 (9:36-11:32)	25	3 (7)	Leica ADS80	Mid-July - Mid-August	381 (383)
2009	20/05	10:42-11:32 (10:18-11:51)	25	4 (7)	Leica ADS80	Mid-July - Mid-August	617 (621)
2012	08/08	10:55-12:01 (10:35-12:20)	10	10 (16)	Leica ADS80	Mid-July - Mid-August	624 (624)
2014	07/06	8:11-8:39 (7:54-8:39)	25	4 (6)	Leica ADS80	Mid-July - Mid-August	619 (622)

2018	20/04	12:53-13:59 (12:40-13:59)	10	7 (9)	Leica ADS100	Mid-July - Mid-August	619 (682)
2021	31/05	8:12-9:00 (8:04-9:12)	10	6 (8)	Leica ADS100	Mid-July - Mid-August	612 (681)

2.2 Ground observations

Ground observations were provided by the «*Institut für Angewandte Pflanzenbiologie* (IAP)», an institute that since 1984, among other things, annually collects observations about the conditions of the trees in the city of Basel using ground surveys done between Mid-July and Mid-August (IAPa, 2022; IAPb, 2022; IAPc, 2022).

Observations about the conditions of the trees were provided, in particular the defoliation, as well as for most of the trees also coordinates and other information (i.e. species).

The defoliation acquired by the IAP is a percentage value between 0 and 100 in steps of 5%, with high values corresponding to high defoliation (IAP, 2021).

However, according to some explanations given by the IAP, their observations differ from the observations of the ICP, because they tend to estimate lower values for the defoliation, and they do not have a reference tree, but it is only an optical estimation of the actual state of the tree (IAP d, 2022).

The observed trees lie in different areas of the city of Basel, and starting from 2009 the number of observed species has increased (IAPa, 2022).

The information about the number of observations per year is summarized in Table 1 for the trees resulting after the processing steps and for the trees of interest.

As an additional source of information about the trees also the tree inventory provided by the Stadtgärtnerei of Basel is used. Compared to the tree register from the IAP, this tree register contains more trees, all the trees managed by the city of Basel (BS a, 2022), but without the observation about the defoliation. This dataset was used in some cases to complete the missing coordinates in the IAP dataset. The dataset was lastly actualized on 10/03/2022 (BS b, 2022).

In the following part of the chapter some considerations about the distribution of the number of ground observations in the different defoliation classes and for the different species are done. This is done considering all the ground observations for all the trees in each year which remain after the processing steps described in Chapter 3.3. It is important to note, that these are not the observations which are actually used in the next steps of the work, where it is worked considering only a subpart of these observations: only the observations corresponding to the trees of interest. The trees of interest are for each year the trees that can be found also in the year 2012.

Figure 4 shows the distribution of the number of ground observations for the different defoliation classes when the data of all years are considered together.
It is clear to see that the data are not uniformly distributed over the whole range of possible values, in fact, there is a peak for 20% defoliation. 15% and 25% defoliation have both more than 700 observations, and all the other values have clearly fewer observations. Not for all possible defoliation classes, there are observations, in particular observations for values higher than 45% are rare.
By considering the same type of plots, but for the single years (see supplementary material 7.1), it is possible to see that the peak with the highest number of observations is almost always corresponding to the 20% defoliation class (with the only exception for 2008 where it is 25%). In general, the other two categories with the major number of observations are 15% and 25%. Only for the years 2008 and 2009, this is slightly different because these two secondary categories are 20% and 30% (for 2008), and 25% and 30% (for 2009).

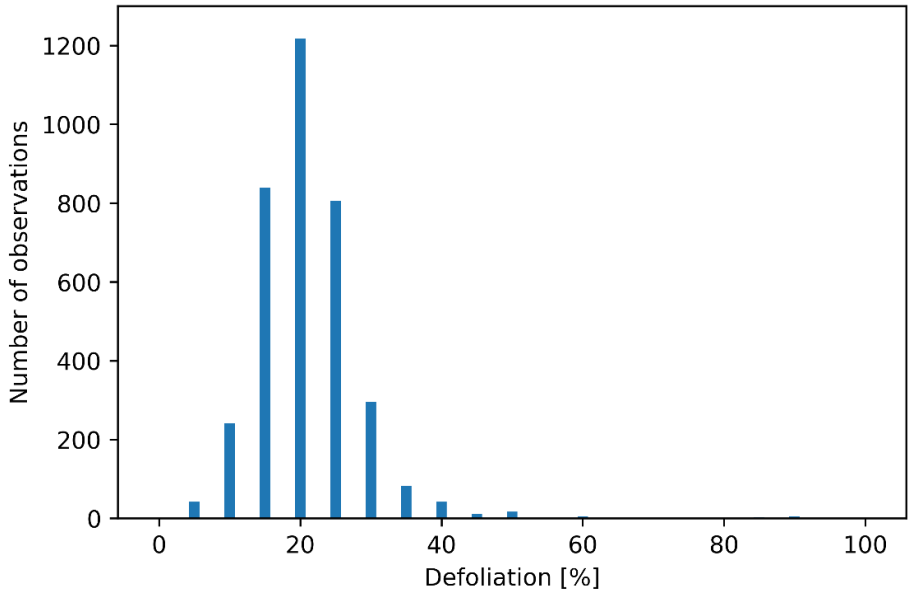


Figure 4: Number of ground observations for the different defoliation classes considering all years together in the region of the city of Basel. The shown observations are the ground observations of the trees which remain after the processing steps described in Chapter 3.3. During the work, only a subpart of these observations is used.

Also the number of observations per specie can be considered, as shown in Figure 5 for all years together. There are species with only a few observations and others with more than 500 observations.

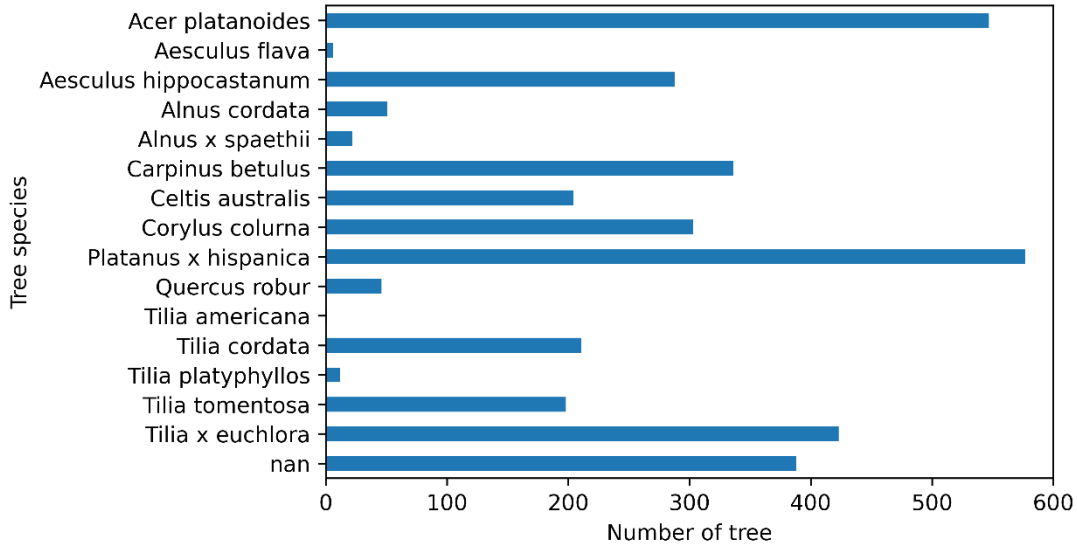


Figure 5: Number of ground observations for the different tree species considering all years together in the region of the city of Basel. Trees for which no information about the specie is known are marked as nan. The shown observations are the ground observations of the trees which remain after the processing steps described in Chapter 3.3. During the work, only a subpart of these observations is used.

3 Methods

3.1 General Overview

Figure 6 shows a flowchart summarizing the general workflow applied in this work. The input data are aerial images and ground observations. The images are provided already processed and the ground observations should firstly be processed. Then the image of only the area of interest around the single trees is extracted. Combining the defoliation value as label as well as the images of the single tree, the dataset is built. The dataset contains data from only one year or data from more years together, according to different investigations done. The dataset is further divided into a train-validation part and a test part. Because a 5-fold cross-validation is applied, the train-validation dataset is further divided into 5 folds. The data are then used as input for a CNN. As architecture a RESNET 50 is used, which solves a regression problem, estimating a defoliation value for each single input image. Different experiments are done using the data from the year 2012 to find a good setup (hyperparameters and structure) for the model. The setup is then applied to do various investigations to better understand the applied approach, considering as input data also data from different years.

The applied workflow is explained more in detail in the next part of the report.

For the computations, Python is used with various libraries. The most important are: pandas, in particular for the data preparation, arcpy, for some GIS operations, scikit-learn, for supporting functions in the machine learning part (Pedregosa et al., 2011), PyTorch, as machine learning framework, and numpy.

All the computations related to deep learning are done using the same computer and working with a GPU of type GeForce GTX TITAN X.

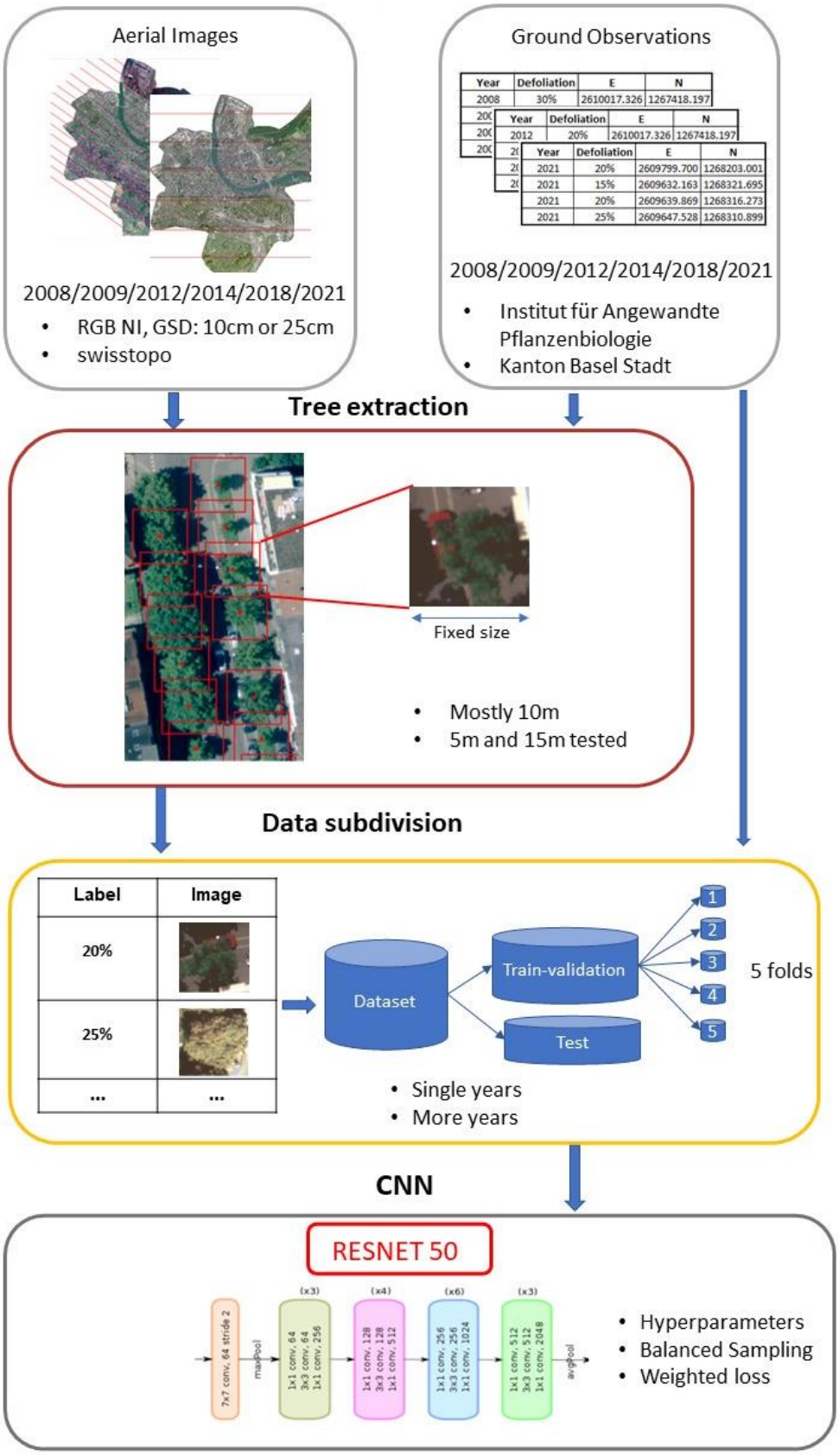


Figure 6: Flowchart summarizing the applied workflow. Starting points are the aerial images and the ground observations. The ground observations should firstly be processed. The input data are used to extract the area of interest around every single tree. The pairs of label-image are used to build a dataset, which is divided into various parts. The dataset is then applied as input to the CNN to do various experiments and investigations. (Images source RESNET schema: (S. Jahromi et al., 2019)

3.2 Preprocessing aerial images

The preprocessing of the aerial images, from the row images provided by swisstopo, was done by a staff member of the remote sensing group at WSL.
The aerial images were clipped to the area of interest and orthorectified. For the aerial images before the year 2017 for every stripe RGB and CIR images are provided, so the next step is to add the near-infrared channel from the CIR to the RGB image.
For the images after the year 2017, they are already provided in 4 channels images. The processing is done for the single stripes for the single years separately.
For the computation of the orthoimages the software SocetGXP 4.3 (BAE SYSTEMS) was used and for the orthorectification the DTM SwissAlti3D for the year 2018 provided by swisstopo. Using a DTM for the processing as the disadvantage that the objects not directly on the surface (like trees) are shown not at the correct location (Valbuena et al., 2008). In addition, also the temporal mismatch between the acquisition of the aerial images and the DTM could decrease the accuracy of the orthorectification.
It is surely possible to generate better orthoimages and the used aerial images are not free of errors or inaccuracies but, for this work, it is assumed that the quality is enough good.
At the beginning of the work, it was also tried to mosaic the single stripes of each year and to work with a yearly single orthomosaic. That did not work, so it was decided to work using single stripes.

3.3 Preprocessing Ground Observations

The raw ground observations provided by the IAP for the 6 years of interest consist of a total of 3740 observations. According to their explanations, the observations refer to the location of the tree rather than to the specific tree. That means, that it is possible that over the years one tree was substituted, but if it was substituted with a tree of the same specie at the same location, the observation was continued with the same identifier number. As a consequence, it is possible that for trees which are considered as same trees over the years, there were actually some changes. However, these possible differences are not considered in this analysis.
Approximately 500 observations do not have coordinates, which is necessary information to be able to use the data. For a part of these trees, no actual identifier number is provided, which can be used to connect the observation with the coordinates. These observations were therefore excluded. For the remaining observations without the coordinates, the actual identifier number is used to extract the coordinates of the tree from the tree register of the City of Basel. This approach works for the majority of these remaining trees and is done under the assumption that the tree in the IAP dataset and in the tree register are the same, and that the coordinates in the tree register can be applied also for observations of previous years. The observations for which this approach does not work are discarded. After these processing steps, there is a total of 3620 observations which are then subdivided into datasets for six single years.
For the six yearly datasets following tests were done to assure the quality of the data: check if there are repetitions of two identifiers numbers, check if all the locations (pairs of x and y coordinates) are unique and check if the defoliation values are multiple of 5. If this is not the case, the repeated observations are manually deleted, and the values of the defoliation are approximated at the nearest multiple of five. After these processing steps, considering all the 6 years together a total of 3613 tree observations are available. The characteristics of these datasets are described in Chapter 2.2.
As ground observations, the value of the defoliation for the single tree as well as their coordinates are used in the next steps of the work.

The above-described dataset is the resulting dataset after the processing steps. It is important to note that it is not worked with all the 3613 observations. In fact, only for the year 2012 all the observations are used. For all the other years only the observations for the trees which can be found also in the dataset of 2012 are used. This results in a smaller dataset with a total of 3472 observations. These trees, which are really used in the computations, are defined as trees of interest.

3.4 Connection between tree observation and most nadir image

Each single tree observation should be connected to the aerial image, which will be used in the next steps of the work, to extract only the pixels of the area of interest around the single trees. Because it is worked using single stripes, for the same tree location, there are potentially multiple images of the same tree of interest. So only the most nadir image acquired for each tree should be used. The connection between single tree and most nadir image is done by considering the coordinates of the single trees and the line of flight of the plane during the image acquisition which is known for each flight.
The analysis is done using the package arcpy, for each tree the distance to each line of flight is computed, and the line of flight corresponding to the smallest distance is taken to determine the most nadir acquisition, which is then associated with the observation.

3.5 Tree extraction

As a next step, pixels corresponding to the area of interest around each tree should be extracted. This area is defined as a bounding-box, which delimits for each tree the region corresponding to the tree crown, to be cropped from the input aerial images. A simple approach is used: centred on each coordinates, which are known, a square of a predefined size, which is assumed as a standard "crown-size" is created. This results in a shapefile with the bounding boxes of a fixed size centred on the coordinates of all the trees of interest. For each tree observation, considering the most nadir image and the defined bounding boxes, the areas of interest are cropped from the total image.
For each image, it is known to which tree it corresponds, and the defoliation value related to it. So, an image-defoliation value pair can be defined for each ground observation in each year and is used in the next steps of the work. It was noticed that for some images a part of the pixels on the boundaries was filled with nan values. It is supposed that this is caused by the used cropping function. To avoid possible problems the images are cropped by eliminating the most external pixels along each boundary.
Because the pre-defined size is fixed and the trees do not have all the same dimension, for a part of the trees the bounding-box will be too big or too small with respect to the tree crown. In addition, when the coordinates of the trees do not exactly match with the center of the crown in the aerial image, the bounding box will not cover perfectly the tree crown.
As an alternative to this approach also a methodology based on a CNN was shortly tested but discarded because of lack of time. More information can be found in Supplementary Material 7.2.
In the work by Kälin et al. (2019) difficulties in the estimation when there are too many trees on the same image with occlusions are shown. It is assumed that for this work it would be ideal to have a bounding box which would perfectly contain the tree crown.

For most of the experiments, a fixed-size of 10 meters is used. The explanation for this choice, which however can be considered as arbitrary, can be found in Supplementary Material 7.3. Also a size of 5 meters and 15 meters were tested (see Chapters 4.5 and 5.6).

3.6 Defoliation estimation using CNN

A CNN approach is used to solve a regression problem to estimate a continuous value for the defoliation of a single tree starting from an input image. A very similar approach as done in the work by Kälin et al. (2019) is applied.
To train, validate and test the model, images of single trees, prepared as described in the previous chapters, and the corresponding labels of the defoliation value are used.
This chapter has the aim to give a general overview of the approaches that can be applied.
Depending on the different experiments done to try to find a good setup for the model, different hyperparameters and structures for the model are applied. Depending on the investigations done to try to get more insights about the approach, different input data differ.
The details about the single approaches applied in the single cases will be discussed in the next sections of the work.
The data (images and labels) are divided into two parts, a train and validation part and a test part. If not differently specified, this is randomly done by taking 80% of the input data as train and validation dataset and considering the remaining 20% as test dataset.

3.6.1 Train and Validation

As network architecture, a ResNet-50 is used. This network was proposed for the first time by He et al., (2015) and has the particularity to use basic residual blocks to develop ultra-deep networks, avoiding the possible problems caused by the vanishing gradient (Alzubaidi et al., 2021). It was decided to start to work with this architecture in particular because it is the same that was used by Kälin et al. (2019). In addition, this architecture should already give a first general idea of the accuracies which can be reached using this approach.
The ResNet-50 implemented in PyTorch expects input images as 3 channels RGB (PyTorch a, 2022). The implementation of PyTorch is used with only some slight modifications, to be able to solve a regression problem with input images with 4 channels: The first convolutional layer is modified so that it can accept 4 channels input images, and the last layer, which is a fully connected layer, is also modified so that it outputs only one single value.
The pretrained network, trained with 3-channels ImageNet images provided by PyTorch is used (PyTorch a, 2022). However, the network cannot be considered completely pre-trained, because for the first convolutional layer and the last fully connected layer the pre-trained weights are not used. These layers are modified and the weights and bias are reinitialized according to the implemented methodology of PyTorch, (PyTorch b, 2022; PyTorch c, 2022).
The input data for the network are labels, as values between 0 and 100 in multiples of 5, and images of the single trees with four channels. To have the input images similar to the ImageNet images, used to pretrain the PyTorch model, the images are resized to 224x224 pixels (PyTorch a, 2022), but the pixel values are not transformed in the range 0 to 1 (as it should be done

according to the official documentation (PyTorch a, 2022)), but they are maintained in the original range of the uint16 datatype.

After resizing, the images are normalized considering the mean and the standard deviation of the pixel values on the four channels over the complete train-validation dataset. All pixels for every single channel are extracted and the mean and the standard deviation are computed for every channel considering all the pixel values of all the images together.

To increase the diversity of the input data, data augmentation is performed after normalizing and resizing the images. If it is not differently specified, as data augmentation only a small rotation, as proposed by Kälin et al. 2019 is applied. As rotation a value in the range minus 15° to plus 15° is applied with a probability of 50%.

To reduce the risk of overfitting, a 5-fold cross-validation is applied (Berrar, 2019).

The data are divided into the single folds using a stratified k-fold based on the defoliation value. For each run, 5 different models are computed (one for each different split). It means that the hyperparameters and the structure of the model are the same, but because for each split the network is reinitialized and retrained, after the complete training process there are 5 resulting models with different weights and biases.

As a loss function, the mean squared error (*MSE Loss)* from PyTorch is applied, and as an optimizer, Adam is used (Kingma and Ba, 2017).

To train the model, the code is divided into a training loop, to actualize the model parameters, and into a validation loop, to track the learning process on an independent dataset. For each split, the training part and the validation part are applied for the defined number of epochs. At the beginning of each split, the model parameters are reinitialized as described above. However, because of the randomness in the initialization of the first convolutional layer and of the last fully connected layer, the initial weights of these layers are not the same among all the five models.

For each split, at the end of each epoch following values are computed: loss training value, loss validation value, overall accuracies for the validation, balanced accuracy for the validation and mean absolute error (MAE) for the validation.

To compute the balanced accuracy (scikit b, 2022; scikit d, 2022), the overall accuracy (scikit a, 2022), and also the MAE (scikit c, 2022) the pre-built functions of scikti-learn are used. The balanced accuracy is defined as the sum of the sensitivity computed on each class, divided by the number of classes (scikit a, 2022).

At each epoch, the parameters of the model are actualized, and the validation loss is used as the criterion to find the best parameters for the model for each split. At the end of each epoch, the value of this loss is compared with the current best value, and if the validation loss is smaller, then the current parameters of the model are considered the current best parameters for this split (and the actual model is considered as the best model for the split). After iterating during all the desired number of epochs, the five above-described metrics, and also the parameters are stored for the best model.

Since model outputs are continuous values, but the true labels are only in multiples of five, before computing the two accuracies, the outputs are rounded to the nearest multiple of five. At this stage it is also checked if there are outputs smaller than 0 or bigger than 100, If this is the case they are then rounded to 0 respectively to 100. For the MAE no rounding steps are done, and the outputs values are directly used.
At the end of the training and validation steps, a mean value and the standard deviation for the four validation metrics are computed, by considering the five values for each metric which are obtained for each fold.
In addition, for some experiments also a learning rate scheduler is applied. A Step LR is used, which after a defined number of epochs, it decreases the learning rate of a predefined value (PyTorch d, 2022).
To ensure the reproducibility of the runs and to easier compare the results, it is worked by fixing random seeds.
The above-described process can be defined as a standard implementation. However, to try to deal also with unbalanced data, two other possible structures of the model are implemented: balanced sampling and weighted loss.
In machine learning, in the case of unbalanced training data, there is a high risk of over-classify the class with a higher number of samples and more often misclassify the class with fewer samples (Johnson and Khoshgoftaar, 2019).

One possibility to deal with this type of data is to up-sample the classes that are under-represented (Alzubaidi et al., 2021) implementing a so-called balanced sampling. An alternative is to work with a weighted loss (Shrivastava, 2020).
For both approaches, a weight for each defoliation class is computed. For each class, the number of samples in the complete train and validation dataset, before the separation in the single folds, is considered. The weight is then computed as the reciprocal of the number of samples for each class (using equation 1) or as the reciprocal of the root square of the number of samples (using equation 2) (Shrivastava, 2020).

$$w_i = \frac{1}{n_i} \qquad \text{(Eq. 1)}$$

$$w_i = \frac{1}{\sqrt{n_i}} \qquad \text{(Eq. 2)}$$

$$with\ n_i\ number\ of\ samples\ in\ the\ defoliation\ classi\ i$$

Balanced Sampling

If for the other approaches all the samples in the training folds are used to train the model, with the balanced sampling this is not the case. The number of samples remains the same, but not all

the samples are used. The samples which come from underrepresented classes, because of the higher weights, are taken with a higher probability. In contrast, the samples from over-represented classes, because of the lower weights, are taken with a lower probability. As a consequence, for samples from underrepresented classes, the same sample is taken multiple times, in contrast for the overrepresented classes, not all the samples are taken for the training. However, in the validation part, balanced sampling is not used for the selection of the samples.

Weighted loss

Using the weighted loss, the loss is weighted differently if the sample comes from an under-represented or from an over-represented class (Shrivastava, 2020)

With this option, when the loss function is computed, a single value for each sample in the batch is computed, and not the mean value among all samples in the batch, as this is the case for the normal approach. These loss function values are multiplied by the weights of each sample. Finally, the mean value of the single weighted loss values inside the batch is computed and used for the backpropagation step. It is important to note that in the case when the option weighted loss is used, the same process is applied also to the validation part of the algorithm, and the values of the metrics and of the parameters corresponding to the best model inside each split are chosen considering a validation weighted loss function.

3.6.2 Test

Additionally, to the train and validation part, also a test part is implemented using the test dataset. On this dataset, no data augmentation is done. The input images are resized to 224x224 pixels and then they are normalized using the mean and the standard deviation which were computed on the train and validation dataset.

The five models with the parameters which were determined during the train validation (one for each split) are used to predict the defoliation value on the test dataset. This result in five predicted values for each single input image. The five predictions are then combined to have only one resulting output value. This is done by weighting equally the five single models, which results in simply computing the mean value of the five single predictions (Brownlee, 2021).

Using the single outputs and the true labels of the defoliation values, the performance of the setup of the model is assessed, by computing the overall accuracy, balanced accuracy and the MAE, in the same way as for the train and validation part.

3.6.3 Confusion Matrix

For the visualization of the results, a confusion matrix (as for example shown in Figure 8) is computed, using the rounded values in multiple of 5 of the predictions. On the first row and on the last column summarizing values are computed: the sum of entries in the rows and in the columns, as well as the sensitivity and the precision. The following formulas are used based on (Gupta, 2022):

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positve}$$

$$sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

4 Experiments and investigations

Using the above-described neural network, various experiments are done with the aim to find good hyperparameters (i.e. batch size, learning rate) and configuration (i.e. weighted loss, balanced sampling, learning rate schedular) for the model,

The experiments are done by changing the hyperparameters, the model configuration or the input data, and applying the CNN on the desired train-validation dataset. This results in 5 trained models (one for each split) and in the three metrics computed during the validation.

Various investigations are done by applying and training the resulting model setup (hyperparameters and structure) on different input data, to try to answer specific research questions.

In this chapter the experiments are described and in the following chapter the results are presented and discussed.

4.1 Model considering only the data of a single year (2012)

The goal of this investigation is to find a good setup (hyperparameters and structure) for the model considering the data only in one single year. As year 2012 is chosen, because in this year the data of the image acquisition is closest to the ground assessments, and it was supposed that this minimal timely difference would provide the best results (Fraser and Congalton, 2021). Additionally, the aerial images from 2012 have a ground sampling distance (GSD) of 10cm, which is also assumed to provide better results than a GSD of 25cm (Fraser and Congalton, 2021). To extract the area of interest around each tree, a dimension of 10 meters for the crown diameter is chosen. After removing the pixels on the boundaries, the resulting input images have a size of 9.80x9.80 meters.

All the experiments are done on the train-validation dataset. The resulting setup, found with these investigations, will be then applied to other data.

To check if a trained CNN would be better than always predicting the defoliation class with the higher number of ground observations, a *baseline model* is used. This pseudo model predicts for all the inputs always a value of 20, which for 2012 is the most present label in the complete dataset, and in the train-validation dataset. The metrics used to evaluate the performance are computed like for all the *normal models*.

All the experiments, except one, are done using 50 epochs. In Table 2 the experiments done are listed, with the different options used. In following the different experiments, grouped by similar goals are shortly discussed.

Table 2: Table showing the different setups (hyperparameters and structure) for the model applied in the experiments done considering only the data from 2012 and using a bounding box size of 10 meters around the trees.The experiments are grouped by similar objectives.

Experiment Number	**Epoch**	**Batch size**	**Learning rate**	**Stratified k-fold**	**Weighted loss**	**Balanced sampling**	**Weight**	**Learning rate scheduler (step_size,gamma)**	**Additional data augmentation**	**Special**
2	100	128	0.001	Yes	No	No	-	No	No	
3	50	64	0.001	Yes	No	No	-	No	No	
4	50	32	0.001	Yes	No	No	-	No	No	
5	50	16	0.001	Yes	No	No	-	No	No	
6	50	64	0.01	Yes	No	No	-	No	No	
7	50	64	0.0001	Yes	No	No	-	No	No	
8	50	64	0.001	Yes	Yes	No	$\frac{1}{\sqrt{n}}$	No	No	
9	50	64	0.001	Yes	Yes	No	$\frac{1}{n}$	No	No	
10	50	64	0.001	Yes	No	Yes	$\frac{1}{\sqrt{n}}$	No	No	
11	50	64	0.001	Yes	No	Yes	$\frac{1}{n}$	No	No	
12	50	64	0.001	Yes	No	No	-	Yes (25,0.1)	No	
13	50	64	0.001	Yes	No	Yes	$\frac{1}{\sqrt{n}}$	Yes (25,0.1)	No	
14	50	64	0.001	Yes	No	No	-	Yes (25,0.1)	Yes	

15	50	64	0.001	Yes	No	Yes	$\frac{1}{\sqrt{n}}$	Yes (25,0.1)	Yes	
16	50	64	0.001	Yes	No	No	-	No	No	Momentum in BatchNorm 2d layer set to 0.99
17	50	64	0.001	Yes	No	No	-	No	No	Outlier removal
19	50	128	0.001	Yes	No	No	-	No	No	Outlier removal
20	50	64	0.001	Yes	No	No	-	No	No	Outlier removal without BatchNorm 2d layers

Batch size and learning rate

Four experiments are done to find an appropriate batch size (experiments 2 to 5 in Table 2). This is done using the simple standard model and as a learning rate value of 0.001.

The options of experiment 3 which lead to the best results considering only the mean value of the three validation metrics are used also for the following two experiments, to test two additional learning rates: one bigger and one smaller (experiments 6 and 7 in Table 2).

Weighted loss and balanced sampling

After fixing the batch size and the learning rate, always by considering the mean value of the three validation metrics, the focus is moved to applying the weighted loss and the balanced sampling. For each approach, two different possibilities for computing the weights are applied.

Learning rate scheduler

As an additional approach to try to improve the results, a learning rate scheduler is used. This is applied to experiments done with the same options as experiments 3 and 10. As gamma parameter, the standard parameter of 0.1 is used (PyTorch c, 2022) and as step size, 25 epochs are applied. This is chosen according to the plot of the training loss for experiment 3 (see Supplementary Material **Error! Reference source not found.**), 25 is approximated as the number of epochs where it seems that the trend of the curves tends to change. These two experiments are the number 12 and 13 in Table 2.

Additional data augmentation

In addition to the standard data augmentation, also an extra data augmentation on the training data is applied. This additional data augmentation is applied with a probability of 50% and consists of a random gaussian blur with values for the kernel size between 5 and 9 and values for the standard deviation between 0.1 and 5 (these parameters are taken from the PyTorch tutorial (PyTorch e, 2022)). Because it is a random gaussian blur, whenever it is applied, the gaussian blur will be different (PyTorch e, 2022). This approach is applied to the experiments done with the same options as for experiments 12 and 13, and they are summarized with the numbers 14 and 15 in Table 2.

Experiments to deal with the not plausible validation loss

Additional experiments are done to try to deal with the validation loss function, which shows a trend not plausible (see Chapter 5.1.1 for more information). A possible cause could be the different behaviour of the mean and standard deviation of the batch normalization layer, in the training phase and in the validation phase (PyTorch f, 2021; PyTorch g, 2022).

This is done by changing the momentum of the batch normalization layers to 0.99, experiment 16 in Table 2, (PyTorch f, 2021; PyTorch i, 2017), by applying an outlier removal on the input images (experiments 17, 19 and 20) and changing the batch size to 128 (experiment 19).

The outlier removal was done considering all the images in the train-validation dataset together, and for each channel the 0.5% and the 99.5% quantile are determined. For each channel, the pixels with a value lower than the 0.5% quantile (respectively higher in the case of the 99.5%) are set to these two extreme values. The rest is done using the standard approach. An additional experiment is done by completely removing the batch normalization layers (experiment 20) from the CNN.

4.2 Comparison of the model performance in the different years

The temporal mismatch between the ground observations and the aerial images acquisition might influence the results (Fraser and Congalton, 2021). To test this hypothesis and to try to quantify this influence, the same setup for the model (hyperparameters and structure) as in experiment 3 done with the dataset of 2012 (see Table 2) which is considered the best model (see Chapter 5.1) is applied to the data of the single year.

To make the comparison as robust as possible for all the years only the trees which can be found in the year 2012 (trees of interest) are used and the separation between the train-validation dataset and test dataset is done for all the years according to 2012. So that also for different years the same tree is either in the train-validation dataset or in the test dataset, but not in both. However, the subdivision in the single folds is not the same over the years. For the year 2012, the subdivision is done randomly by applying a 20-80% split. As a bounding box size to generate the images of the single tree a dimension of 10 meters is used. For every single year, the CNN is trained and validated on the train-validation dataset, applying a 5-fold cross-validation and using only the data of the single years. The validation metrics are computed and compared between the different years. To quantify the temporal gap, it is assumed that all the field observations were done on the 1st of August. In fact, the exact days of the acquisition of the ground observations are not known, so it is approximated to the day in the middle of the acquisition period according to the IAP (IAPc, 2022).

To consider in the analysis also the time when the aerial images were taken, also a mean time pro acquisition date is computed, considering the mean between the earliest and the latest image which covers the areas with the ground observations.

4.3 Generalization over new trees

An additional investigation is done about the capability of the CNN to generalize well on new trees using data from more years together. Considering more years together increases the number of training samples and their variability and can also help to understand if that is beneficial for the model performance or not.

The data for the years 2008 and 2018 are not considered, because they were taken at the two extremes of the acquisition period and according to the results shown in Chapter 5.3 they could maybe lead to poor results.

For the subdivision in the train-validation and test dataset, the data from the years 2009, 2012, 2014 and 2021 are considered. For each year only the trees that can be found in the year 2012 are used (trees of interest). The data are divided for the single years into two parts according to the same separation as for 2012. Then, all the train-validation data of the single years are taken together in a combined train-validation dataset, and the same is done for the data in the test dataset which are combined. The subdivision in the single folds for the train-validation dataset

is stratified considering only the defoliation value. So the same tree for different years can be in different folds. It is possible that there are a different number of observations for a specific year or tree inside each fold.
In the two datasets there are data from all four years per a total of 1975 samples in the train-validation and 497 in the test. For the extraction of the single images of the trees, a bounding-box size of 10 meters is used.
The same setup for the model (hyperparameters and structure) as in experiment 3 done with the dataset of 2012 (see Table 2), which is considered the best model (see Chapter 5.1), is used. The CNN is trained and validated on the train-validation dataset and the five resulting models are then applied to the test dataset.
To understand if the training process is helpful, also a *baseline model* is used. This model predicts for each input data a defoliation value of 20, which is the class with the higher number of samples considering only the train-validation dataset and considering only the test dataset. This model is applied to the test dataset and the resulting metrics are compared with the metrics resulting after the training process.

4.4 Generalization over time

A third investigation is done by considering the capability of the CNN to generalize well over time, it means to predict the defoliation values of trees, after being trained using the same trees but in previous years. Also in this case the data from the years 2009, 2012, 2014 and 2021 are considered, and the data from 2008 and 2018 are discarded for the same above-described motivations (see Chapter 4.3). To work as much as possible with the same trees, the trees of interest are used. That means, for the years 2009, 2014 and 2021 only the same trees which can be found also in the year 2012 are considered. All the resulting data in the years 2009, 2012 and 2014 are considered in the train-validation dataset, and the data in the year 2021 are used as test dataset. The separation in the single folds in the train-validation dataset is stratified, considering only the defoliation value. That means that the information about the year (temporal dimension) and about the tree id is not considered to divide the trees in single folds. To extract the single images of the trees a bounding-box size of 10m is used. The same setup for the model (hyperparameters and structure) as in experiment 3 done with the dataset of 2012 (see Table 2), which is considered the best model (see Chapter 5.1), is used. Also in this case the CNN is trained and validated on the train-validation dataset and the five resulting models are then applied to the test dataset. Also, a baseline model which always predicts 20% defoliation is computed on the test dataset and used to compare the results.

4.5 Comparison of different bounding box sizes

In Chapter 3.5 the process of extracting the images of the single trees with a bounding box size of a fixed dimension of 10 meters was described. However, this dimension is still considered arbitrary. The goal of this investigation is to try to understand if the use of different sizes for the definition of the bounding boxes influences the performance of the defoliation prediction. Only the data from the year 2012 are considered. For this year the images of the trees are extracted using 3 different sizes of the bounding box: 5, 10 and 15 meters. The data with the different bounding box sizes are considered independently. For each size, the data are divided into the same training-validation dataset and test dataset.

The same hyperparameters and structure of experiment 3,(see Table 2) which is considered the best model (see Chapter 5.1), were applied to all three different scenarios, to enable a comparison between the three different sizes.

In the three cases, the models are trained on the train-validation dataset and then they are applied to the test dataset.

5 Results and Discussion

5.1 Experiments model considering only the data of a single year (2012)

Table 3: Table showing the mean absolute error (MAE), the overall accuracy and the balanced accuracy for the different experiments done to find the good hyperparameters and the good structure for the model. The experiments are the same which are listed in Table 2, where the details about the single setups can be found. Only the data for the year 2012 are considered. In bold the results which correspond to the experiment with the setup that is defined as *best model* is marked.

Experiment numbers	MAE [%]	Overall Accuracy [%]	Balanced Accuracy [%]
Baseline model	4.73	30	13
2	4.57	36	18
3 (*Best Model*)	**4.37**	**38**	**18**
4	4.54	34	16
5	4.47	36	18
6	4.55	33	16
7	4.56	35	18
8	5.66	28	17
9	6.27	27	18
10	4.76	33	18
11	5.21	30	19
12	4.37	38	18
13	4.80	35	19
14	4.45	36	18
15	4.80	35	19
16	4.36	39	19
17	4.47	36	19
19	4.55	34	19
20	5.08	29	13

Table 3 shows the results of the different experiments done to find the good setup (hyperparameters and structure) for the model, considering the data of the year 2012. The experiments are described in Chapter 4.1 and in Table 2.

To compare the results three metrics are used. Because of the approach of the 5-fold cross validation, for each experiment five models one for each single split, with the same setup, are trained and then are validated. So, the values shown for each metric in the table are mean values, computed considering five values. For each mean value there is also a related standard

deviation, which however is not considered in this analysis. To define which setup for the model leads to the best results, only the mean value is considered.
To get information about the ability of the model to be able to learn, the results are computed also for a baseline model.
To choose the best model between the 18 experiments, the first criterion is that the metrics should be better than the results achieved with the baseline model. And then the second criterion is that both accuracies should be as high as possible and the MAE as small as possible.
Considering these two criterions the best results are achieved with the experiment 16. In this setup the parameter momentum was increased in order to try to deal with the not plausible test loss results (PyTorch f, 2021; PyTorch i, 2017).
But increasing the momentum is also not recommended (PyTorch h, 2019). However, it should be noted that these are not reviewed sources (are only forums) and are not exactly for the same approach like in this work. According to various forum discussions it seems that decreasing the momentum would be more beneficial. So further investigation about the advantage (or disadvantage) of changing the momentum should be done, and more parameters should be tried. However, to stay on the safe side, it was decided to do not use this setup.
Considering the results of the other experiments (without the experiment 16), there are only two other experiments with the best results for two metrics and the second-best result for the remaining metric: they are the experiments 3 and 12. The only difference between these two experiments, is that for the experiment 12 a learning rate scheduler is used. But in this case that does not help to further improve the results.
So, as *best model*, the setup (hyperparameters and model structure) corresponding to the experiment 3 is chosen. This is a simple setup, with no weighted loss or balanced sampling. Theses hyperparameters and model structure will then be applied for all the other investigation done in this work, every time by training the model with the corresponding dataset.
The accuracies reached with the experiment 3 are 1% lower and the MAE 0.01% higher than the metrics reached with the experiment 16.
Considering the results of all experiments, they are in general similar and the differences relative small. It still remains the question how significant the differences are.

5.1.1 Validation loss curve

Considering the results of the experiments done using only the data of 2012 (see Chapters 4.1 and 5.1), it was noticed that most of the validation loss curves follow a not plausible behaviour. The curves tend to have extreme high peaks in the first epochs and then tend to stabilize on relative low values and stay approximately constant. Also smaller peaks, compared to the peaks at the beginning, are observed on the more constant part. Figure 7 shows an example of this type of curve for experiment number 3 (which is considered the best model), and experiment 10.

A possible explanation for this behaviour is that the mean and the standard deviation computed in the training part for the batch normalization layers, and then applied during the validation part are not enough representatives for the activation, which can be explained by different distributions in the data (PyTorch f, 2021; PyTorch m, 2019; PyTorch n, 2017).

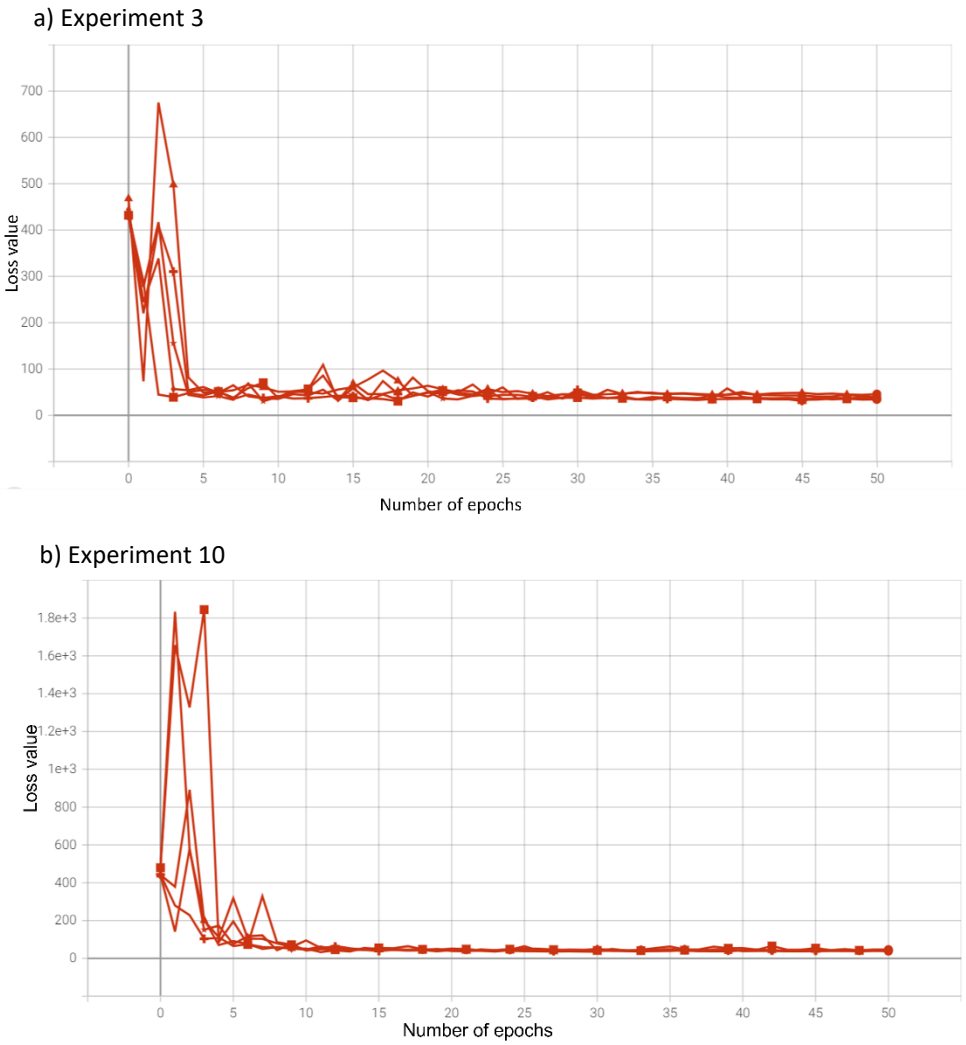


Figure 7: Figure showing the not plausible behaviour of the loss function on the validation dataset for experiment 3 (subplot a), and for experiment 10 (subplot b). The single curves in the plots correspond to the results for the single folds.

It could be that some of the experiments do not show this behaviour, but they have lower performance than the best model.

As described in the previous chapters, some attempts were done to try to avoid this behaviour, but because it was not solved, or because the performances of the model were not improved, or because of the motivations given in Chapter 5.1 about the modification of the momentum, these attempts were discarded. It was decided to continue to work considering the model from experiment 3 as the best model, and to apply it also to other investigations, also if the curve of the loss function is not plausible (see Figure 7 subplot a) and could influence the results. Further

investigations to try to better understand the causes and to solve this problem are needed. Like for example by using a smaller momentum for the batch normalization (PyTorch o, 2019). Also the behaviour of the validation loss curves for the investigations done to answer the various research questions (see Chapters 4.2 to 4.5) should be further investigated.

5.2 Model considering only data of a single year (2012)

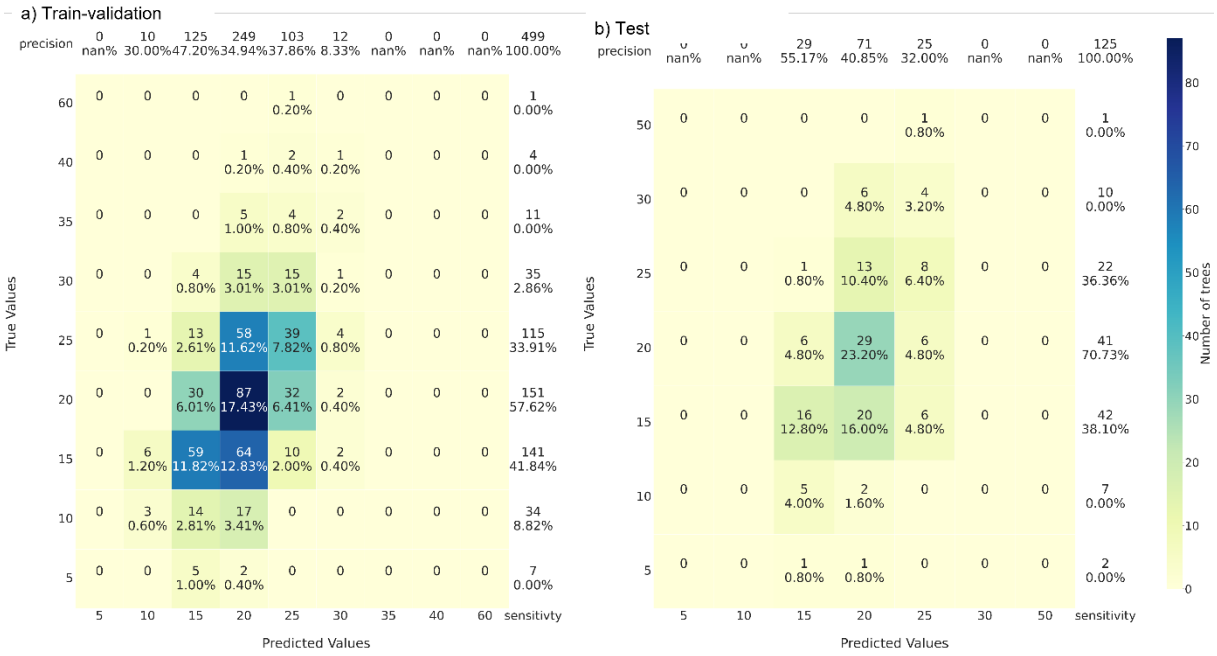


Figure 8: Confusion matrices resulting from the best model applied a) on the train-validation dataset, and b) on the test dataset, considering the data of 2012. On the x-axis, the predicted defoliation values by the CNN are shown and, on the y-axis, the true label values. Inside the matrix, the absolute value refers to the number of trees for each combination of predicted-true values, the percentages are computed considering all the samples in the dataset (499 and 125). The values in the last column in both subplots show the row-wise sum and the sensitivity values, the first row shows the column-wise sum and the precision values.

The setup for the model (hyperparameters and structure) as in experiment 3 (see Table 2) is considered the best model for the year 2012, and the results on the train-validation dataset and on the test dataset are shown in the confusion matrices in Figure 8. The correct predictions are the values on the diagonal.

Considering the train-validation dataset (subplot a), in total 499 predictions are done in the process of validation, and 50% of them are done for the defoliation class 20%. This is also the class with the highest number of ground observations (30% of the total observations). However, the defoliation class 15% has only 10 observations less than the class 20%, and 25% of all the predictions are done for this defoliation class.

Most of the predictions of the model (96%) are done for the defoliation classes 15%, 20%, and 25%, classes which together contain 82% of the observations. The predictions for the classes 10% and 30% correspond only to 4% of the total predictions. For the other classes (5,35,40 and 60), for which there is at least one ground observation, no prediction is done.

By considering the confusion matrix for the test dataset (subplot b), the defoliation class with the higher number of predictions is the 20% defoliation class, with 57% of the total 125

predictions. This class has 33% of the ground observations and is the second represented class with only one sample less compared to the 15% defoliation class, the class with the highest number of observations. For the defoliation class 15% (with 33% of the ground observation), 23% of the total predictions are done. All the values are predicted for the defoliation classes 15, 20, and 25%. For classes 5, 10, 30, and 50, which have at least one ground observation, no prediction is done.

Table 4: Table summarizing the values of the 3 metrics: overall accuracy, balanced accuracy and mean absolute error (MAE) for the baseline model, and for the best model applied on the train-validation dataset and on the test dataset for the year 2012. The mean and the standard deviation for the train-validation result from the single models in each fold, for the dataset only a single value is computed.

	Overall accuracy [%]	Balanced Accuracy [%]	MAE [%]
Baseline model on the train-validation dataset	30 (σ=0.45)	13 (σ=1.10)	4.73(σ=0.16)
Best model on the train-validation dataset	38 (σ=3.35)	18 (σ=1.41)	4.37(σ=0.15)
Best model on the test dataset	42	21	4.01

The resulting metrics for the baseline model and for the best model firstly applied on the train-validation dataset and then on the test dataset, are summarized in Table 4. The best results are reached on the test dataset, and the worse using the baseline model.

When considering these results, it seems that the performance is relatively low, because of the accuracy values which maximal reach 42%. However, the performances achieved with a trained model are better than with the baseline model. That suggests that the learning process somehow, at last partially, works.

For the trained model, the performance on the test dataset is slightly better for all three metrics, than for the train-validation dataset. One possible explanation could be that the results on the test datasets are computed with an *ensemble model* by combining the outputs of 5 single models, which in machine learning can often generate better results than considering only one single model (Džeroski et al., 2009).

Most of the predictions in the train-validation dataset, and all predictions in the test dataset, are done for the defoliation classes 15, 20, and 25%. These are also the classes with the highest number of ground observations. It is reasonable and reflects the reality: it can be expected that in a city where the trees are managed (Vogt, 2020), it is difficult to get observations with high defoliation values. But this is problematic because to generate a model which can predict defoliation values on a large range of defoliation classes, ideally on the complete spectrum from 0 to 100%, enough ground observations for all the classes are needed.

The results show that a prediction on the test dataset is possible only for defoliation classes with at least 115 ground observations in the training-validation dataset. With the 5-fold approach

during the training only a part of the samples is used, which additionally decreases the number of data. When considering only the prediction done on the train-validation dataset, at least 34 ground observations are needed to be able to make at least one prediction.

In the beginning, it was hypothesized that the model would be able also to predict enough values also for underrepresented classes (or even for classes with any samples), but that is not the case. Considering more in detail the difference inside the three defoliation classes with the highest number of samples and predictions (15,20 and 25%), it is possible to see that in the train-validation dataset for the defoliation class 20% there are double predictions as for the defoliation class 15%. That does not respect the difference in the number of samples (where the defoliation class 15% has only 10 samples less than the defoliation class 20%). The same can be observed also considering the test dataset. In this case, the defoliation class 15% has only one sample more than the defoliation class 20%, but for the defoliation class 20% there are more than double the predictions as for the 15% class defoliation.

Therefore, it seems that the CNN tends to overpredict the defoliation class 20%. It is supposed that this overprediction is not only caused by the high number of observations in the defoliation class 20% but also because the model learned during the training process that it is most convenient to predict values around 20 (which are then rounded in the class 20%), to minimize the loss value. Because these predictions will lie approximately between the other two classes with a high number of samples, and near the class with 20% defoliation.

To limit this effect one possibility would be to try also other loss functions, or try to solve a classification problem, instead of a regression problem, but considering the order of the classes. Considering the results obtained in the work by Kälin et al. (2019), the order of magnitude for the MAE seems to be comparable. They achieved an averaged MAE across the five test splits which ranged from 5.5% (with σ = 0.7%) to 14.8% (with σ = 6.9%). In this work values of 4.37% (with σ = 0.15%), on the train-validation dataset, and 4.01% (see Table 4), on the test dataset, are achieved. However, particular attention should be paid when comparing the results, because both approaches are similar but not identical, so it is not clear if the results are directly comparable, and methodological differences could also explain the different results. Additionally, in the work by Kälin et al. (2019), they do not always have more images for the training, but they have observations on a larger range of defoliation classes as expected under natural forest conditions. That could explain their ability to predict over a larger range of defoliation values, but that could maybe also lead to their higher MAE.

An important point to consider is the difference between the point of view used for the visual assessments and the point of view of the aerial images. The defoliation values were acquired by observing the tree from the ground when instead the model uses aerial images which have a top to bottom view. This is a substantial difference from the work done by Kälin et al. (2019), where the point of view during the visual assessment is similar to or is the same as the point of view of the images which are seen by the model.

According to the investigations done by Kälin et al. (2019) about the observer bias, considering six different observers, they agree when they consider defoliation classes at the extreme, but they tend to disagree in the central classes (approximately between 15% and 60% defoliation), the average MAE of the six observers was of 4.6% with a standard deviation of 0.5%. This shows that the central classes could be challenging to be estimated. Considering the data from 2012, most of the observations are in the classes 15, 20, and 25% defoliation and it is possible that also the CNN could have some additional difficulties distinguishing by nearby central classes.

In the implementation, no specific error tolerance for the evaluation of the predictions is introduced, as was for example shown in (Kälin et al., 2019). Only the approximation done when rounding the predictions to the nearest multiple of five for the computations of the accuracies and of the confusion matrix (for the MAE this is not done), can be considered as a sort of error tolerance. Introducing an error tolerance, for example of 10%, in the model predictions, would lead to an increase in the overall accuracy.

One other possible approach which could maybe help to improve the results is to work with fewer defoliation categories. For example, Fraser and Congalton (2021) and Lehmann et al. (2015) considered three healthy classes for the trees, and Otsu et al. (2019), considered only two categories: trees with and without defoliation caused by the pine processionary.

For this work, one possibility would be to set a threshold to subdivide the trees into only two classes of vitality, according to the defoliation value. However, because of the distribution of the data for 2012, it is possible that according to how the threshold is chosen, almost all the samples will be in the same class. Leading to problems for the training and the risk of performance of nearly 100%.

In Kälin et al. (2019) it was mentioned that having a specie specific model would lead to better results. They discarded this option because of an insufficient number of samples. Also in this work, as shown above, the number of samples plays a fundamental role in the training process and considering the single species separately would decrease the number of samples, probably causing a decrease in the performance. So, this approach was discarded.

To further improve the results, it is supposed that doing more experiments, trying more hyperparameters and structure for the model, and doing a better and deeper interpretation of the single metrics and results after each experiment, could be beneficial. Also training for more than 50 epochs could maybe help to improve the results because it seems that after 50 epochs the learning process is not completely finished. It would also be interesting to use different network architectures, to see if that could be beneficial. Additionally, further investigations of the preprocessing steps and the outlier removal could be done to try to improve the performance.

5.3 Comparison of the model performance in the different years

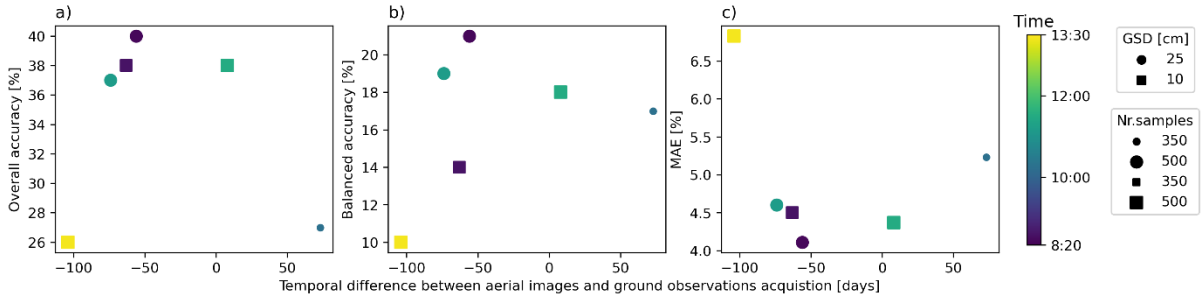


Figure 9: Values of the metrics, a) overall accuracy, b) balanced accuracy and c) mean absolute error (MAE) computed using the same model (hyperparameters and structure fixed for the year 2012), retrained and validated on the train-validation dataset for the single years (trees subdivided according to 2012). The colour shows the time of acquisition for the aerial images and the symbols the ground sampling distance (GSD). The size of the symbols is related to the number of samples. On the x-axis the difference in days between the data of the aerial image acquisition and the mean date of the ground observations acquisition (the 1^{sf} of August) is shown. A negative number of days means that the aerial image acquisition was done before the 1st of August, and a positive number means that it was done after the 1st of August. Each point corresponds to a different year (from left to right in each subplot: 2018, 2009, 2021, 2014, 2012, 2008).

Figure 9 shows the results, achieved on the train-validation dataset, for the investigation done to compare the performance of the CNN when the single six years are considered individually, and to try to understand the influence of the temporal mismatch between aerial images acquisition and ground observations acquisitions. The same hyperparameters and structure of the model, which were defined for the year 2012 (see experiment 3, Table 2) are applied to the dataset of the single years. The CNN is trained and validated for each year singularly on the train-validation dataset. The values for each year shown on the plot are mean values, computed considering the result from each split during the process of the 5-fold cross-validation. There is a related standard deviation, but in this analysis, only the mean value is considered.

The worst overall accuracies (subplot a) are obtained for the two years for which aerial image acquisition was either much earlier (20/04/2018) or much later (12/10/2008) as compared to ground observations. A similar consideration can be done by considering the MAE (subplot c), also in this case the highest error values are achieved for these two acquisitions. The other four acquisitions in between seem to be clustered for both metrics.

The results of the balanced accuracy (subplot b) do not confirm the same trend as in the other two metrics, for example considering the value for 2021, where a higher accuracy value would be expected, and for 2008, where instead a lower value would be expected. However also in this case the worse result is achieved for the year 2018.

When considering the ground sampling distance, it seems that similar results are achieved with 10cm and with 25cm. For the years 2009 to 2021, there are between 488 and 499 samples,

which can be considered as approximately the same number, for the year 2008 there are only 301 observations. For this year for two metrics, the results are the second worse.
The worse result is always achieved with the acquisition done around 13:25, for two metrics (overall accuracy and MAE) the results are similar for acquisitions done around 8:30 and 11:15. For the balanced accuracy is not so clear. The acquisition taken around 10:15 is for one metric more similar to the acquisition around 13:25 and for the other two metrics similar to the other acquisition times.
The hyperparameters and the structure of the model were defined considering only the year 2012. So, this year it is somehow "advantaged" and that could partially explain the high performance of this year (compared to the others year). Nevertheless, the best results are not achieved in 2012, as could be expected. A possible explanation could be because the CNN is however retrained for each year.
The results shown in the subplots a) and c) seem to suggest that the temporal mismatch between the aerial images acquisition and the field observations acquisition could play a role, because for the two years with the areal images taken either much earlier (20/04/2018) or much later (12/10/2008) the worse results are achieved.
The importance of the data acquisition at the correct time is introduced also in Kälin et al. (2019), and in the work by Fraser and Congalton (2021).
According to these results, it also seems that acquiring the images after the field observations is more problematic that acquiring them before. For the years 2008 and 2009, the temporal difference in the absolute number of days is practically the same, but in 2008 the acquisition was done 73 days after the ground observations (which leads to worse results) and in 2009 74 days before. This can probably be explained by the different states of the trees in these two periods, for example, in October there are brown leaves, were in Mai they are green.
However, it is not clear how well these considerations can be generalized. In particular because the results of the balanced accuracy (subplot b)) do not completely confirm the other results.
In addition the results reached on the test dataset (see Supplementary Material 7.4) should be further investigated to see if they can at least partially confirm the results achieved on the train-validation dataset or not (see Figure 9). Further investigations are still needed.
To make the analysis as robust as possible, for each year the same trees as in the year 2012 are used. However, some trees may have been substituted during the years, but because the location remained the same, they are considered the same tree. This subtle difference is not considered in this investigation. In addition, in 2008 not all the trees which were observed in 2012 were already been monitored.
Another possible point which could maybe partially explain the differences in the results over the different years could be given by differences in the single years, for example, if the vegetation development started earlier or later. In addition, it should be considered that the 1st of August is only an approximation of the real field acquisition date. It could be that for some

years the acquisition was done earlier or later, and that would have as a consequence that the temporal difference would be different as shown in the figure.
Finally it is can that also the different data distribution in the defoliation classes, which are not exactly the same in the different years, could also partially explain the different results that are achieved in the different years.

5.4 Generalization over new trees

Table 5: Table summarizing the values of the 3 metrics: overall accuracy, balanced accuracy and mean absolute error (MAE) for the baseline model and the best model applied on the train-validation dataset and on the test dataset. The train-validation dataset contains trees from the years 2009, 2012, 2014 and 2021. The test dataset contains other trees, always from the same four years. In the first two rows, the mean and the standard deviation are shown, and in the last two rows only the resulting value.

	Overall Accuracy [%]	Balanced Accuracy [%]	MAE [%]
Baseline model on the train-validation dataset	36 (σ=0)	9 (σ=0)	4.61 (σ=0.03)
Best model on the train-validation dataset	36 (σ=3.05)	13 (σ=0.45)	4.56 (σ=0.21)
Baseline model on the test dataset	35	11	4.19
Best model on the test dataset	39	16	3.92

Table 5 shows the results of the computations done to check the ability of the model to generalize well on new trees considering data from more years together. For two of three metrics, the model trained and validated on the train-validation dataset outperforms the baseline model. The mean value of the overall accuracy is instead the same, considering the standard deviation the trained model is even worse. When the 5 models generated from the train and validation process are applied to the test dataset the results show higher accuracies for all the three metrics. Considering only the results achieved on the test dataset, the trained model reach better results for all the metrics as the baseline model.

For this investigation trees from different years, with aerial images acquired with different GSD, at different times and with different sensors are used together. This increases the variability of the used data.

The goal is to have a CNN which can generalize well on new trees. To check this capability, the separation of the data between the train-validation and test dataset was done so that the same tree, also in different years, is either in the test dataset or in the train-validation dataset.

But this is not the case for the separation of the tree in the train data and in the validation data using the 5-folds. In fact, the separation in the single folds is a shuffled stratified separation considering only the defoliation values.

In the train-validation dataset the observations of the same tree in the different years can be in different folds. One consequence for example is that, the same tree for some years can be in the train dataset and the same tree but for other years can be in the validation dataset, which is not the idea of this investigation. Another consequence is that the same trees are seen a different number of time during the training process.

The so-trained five models are then applied to the test dataset (on completely new trees).

Two points can be deduced. Firstly, the analysis of the results of the model applied to the train-validation data does not give information about the main goal of this investigation (generalization over new trees), it gives only information about the effect of considering more years together and a dataset with higher variability. Secondly is questionable if the applied approach of training the five models on datasets with mixed trees and then applying to the test dataset make sense or not, because they were trained and validated on data which do not correspond to the research question. It should be further investigated if this subdivision can be still considered correct or not.

A better alternative would be probably to separate the train-validation data also considering same trees (so that the same tree in different years is either in the train dataset o in the validation dataset, but not in both), and ideally also the defoliation level. Or another alternative would be to train on the complete train-validation dataset without any subdivision in the single folds.

For this computation, it is assumed that the defoliation is independent of the value of the precedent years. In contrast to the two above-described investigations for the year 2012 and for the single years, the used dataset is bigger (1975 trees for the train-validation dataset and 497 for the test dataset). It was supposed that with more samples, the model would be able to be trained with a higher diversity and generalize better.

By comparing the results, only on the test dataset, for this investigation (see Table 5) with the results of the investigation done considering only the year 2012 (see Table 4), the overall accuracy is slightly higher (3%) for the case of the year 2012, the balanced accuracy is also 5% higher, but the MAE is 0.09% lower for this investigation considering more years together.

An at least partially explanation for the lower balanced accuracy when considering more years together can be the higher number of defoliation classes in the test dataset compared to only the data of 2012. It is difficult only with these results to say if considering the data of more years together is helpful or not because for the overall accuracy it seems not to be beneficial but for the MAE better results can be achieved. Further investigations are needed.

An further possible explanation for these differences in the results could be the different distribution of the number of samples in the different classes in the different datasets.

It should also be considered that the higher variability could also be problematic, in fact, the images were acquired at different times, and it could be that there is too big variability inside the same defoliation class, in the sense that trees (maybe also the same tree in different years) with the same defoliation value, has very different images, which can “confuse” the algorithm and lead to poor results.

In addition, for this experiment, the hyperparameters and the model structure were not optimized, but the best model (model 3 defined for the year 2012) is reused. It is supposed that searching for specific hyperparameters and model structure also for this experiment could improve the results.

Considering only the results on the test dataset for this investigation, it seems that the model is able to learn something during the training process, because the performances achieved with the trained model are higher than the performances on the baseline model

5.5 Generalization over time

Table 6: Table summarizing the values of the 3 metrics: overall accuracy, balanced accuracy and mean absolute error (MAE) for the baseline model and for the best model applied on the train-validation dataset and on the test dataset. As the train-validation dataset, the trees from the years 2009, 2012 and 2014 and for the test dataset the trees from 2021 are used. For each year only the same trees as in the year 2012 are considered. In the first two rows, the mean and the standard deviation are shown, and in the last two only the resulting single value.

	Overall Accuracy [%]	Balanced Accuracy [%]	MAE [%]
Baseline model on the train-validation dataset	35 (σ=0.55)	10 (σ=1.14)	4.60 (σ=0.08)
Best model on the train-validation dataset	38 (σ=3.03)	15 (σ=1.95)	4.29 (σ=0.10)
Baseline model on the test dataset	39	9	4.29
Best model on the test dataset	39	11	4.47

Table 6 shows the results of the investigation done to explore the capability of the proposed approach to generalize well during the time. The model trained and validated on the train-validation dataset leads to better results for the balanced accuracy and for the MAE as the baseline model. The balanced accuracy and the MAE achieved with the five trained models applied on the test dataset are worse than the trained and validated model on the train-validation dataset. The performance of trained model applied on the test dataset compared to the baseline model for the same dataset is: for the balanced accuracy better, for the MAE worse and for the overall accuracy exactly the same.

It should be noted that the results achieved on the train-validation dataset are not representative of the research question for this investigation. In fact, during the training and the validation the data, from the years 2009,2012 and 2014 are used. But the subdivision in the single folds, and as consequence also in the train dataset and in the validation dataset, is shuffled and stratified only based on the defoliation values and without any consideration about the temporal dimension. So, the subdivision do not actually respect the research question of this investigation. It should be further investigated if this subdivision can be still considered correct or not, and if the model can take advantage of it or not. The so-trained five models are then applied to the test dataset (which contains only the data for the year 2021).

A possible alternative approach would be to directly use all the data of 2009, 2012 and 2014 for the training, without any subdivision in the single folds. Or to considered also, in addition to the defoliation, for each single tree, when the subdivision in the single folds is done.

For these reasons it makes more sense to consider the results on the test dataset.

By comparing the results on the test dataset achieved considering only the data of 2012 (see Table 4) and for the model that generalizes on new trees (see Table 5), the same overall accuracy is reached as is the case for the "generalization over new trees". This is slightly worse than

considering only the year 2012. Instead, the balanced accuracy and the MAE are worse in both cases. However, a possible explanation for the lower value in the balanced accuracy is the number of defoliation classes in the different test datasets: in particular in the investigation of "generalization over time" there are 11 classes with at least one sample. This is a higher number compared with the other two investigations where there are 7 classes (in the case of only 2012), respectively 9 classes. On one side probably a lot of them are not predicted correctly because of the lower number of samples, and second because having a high number of classes with bad individual predictions can cause a lower value in the balanced accuracy. In addition, in this case, in contrast with what was done for the year 2012, the hyperparameters and the structure of the model were not optimized, which could maybe help to even reach better results.

By considering also the results on the test dataset for the baseline model and the trained model for the generalization over new trees (see Table 5), it is possible to see that there is an improvement for all the metrics when the model is trained. This is not the case for the generalization over time. That suggest that with the proposed approach it is easier to generalize over new trees in the same years instead to generalize over a new year but considering the same trees.

5.6 Comparison of different bounding box sizes

Table 7: Table showing the three metrics computed using the best model trained and validated on the data for the year 2012 and applied on the test dataset. The shown results are reached on the test dataset. This is done for three different types of bounding box sizes used to generate the images of the area around the trees.

Bounding Box size [m]	Overall accuracy [%]	Balanced Accuracy [%]	MAE [%]
5	34	15	4.35
10	42	21	4.01
15	38	18	3.91

Table 7 shows the results of the same model applied to images extracted with a bounding box size of 5, 10 and respective 15 meters. For the model, the same hyperparameters and structure of the best model (model 3 shown in Table 2) are used. The worse results are reached with a size of 5 meters. For 10-meter size for two metrics, the best results are reached. Only for the MAE, the best result is achieved with a size of 15 meters.

In Chapter 3.5 it was already described how the single areas around the trees are cropped from the aerial images. Until this subchapter, all the described investigations and experiments were done using a size of the bounding box of 10 meters. However, the choice of 10 meters can be considered as arbitrary.

From the results, it is difficult to say if it would had be better to work with a size of 15 meters or 10 meters. It seems that working with 15 meters would be a good alternative, and also maybe better than 10 meters. In fact, with 15 meters the MAE is smaller than for 10 meters.

The two accuracies are better for 10 meters, but it should not be forgotten that the hyperparameters and the model structure were optimized for this size, and they were directly applied to 5 and 15 meters, without further investigations.

So, it could be that optimizing the model also for 15 meters would lead to better results also for the two accuracies. The comparison is done on the test dataset to try to at least limit this difference in the optimization of the model. A specific optimization for 5 meters could help to achieve better results also for this size. Further investigation seems to be needed to find the optimal bounding box size.

The approach of using a fixed size (for example 10 meters) has the advantage that it is possible to easily generate bounding boxes of the desired dimension, but that is possible only if the coordinates of the trees are known and are precise.

With this approach, the crown diameter is fixed and is the same for all the trees, when in the reality this characteristic varies. As a consequence, a fixed size could match the crown size for some trees, but for a part of the trees would be too big or too small. If the bounding box size is

too big, also part of other trees will be considered in the image, and also part of the surroundings (i.e. ground, vehicles). In contrast, when the bounding box size is too small, not all of the tree crown is shown in the cropped image.

In the data of this work, the coordinates of the tree are not always perfectly centred on the tree crown as seen in the aerial image. However, it is supposed that for this work the quality is enough good.

Figure 10 shows the resulting cropped images for the same tree if three different bounding box sizes are used.

Figure 10: Example of three different images for the same tree of interest. Three sizes for the bounding box are used to extract the region of interest around the tree. From left to right, a) 5 meters, b) 10 meters, c) 15 meters. For better visualization, the size of the images in this document was increased by 1.5, and the colours are also not the original which are seen by CNN. However, it should give an idea of the main differences between the used images. The tree for which the defoliation value is associated with the image is the tree in the centre of the images.

To extract more flexible bounding boxes, which are adapted to each single tree crown and also for trees without known coordinates, an approach based on deep learning could be used. For example using the package deep forest (Weinstein et al., 2019), as it was also shortly tested during this work (see Supplementary Material 7.2). This approach would probably also have the advantage to be able to extract only the part of the crown if they are partially covered for example by the facades of buildings.

Another alternative, which however would also work only for trees for which the location, as well as additional information about the tree condition like the diameter at breast height (DBH), is known, is to use allometric equations. With these formulas, like for example (McPherson et al., 2016) or (Franceschi et al., 2022), it is possible to compute an approximation of the dimension of the tree crown for each individual tree.

In Kälin et al. (2019) it was shown using activation maps some useful parts of the images for the prediction. A possibility to see in the case of aerial images which regions of the images are most important for the CNN, i.e. only the centre or the complete image could be for example applying a Gradient-weighted Class Activation Mapping (Selvaraju et al., 2020). This could be

helpful to understand when more trees are shown on the same image if the CNN is able to focus on the desired tree or not, because as shown in Kälin et al. (2019) having more trees on the same image is problematic for the model.

6 Conclusions and Outlook

This work has the goal to investigate the estimation of the defoliation of urban trees at the single tree level in an urban environment using a deep learning approach. Ground observation of the defoliation of the trees and 4-channels aerial images acquired with a manned aircraft are applied. Data from the year 2012 are mostly used, but for some investigations also data from other years. The extraction of the area around the single tree is done using a simple approach based on a bounding box of a predefined and fixed tree crown size, usually 10 meters. As CNN a RESNET 50 is used to predict the defoliation.

For the experiments done to find a good model considering only the year 2012, the resulting model shows the ability to learn something. On the test dataset, an overall accuracy of 42% a balanced accuracy of 21% and a MAE of 4.01% are achieved. The model seems to overpredict the defoliation class 20%, and having enough observation in single defoliation classes is important to be able to make predictions for the corresponding defoliation class.

Considering the investigation done to better understand the effect of the temporal difference between aerial image acquisition and ground observation, by comparing the results computed on the single years individually, a part of the results suggests that acquiring the images too earlier or too late in the season could be problematic. However, this result is not confirmed by all the computations, so it is possible that there are other factors that explain the different results, for example maybe the different distribution of the defoliation classes in the used dataset for the single years.

For the investigation done to find a model which generalizes well over new trees or over time, combining data from more years together, it seems that the generalization over new trees is easier than the generalization over time. However, the validity of these results should be further investigated, because of the subdivision of the data in single folds for the train-validation part of the process, which is not completely correct.

In general, it seems that the data are one of the limiting factors, and further investigations are needed to learn more about the approach, because there are some points which are not fully understood. With the actual results, it seems difficult to use this approach in the practice, but it is supposed that the approach can be further improved.

On possibility to improve the approach, would be to work with a bigger dataset with a higher defoliation variability, having more samples per defoliation classes, and to see if the CNN would be able to better predict on a wider range of defoliation classes.

In addition the best setup for this type of problem should be further investigated: more experiments to find the most appropriate hyperparameters and model structure should be done, as well to test more architectures. The codes which are used in this work can be further improved, on one side they can be optimized, and they can be controlled to ensure that there

are no bugs which could potentially limit the performance. Additionally, also the not plausible behavior of the validation loss function should be further investigated, to try to find a solution, which could help the performance of the model.
To extract the area of interest around the single tree it would be interesting to apply a deep learning approach, which can automatically detect the tree crown from aerial images, and to see if that could be helpful for the model. This approach would make possible to apply the trained model also on regions were the coordinates of the trees are not known. To acquire the images also an UAV could be used to see if a GSD smaller than 10cm could further improve the results. Using data acquired in other cities could be useful to check if the approach would be scalable on a other region than only Basel, this could be also helpful to improve the variability and the number of the data that can be used to train the model.

7 Supplementary material

7.1 Ground observations per single year

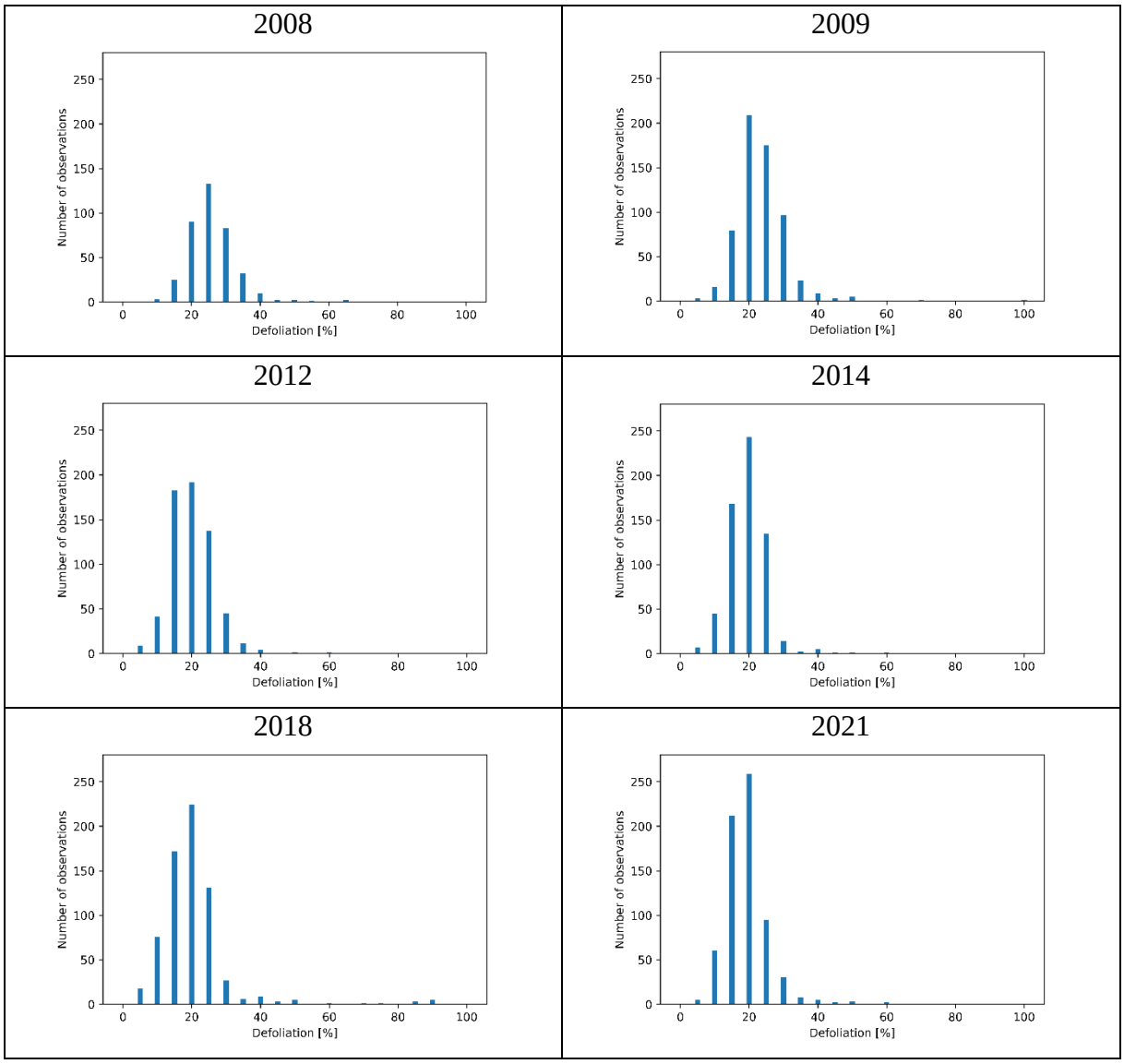


Figure 11: Number of observations for the different levels of defoliation for each year. The shown ground observations are for all the trees in each year which remain after the processing steps described in Chapter 3.3.The observations of the trees of interest, which are used in the work, are a subpart of the shown ground observations. The trees of interest are for each year the trees that can be found also in the year 2012.

7.2 CNN approach to extract single trees

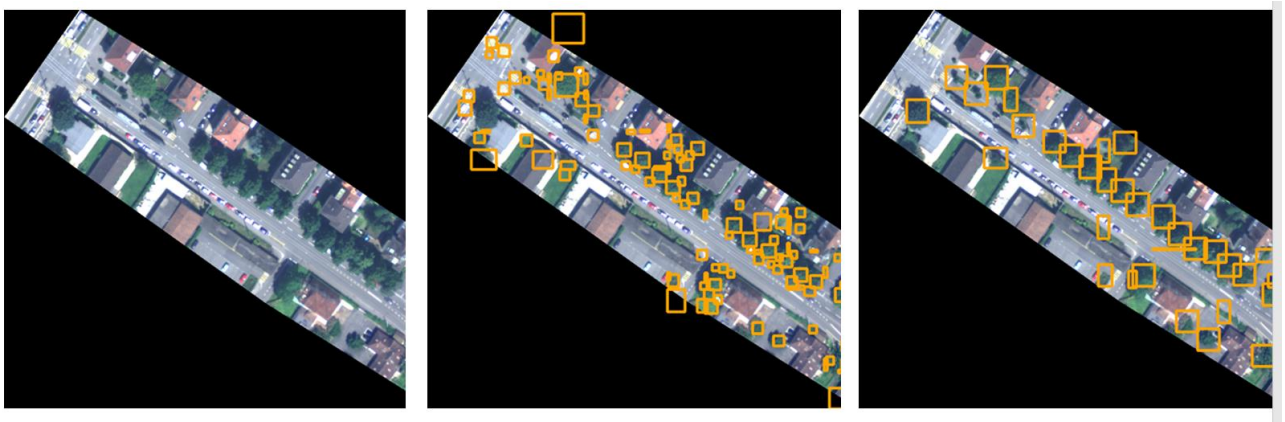


Figure 12: Example of the test done using DeepForest (Weinstein et al., 2019), a Python package, for the tree extraction. Left the: aerial image of a street in Basel, in the center the results using the pretrained model. Right the results achieved after re-train the pre-trained model using the trees and the images of Basel.

7.3 10 Meters size as tree crown dimension

To define a size for the bounding boxes a simple analysis is done. The goal is to have a bounding box size which match as good as possible as many as possible trees. The trees were geographically divided into 17 areas. For each aera a single tree was chosen mostly random. Using the aerial images and ArcGIS Pro the diameter of the tree crown was measured.
In at last one case, the defined crown size was not from the randomly picked tree but from a more representative crown size of trees in the area.
Measuring the size of the tree crown was not easy, and approximations were done. Because of these difficulties and of the not complete coherent approach, it is possible that there are inconsistencies in the measured crown sizes.
The measured crown sizes are shown in Table 8. According to these measurements a size of 10 meters was chosen to use as bounding box size to extract the areas of interest around each tree. So it can be affirmed, that this approach is not robust and the choice of 10 meters can be considered as arbitrary. It is however assumed that this approach gives a first general and approximative overview of different tree crown sizes which can be found in the city.

Table 8: Table with the tree number and the crown size manually measured in Arc GIS Pro.

Tree Number	Measured Tree Crown Size [m]
BS015104	2.35
BS016296	5.0
BS005446	5.30
BS005468	5.60
BS019889	6.30
BS018308	7.30
BS004971	7.5
BS009669	7.6
BS003834	8.0
BS005464	8.30
BS008419	9.0
BS009913	9.60
BS005170	9.60
BS020415	10.50
BS007644	12.0
BS029136	12.0
BS000875	12.0

7.4 Comparison of the single years on the test dataset

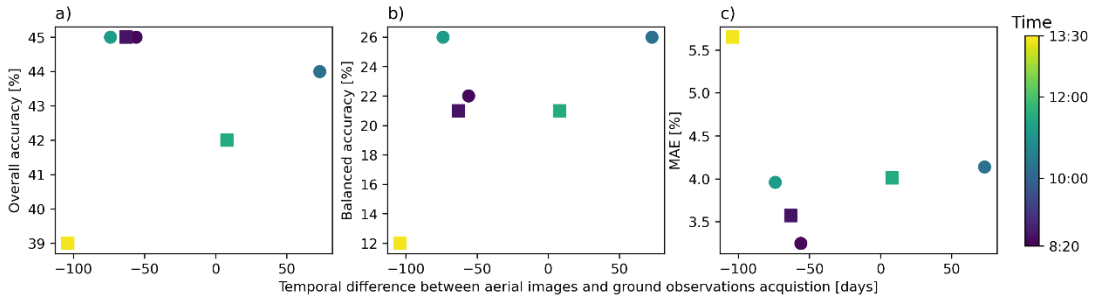


Figure 13: Figure showing the same analysis as done in Chapter 5.3 but on the test dataset. In this case the size of the symbols is not proportional to the number of samples. A circle means a GSD of 25 cm, a square a GSD of 10cm

7.5 Analysis training loss

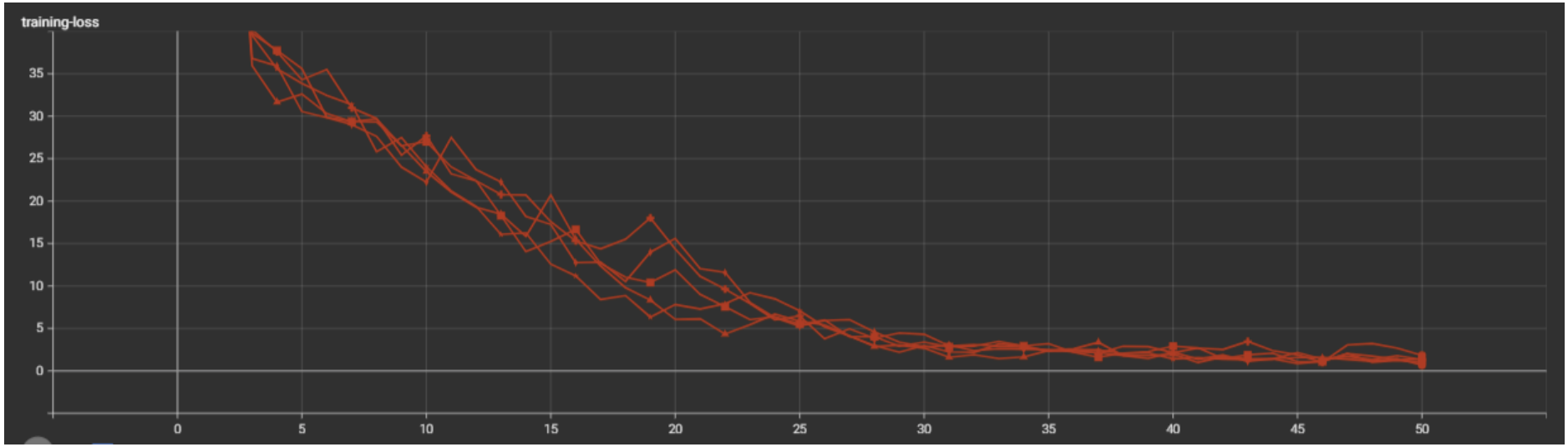


Figure 14: Training loss computed during the experiment 3, (the outliers are ignored in the chart scaling). 25 epochs is taken as approximative value where the slope of the curve tends to change.

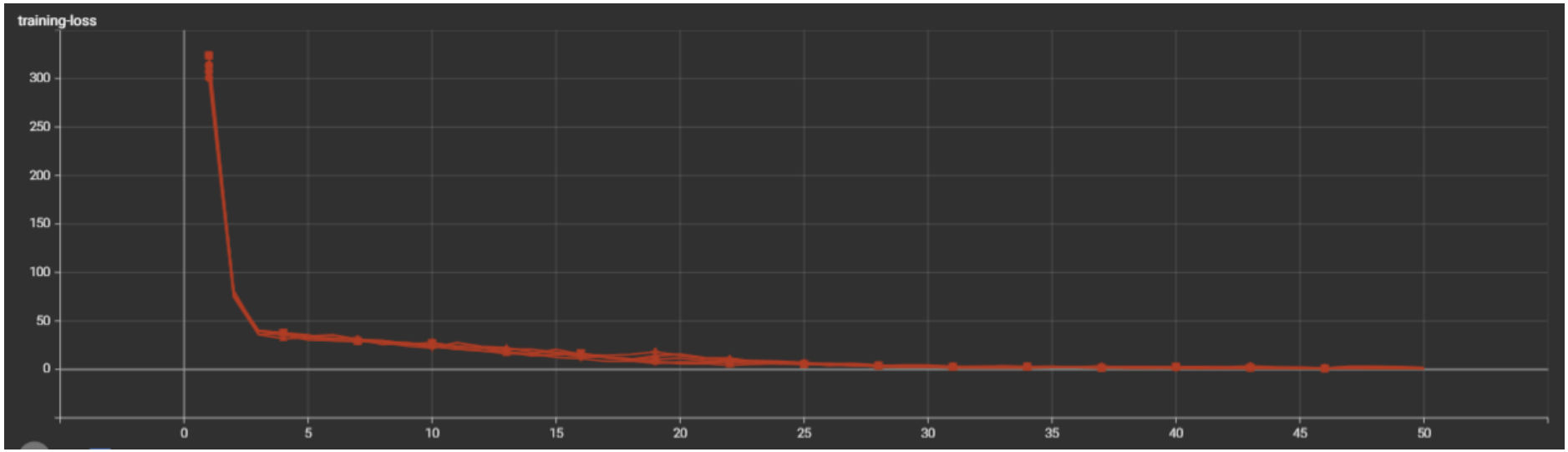


Figure 15: Same plot as in the figure above, but without ignoring the outliers in the chart scaling

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten Semester-, Bachelor- und Master-Arbeit oder anderen Abschlussarbeit (auch der jeweils elektronischen Version).

Die Dozentinnen und Dozenten können auch für andere bei ihnen verfasste schriftliche Arbeiten eine Eigenständigkeitserklärung verlangen.

Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuer und Betreuerinnen der Arbeit.

Titel der Arbeit (in Druckschrift):

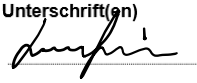
Monitoring Vitality of Urban Trees Combining Airborne Imagery and Deep Learning

Verfasst von (in Druckschrift):
Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n):	**Vorname(n):**
Gaia	Luca

Ich bestätige mit meiner Unterschrift:
- Ich habe keine im Merkblatt „Zitier-Knigge“ beschriebene Form des Plagiats begangen.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert.
- Ich habe keine Daten manipuliert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Plagiate überprüft werden kann.

Ort, Datum	**Unterschrift(en)**
Zürich, 01.07.2022	

Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.

8 References

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53. https://doi.org/10.1186/s40537-021-00444-8

Berrar, D., 2019. Cross-Validation, in: Encyclopedia of Bioinformatics and Computational Biology. Elsevier, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Brang, P., 1998. Sanasilva-Bericht 1997, Zustand und Gefährdung des Schweizer Waldes - eine Zwischenbilanz nach 15 Jahren Waldschadenforschung. Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft.

Brownlee, J., 2021. How to Combine Predictions for Ensemble Learning. Machine Learning Mastery. URL https://machinelearningmastery.com/combine-predictions-for-ensemble-learning/ (accessed 7.4.22).

BS a, 2022. Baumkataster [WWW Document]. URL https://www.stadtgaertnerei.bs.ch/stadtgruen/stadtbaeume/baumkataster.html (accessed 6.12.22).

BS b, 2022. Zusatzinformationen zur Geodaten-Shop Lieferung.

Chiang, C.-Y., Barnes, C., Angelov, P., Jiang, R., 2020. Deep Learning-Based Automated Forest Health Diagnosis From Aerial Images. IEEE Access 8, 144064–144076. https://doi.org/10.1109/ACCESS.2020.3012417

Dai, A., 2013. Increasing drought under global warming in observations and models. Nature Clim Change 3, 52–58. https://doi.org/10.1038/nclimate1633

Dobbertin, M., 2005. Tree growth as indicator of tree vitality and of tree reaction to environmental stress: a review. Eur J Forest Res 124, 319–333. https://doi.org/10.1007/s10342-005-0085-3

Dobbertin, M., Hug, C., Schwyzer, A., Borer, S., Schmalz, H., 2016. Aufnahmeanleitung. Kronenansprachen auf den Sanasilva- und den LWF-Flächen.

Džeroski, S., Panov, P.E., Ženko, B., 2009. Machine Learning, Ensemble Methods in. Machine Learning 542.

Eichhorn, J., Roskams, P., Potočić, N., Timmermann, V., Ferretti, M., Mues, V., Szepesi, A., Durrant, D., Seletković, I., Schröck, H.-W., Nevalainen, S., Bussotti, F., Garcia, P., Wulff, S., 2020. Part IV: Visual Assessment of Crown Condition and Damaging Agents. Version 2020-3. In: UNECE ICP Forests Programme Co-ordinating Centre (ed.): Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests. Thünen Institute of Forest Ecosystems, Eberswalde, Germany.

Franceschi, E., Moser-Reischl, A., Rahman, M.A., Pauleit, S., Pretzsch, H., Rötzer, T., 2022. Crown Shapes of Urban Trees-Their Dependences on Tree Species, Tree Age and Local Environment, and Effects on Ecosystem Services. Forests 13, 748. https://doi.org/10.3390/f13050748

Fraser, B.T., Congalton, R.G., 2021. Monitoring Fine-Scale Forest Health Using Unmanned Aerial Systems (UAS) Multispectral Models. Remote Sensing 13, 4873. https://doi.org/10.3390/rs13234873

Gottardini, E., Cristofolini, F., Cristofori, A., Pollastrini, M., Camin, F., Ferretti, M., 2020. A multi-proxy approach reveals common and species-specific features associated with tree defoliation in broadleaved species. Forest Ecology and Management 467, 118151. https://doi.org/10.1016/j.foreco.2020.118151

Gupta, M., 2022. Calculating Precision & Recall for Multi-Class Classification. Data Science in your pocket. URL https://medium.com/data-science-in-your-pocket/calculating-precision-recall-for-multi-class-classification-9055931ee229 (accessed 6.22.22).

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition.
IAP, 2021. Bonitierung der Stadtbäume: Erläuterung zum Bonitierungsbogen.
IAP d, 2022.
IAPa, 2022. IAP - Stadtbäume [WWW Document]. URL https://www.iap.ch/stadt.html (accessed 6.10.22).
IAPb, 2022. IAP - Stadt Slaz [WWW Document]. URL https://www.iap.ch/stadt_salz.html (accessed 6.10.22).
IAPc, 2022. IAP - Stadt Gesundheit [WWW Document]. URL https://www.iap.ch/stadt_gesundheit.html (accessed 6.10.22).
ICP, 2022. Large-scale forest condition monitoring (Level I) [WWW Document]. URL http://icp-forests.net/page/largescale-forest-condition (accessed 6.11.22).
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J Big Data 6, 27. https://doi.org/10.1186/s40537-019-0192-5
Kälin, U., Lang, N., Hug, C., Gessler, A., Wegner, J.D., 2019. Defoliation estimation of forest trees from ground-level images. Remote Sensing of Environment 223, 143–153. https://doi.org/10.1016/j.rse.2018.12.021
Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization.
Lehmann, J., Nieberding, F., Prinz, T., Knoth, C., 2015. Analysis of Unmanned Aerial System-Based CIR Images in Forestry—A New Perspective to Monitor Pest Infestation Levels. Forests 6, 594–612. https://doi.org/10.3390/f6030594
Leica, 2016. Leica ADS100 -Airborne digital sensor – airborne evolution.
Leica, 2011. Leica ADS80 - Airborne Digital Sensor - Digital Airborne Imaging Solution.
Lottering, R., Mutanga, O., Peerbhay, K., Ismail, R., 2019. Detecting and mapping Gonipterus scutellatus induced vegetation defoliation using WorldView-2 pan-sharpened image texture combinations and an artificial neural network. J. Appl. Rem. Sens. 13, 1. https://doi.org/10.1117/1.JRS.13.014513
Lüttge, U., Buckeridge, M., 2020. Trees: structure and function and the challenges of urbanization. Trees s00468-020-01964–1. https://doi.org/10.1007/s00468-020-01964-1
McPherson, E.G., van Doorn, N.S., Peper, P.J., 2016. Urban tree database and allometric equations (No. PSW-GTR-253). U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, Albany, CA. https://doi.org/10.2737/PSW-GTR-253
Otsu, K., Pla, M., Duane, A., Cardil, A., Brotons, L., 2019. Estimating the Threshold of Detection on Tree Crown Defoliation Using Vegetation Indices from UAS Multispectral Imagery. Drones 3, 80. https://doi.org/10.3390/drones3040080
Patel, R., Patel, S., 2020. A Comprehensive Study of Applying Convolutional Neural Network for Computer Vision. International Journal of Advanced Science and Technology 29, 15.
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., 2011. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON 6.
PyTorch a, 2022. Models and pre-trained weights — Torchvision 0.12 documentation [WWW Document]. URL https://pytorch.org/vision/0.12/models.html (accessed 7.3.22).
PyTorch b, 2022. Conv2d — PyTorch 1.11.0 documentation [WWW Document]. URL https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html (accessed 6.13.22).
PyTorch c, 2022. Linear — PyTorch 1.11.0 documentation [WWW Document]. URL https://pytorch.org/docs/stable/generated/torch.nn.Linear.html (accessed 6.13.22).

PyTorch d, 2022. StepLR — PyTorch 1.11.0 documentation [WWW Document]. URL https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.StepLR.html (accessed 6.19.22).
PyTorch e, 2022. Illustration of transforms — Torchvision 0.12 documentation [WWW Document]. URL https://pytorch.org/vision/stable/auto_examples/plot_transforms.html#sphx-glr-auto-examples-plot-transforms-py (accessed 6.19.22).
PyTorch f, 2021. Trained ResNet doesn't work in eval mode, behaves strangely - vision [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/trained-resnet-doesnt-work-in-eval-mode-behaves-strangely/121242 (accessed 6.21.22).
PyTorch g, 2022. BatchNorm2d — PyTorch master documentation [WWW Document]. URL https://pytorch.org/docs/master/generated/torch.nn.BatchNorm2d.html#torch.nn.BatchNorm2d (accessed 6.21.22).
PyTorch h, 2019. Performance highly degraded when eval() is activated in the test phase [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/performance-highly-degraded-when-eval-is-activated-in-the-test-phase/3323/33 (accessed 7.2.22).
PyTorch i, 2017. Model.eval() gives incorrect loss for model with batchnorm layers [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/model-eval-gives-incorrect-loss-for-model-with-batchnorm-layers/7561/3 (accessed 7.2.22).
PyTorch m, 2019. Model.eval() gives incorrect loss for model with batchnorm layers [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/model-eval-gives-incorrect-loss-for-model-with-batchnorm-layers/7561/41 (accessed 7.4.22).
PyTorch n, 2017. Performance highly degraded when eval() is activated in the test phase [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/performance-highly-degraded-when-eval-is-activated-in-the-test-phase/3323/2 (accessed 7.4.22).
PyTorch o, 2019. Model.eval() gives incorrect loss for model with batchnorm layers [WWW Document]. PyTorch Forums. URL https://discuss.pytorch.org/t/model-eval-gives-incorrect-loss-for-model-with-batchnorm-layers/7561/43 (accessed 7.4.22).
S. Jahromi, M.N., Buch-Cardona, P., Avots, E., Nasrollahi, K., Escalera, S., Moeslund, T.B., Anbarjafari, G., 2019. Privacy-Constrained Biometric System for Non-Cooperative Users. Entropy 21, 1033. https://doi.org/10.3390/e21111033
scikit a, 2022. sklearn.metrics.accuracy_score [WWW Document]. scikit-learn. URL https://scikit-learn/stable/modules/generated/sklearn.metrics.accuracy_score.html (accessed 6.13.22).
scikit b, 2022. sklearn.metrics.balanced_accuracy_score [WWW Document]. scikit-learn. URL https://scikit-learn/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html (accessed 6.13.22).
scikit c, 2022. sklearn.metrics.mean_absolute_error [WWW Document]. scikit-learn. URL https://scikit-learn/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (accessed 6.13.22).
scikit d, 2022. 3.3. Metrics and scoring: quantifying the quality of predictions [WWW Document]. scikit-learn. URL https://scikit-learn/stable/modules/model_evaluation.html (accessed 7.4.22).
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7
Shrivastava, I., 2020. Handling Class Imbalance by Introducing Sample Weighting in the Loss Function. GumGum Tech Blog. URL https://medium.com/gumgum-tech/handling-

class-imbalance-by-introducing-sample-weighting-in-the-loss-function-3bdebd8203b4 (accessed 7.4.22).
Valbuena, R., de Sevilla, T.F., Mauro, F., Pascual, C., García-Abril, A., Martín, S., Manzanera, J.A., 2008. Lidar and true-orthorectification of infrared aerial imagery of high Pinus sylvestris forest in mountainous relief 11.
Velasquez-Camacho, L., Cardil, A., Mohan, M., Etxegarai, M., Anzaldi, G., de-Miguel, S., 2021. Remotely Sensed Tree Characterization in Urban Areas: A Review. Remote Sensing 13, 4889. https://doi.org/10.3390/rs13234889
Vogt, J., 2020. Urban Forests: Biophysical Features and Benefits, in: Encyclopedia of the World's Biomes. Elsevier, pp. 48–57. https://doi.org/10.1016/B978-0-12-409548-9.12404-2
Weinstein, B.G., Marconi, S., Bohlman, S., Zare, A., White, E., 2019. Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. Remote Sensing 11, 1309. https://doi.org/10.3390/rs11111309
WSLa, 2022. Sanasilva forest health inventory - WSL [WWW Document]. URL https://www.wsl.ch/en/forest/forest-development-and-monitoring/sanasilva-forest-health-inventory.html (accessed 6.11.22).
Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307
Zhuo, X., Koch, T., Kurz, F., Fraundorfer, F., Reinartz, P., 2017. Automatic UAV Image Geo-Registration by Matching UAV Images to Georeferenced Image Data. Remote Sensing 9, 376. https://doi.org/10.3390/rs9040376