Geo-Information Science and Remote Sensing
Thesis Report GIRS-2019-24

Aerial Plant Recognition Through Machine Learning

A deep learning approach to identifying marsh marigold (*Caltha palustris*) from UAV imagery in Biesbosch National Park, the Netherlands

Sake Alkema

July 2019

WAGENINGEN
UNIVERSITY & RESEARCH

Aerial Plant Recognition Through Machine Learning

A deep learning approach to identifying marsh marigold (*Caltha palustris*) from UAV imagery in Biesbosch National Park, the Netherlands

Sake Alkema

Registration number: 930901011040

Supervisors:

dr. ir. Sander Mücher [1]

prof. dr. Devis Tuia [2]

dr. Sylvain Lobry [2]

A thesis submitted in partial fulfillment of the degree of Master of Science

at Wageningen University and Research Centre,

The Netherlands

July 2019

Wageningen, The Netherlands

Thesis code number: GRS-80436

Thesis report: GIRS-2019-24

[1] Wageningen Environmental Research (WENR)

[2] Laboratory of Geo-Information Science and Remote Sensing (GRS)

WAGENINGEN UNIVERSITY & RESEARCH

Summary

Proper nature management requires a good overview of the plant and animal species that are present in the area. While conventional forest inventories are able to provide accurate and detailed field information on biodiversity and species abundance, they are often time-consuming and costly. This is especially the case for areas that are hard to access like swamps or steep slopes. Recent improvements in hardware and software potentially allow partial automation for these practices. The professional deployment of unmanned aerial vehicles (UAVs) with high resolution sensors and developments in computer vision have made it possible to capture and analyze detailed images in ways that were previously impossible. This study attempts to use a convolutional neural network (CNN) for species recognition from UAV images, to potentially assist or replace field inventories. The target species for this pilot study was the marsh marigold (*Caltha palustris*), which grows in the Biesbosch National Park, the Netherlands. Its bright yellow flowers and reflective leaves allow for relatively easy recognition in the field, and as an indicator species its presence or absence gives insight in the status of the surrounding swampy habitat. Five plots of 70x70m were selected from UAV flights over a patch of grasslands (two plots) and willow forests (three plots) in the reserve. A total dataset of 405 image samples was extracted from these plots. Without ground-truth data of marigolds inside the plots, the images were annotated manually based on visual inspection. To allow comparison, two annotation sets were created: one for flowers and the second which also included the leaves. Two models with different prediction methods were used for comparison. One produced single classifications of marigolds per sample, while the other predicted the presence of marigolds on a 16x16 grid inside each sample, for better localization.

After training the two models for 1000 dataset iterations (epochs), the single predictions were clearly outperforming the grid predictions. With a poor Matthew's Correlation Coefficient (MCC) of 0.2 the grid prediction model was not used for further experiments. The single prediction model was able to reach an MCC of 0.62 (recall: 0.85 / precision: 0.72) after 1500 epochs, on the flower dataset. The dataset that included the leaves of the marsh marigolds performed slightly worse at an MCC of 0.52. Lastly, differences in the flower detection were assessed by separately predicting marigolds in grassland and forest plots. The forest samples showed similar results to the full dataset, but the model failed to detect marigolds in the grasslands. The largest limitation was the small sample size, which is likely the main cause for the poor grid predictions. The grid predictions model contained many parameters, which all had to be adjusted in the training process. In order to properly alter all these parameters, a lot of data are required. This model therefore proved too complex for a dataset with this low level of variation and was thus unable to provide reliable results. The lower number of parameters led to a lower complexity of the single prediction model, which enabled better results for this small dataset. The lower prediction accuracy of the whole plants versus the flowers was likely caused by annotation errors. The manual annotation of the leaves was much harder than flowers and thus more prone to misclassifications. The model very confidently predicted the marsh marigolds in darker patches of the forest plots. Often this was correct, but it also led to a number of false positives. Since the grasslands do not show any dark areas, the model consistently produced false negatives there. This indicates the model was likely to focus on the growing locations of the marsh marigolds, instead of the actual characteristics of the target species. It is premature to propose using this model as a replacement for in situ inventories, as lack of accuracy figures and ground-truth data for conventional practices prevent a direct comparison of these methods. Despite this, the model demonstrated the ability to reliably predict the presence of marsh marigolds in the forest samples, within the confines of this study. This pilot study shows that current techniques can already assist in current field practices. However, future research with reliable ground-truth data and a larger sample size is needed to explore the full potential of using neural networks for species recognition.

Table of Contents

1. Introduction

1.1 Problem Statement

Throughout the past decades the importance of biodiversity and dedicated measures according to the present vegetation types have been increasingly emphasized in sustainable nature management. It is widely acknowledged that proper management of a nature reserve requires an extensive overview of the presence and abundance of plant and animal species (Magurran, 2004). This is especially the case in areas where rare species or habitats are present. The European Union instated a program dedicated to identifying valuable natural habitats and species and assigning these a Natura 2000 status (European Commission, 2008). The Natura 2000 policy aims to restore and maintain characteristic and vulnerable habitats throughout Europe by mandating flora and fauna inventories at six year intervals to provide tailored management strategies (Vanden Borre et al., 2017). Covering over 18% of the terrestrial European Union area, these habitats provide an extensive database of species' presence and abundance (European Commission, 2018).

Despite the wealth of knowledge regular monitoring offers, thorough vegetation assessments are often hard to accomplish due to financial and time constraints. Flora and fauna inventories are still largely dependent on manual labor by visual inspection in the field, making it a cost-intensive operation (Förster et al., 2008; Vanden Borre et al., 2011). This is especially the case for areas that are relatively poorly accessible like swamps and steeply sloping areas (BIJ12, 2018; Schaminée et al., 1995). Conventional inventories in areas like these are conducted by walking line transects to give a global indication of presence of species within a rough proximity of the observer. This often results in a coarse point-grid of roughly 50x50 meters. This procedure does usually provide sufficient information, since the exact number of individual plants is less crucial than the knowledge that these species occur within the study area. Still this way of monitoring is very prone to error and highly time-consuming, leaving much room for improvement.

Automation of species assessment has high potential to make nature monitoring more efficient on both large and small scales. To some extent, remote sensing techniques have offered a means to supplement or replace manual vegetation mapping in the field. Digital classification methods have existed for decades, and have been widely used to classify land use, climatic zones and other environmental characteristics like vegetation types and agricultural crops based on satellite or aerial imagery (Rogan & Chen, 2004). Despite the successes on a large scale, multiple attempts at automating habitat classification on a more detailed and smaller scale have shown mixed results in the past (Lucas et al., 2015; Mücher et al., 2015; Vanden Borre et al., 2011). However, in recent years large developments in object-based classification offer new ways of identifying habitats, which in some cases already prove more reliable than previously assessed methods (Haest et al., 2017; Kampichler et al., 2010; Kooistra et al., 2014). Although habitat classification is certainly promising for nature management practices in general, detection of individual species is often necessary to devise suitable conservation measures for nature reserves. Therefore, habitat-based recognition might not always suffice as species-specific detection is needed, especially in cases where rare species or indicator species are crucial for proper management strategies (Prendergast et al., 1993). The step to an even more localized and detailed scale is needed to automate detailed measurements and mapping of these vegetation characteristics. Naturally, this level of analysis requires a matching level of high resolution source imagery at a detail level that has been unavailable or too expensive for most these purposes up until recent years.

Improvements in camera sensors and increased interest in local airborne monitoring and transport of goods have resulted in rapid technological advancements and increased affordability for unmanned aerial vehicles (UAVs). While versatile in its applications, these cost reductions have especially led to a

steep increase in the deployment of UAVs for monitoring purposes and aerial photography (Anderson & Gaston, 2013; Bryson et al., 2014). Relatively low cost, compact and high resolution sensors have made it possible to obtain images of study areas in <1 cm pixel resolution to analyze areas on a much higher detail level compared to satellite images or conventional aerial images (Linchant et al., 2015). This increase in spatial detail does, however, reduce the options for spectral imagery, since these sensors are currently only deployed on lower resolution sensors. This tradeoff between spatial resolution and spectral information shifts the focus for remote sensing to the textures in the image, instead of the spectral signatures.

Aside from the hardware improvements on the side of airborne photography, software development has also shown large potential in automation of image analysis by computer vision. Computer vision technology has been especially rapidly developing as a result of improvements in machine learning algorithms and the ability to afford and efficiently use more processing power than ever before. A wide variety of machine learning algorithms and methods with differing use cases and limitations have emerged, some of which have already been applied in ecology (Cutler et al., 2007; Peters et al., 2014). The past few decades, supervised machine learning techniques like support vector machines (SVMs), decision trees and random forest classifications have repeatedly shown increases in prediction accuracy in the environmental remote sensing field over more established and basic classification techniques (Camps-valls et al., 2013; Delalieux et al., 2012; Kooistra et al., 2014). An emerging branch of machine learning is the development of neural networks, which has been applied to tackle a wide variety of tasks with analyzing, predicting and generating different data types like text, audio and imagery. Convolutional neural networks (CNNs) have become a versatile tool for object recognition purposes due to their ability to recognize complex patterns in images when a sufficient number of training samples is provided (Ciresan et al., 2011; He et al., 2016). This makes CNNs highly suitable for remote sensing applications, which often involve large image datasets and address complex questions that cannot always be answered by traditional classification approaches (Zhu et al., 2017). Within ecology CNNs have successfully been deployed for species recognition from camera trap images (Chen et al., 2014) and for mobile applications that allow users to identify plants from their smartphone (Dyrmann et al., 2016). In the field of remote sensing these neural networks have found their use in detection and monitoring of agricultural crops (Krogh Mortensen et al., 2016), detecting land use change (Lyu et al., 2016) and vegetation classification (Kussul et al., 2017), among many other use cases and data types (Zhu et al., 2017).

These rapid developments in high resolution aerial imagery and computer vision algorithms offer many opportunities for increased automation of ecological studies and environmental monitoring (Dell et al., 2014). While some aspects of monitoring by machine learning have been examined by habitat classification, species-specific recognition of plants has been relatively under-explored as a nature management tool. Where image resolution was previously a limiting factor in the scale on which objects could be recognized, the highly detailed contemporary UAV images should allow for individual plant recognition. Adopting and combining the technological advancements could potentially allow for increased efficiency in vegetation surveys and could prove a valuable addition to conventional ecological monitoring.

1.2 Study objectives

A field campaign from 17 April 2018 using a UAV in the Netherlands resulted in high-resolution imagery on which marsh marigolds (*Caltha palustris*) are visible (Figure 1). This plant species is native to the Dutch wetlands and blooms in early spring with distinct bright yellow flowers, while most tree species are still leafless

Figure 1: Marsh marigolds (Caltha palustris) as seen in the field (top) and on the UAV images (bottom)

(Van Steenis, 1971). This makes it an ideal target species for an exploratory study to assess the feasibility of using machine learning algorithms to identify separate plant species from UAV images.

This pilot study aims to identify these marsh marigolds through machine learning by applying a CNN-based model, without assisting field information. A wide a variety of CNNs exist, so this study explores which fits its purposes. The individual plants in the images were found in two vegetation types with varying levels of visibility: open wet grasslands and willow forests. Two datasets were used for training: one only including flowers and the second dataset also including ground-truth information on the non-blooming marsh marigolds by including the leaves. An increase in localization precision usually makes for a harder prediction task. This study therefore explores the boundaries of localized flower predictions by testing predictions on two scales, by adopting two separate models. The first model focuses on a single classification per sample, while the second is used for localized predictions on a grid within the sample. The models are trained and tested using different parameters and configurations to determine the most suitable approach for recognizing *Caltha palustris* from the aerial images. This project intends to quantify and compare the performances of the two models, based on these different vegetation types, annotation datasets and (hyper)parameters.

The final results are assessed for their usability in the field to see whether these techniques might complement or replace manual forest inventories. Ultimately, this pilot study could function as a means to identify strengths and weaknesses of using deep learning as a tool for future vegetation monitoring in other areas or on a variety of different flora and fauna.

1.3 Research questions

To what extent is a convolutional neural network able to classify and localize marsh marigolds (*Caltha palustris*) from UAV images in Biesbosch National Park?

- Which deep learning architecture is most suitable for the purposes of this study?
- How do localization scales and model parameters influence model performance?
- To what extent does the detection accuracy differ between flowering and non-flowering plants?
- How is the identification accuracy of *Caltha palustris* affected by different vegetation types?

2. Data

2.1 Study area and materials

This study was conducted in the Biesbosch National Park, one of the largest Dutch national parks, located in the West of the Netherlands. The reserve is a floodplain area at the intersection between the Meuse and Rhine, two large rivers. With a total area spanning over 90 square kilometers it is the largest fresh water tidal zone in Europe (Struyf et al., 2009). In the past, the very strong tidal forces have transformed the Biesbosch into a network of small creeks, streams and ponds. As delta management projects cut off some connections to the coast, the tidal amplitudes have been reduced to a few decimeters, causing many of the wettest areas to transform into drier vegetation (van Emmerik et al., 2009). Despite this, the Biesbosch is a Natura 2000 site with different habitats which still mainly consists of wetlands, filled with mostly willow forests (H91EOA), swampy grasslands (H6120) and fields of reed (Dutch Ministry of Agriculture Nature and Food Quality, 2018). The rarity and size of these habitats causes it to contain a high biodiversity in both flora and fauna (Weeda et al., 2003). As the reserve has a protected status, the public is not allowed in most terrestrial sections, causing the area to mostly be traversed by boat and only few walking routes throughout the area. These characteristics of high natural value and poor accessibility make it both valuable and challenging to conduct a reliable vegetation inventory there.

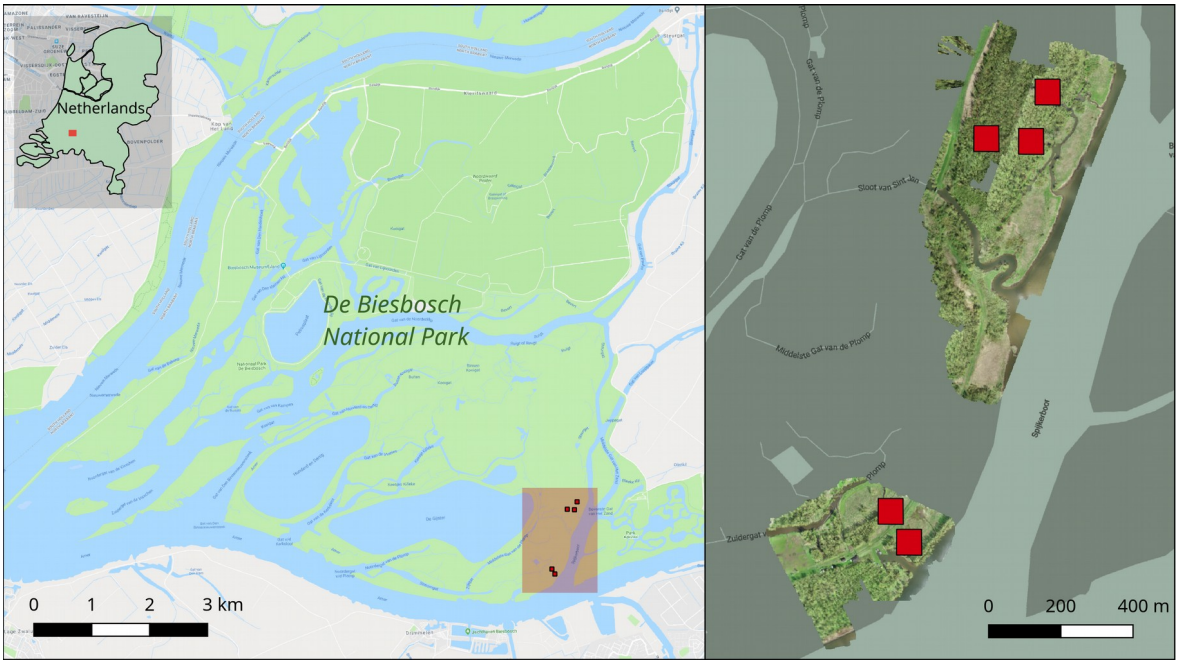


Figure 2: Map of the study area in Biesbosch National Park, The Netherlands, showing the UAV orthomosaics and selected plots used for sampling (red squares).

The study area contains mostly alluvial willow forest, combined with some open grasslands and fields of reed. The most dominant tree species in the forested areas are willow species *Salix alba* and *Salix viminalis*, combined with black poplar (*Populus nigra*), all thriving in partially flooded areas (Hennekes et al., 2010). The marsh marigolds are mostly found in large numbers alongside creeks that run through the forest. The grasslands found in the Biesbosch are known to contain a wide variety of grasses, sedges and herbs due to the moist soil and high influx of nutrients from the rivers into the reserve. Although generally easier to traverse, compared to the forests and reedlands, the swampy grasslands can occasionally be flooded as well. In the study area the marsh marigolds grow sparser in these grasslands, compared to the forests. The more homogeneous circumstances of the grasslands cause the distribution of the marigolds to be more spread out, often just showing single plants instead of the clusters found in the willow forests.

On the 17th of April 2018 an area of 46 ha was flown in the south-east of the reserve with a UAV (DJI Phantom 4 Pro+) at a flying altitude of 50 meters (Figure 2). This height ensured a safe margin above the highest treetops of roughly 30 meters. The images from four flights were processed and stitched in Agisoft Photoscan (Agisoft LLC, 2018) to produce orthomosaics with 1.5 centimeter pixel size. Flying in early spring ensured the canopy was still open in the forested areas, while the marsh marigolds would already bloom at most points, making them easier to detect for annotation. Despite this, many willows did already show their first leaves, complicating the visual inspection of marigolds in the images. The combination of low flying altitude and the tree cover did also result in a large number of visual distortions and stitching errors. Due to these restrictions in data quality, five plots of 70x70 meters were manually selected in QGIS3.0 (QGIS Development Team, 2017) in areas where marsh marigolds were visible and the data were not too severely distorted in the stitching process. Three of these plots were located in the willow forests and two in the grasslands (Figure 2), totaling a sample area of ~2.5ha. Although still containing some visual errors, these plots were the most promising and representative to serve as sample data for the models.

2.2 Annotation of marsh marigold

The models in this study were supervised and thus learn from examples that include marsh marigolds. However, no field data on the locations of marsh marigolds were available for quick selection of training and validation samples in the study area. In order to create the two sample datasets, manual annotation was necessary by systematic visual inspection (marigold leaves) and color-thresholding (marigold flowers). .

2.2.1 Flower annotation

To identify the individual blooming plants, image thresholding was applied by combining two filters on the plots. First, a yellowness index was used to filter out all non-yellow pixels by taking the average of red and green values and subtracting the blue pixel values (Equation 1). Since in the RGB spectrum yellow colors are a combination of high red and green, while having relatively low blue values. Only selecting the pixels exceeding 0.7 of this the yellowness index were selected and used for additional filtering. The yellowness index was normalized, which made it more robust against small differences in light circumstances between the different orthomosaics.

$$YI = \frac{Red + Green}{2} - Blue$$

Equation 1 Yellowness Index (YI) from RGB pixel values

The remaining thresholding steps were based on HSV, a different color space which expresses colors in hue, saturation and value. By iteratively altering these hue, saturation and value thresholds manually, only the bright yellow pixels were selected. The HSV threshold made it possible to locate the brightest flowers in the images, whereas the yellowness index showed a better ability to select more obscured flowers. Adjacent selected pixels were clustered and subsequently filtered by size by morphological opening (erosion & dilation operation) to exclude noise from surrounding trees or single selected pixels throughout the images. The centroids of the remaining segments were extracted for visual inspection to avoid errors in the annotation set. Over the five plots a total of 14699 points were selected by thresholding, of which 11,1% (1628) remained after manual filtering.

2.2.2 Leaves annotation

The annotation for the non-blooming plants could not be assisted by color-thresholding, since the leaves of the marsh marigolds did not have a distinct RGB color range, compared to other plant species within the study area. To systematically search the plots for marigolds a tight grid of points (32 pixels / ~0.5m distance) was used as an overlay on the plots. This grid allowed for faster and systematic manual selection and deletion of the areas that did not include any flowers or leaves of the marsh marigold, based on visual inspection. In total 104 thousand points were manually filtered with this procedure over the five vegetation plots.

2.2.3 Sample preprocessing

The two sets of remaining spatial points were converted to a ground-truth raster for each of the five plots. The UAV images of the plots and rasters were divided into smaller sample images of 512x512 pixels (59.0 m^2), which then served as the two training sets for the two prediction models (Figure 3). For the grid prediction model, the ground-truth images were down-sampled to 16x16 binary grids that indicated presence and absence of marigolds. The single prediction model simply needed one single binary value per sample image as ground-truth.

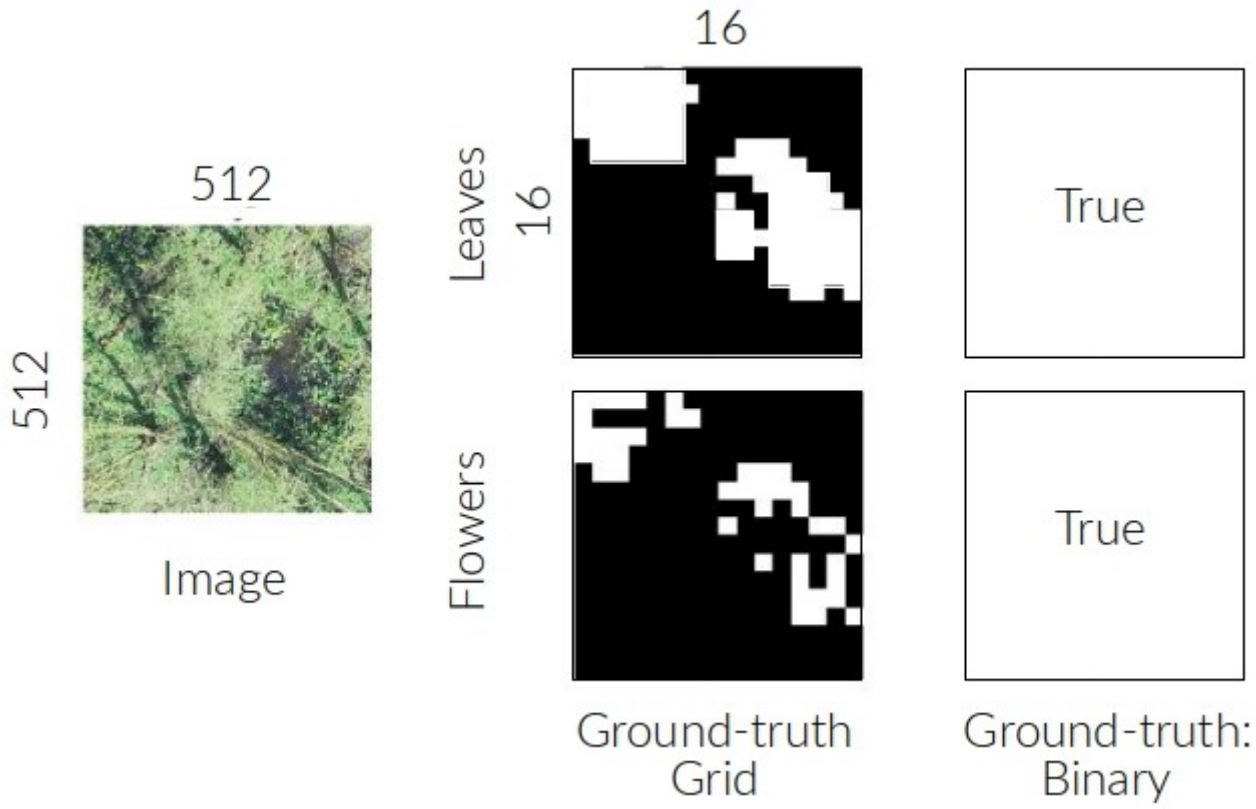


Figure 3: Example of a sample image and corresponding ground-truth sets for both the leaves as well as the flowers of the marsh marigolds.

This resulted in a total dataset of 405 sample images, from which 40,7% contained leaves and 38,8% contained at least one flower. However, when looking at the cell level of the grid samples only 2,3% percent of the cells showed leaves and just 0.9% contained flowers. Especially for the grid predictions this indicated a very large class imbalance, with an extreme disparity between background and marigolds.

2.3 Dataset division

For most experiments of this study, the dataset was randomly split into a training and validation set, which ensured the marigold/background ratio in the subsets would be roughly similar to the full dataset. In the training phase of the analysis, 70% of the total dataset were ran through the CNN to teach the model to classify marigolds. The remaining 30% of the samples was assigned to a validation set for the final predictions and independent model evaluations. After the training epoch (one training iteration of the full training dataset) has been completed, this validation set was used to assess the performance of the model using samples it is not familiar with. For the grid predictions model, entering these validation samples then results in a prediction grid of 16x16 cells where areas were flagged for presence and absence of *Caltha palustris*. The single prediction model only produces a binary value per sample. While this random division dataset was used in almost every step op analysis, some experiments required these data to be temporarily restricted or divided into non-random training and validation sets. This will be treated further in Chapter 2.3.2.

2.3.1 Curriculum learning

To enable the model to gain balanced knowledge of marigold and background characteristics, curriculum learning was applied for the grid predictions (Sung, 1996). The first stages of the training are very decisive for the model to be optimized for the specific classification task. Due to the class imbalance of the grid prediction dataset, there was a probability that the first training samples would not contain any marigolds, if the dataset would be entered randomly from the start. Curriculum learning is a method that essentially focuses on assisting the training by providing structured data that are either less complex or, in this case, more balanced in class distribution (Bengio et al., 2009). For the grid predictions, experiments with curriculum learning were conducted where the model was initially trained on images that contained at least one flower (Kellenberger et al., 2018). If necessary, this could allow the model to gain a better perception of marigold characteristics, instead of background. After this initialization, the rest of the full dataset would then be entered in a random order to adapt the model to a larger variety in background and a smaller percentage of flowers to reduce the number of predicted false positives.

2.3.2 Performance per vegetation type

The other case where the random division of the samples was not used, was for the analysis of the model performance of the different vegetation types. To fairly assess the accuracy of the targeted vegetation type and derive any statistically sound conclusions, sufficient samples of those plots needed to be located in the validation set. Therefore, a non-random division of the samples was chosen by assigning one entire plot with the corresponding vegetation type as validation set. In these cases, the ratios of training and validation samples were still close to 70% and 30%, respectively. This approach of using a single plot for validation also allowed easier interpretation of the predictions, since the outputs of the models could be stitched into a larger area and overlayed on the UAV orthomosaic, instead of single images.

3. Methodology

3.1 Framework and architecture

Neural networks have allowed classify and recognize classes from imagery, based on their ability to learn to recognize features, instead of relying on predetermined and programmed rules. This characteristic makes them very versatile and able to tackle many different visual problems. Neural networks get their name for their slight similarity to the functions of neurons in the human brain. While this comparison is not very accurate in reality, the main connection is the ability to improve solving computational tasks when more examples (samples) are observed. In this respect neural networks have shown the ability to distinguish many different objects in images if sufficient training data is fed into them, as mentioned in Chapter 1.1. In this study a convolutional neural network (CNN) is used. CNNs are especially well-equipped for handling visual recognition tasks, due to their convolutional layers. These layers are responsible for using filters to learn features and checking where these features occur within images (Goodfellow et al., 2016; Kellenberger et al., 2018). In the last step of a typical convolutional layer, pooling functions collect these filters and use these to downsample the images by only maintaining a summary statistic for each location in an image. This helps reduce the computational load of the following layers and makes it more robust to small changes in the image features. The more series of convolutional layers a CNN contains, the more complex a model becomes (often called deep neural networks), with a higher theoretical number of features to extract. The final part of a CNN consists of one or multiple fully-connected layers, which link the presence of the most prominent extracted features to a class and are therefore responsible for the classification of the image (Goodfellow et al., 2016).

Neural networks contain a large number of parameters to train, depending on their complexity. To properly generalize to unseen data, they are usually heavily dependent on a large training dataset with sometimes thousands of images per label class to properly adjust these parameters. As the image dataset for this project was very small, this low total number of samples on itself would be insufficient for the model to train properly. A large risk of such a small sample size in training a neural network is overfitting. This occurs when the model is trained not only to recognize the general sample features, but also the outliers and noise in the samples. This results in very high prediction accuracy on the training set, but poor performance on the validation set. To mitigate this issue data augmentation is often used, which artificially increases the variation in the data. The samples in this study were also subject to data augmentation by randomly rotating (50% chance) the sample images in four directions (90 degree turns) and flipping those horizontally. Additionally, the images had a 75% chance of having the hue, saturation and contrast enhanced or diminished. This made for a total of 80 potential variations of every single sample image. Despite these data augmentation operations, the variation in the sample dataset would remain relatively small and additional measures are necessary to avoid overfitting of the model.

As this study served as a pilot for species recognition in nature management, the implementation should be relatively straight-forward. The most commonly-used programming language in applied geoinformatics is Python and was therefore most suitable for this study. A number of deep learning frameworks are available for Python of which PyTorch and TensorFlow are the most commonly used. Until several years ago, Google TensorFlow has been the dominant framework in commercially deployed deep learning applications. In academia, however, PyTorch is now often seen as a preferred alternative due to its easier implementation and flexibility (Ketkar, 2017). To allow relatively simple alteration of the used models for potential future studies, PyTorch was adopted as framework for this study.

In computer vision technology there has been a large focus on the development of (near) real-time detection, resulting in architectures like You Only Look Once (YOLO)(Redmon et al., 2016) and Faster R-CNN (Ren et al., 2015) for applications in autonomous systems and analysis of live sensor feeds. For

these systems a high processing efficiency and therefore fast detection is crucial. This priority on speed allows videos to be rendered in near real-time, but is offset by a lower prediction accuracy compared to more computationally heavy architectures. For the purposes of this study prediction accuracy was deemed more important than speed, due to the static nature of the data (UAV images). Despite this, computation restraints did still apply as access to powerful processing clusters should ideally not be required for the training process. This meant deeper neural networks could be explored in this study for better model performance, as long as model optimization methods were considered to remain within the capabilities of single-GPU desktops.

Prediction methods differ quite strongly, based on the applications of the neural networks. The most basic image prediction method is classification, where a single object class is predicted over an entire sample image (Campos-Taberner et al., 2016; Penatti et al., 2015). Although more complex, bounding box predictions offer better localization, by classifying and pinpointing objects within an image. More recently image segmentation has been explored for very precise localization tasks by accurately outlining the classified objects (Maggiori et al., 2017; Volpi & Tuia, 2017). However, when specific locations or concrete outlines of the detection classes are not necessary, more basic prediction methods are simpler and faster to implement. For the detection of *Caltha palustris* just one class had to be recognized. Segmentation or other very precise prediction methods was unnecessary and probably unfeasible, given the lack of accurate ground-truth data. This simplified the detection process as simply predicting presence or absence of the target species would suffice, which prevented over-complication of using more specific but computationally heavy and complex detection methods.

A study with comparable goals to this project was aimed at automatic detection of wild mammals from UAV imagery in Namibia (Kellenberger et al., 2018). Similar to the marsh marigolds, the mammals did not have to be segmented or marked by a bounding box, as well as having a relatively small dataset available for training. That particular study focused on the problems that arise when a dataset shows a heavy imbalance in classes, where the background class is much more frequently encountered than the animals. This same disproportion of classes was also found in the marsh marigold dataset as these plants often grow in clusters, causing large parts of the images to not contain any marigolds. Only 2.3% of the ground-truth pixels contained leaves and only 1,0% contained flowers. This complicates predictions on a grid level, as there are relatively few target plants for the model to train on. For single predictions per sample image this class imbalance is much less severe. On an image level the flowers occurred in 38,8% of the samples and 40,7% contained leaves.

3.2 Models and components

The detection models developed and used by Kellenberger et al. (Kellenberger et al., 2018) were partially applied in this study. This approach relied on an adapted residual neural network (ResNet), which allows deeper neural networks to be trained without suffering from a number of optimization issues, as seen with conventional “plain” networks (He et al., 2016), as explained below.

In theory adding more layers enables CNNs to recognize features of higher complexity, thus making them more powerful predictors. In practice, however, deeper networks (>10 layers) were found to be subject to larger training and validation errors. A cause for this are vanishing or exploding gradients, which refers to the issue of weight factors becoming infinitely small or large, due to the large number of gradient adjustments as result of the many hidden layers (Bengio et al., 1994). Another reason for high errors in deeper networks is a degradation problem, where optimization becomes substantially harder, thus hampering the model performance (R. K. Srivastava et al., 2015). The ResNet architecture was developed to avoid this problem by reintroducing activations from shallow layers into the deeper layers, thus preventing the CNN from ignoring earlier results and avoiding exploding/vanishing gradients. The applied architecture for this study was based on ResNet-18, which consists of 18 layers. This ResNet-model was pre-trained on the commonly used ImageNet classification dataset, which includes 1.3 million

images from 1000 object categories not specifically related to environmental sciences or remote sensing (Russakovsky et al., 2015). This pre-training ensured the CNN was already capable of distinguishing rough and distinct features in images, without specific object classes. This also lowered the risk of overfitting on training data and was faster than fully training the CNN from self-provided datasets (Yosinski et al., 2014). To this pre-trained model a number of final layers was added. The study-specific marigold class was subsequently taught in these final layers of the model (Goodfellow et al., 2016), based on the marigold ground-truth datasets for this study. These so-called fully-connected layers are trained for classifying the features found by the pre-trained ResNet. A first fully-connected layer (FC1) with dropout regularization and non-linear activation was introduced after the ResNet (Figure 4). Dropout regularization is a technique that causes each neuron to be left out with a given probability (in this study 50%) for every epoch during training (N. Srivastava et al., 2014). By including this technique, the risk of overfitting became smaller as only robust features remain consistent predictors when neurons were randomly dropped throughout the training cycles. Activation functions ensure the outputs of a layer stay within a certain value range. A non-linear activation called the rectified linear unit (ReLU) is currently the most-used function in deep learning (Gulcehre et al., 2016). ReLU activations simply ensure all outputs are converted into values above 0, in a way that allows the values to be adjusted again in a later stage. This makes it a very simple, but effective and fast activation function. The second and final fully-connected layer (FC2) of the CNN has a sigmoid activation to ensure all outputs fell in a range between 0 and 1, corresponding with absence and presence of marigolds. Dimensions and number of parameters for the two models are shown in Appendix A.

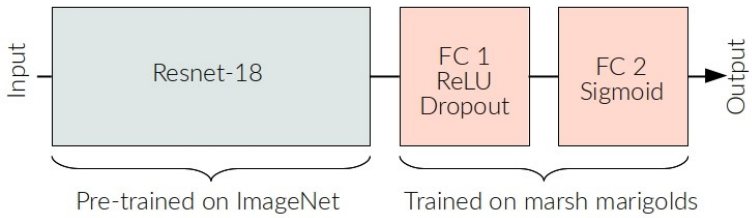


Figure 4: Schematic overview of the CNN model used for the classification of marsh marigolds.

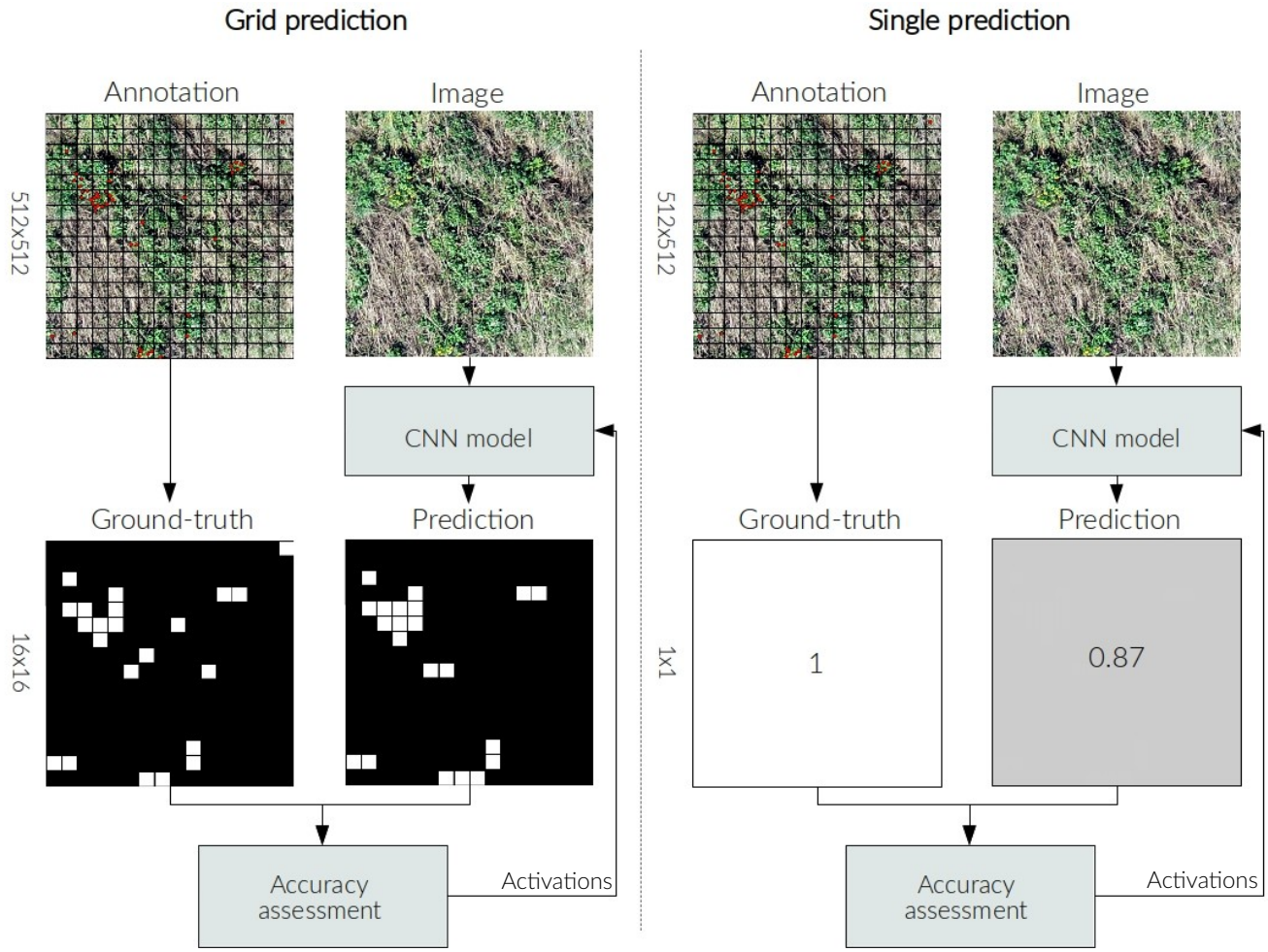


Figure 5: Schematic depiction of the two models for the detection of marsh marigolds using sample images and ground-truth for 16x16 prediction grids and single predictions.

The samples were passed through the model in batches of 24 images, which is faster than entering single samples at a time. This also enables more robust changes to the parameters of the model, since there is more variation in a batch of samples compared to single samples. After the final fully-connected layer of the model (FC2) each of these samples was assigned either a binary prediction value or a grid of predictions, depending on the model configuration (Figure 5). Due to the sigmoid activation of the final layer, the predictions would always be a value between 0 and 1, based on the certainty of the model that marigolds were present or absent in the assessed sample. A high prediction would correspond with a high probability of marigolds occurring in that sample, while values close to 0 would indicate a high chance of absence. In this study, the model would start out with predictions close to 0.5 in the early training stages, since there was still a large uncertainty on which characteristics it should base its predictions on. In later epochs the model should be able to predict more reliable values.

During training, the internal assessment of model performance is evaluated through errors in the predictions, which constitute what is called a loss function. Higher discrepancies between prediction and ground-truth lead to a higher loss. After a training epoch, this loss is then used to adjust the model parameters in an attempt to reach a lower loss in the next epoch. This iterative process enforces the model to focus on the right characteristics of the marigolds in order to achieve the lowest error rates and thus the lowest loss. In this study the loss is calculated by taking the binary cross-entropy (BCE-loss) between prediction and ground-truth (Equation 2).

$$Loss = W \cdot (-(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y}))$$

Equation 2: Loss calculation of binary cross-entropy between prediction (ŷ) and ground-truth (y), W is for class weights.

Most of all the binary cross-entropy loss suited the single-class predictions of this study, as only absence or presence of marigolds had to be predicted. Another reason to choose this loss function was the possibility of adding class weights to the loss calculation. Since there was a large imbalance in presence and absence of marigolds in the grid samples, the model was likely to overtrain on the background class and therefore be prone to predict substantially more false negatives. To overcome this imbalance, Kellenberger et al. (Kellenberger et al., 2018) propose using the inverse of the class frequency as a weight factor in the loss function of the CNN. This weight factor is calculated by taking the number of marigold instances and dividing this by the total number of background instances. These weights ensure a realistic proportion of plants and background samples can still be used for training, but the model is penalized harder for missing marigolds (false negatives) than errors in background predictions (false positives). This procedure should allow the model to better train on characteristics and patterns of the target plants instead of focusing on surrounding vegetation.

The loss of the predictions is then used to adjust the internal parameters of the CNN, after which the next batch of samples will be passed through the model with the new parameters. This process was repeated until the full dataset has been analyzed, after which the next epoch starts and the cycle repeats. The total number of epochs was based on the increments in prediction accuracy. If the model showed no further improve after several epochs, the training would be stopped and the final results were taken from the predictions on the validation set.

3.3 Validation metrics

In the validation phase, the internal parameters of the model were not adjusted to avoid teaching the model to recognize specific details of the set, instead of general characteristics of the marigolds. In this phase, the Matthew's Correlation Coefficient (MCC) was used as the main metric to assess the performance of the model. In machine learning this coefficient is often preferred over regular confusion matrices or accuracy percentages in the case of a class imbalance (Boughorbel et al., 2017). A confusion matrix counts all correct (true positives and true negatives) as well as all false predictions (false negatives and false positives). If the model were to predict absence of marigolds, while they were actually present, this would count as a false negative. The inverse situation would count as a false positive. The MCC takes all components of a confusion matrix into account to compensate for the large proportion of background and provide a single metric that allows for easy comparison to other configurations of the models (Equation 3). The MCC is a value on the scale from -1 to 1, where 0 represents random predictions. The aim is to reach an MCC as close as possible to 1, which corresponds with a set of perfect predictions.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Equation 3: Matthew's Correlation Coefficient (MCC) from true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN)

However, the ratio between false positives and false negatives is very important in order to make a fair comparison with conventional vegetation monitoring. Since the MCC only consists of a single value, it is not possible to derive these underlying metrics for its calculation from this value alone. A common way to

assess the performance of a prediction model in machine learning is to look at the recall and precision of the outputs (Davis & Goadrich, 2006). Recall is the ratio of instances where marigolds were correctly predicted, i.e. the percentage of flowers that were recalled by taking the number of true positives and dividing this by the total number of marigolds from the ground-truth (Equation 4). The precision is the metric that assesses the reliability of the flower predictions by taking the true positives and dividing that by the total of instances of marigold predictions. Oftentimes the final results from a prediction model show a trade-off between recall and precision, where high recall leads to many false positives and a good precision leads to more false negatives (Powers, 2007).

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP}$$

Equation 4: Calculation of prediction Recall and Precision, based on true positives (TP), false negatives (FN) and false positives (FP)

By default, outputs which contained high prediction values (>0.5) were marked as marsh marigolds (1) during training, whereas images where these plants are absent served as the background class (0). To further assess the model performance, this threshold was then shifted to every value between 0 and 1 to influence the precision and recall of the predictions. Plotting these values into a precision-recall curve gives further insight into the robustness of the model predictions.

.

4. Results

4.1 Model performance using different prediction methods and parameters

To assess which configuration of the model would obtain the best prediction results, a variety of model parameters was used for both single predictions and 16x16 prediction grids. The model was trained in a series of different parameter combinations to assess the impact of each variable and selecting the best performing model. Due to the complexity differences in the model architectures of single predictions and grid predictions (Figure 5), this parameter calibration was performed separately for both detection methods.

First, to compensate for the larger class imbalance for the grid predictions, the effect of curriculum learning for that model was assessed. With curriculum learning, the training set was reduced to only include samples that at least contained one flower to ensure the model would potentially accelerate the learning process of the first epochs. Inclusion of curriculum learning, however, did not result in better results, compared to training on the full dataset as the use of class weights seemed to sufficiently outweigh the class imbalance for the training (Appendix B). The more balanced ratio (158/247) between flowers and background made testing curriculum learning unnecessary for the single predictions.

The remaining tested parameters for both single and grid predictions were learning rate, class weighting and different detection thresholds. After calibrating these parameters a direct comparison of both prediction models was made to determine whether both models perform well enough to continue testing the remaining research questions. The parameters were tested over 1000 epochs to make a fair assessment of the performance increments throughout the training.

4.1.1 Learning rate

The learning rate is the parameter that affects how much the internal weights of the model are adjusted during the training. A higher learning rate enables the model to make large adjustments to allow for a quick learning process. On the other hand, these large changes in weights might become too coarse in later stages of training, thus never reaching the optimal accuracy. This trade-off is a trial and error process that differs for each model and dataset. The effects of learning rates on the model performance for the both prediction methods were tested in four variations. The tested learning rates for the models were 1e-3, 1e-4, 1e-5 and 1e-6 (Figure 6). Default inverse frequency class weights (single: 2.5 / grid: 106.0) were used in these experiments.

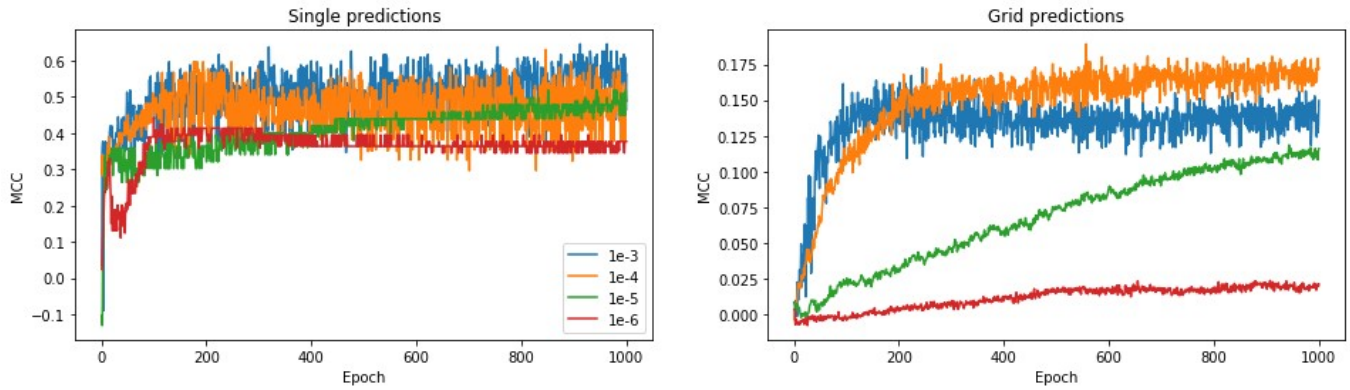


Figure 6: MCC curves over 1000 epochs of training using different learning rates (1e-3, 1e-4, 1e-5, 1e-6). Learning curves of these four rates are shown for both the single prediction (left) and grid prediction models (right).

For the single prediction model, the highest learning rates of 1e-3 and 1e-4 showed very similar results with an to ability to reach high MCC values of above 0.6. However, these learning rates do show an extreme variability in the prediction performance throughout the training process. Although the training using a learning rate of 1e-5 took was slower, the performance was much more stable and managed to reach a maximum MCC of 0.52 after 1000 epochs, while still showing potential for further increase in prediction accuracy. The lowest learning rate initially showed this same stability in the first 200 epochs, but the MCC leveled off and even performed a bit worse in the later epochs with an MCC of 0.36. Considering these results, the learning rate of 1e-5 proved the most stable and reliable for the remaining experiments.

Contrary to the single predictions model, the grid predictions model performed optimally at higher learning rates. The model reached the best results with learning rates of 1e-3 and 1e-4, both peaking at an MCC of approximately 0.16. Experiments with learning rates of 1e-5 and 1e-6 were not able to obtain the same results as the higher learning rates, even after prolonged training, staying behind at MCCs of 0.12 and 0.02, respectively. While the 1e-3 setup leveled off earlier in the training process, the graphs show a more balanced training process for the learning rate of 1e-4, with less outliers throughout the training.

4.1.2 Class weights

Class weights were introduced to account for the class imbalance between marigolds and background in the dataset, in addition to curriculum learning. This was necessary for the grid predictions, since leaving out the class weights resulted in a model that was completely unable to identify any marigolds and predicted background in every location. Initially the inverse of the marigold class frequency was used as the class weight, resulting in a 106 fold multiplication for grid predictions. Especially for the grid prediction model this would ensure misclassifications of marigolds were penalized much harder, compared to background misclassifications. However, the severity of the imbalance caused false positives to result in a lower loss, causing the model to greatly overpredict the number of flowers and misrepresent their presence. This factor thus overcompensated for the class imbalance. A number of less strict class weights were tested to assess the impact of this variable on the performance. The inverse of the class frequency was divided by factor 2 and 3, in an attempt to reduce the number of false positives of the default weights.

By dividing the class weights by factor 2 (53.0), the maximum MCC of the grid predictions stayed roughly the same around 0.2 (Table 1). The reduced class weights only led to a very minor improvement in precision at the expense of a small decline in recall. Using the lowest class weights (35.5) did not lead to any additional improvements in model performance. The altered class weights did not result in sufficient improvement to prefer these parameters for further experiments for the grid predictions (Appendix C).

Table 1: Performance metrics of the grid predictions model using three class weighting factors. The inverse of the marigold frequency in the samples was used, alongside this value divided by 2 and 3.

Grid prediction model	Class weights		
	Inverse frequency	2x	3x
MCC	0.19	0.21	0.2
Recall	0.68	0.65	0.59
Precision	0.07	0.09	0.09

Although the class imbalance was very subtle for the single predictions, the differences of using class weights were tested to assess if further improvements in the recall and precision were possible using this model (Table 2). For the single prediction the regular class frequency (2.5), halved weights (1.26) and no class weights (1) were tested. While using the higher class weights were able to slightly improve the precision of the single predictions, it did not outweigh the negative impact on the recall. Also in the case of this model using the inverse class frequency as class weights offered the best balance between recall and precision.

Table 2: Performance metrics of the single predictions model using three class weight settings. The inverse of the marigold frequency in the samples was used, alongside this value divided by 2. In a third scenario, no class weights were assigned.

Single prediction model	Class weights		
	Inverse frequency	2x	No weights
MCC	0.52	0.46	0.40
Recall	0.79	0.64	0.52
Precision	0.67	0.70	0.71

4.1.3 Detection threshold and prediction methods

Another way of altering the ratio between recall and precision is by changing the detection threshold. Generally in binary classification a detection threshold of 0.5 is applied, meaning all predictions above that value are considered a marigold prediction and all values below belong to the background class. By adopting a higher threshold, only the values with higher prediction certainty are considered as flower predictions. This approach would generally lead to a lower number of false positives, since the uncertain predictions would then be moved to the background class.

The predictions of the best performing configurations of both the single and grid models were extracted and compared to the ground-truth values. The detection threshold was then shifted by steps of 0.01 between 0 and 1 to get one hundred instances of precision and recall throughout these threshold values. Plotting these datapoints produced a precision-recall curve for each of the two models, which offers an overview of how these two prediction approaches perform (Figure 7).

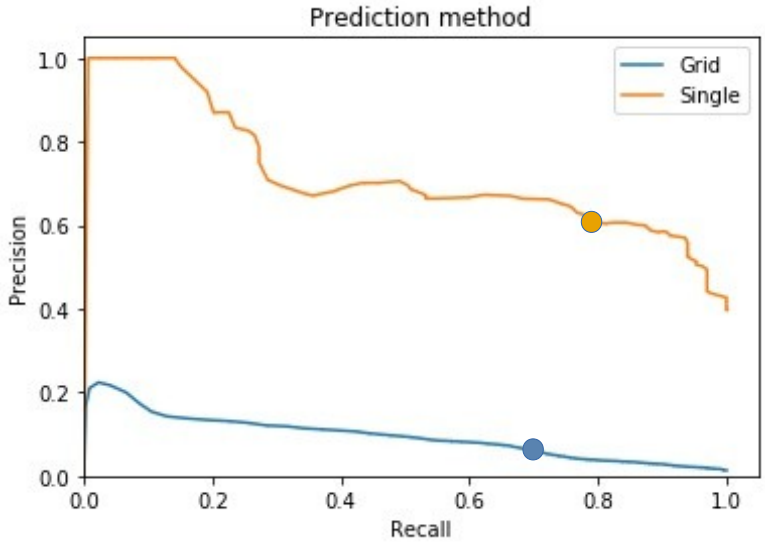


Figure 7: Precision-recall curves comparing the two models for grid (blue) and single (yellow) predictions of marsh marigolds. Points show a detection threshold of 0.5.

The bad fit of the grid prediction model became very clear in the precision-recall curve as even the highest thresholds were not able to exceed a precision of 0.23. Over the entire curve the model predictions resulted in a very large number of false positives. As such, there was not an ideal detection threshold where the grid predictions would produce a reliable assistance in the detection of marsh marigolds with this model.

Given the difference in model complexity, it was expected the single predictions model would outperform the grid predictions at every step. The differences in performance between the models was very high, with large consequences for their usefulness as predictors. Whereas the grid predictions model was very prone to overpredict the abundance of marsh marigolds, the single predictions consistently offered both a higher precision and recall. Even a recall rate of 100% still produced a precision above 0.4.

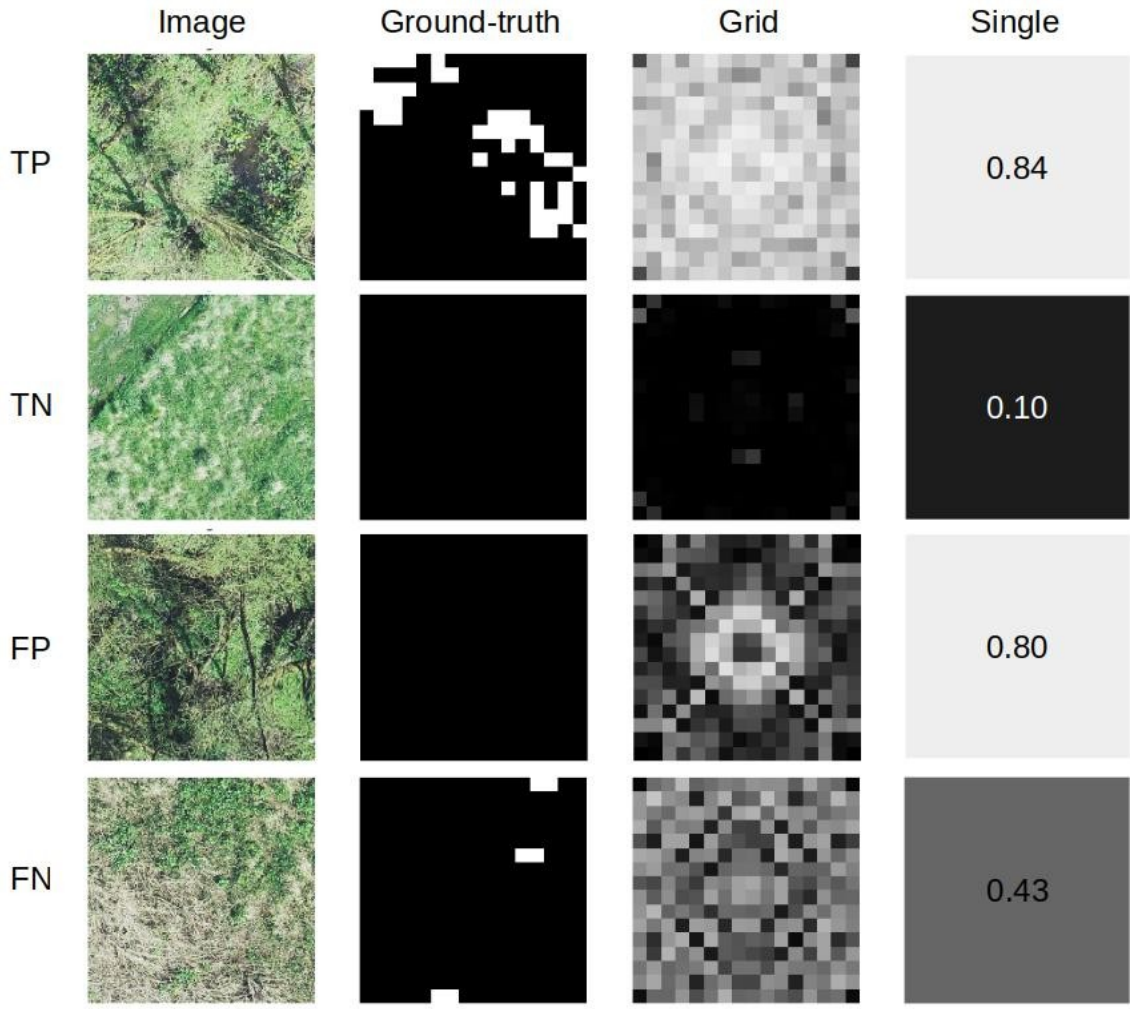


Figure 8: Examples of correct and false predictions of the grid (3rd column) and single prediction models (4th column). True positive (TP), true negative (TN), false positive (FP) and false negative (FN) outputs are depicted next to the corresponding UAV images and ground-truth masks, given a threshold of 0.5.

This performance difference becomes very clear when the predictions are visualized (Figure 8). The grid prediction model defaults to a pattern that is not able to offer any form of localization of the marigolds. The relatively simple model for the single predictions is therefore able to produce much more accurate results, at the expense of less precise localization of the marsh marigolds.

Given the substantial decrease in prediction accuracy of the grid prediction model compared to the single prediction model, the remaining research questions were answered by only using the single predictions.

4.2 Prediction accuracy using flowers or whole plants

The sample images were annotated twice to create a dataset for merely the flowers of the marsh marigold and one including the leaves as well. These two different annotation sets were used separately to training the model for 1500 epochs on the full dataset. For both annotation datasets the same training parameters were used: a learning rate of 1E-5 using default class weighting.

The model trained on the flowers still managed to improve its performance, compared to the results after 1000 epochs as presented in the previous paragraph. With these additional 500 epochs, the flower predictions reached a maximum MCC of 0.62, with a recall of 0.85 and a precision of 0.72 at a detection threshold of 0.5.

The training on the whole plants performed fairly similar to the flowers in terms of recall (0.79) and precision (0.72). However, inclusion of the whole plants did not manage to reach the same MCC after leveling off at 0.52 (Appendix D). This difference occurs due to a slightly lower overall accuracy (69% versus 81% for flowers), since the recall and precision of the background class are taken into account for the calculation of the MCC as well.

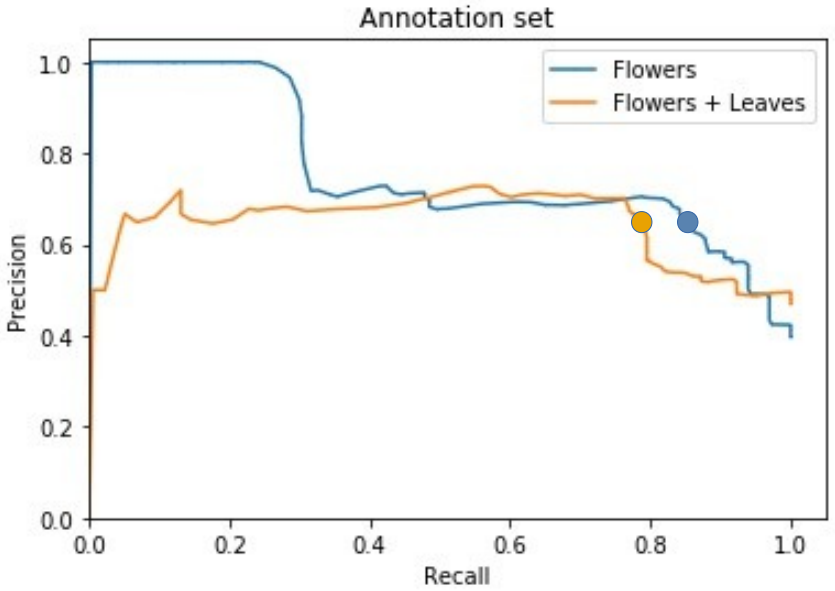


Figure 9: Precision-recall curves comparing the predictions of flowers (blue) and whole plants (yellow) of marsh marigolds. Points show a detection threshold of 0.5.

The precision-recall curve clearly shows that recall and precision are fairly comparable in the top half of the recall range (Figure 9). If precision rates above 0.8 are required to compete with field inventories, the flower dataset is able to reach these accuracies, given a lower recall. At recall rates below 0.4 the flower annotation jumps to maximum precision, showing the model is able to very accurately predict one-third of the flower samples. The annotation set of the entire plants is not able to provide this same baseline and actually has a lower precision in the lower recall ranges. This indicates the predictions of the model are more robust when trained on the flower samples than on both flowers and leaves, showing flowers are probably more distinct to the model than the leaves. The problem for recognition of distinguishing the leave structure from the background is most likely too complex, compared to the flowers. At a recall of

80% the model is able to classify flowers with a precision of above 70%, while the other annotation set stays below 60%.

4.3 Effect of vegetation types on accuracy

The forest and grassland plots were tested independently after the model trained on the entire flower dataset. Ideally, the model would be trained on a single vegetation type, but the small sample size of this study did not allow further division of the dataset. In both instances one of the five complete plots was used as the validation set. The best predictions for both vegetation types were stitched and visualized alongside the ground-truth points on a scale from 0 (background) to 1 (flower).

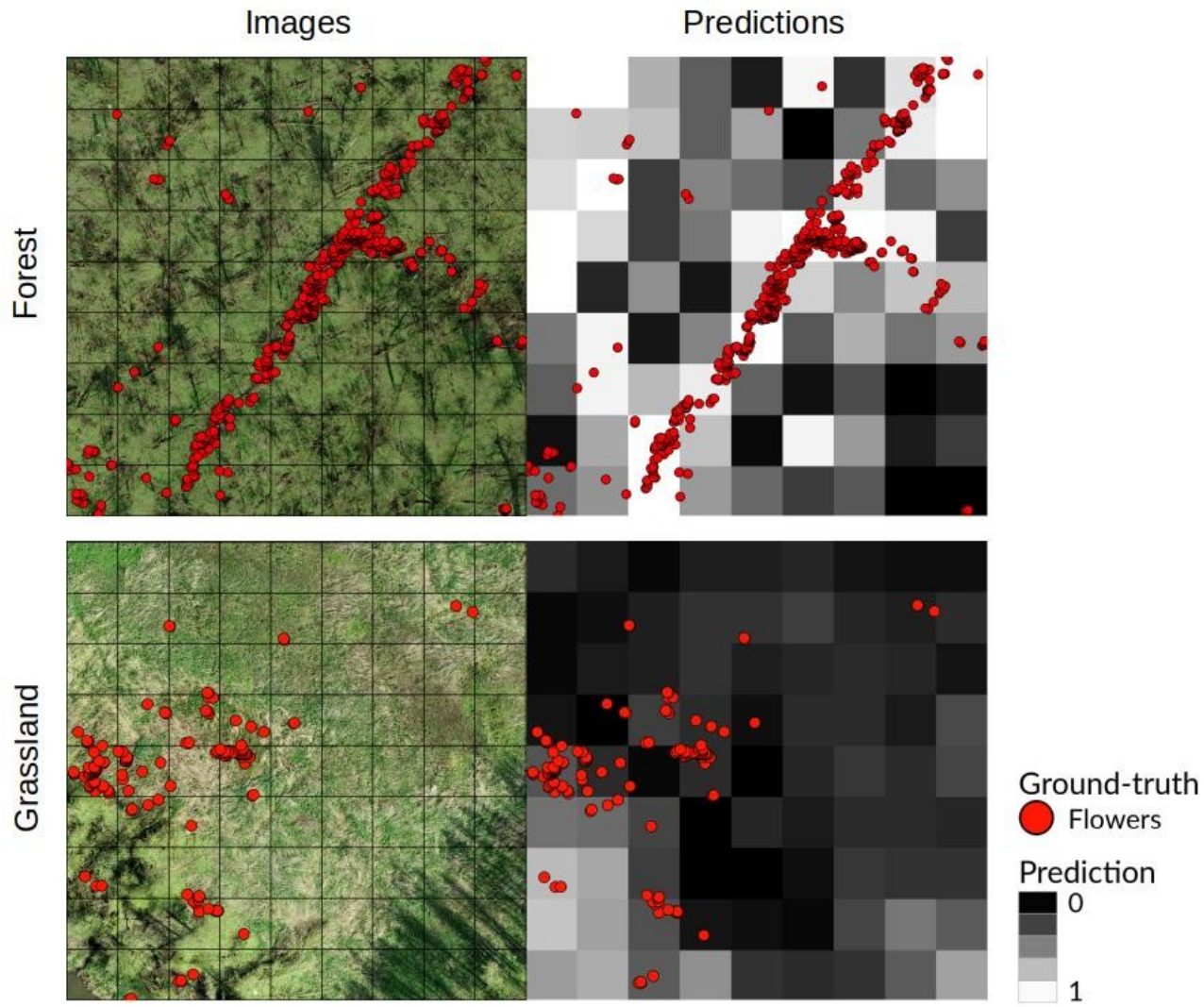


Figure 10: Stitched predictions and ground-truth data of flowers of marsh marigolds in a willow forest plot (top) and grassland plot (bottom)

The forest plot performed well with a MCC of 0.55 (Appendix E). The recall (0.87) and precision (0.7) of the forest predictions were comparable to the results of the full dataset. The tiles containing large numbers of flowers are consistently predicted correctly (Figure 10). The patches where less flowers are present show a less reliable accuracy, as seen in the lower left corner of the forest plot. The upper left corner shows the model is still prone to overpredict in some cases as well, leading to a number of false positives.

Contrary to the forest plot, the model was unable to generate any reliable grassland predictions. After the first few epochs the model very quickly defaulted to a high rate of false negatives, with the exception of the treeline in the bottom corners of the grassland plot. This caused the MCC to fluctuate around 0, which corresponds with random predictions. These values thus indicate that the model was unable to train on any marigold characteristics in the grasslands. Based on these results the CNN shows a very clear difference between the two vegetation types in terms of prediction accuracy, with only forest plots providing valuable outputs.

5. Discussion

5.1 Model configurations and prediction methods

Throughout the tuning of the training parameters, the grid prediction model was unable to match the results of the model that produced single predictions. While a drop in overall prediction accuracy was expected in exchange for better localization, the grid predictions did not improve the exact localization of the marsh marigolds. The number of internal parameters within the neural network was most likely far too high for the small sample size available in this study, thus overcomplicating the prediction model for the grid predictions. By opting to go for a single prediction, the number of adjustable parameters within the model was divided by 256, thus making the model much less complex. This difference certainly helped improve the single predictions to reach an MCC of 0.52, while the grid predictions leveled off at 0.21.

With a recall of 0.68 and a precision of 0.07 at a detection threshold of 0.5, the grid predictions were not able to offer any form of improved localization of the marigolds in the plots, compared to the single predictions. Moreover, the vast number of false positives caused the grid predictions model to perform even slightly worse than the initial RGB thresholding operation for the annotation. This simple combination of color selection, clustering and filtering was already able to outperform the grid predictions in both recall and precision, scoring 0.98 and 0.11, respectively. This comparison, however, only holds true for the flower annotation, since RBG thresholding was not possible for the selection of the leaves of the marsh marigolds. Nevertheless, it showcases the poor performance from the grid prediction model in this experimental setup and severely limited sample size.

Machine learning models are often restrained by the size of the training dataset, which is dependent on the number of parameters to learn. Not having access to the sufficient training samples to is known to strongly limit the accuracy and feasibility of training neural networks. Comparable to this study, Kellenberger et al (Kellenberger et al., 2018) were able to make reasonable grid predictions with a sample size of 654 images. With 405 samples of which only 153 included flowers of the marsh marigold, this was most likely a major factor in the low predictive power of the grid prediction model in this study. While data augmentation was used to effectively multiply the sample size by factor 40 and avoided overfitting, this could not fully compensate for a larger dataset as the total variation between samples was much lower compared to unique samples. Grid predictions might still become be a viable improvement over the single predictions if more UAV images with marsh marigolds become available, but additional flights and analyses would be necessary to fairly assess these possibilities. A quick assessment of the grid predictions model was performed by using a larger, roughly annotated, dataset of cows in grassland (N=4548). After training for 100 epochs (learning rate 1E-3) the model was able to produce substantially higher prediction accuracies, compared to the marsh marigold dataset (Figure 11; Appendix F). A maximum MCC of 0.32 was reached, at a recall rate of 0.85 and a precision of 0.12.

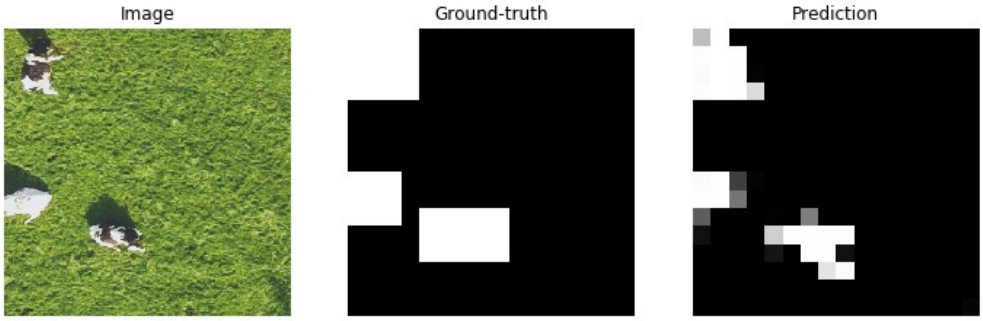


Figure 11: Example of the 16x16 prediction grid model on a dataset of cows.

These predictions do prove the model itself was valid and able to provide a better localization, compared to the single predictions model. These results emphasize the need for a larger sample size to train the larger number of variables for the grid predictions.

Despite not being able to produce a precise localization within the samples, the single predictions model performed much better. Given the size of the 512x512 pixel samples, the flowers were still localized in an area of less than 8 by 8 meters in the field. The precision-recall curve shows that even at the highest recall rates the precision never seems to fall below 0.4. Likewise, the model is able to detect 30% of the flowers with a perfect precision. This simplified model was therefore a much more promising tool as an alternative or addition to current species inventories in the field.

5.2 Annotation methods

In terms of precision and recall the two annotation sets performed similarly. The flowers annotation set showed a higher maximum MCC (0.62), compared to the annotation that included the entire plants (0.52). For human observers the flowers are much easier to recognize than the leaves, and this could therefore also be the case for the neural network. This increase in total prediction accuracy might therefore simply be caused by a simpler identification task of the bright yellow flowers. Although the leaves are highly reflective and show slightly more shadows than surrounding vegetation, the light circumstances and surroundings can cause the leaves to look very differently in different situations.

The same applies to the other annotation set, but the total variation of the flowers is most likely smaller than the leaves. The difference in prediction accuracy between the annotation sets might also be the result of the annotation accuracy. The bright yellow flowers are very distinctive and straightforward to locate, especially with the preselection of points by RGB thresholding. As mentioned in chapter 2.2, the annotation task for the entire plants was much more difficult and therefore very prone to bias from the observer. The combination of the image resolution, surrounding vegetation and stitching distortions especially made annotation of the leaves substantially more challenging. The ecological expertise of the observer responsible for the image annotation could have large consequences on the introduced errors, especially in the grasslands. Any misclassifications inside the annotation set will inevitably have led to a less robust prediction model, thus resulting in reduced accuracy for the whole plants compared to the flowers. A solution to this would have been to only introduce samples where visual interpretation of marsh marigold was deemed most reliable, but this would have severely reduced the already low number of samples available and affect the variability in the data.

5.3 Grassland and forest predictions

Validating the test results on the two vegetation types in the study plots, made it possible to evaluate the differences in detection of marsh marigolds between grasslands and willow forests. The separate validation set for the forests performed similar to the mixed dataset, with a maximum MCC of 0.55, a recall of 0.87 and precision of 0.7. The trained neural network, however, was not able to retain any marigold characteristics from the flowers in the grasslands. With the more homogeneous surroundings of the low vegetation, the MCC did not increase above the baseline of 0 after prolonged training. These results clearly show the inability of this model to identify the flowers in these plots.

These poor results could simply indicate a more challenging recognition task of the flowers in grasslands, compared to forest. The higher light intensity within the grasslands could make the brightness of the flowers less distinct, as well as more yellow and brown plant species (e.g. reed and moor-grass) in the surrounding vegetation. Based on the sampling plots, the marigold flowers do seem to grow in cluster more often in forests than in the grasslands, possibly due to more homogeneous growing circumstances in the open fields. Samples containing more flowers have a higher chance to be correctly identified by a neural network, so this might make recognition in the grasslands more complex. This seems somewhat contrasting to conventional fieldwork, where marigolds are more easily spotted in

grasslands, compared to forests. This can be attributed to the difference in perspective, as swamp forests are generally harder to access and offer a more closed habitat, which obscures the view. While the view can still be more obscured beneath canopies in the UAV images, the accessibility and line of sight is much less of a limitation in spotting the marigolds.

In the predictions for the grassland plot, as seen in Figure 10, the model consistently predicts background in the open meadows and only predicts flowers where trees or their shadows were present. These predictions are most likely the result of the training set, which largely consisted of forest plots in this case. As only two out of the five study plots contained grassland, just one plot was available for the training samples, while the other served as the validation set. This meant merely 25% of the training samples contained grassland, which probably made the characteristics of the flowers in the forest plots much more dominant factors when adjusting the parameters of the model. For the forest predictions this balance was a bit better, with 50% of the training set still consisting of forest samples. Ideally, to make a fair comparison between vegetation types, the model would train on samples of that specific vegetation type only. The very limited sample size for this study, however, did not allow this, as only 162 grassland samples and 243 forest samples were available in total. Fully leaving out the forest plots in the training process of the grasslands caused the model to immediately overfit from the first epochs. The same occurred for the training of the forest predictions, when both grassland plots were left out. Additional data would be needed to fairly assess whether the difference in prediction accuracy between the vegetation types was caused by the difficulty of the recognition task or whether the sample size was the main effect of these results.

5.4 Putting best results into perspective

The best results in predicting marsh marigolds in the UAV images of the Biesbosch were achieved with single predictions for the flower dataset, with a total precision of 0.72. At a detection threshold of 0.5 the model was able to correctly locate 85% of the marigolds in the samples. Given the limited sample size and lacking data quality these results are very positive and certainly show potential for future application in vegetation monitoring. Despite this, the CNN model still showed its flaws and the trade-off in recall and precision is still a large consideration for the interpretation of the predictions. If a near-perfect accuracy is required for deployment in practice, this model will be able to locate a third of all the flowers in its current configuration. On the other hand, if practically none of the flowers are allowed to be left out of the predictions, more than half of the predictions would consist of false positives.

When looking at the forest predictions as presented in Figure 10, the model is often able to find the locations of large clusters of marsh marigolds. Nevertheless, some very confident predictions were produced by the model in dark forest patches where marsh marigolds were completely absent. These strong false positives consistently seemed to occur in the darker forest patches, whereas false negatives seemed to be located in brighter samples. This might indicate that the model was not simply focusing on the characteristics of the marsh marigolds themselves, but on their natural growing conditions. Since most of the marigolds in the forest patches were located in and around the small streams that cross the forest, the model is likely to have picked up on this relation to some extent. This focus on spots of dark canopy gaps could partially explain the poor performance of the model in the grasslands, since no streams and bare soil were present there. In that case, the only darker spots were the shadows from the treeline in the corners of the grassland plot, which did trigger a number of false positives. This “black box” effect is an often-used criticism to the application of neural networks, as there is regular uncertainty on which features the model has trained. The simplest ways to influence this is by interpreting the predictions and adjusting the training samples accordingly to steer the model into training on the right features (Koh & Liang, 2017; Olden & Jackson, 2002). To some extent, the poor localization power of the grid prediction model might have also been affected by this focus on growing conditions, instead of

marigold characteristics. Nevertheless, the grid predictions showed such a large discrepancy from the ground-truth that this was unlikely to been the main obstacle for that model.

Since the grid predictions model was not able to provide any improved localization within the samples, the best scale at which the marigolds are predicted is roughly 8x8m. This spatial scale should be comparable to the current accuracy at which target species are often recorded in vegetation monitoring, given these are usually mapped using handheld GPS devices. These measurements are usually subject to an accuracy of between 5 and 9m below forest canopies (Rodríguez-Pérez et al., 2007). This also means that regular field measurements probably would not have sufficed as ground-truth for image recognition purposes, without the need for visual interpretation and annotation of the images. On the other hand, these data points would have been helpful as support and validation for the manual annotation. For ground-truth data to have actually been a replacement to the annotation in this study, the measurements would have to be taken by high-precision GPS devices or visually delineated in the field for easier observation in the images.

This study was completely conducted without any available field information for accurate ground-truthing, meaning all annotation tasks were completed based on visual interpretation of the UAV images. Aside from influencing the robustness of the model by inevitably introducing errors in the annotation set, this manual annotation also has implications for uncertainties in the presented results. This sampling method is undeniably subject to a selection bias where only the clearly visible plants are identified and used as training and validation data. More obscured flowers either by tree canopy or other surrounding vegetation might would have been missed in the visual interpretation of the images, while field inventories might still be able to locate it. This observer bias does not directly affect the performances of the neural network, so within the scope of this pilot study this is acceptable. It does, however, restrict the ability to make a direct comparison with in situ measurements. This effect could have been partially mitigated by capturing the UAV images a month earlier, when the willow trees would still have been totally bare. Since the marsh marigold is known to bloom very early in spring, most of the plants would have already been blooming in March. This could have made the annotation task easier, whilst also showing individual marigolds that were now already obscured by the tree canopy. However, this would have also led to longer shadows due to the lower angle of the sun, potentially complicating the recognition task for the model.

Another factor that makes it difficult to balance species detection by deep learning against field measurements, is a lack of concrete accuracy figures for conventional field visits. Regular are generally deemed very accurate on a small scale, but without revisits by different observers in the same season, it is impossible to express these measurements in recall and precision rates. Fitzpatrick et al. (2009) found that up to 25% of target species could either be missed or falsely classified by experts in the field, which indicates the model does not need to produce perfect predictions in order to replace field studies. These field errors would also have consequences if they would have been used as ground-truth data for the model (Carlotto, 2009; Foody, 2009). In current practice for species assessment, field observations are often extrapolated to a much broader scale (50x50m grids) to indicate presence and absence of species within the study area. Given these heavy extrapolations and the density of the occurring flowers, it would generally be preferred to miss an individual plant than to risk falsely identifying it, in certain cases. For the results of the neural network this leads to a situation where false negatives would potentially be preferred over false positives, thus favoring a high precision over a high recall. Given the uncertainty of the accuracy of field inventories, the detection threshold will need to be shifted to find the optimal use for the model as assistance or replacement in the field.

6. Conclusions

6.1 Implications

The main findings of this study are answered with regards to their respective research questions. This will elaborate on the extent to which this CNN-based model is able to classify and localize marsh marigolds and what its implications are for conventional field monitoring.

How do localization scales influence model performance?

Given the results of this study, it is feasible to use this ResNet-based model as a reliable predictor for the presence of marsh marigolds in UAV images. Despite that grid predictions are not able to provide any detailed localization of the marigolds within the samples in this study, the single predictions per sample look very promising. Based on this work, the model can give an indication on where this species is likely present on a localization scale of roughly 8x8 meters. It is likely that a larger sample size would have made it possible to gain better localization of marsh marigolds from grid predictions.

To what extent does the detection accuracy differ between flowering and non-flowering plants?

Ideally, whether the plants are blooming or not, all individual marsh marigolds should be included in the predictions. However, the annotation set that includes the entire plants performs slightly worse than the model trained on merely recognizing the bright yellow flowers of the marsh marigold. This does not necessarily indicate an inherently more difficult recognition task for the model, but could be related to errors introduced in the annotation of the leaves. It is possible to still use the annotation for the non-blooming plants, but a larger error margin in the predictions needs to be taken into account. Of course, this decline in prediction accuracy does reduce the ability of the neural network to potentially replace field measurements.

How is the identification accuracy of Caltha palustris affected by different vegetation types?

The model shows the best ability to predict the presence and absence of marigold flowers within the willow forests, a vegetation type which is especially hard to access by foot. This therefore is also where the model has the largest potential to contribute to field monitoring of vegetation. The results show a strong tendency to predict marigold presence in dark patches of the samples. This indicates that the model was likely to have been trained on recognizing marsh marigold growing habitats, instead of the marigolds characteristics. This would explain the poor performance of the model within grassland areas, contrary to the forest plots. Due to uncertainty about how well the predictions of the model generalize over a larger area with varying vegetation, it is premature to propose using this model as a full alternative to conventional field inventories. This is also related to the unknown accuracy of current field measurements and the lack of reliable ground-truth data in this study.

Despite the mentioned limitations of this study, the results are a testament to the flexibility of neural networks and show that current technologies can already provide valuable assistance to field practices. By scheduling exploration flights with the UAV, the current model can be able to provide an early selection of areas where the marsh marigolds are most likely to occur (Appendix G). Using this model as a supplement, could therefore allow field visits to become more cost-efficient and less time-consuming.

6.2 Recommendations

Given the constraints and limitations of this study, especially in terms of data quality and quantity, there is a lot of potential to further explore the capabilities of neural networks in species detection and vegetation monitoring. Additional research is necessary to fully explore these possibilities of using machine learning techniques for species recognition from UAV images. This study merely focused on a very small aspect of deep neural networks and machine learning in general. Moreover, even within the

confines of this study many uncertainties require further investigation. Based on the findings of this pilot study, some structural changes in adopted methods and data are recommended for future research.

- The available data for this pilot study is the largest limitation for the training of the model. While the UAV orthomosaics from the flights are fairly large, the amount of usable data for training and testing purposes is very limited. The stitching operation of the original images resulted in many image distortions, which caused a large area of the orthomosaics was to be either unusable or unrepresentative for samples of the marsh marigold. A slower flying speed with more overlap of images might result in better end results if the images are stitched. Instead of an orthomosaic, georectified separate images do also suffice as samples and would leave more usable data for a larger sample size.
- The lack of ground-truth data from the field inevitably introduced errors in the annotation of the images. A way to achieve more and better data would be to establish study plots in the field where thorough ground-truth data are collected, together with accurate spatial information of the marigold locations within the plot. This would allow the model to train on much more accurate field information and make it easier to compare field inventories with the model predictions from UAV images.
- The UAV images also partly consisted of fields of reed, where Marsh marigolds are also found. Considering Biesbosch National Park consists for a substantial part out of reedlands, this would make it an interesting habitat to include in the detection models. While grasslands and willow forests were included in the dataset, the reedlands of the Biesbosch had to be left out, since there were too few marigolds to thoroughly test the accuracy of the model in this vegetation type. Reedlands are, however, very hard to monitor in the field, so it could be valuable to include more plots to further assess the model.
- Since the marigolds are very small, the 1.5cm resolution of the UAV images were unable to preserve much detail of the plant characteristics. A lower flying altitude could make the training of the neural network more robust by allowing it to train on higher resolution samples for easier distinction between the marigolds and surrounding vegetation. However, this lower flying altitude (<50 m) might not be possible for some areas when a safe margin above the highest treetops cannot be guaranteed. Wherever this would not be a concern a higher resolution could be beneficial to the recognition task and training of the neural network. Also if other, even smaller, plant species are targeted, this might be necessary.
- The images used in this study were limited to red, green and blue bands. In vegetation monitoring other bands within the infrared spectrum are often included in the imagery. Including these bands in the imagery of future studies on species recognition might assist the model in distinguishing separate species. This might also help with the annotation of the target species, if its spectral signature deviates from others. Additional other data types like Light Detection And Ranging (LiDAR) data could also prove useful to include vegetation structure as a factor in the classification of larger plant species or species confined to specific habitats. For small species like the marigold which grow in multiple habitats, however, this would probably not be of added value.
- While neural networks are very promising and currently applied in a wide range of applications, machine learning offers completely different techniques as well. It might be interesting to explore other classification methods like Random Forest models on this same marigold dataset to make a comparison of the prediction results with the deep learning model in this pilot study.

- This pilot study only explores the possibilities of using a single architecture and methodology that is deemed most appropriate given the available data, time and resources. This study is only focused on the recognition and classification task of the marsh marigold. Other deep neural networks are often adopted to provide more complicated segmentation tasks to provide more detailed localization of their target subjects. Segmentation, however, is a harder detection problem and would require more and higher quality data. Thus, if more images and reliable ground-truth data become available, segmentation might become a very interesting option to improve the current results. Another option might be to train a model on near real-time detection of the marigolds to give instant feedback to the field personnel, if enough training samples would be available. This method would require a completely different architectures like YOLO or Faster-RCNN with bounding box detection.

Acknowledgements

First of all, I would like to thanks Sander Mücher for the chance to conduct this thesis at Wageningen Environmental Research and his valuable supervision throughout the project. I am very grateful for the time and effort Sylvain Lobry invested into me to help me with my first steps into deep learning; this thesis would not have been possible without the technical advice and guidance he and Devis Tuia offered. Thanks to Jappe Franke, Henk Kramer and Benjamin Kellenberger for the knowledge exchange and the valuable remarks on my methodology and results. This thesis was facilitated by the SPECTORS project, which enabled data acquisition and knowledge exchange between WENR and Bureau Waardenburg. Lastly, I really appreciate the help from Xadya for providing the necessary mental support during this and previous theses.

References

Agisoft LLC. (2018). Agisoft PhotoScan. *Professional Edition, Version 1.3.5*. Retrieved from http://www.agisoft.com/downloads/installer/

Anderson, K., & Gaston, K. J. (2013). Lightweight unmanned aerial vehicles will revolutionize spatial ecology. *Frontiers in Ecology and the Environment*, *11*(3), 138–146. http://doi.org/10.1890/120150

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (pp. 1–8). http://doi.org/10.1145/1553374.1553380

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. http://doi.org/10.1109/72.279181

BIJ12. (2018). Natuurkwaliteit Natuurnetwerk beoordelen. Retrieved September 24, 2018, from https://www.bij12.nl/onderwerpen/natuur-en-landschap/monitoring-en-natuurinformatie/subsidiestelsel-natuur-en-landschap/beoordeling-natuurkwaliteit-natuurnetwerk/

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*, *12*(6), 1–17. http://doi.org/10.1371/journal.pone.0177678

Bryson, M., Reid, A., Hung, C., Ramos, F. T., & Sukkarieh, S. (2014). *Cost-effective mapping using unmanned aerial vehicles in ecology monitoring applications*. *Springer Tracts in Advanced Robotics* (Vol. 79). http://doi.org/10.1007/978-3-642-28572-1_35

Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Saux, B. Le, … Tuia, D. (2016). Processing of Extremely High-Resolution LiDAR and RGB Data: Outcome of the

2015 IEEE GRSS Data Fusion Contest–Part A: 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(12), 5547–5559. http://doi.org/10.1109/JSTARS.2016.2569162

Camps-valls, G., Tuia, D., & Bruzzone, L. (2013). Advances in hyperspectral image classification. *IEEE Signal Processing Magazine*, *31*(1), 45–54. http://doi.org/10.1109/MSP.2013.2279179

Carlotto, M. J. (2009). Effect of errors in ground truth on classification accuracy. *International Journal of Remote Sensing*, *30*(18), 4831–4849. http://doi.org/10.1080/01431160802672864

Chen, G., Han, T. X., He, Z., Kays, R., & Forrester, T. (2014). Deep convolutional neural network based species recognition for wild animal monitoring. In *2014 IEEE International Conference on Image Processing, ICIP 2014* (pp. 858–862). http://doi.org/10.1109/ICIP.2014.7025172

Ciresan, D., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, High Performance Convolutional Neural Networks for Image Classification. *International Joint Conference on Artificial Intelligence (IJCAI) 2011*, 1237–1242. http://doi.org/10.5591/978-1-57735-516-8/ijcai11-210

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. http://doi.org/10.1890/07-0539.1

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). http://doi.org/10.1145/1143844.1143874

Delalieux, S., Somers, B., Haest, B., Spanhove, T., Vanden Borre, J., & Mücher, C. A. (2012). Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers. *Remote Sensing of Environment*, *126*(1), 222–231. http://doi.org/10.1016/j.rse.2012.08.029

Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., de Polavieja, G. G., Noldus, L. P. J. J., … Brose, U. (2014). Automated image-based tracking and its application in ecology. *Trends in Ecology and Evolution*, *29*(7), 417–428. http://doi.org/10.1016/j.tree.2014.05.004

Dutch Ministry of Agriculture Nature and Food Quality. (2018). Biesbosch. Retrieved March 25, 2019, from https://www.synbiosys.alterra.nl/natura2000/gebiedendatabase.aspx?subj=n2k&groep=9&id=n2k112

Dyrmann, M., Karstoft, H., & Midtiby, H. S. (2016). Plant species classification using deep convolutional neural network. *Biosystems Engineering*, *151*(1), 72–80. http://doi.org/10.1016/j.biosystemseng.2016.08.024

European Commission. (2008). *NATURA 2000: protecting Europe's biodiversity*. (S. Wegefelt, Ed.) (1st ed.). Brussels: European Commission.

European Commission. (2018). Natura 2000. Retrieved September 24, 2018, from http://ec.europa.eu/environment/nature/natura2000/index_en.htm

Fitzpatrick, M. C., Preisser, E. L., Ellison, A. M., & Elkinton, J. S. (2009). Observer bias and the detection of low-density populations. *Ecological Applications*, *19*(7), 1673–1679. http://doi.org/10.1890/09-0265.1

Foody, G. M. (2009). The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, *30*(12), 3275–3281. http://doi.org/10.1080/01431160902755346

Förster, M., Frick, A., Walentowski, H., & Kleinschmit, B. (2008). Approaches to utilising QuickBird data for the monitoring of NATURA 2000 habitats. *Community Ecology*, *9*(2), 155–168. http://doi.org/10.1556/ComEc.9.2008.2.4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. *MIT Press*. http://doi.org/10.1038/nmeth.3707

Gulcehre, C., Moczulski, M., Denil, M., & Bengio, Y. (2016). Noisy activation functions. In *International conference on machine learning* (pp. 3059–3068).

Haest, B., Vanden Borre, J., Spanhove, T., Thoonen, G., Delalieux, S., Kooistra, L., … Kempeneers, P. (2017). Habitat mapping and quality assessment of NATURA 2000 heathland using airborne imaging spectroscopy. *Remote Sensing*, *9*(3), 266. http://doi.org/10.3390/rs9030266

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://doi.org/10.3389/fpsyg.2013.00124

Hennekes, S., Smits, N. A. C., & Schaminée, J. H. J. (2010). SynBioSys Nederland. Alterra, Wageningen UR. Retrieved from https://www.synbiosys.alterra.nl/synbiosysnl/help/synbiosys.pdf

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics*, *5*(6), 441–450. http://doi.org/10.1016/j.ecoinf.2010.06.003

Kellenberger, B., Marcos, D., & Tuia, D. (2018). Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, *216*, 139–153. http://doi.org/10.1016/j.rse.2018.06.028

Ketkar, N. (2017). Introduction to PyTorch. In *Deep Learning with Python*. http://doi.org/10.1007/978-1-4842-2766-4_12

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 1885–1894). JMLR. org.

Kooistra, L., Kuilder, E. T., & Mucher, C. A. (2014). Object-based random forest classification for mapping floodplain vegetation structure from nation-wide CIR and LiDAR datasets. In *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (Vol. June, pp. 1–4). http://doi.org/10.1109/WHISPERS.2014.8077590

Krogh Mortensen, A., Dyrmann, M., Karstoft, H., Nyholm Jørgensen, R., & Gislum, R. (2016). Semantic Segmentation of Mixed Crops using Deep Convolutional Neural Network. In *International Conference on Agricultural Engineering* (p. 259). Retrieved from www.elementar.de

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 778–782. http://doi.org/10.1109/LGRS.2017.2681128

Linchant, J., Lisein, J., Semeki, J., Lejeune, P., & Vermeulen, C. (2015). Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, *45*(4), 239–252. http://doi.org/10.1111/mam.12046

Lucas, R., Blonda, P., Bunting, P., Jones, G., Inglada, J., Arias, M., … Mairota, P. (2015). The earth observation data for habitat monitoring (EODHaM) system. *International Journal of Applied Earth Observation and Geoinformation*, *37*, 17–28. http://doi.org/10.1016/j.jag.2014.10.011

Lyu, H., Lu, H., & Mou, L. (2016). Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sensing*, *8*(6), 506. http://doi.org/10.3390/rs8060506

Maggiori, E., Charpiat, G., Tarabalka, Y., & Alliez, P. (2017). Recurrent Neural Networks to Correct Satellite Image Classification Maps. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(9), 4962–4971. http://doi.org/10.1109/TGRS.2017.2697453

Magurran, A. E. (2004). *Measuring of Biological Diversity*. *Blackwell science*. http://doi.org/10.2989/16085910409503825

Mücher, C. A., Roupioz, L., Kramer, H., Bogers, M. M. B., Jongman, R. H. G., Lucas, R. M., … Blonda, P. (2015). Synergy of airborne LiDAR and Worldview-2 satellite imagery for landcover and habitat mapping: A BIO SOS-EODHaM case study for the Netherlands. *International Journal of Applied Earth Observation and Geoinformation*, *37*, 48–55. http://doi.org/10.1016/j.jag.2014.09.001

Olden, J. D., & Jackson, D. A. (2002). Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*(1–2), 135–150. http://doi.org/10.1016/S0304-3800(02)00064-9

Penatti, O. A. B., Nogueira, K., & Santos, J. A. dos. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 44–51). http://doi.org/10.1109/CVPRW.2015.7301382

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., & Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, *5*(6), 1–15. http://doi.org/10.1890/ES13-00359.1

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. http://doi.org/10.1.1.214.9232

Prendergast, J. R., Quinn, R. M., Lawton, J. H., Eversham, B. C., & Gibbons, D. W. (1993). Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature*, *365*(6444), 335–337. http://doi.org/10.1038/365335a0

QGIS Development Team. (2017). QGIS Geographic Information System 2.18.14. *Qgisorg*. http://doi.org/http://www.qgis.org/

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788). http://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in neural information processing systems* (pp. 91–99). http://doi.org/10.1109/TPAMI.2016.2577031

Rodríguez-Pérez, J. R., Álvarez, M. F., & Sanz-Ablanedo, E. (2007). Assessment of Low-Cost GPS Receiver Accuracy and Precision in Forest Environments. *Journal of Surveying Engineering*, *133*(4), 159–167. http://doi.org/10.1061/(asce)0733-9453(2007)133:4(159)

Rogan, J., & Chen, D. M. (2004). Remote sensing technology for mapping and monitoring land-cover and land-use change. *Progress in Planning*, *61*(4), 301–325. http://doi.org/10.1016/S0305-9006(03)00066-7

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. http://doi.org/10.1007/s11263-015-0816-y

Schaminée, J. H. J., Weeda, E. J., & Westhoff, V. (1995). *De vegetatie van Nederland. Deel 2: Plantengemeenschappen van wateren, moerassen en natte heiden*. Uppsala: Opulus.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. http://doi.org/10.1214/12-AOS1000

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training Very Deep Networks. In *Advances in neural information processing systems* (pp. 2377–2385). Retrieved from http://arxiv.org/abs/1507.06228

Struyf, E., Jacobs, S., Meire, P., Jensen, K., & Barendregt, A. (2009). *Plant communities in European tidal freshwater wetlands*. *Tidal freshwater Wetlands*. Backhuys Publisher Leiden.

Sung, K. (1996). *Learning and Example Selection for Object and Pattern Detection*. *PhD thesis*. Massachusetts Institute of Technology.

van Emmerik, W. A. M. ., de Nie, H. W. ., Peters, J. S. ., Kroon, J. W. ., & Quak, J. (2009). Natura 2000-gebied 112-Biesbosch, doelsoorten zoetwatervis: Habitatgebruik en -eisen, knelpunten en trends, 74.

Van Steenis, C. (1971). De zoetwatergetijde-dotter van de Biesbosch en de Oude Maas: Caltha palustris L. var. araneosa, var. nov. *Gorteria*, *5*, 213–219. Retrieved from citeulike-article-id:10499563

Vanden Borre, J., Paelinckx, D., Mücher, C. A., Kooistra, L., Haest, B., De Blust, G., & Schmidt, A. M. (2011). Integrating remote sensing in Natura 2000 habitat monitoring: Prospects on the way forward. *Journal for Nature Conservation*, *19*(2), 116–125. http://doi.org/10.1016/j.jnc.2010.07.003

Vanden Borre, J., Spanhove, T., & Haest, B. (2017). *Towards a Mature Age of Remote Sensing for Natura 2000 Habitat Conservation: Poor Method Transferability as a Prime Obstacle*. *The Roles of Remote Sensing in Nature Conservation*. http://doi.org/10.1007/978-3-319-64332-8_2

Volpi, M., & Tuia, D. (2017). Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), 881–893. http://doi.org/10.1109/TGRS.2016.2616585

Weeda, E. J., Neut, J. J. M., Boesveld, A. A. M., & Weel, B. A. M. (2003). *Nationaal park De Biesbosch: schatkamer van de wilde flora; een overzicht van zeldzame en bedreigde vaatplanten*. Staatsbosbeheer. Retrieved from https://library.wur.nl/WebQuery/wurpubs/reports/320705

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

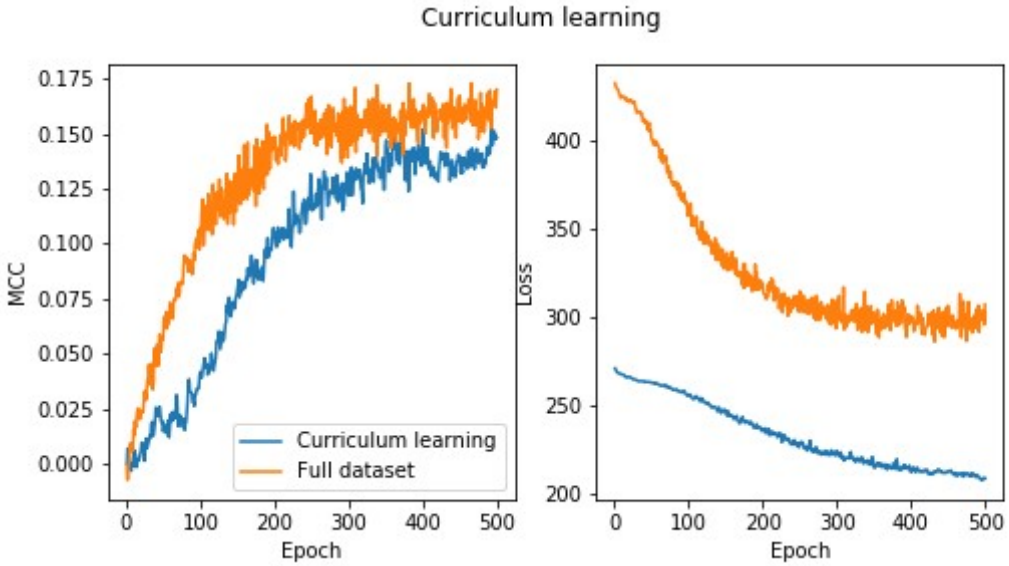
Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. http://doi.org/10.1109/MGRS.2017.2762307

Appendices

Appendix A: Dimension and trainable parameters of the models

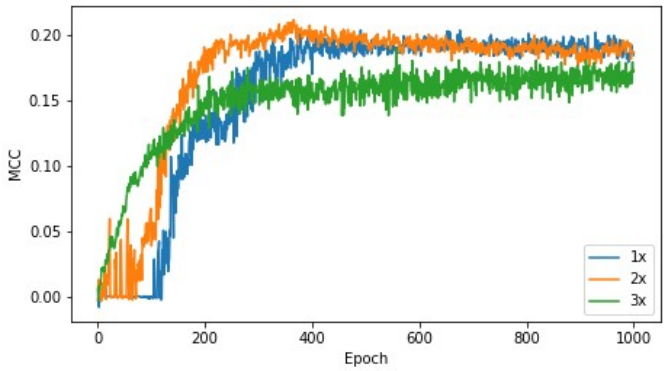
Model / Layer	**Dimensions**	**Parameters to learn**
Single prediction		
FC1	512x1	512 x 512 + 512 = 262656
FC2	1x1	512 x 1 + 1 = 513
	Total	263169
Grid prediction		
FC1	512x1	512 x 512 + 512 = 262656
FC2	256x1	512 x 256 + 256 = 131328
	Total	393984

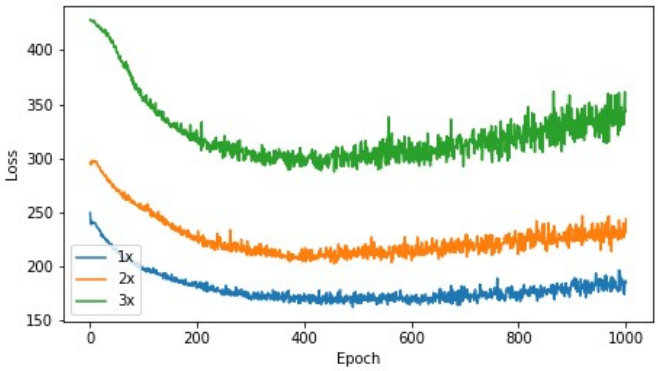
Appendix B: Learning curves with and without curriculum learning



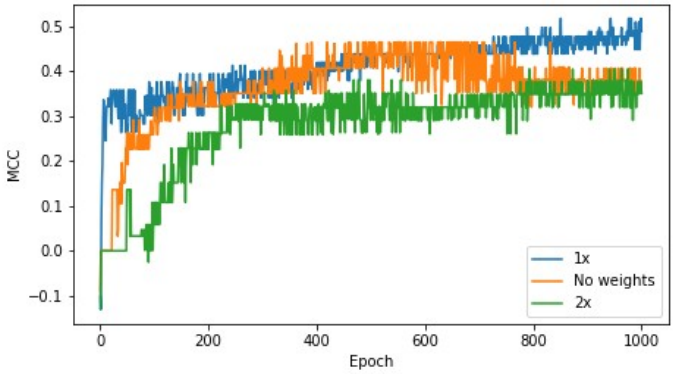
Appendix C: Learning curves for class weights

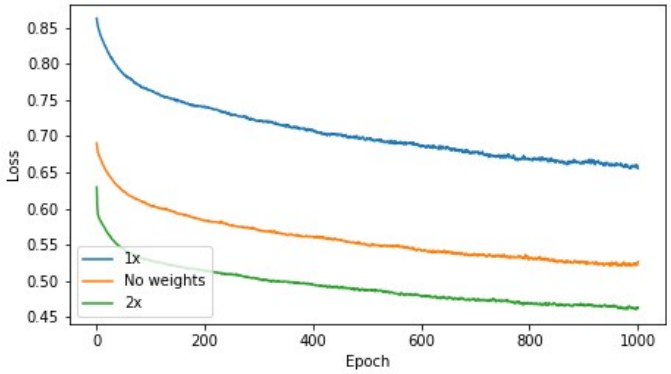
Grid Prediction Model





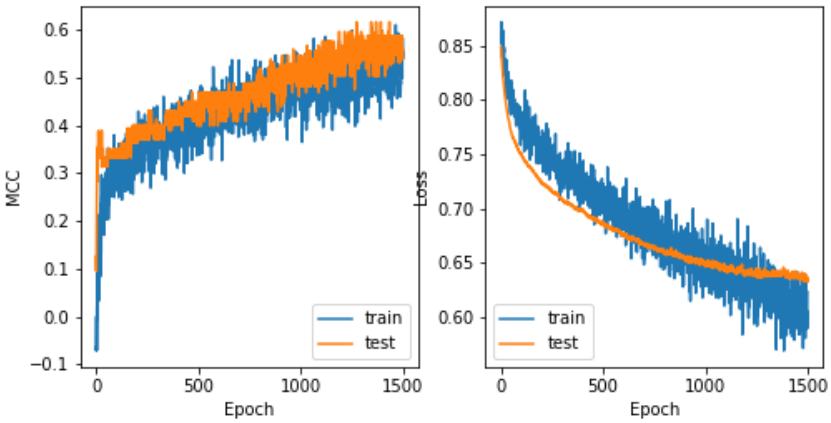
Single prediction model



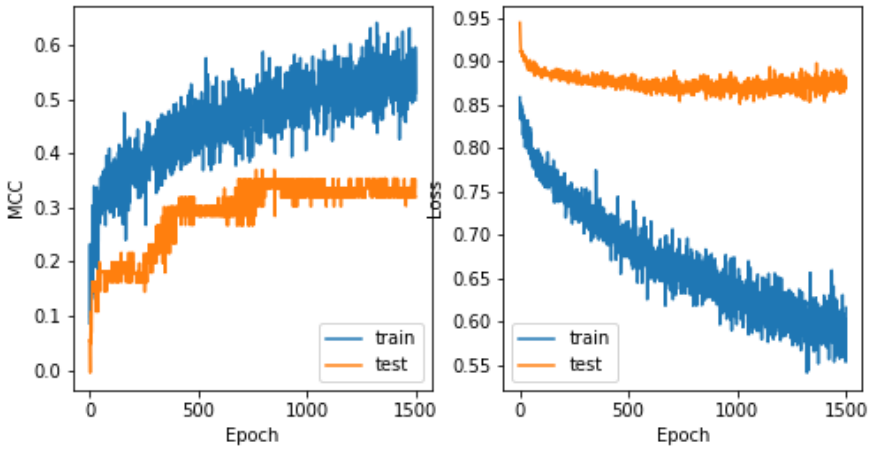


Appendix D: Learning curves for flowers and leaves datasets

Flowers dataset

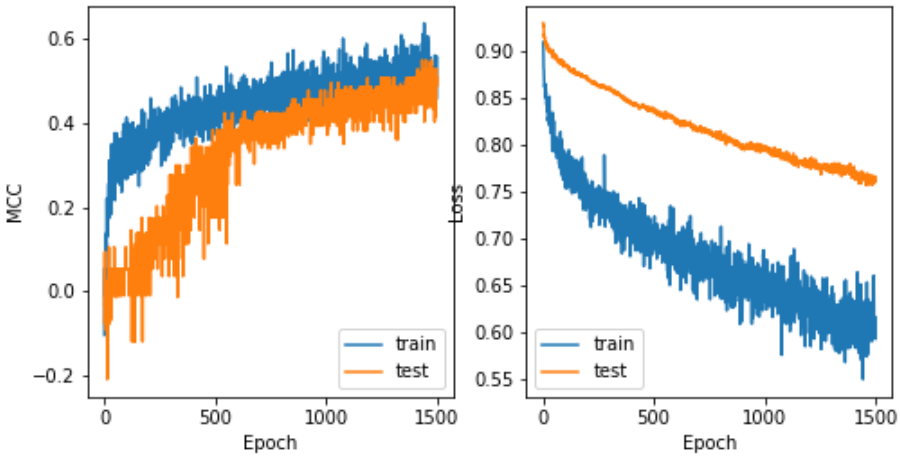


Leaves dataset

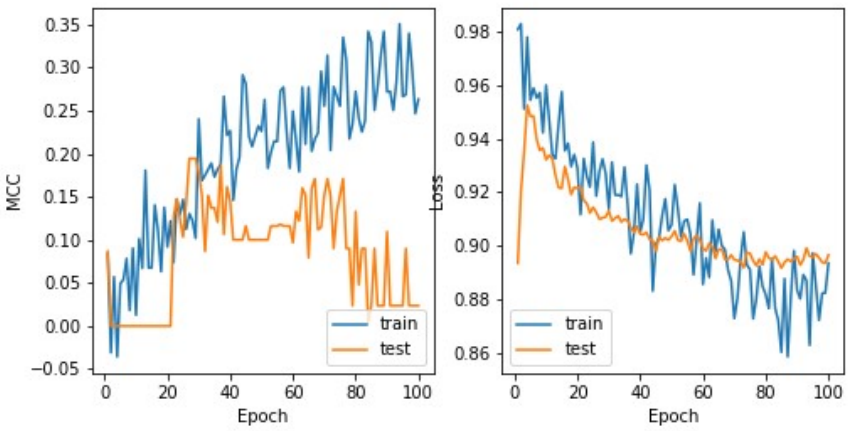


Appendix E: Learning curves for forest and grassland plots

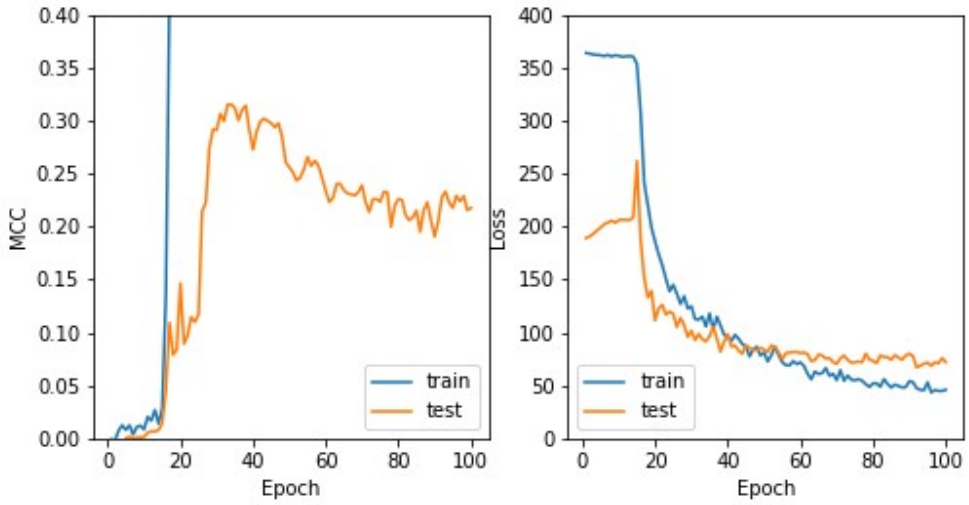
Forest plots



Grassland plots



Appendix F: Learning curve for cow dataset



Appendix G: Predictions of best performing model on total study area