







# Towards Confirmable Automated Plant Cover Determination

Matthias Körschens<sup>1</sup>, Paul Bodesheim<sup>1</sup>, Christine Römermann<sup>1,2,3</sup>,  
Solveig Franziska Bucher<sup>1,3</sup>, Josephine Ulrich<sup>1</sup>, and Joachim Denzler<sup>1,2,3</sup>

<sup>1</sup> Friedrich Schiller University Jena, Jena, Germany

<sup>2</sup> German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,  
Leipzig, Germany

<sup>3</sup> Michael Stifel Center Jena, Jena, Germany

{matthias.koerschens,paul.bodesheim,christine.roemermann,  
solveig.franziska.bucher,josephine.ulrich,joachim.denzler}@uni-jena.de

**Abstract.** Changes in plant community composition reflect environmental changes like in land-use and climate. While we have the means to record the changes in composition automatically nowadays, we still lack methods to analyze the generated data masses automatically.

We propose a novel approach based on convolutional neural networks for analyzing the plant community composition while making the results explainable for the user. To realize this, our approach generates a semantic segmentation map while predicting the cover percentages of the plants in the community. The segmentation map is learned in a weakly supervised way only based on plant cover data and therefore does not require dedicated segmentation annotations.

Our approach achieves a mean absolute error of 5.3% for plant cover prediction on our introduced dataset with 9 herbaceous plant species in an imbalanced distribution, and generates segmentation maps, where the location of the most prevalent plants in the dataset is correctly indicated in many images.

**Keywords:** Deep learning · Machine learning · Computer vision · Weakly supervised segmentation · Plants · Abundance · Plant cover

## 1 Introduction

In current times the effect of anthropogenic activities is affecting ecosystems and biodiversity with regard to plants and animals alike. Whereas poaching and clearing of forests are only some of the smaller impacts of humans, one of the biggest is the anthropogenic effect on climate change.

Plants are strong indicators of climate change, not only in terms of phenological responses [37,31,7,32,9,13], but also in terms of plant community compositions [37,27,29]. However, these compositions do not only reflect changes in climate, but also in other aspects, like land use [14,2] and insect abundance [39]. Hence, plant community compositions are a valuable metric for determining environmental changes and are therefore focus of many experiments [27,14,39,6].

In the last years, technology enabled us to develop systems that can automatically collect images of such experiments in high resolution and high frequency, which would be too expensive and time-consuming if done manually. This process also creates large masses of data displaying complex plant compositions, which are also hard to analyse by hand. As we are missing methods to survey the data automatically, this is usually still done manually by biologists directly in the field. However, this process is bound to produce subjective results. Therefore, an automated, objective method would not only enable fast evaluation of the experimental data, but also greatly improve comparability of the results.

Krizhevsky et al. [24] showed in the ILSVRC 2012 challenge [38] that convolutional neural networks (CNNs) can be used to analyze large numbers of images by outperforming alternative approaches by a large margin. Following this, deep learning became a large area of research with many different developments, but only a small number of approaches deal with the analysis of plants and therefore there are only few existing solutions for the very specific problems in this area.

With our approach we propose a system for an important task: the analysis of plant community compositions based on plant cover. The plant cover, i.e., the amount of ground covered by each plant species, is an indicator for the plant community composition. The information on the spatio-temporal distribution of plant communities leads to a better understanding of effects not only related to climate change, but also concerning other environmental drivers of biodiversity [6,5,44]. We present an approach using a custom CNN architecture, which we train on plant cover percentages that are provided as annotations. We treat this as a pixel-wise classification problem known as semantic segmentation and aggregate the individual scores to compute the cover predictions.

CNNs are often treated as black boxes, returning a result without any information to the user what it is based on. To prevent trust issues resulting from this, we also focus on providing a segmentation map, which the network learns by training on the cover percentage labels only. With this map, the user can verify the network detections and whether the output of the network is reasonable. For implausible cases or manual inspections of random samples, the user can look at the segmentations. If detections of the network are deemed incorrect, a manual evaluation of the images can be suggested in contrast to blind trust in the output of the network. To the best of our knowledge, we are first in applying CNNs to plant cover prediction by training on the raw cover labels only and using relative labels (cover percentages) to train a network for generating segmentation maps.

In the next section we will discuss related work, followed by the dataset we used and its characteristics in Section 3. In Section 4 we will then present our approach, with the results of our experiments following in Section 5. We end the paper with a conclusion and a short discussion about future work in Section 6.

## 2 Related Work

An approach also dealing with plant cover is the one by Kattenborn et al. [20], who developed their approach using remote sensing image data, specifically im-

ages taken from UAVs. They developed a small convolutional neural network (CNN) architecture with 8 layers to determine the cover of different herb, shrub and woody species. In contrast to our approach, their network was trained on low-resolution image patches with delineations of tree canopies directly in the images. In addition to this, their approach was mostly concerned with the distinction of 2-4 tree species with heterogeneous appearances, which makes the classification easier as compared to our problem.

While, to the best of our knowledge, the aforementioned approach appears to be the only one dealing with plant cover, there are many methods which tackle plant identification in general, e.g. [48,4,25,15]. One example for such a method is the one by Yalcin et al. [48], who applied a pre-trained CNN with 11 layers on fruit-bearing agricultural plants. Another, more prominent project concerned with plant identification is the Flora Incognita project of Wäldchen et al. [45], in which multiple images of a single plant can be used for identification. These approaches, however, are usually applied on one or multiple images of a single plant species in contrast to pictures of plant communities with largely different compositions like in our dataset.

Weakly-supervised segmentation, i.e., the learning of segmentation maps using only weak labels, is an established field in computer vision research. Therefore, we can also find a multitude of different approaches in this area. Some of them use bounding boxes for training the segmentation maps [10,21] while others use merely image-level class annotations [3,18,23,33,46], as these are much easier to acquire than bounding box annotations. However, most of these approaches are only applied on images with mostly large objects like the PASCAL-VOC dataset [12] as opposed to high-resolution images with small fine-grained objects like in our dataset. In addition to this, in our dataset we have a new kind of weak labels: plant cover percentage labels. This type of label enables new approaches for learning segmentation maps, which we try to exploit in this paper.

At first glance the task of predicting the cover percentage appears similar to counting or crowd-counting tasks, which are often solved by training a model on small, randomly drawn image patches and evaluating them on complete images, or also evaluating them on patches and aggregating information afterwards [28,47,19]. This can be done, because only absolute values have to be determined, which are usually completely independent from the rest of the image. However, in our dataset the target values, i.e., the cover percentages, are not absolute, but relative and therefore depend on the whole image. Because of this, we have to process the complete images during training and cannot rely on image patches.

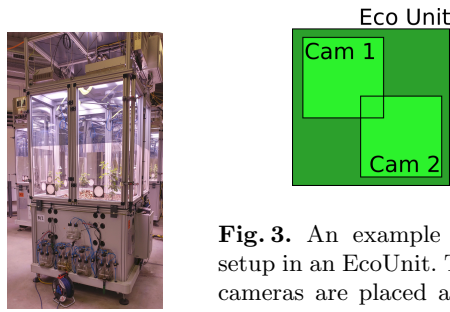
### 3 Dataset

For our experiments we used a dataset comprising images from the InsectArmageddon project<sup>1</sup> and therefore we will refer to this dataset as the InsectArmageddon dataset. During this project the effects of invertebrate density on

<sup>1</sup> [https://www.idiv.de/en/research/platforms\\_and\\_networks/idiv\\_ecotron/experiments/insect\\_armageddon.html](https://www.idiv.de/en/research/platforms_and_networks/idiv_ecotron/experiments/insect_armageddon.html)

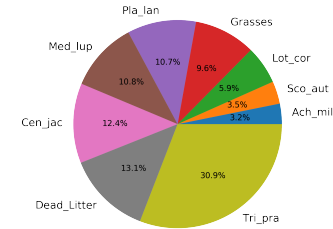


**Fig. 1.** A selection of example images from the image series of a single camera in a single EcoUnit. The complete life cycle is captured in the image series, including flowering and senescence.



**Fig. 2.** An EcoUnit from the Ecotron system.

**Fig. 3.** An example camera setup in an EcoUnit. The two cameras are placed at opposite corners of the EcoUnits and can have an overlapping field of view. In some cases not the complete unit is covered by the cameras.



**Fig. 4.** The mean cover percentages of the plant species over all annotated images in the dataset in a long-tailed distribution. The abbreviations are explained in Section 3.

plant composition and growth were investigated. The experiments were conducted using the iDiv Ecotron facilities in Bad Lauchstädt [11,42], which is a system comprising 24 so-called EcoUnits. Each of these EcoUnits has a base area of about  $1.5\text{ m} \times 1.5\text{ m}$  and contains a small, closed ecosystem corresponding to a certain experimental setup. An image of an EcoUnit is shown in Figure 2.

Over the time span of the project, each of the EcoUnits was equipped with two cameras, observing the experiments from two different angles. One example of such a setup is shown in Figure 3. It should be noted that the cameras have overlapping fields of view in many cases, resulting in the images from each unit not being independent of each other. Both cameras in each unit took one image per day. As the duration of the project was about half a year, 13,986 images have been collected this way over two project phases. However, as annotating this comparatively large number of images is a very laborious task, only about one image per recorded week in the first phase has been annotated per EcoUnit. This is drastically reducing the number of images available for supervised training.

The plants in the images are all herbaceous, which we separate in nine classes with seven of them being plant species. These seven plants and their short forms, which are used in the remainder of the paper, are: *Trifolium pratense* (tri\_pra), *Centaurea jacea* (cen\_jac), *Medicago lupulina* (med\_lup), *Plantago lanceolata* (pla\_lan), *Lotus corniculatus* (lot\_cor), *Scorzoneroidea autumnalis* (sco\_aut) and *Achillea millefolium* (ach\_mil). The two remaining classes are grasses and dead litter. These serve as collective classes for all grass-like plants and dead biomass, respectively, mostly due to lack of visual distinguishability in images.

As with many biological datasets, this one is heavily imbalanced. The mean plant cover percentages over the complete dataset are shown in Figure 4. There, we can see that tri\_pra represents almost a third of the dataset and the rarest three classes, ach\_mil, sco\_aut and lot\_cor together constitute only about 12% of the dataset.

### 3.1 Images

The cameras in the EcoUnits are mounted in a height of about 2 m above the ground level of the EcoUnits and can observe an area of up to roughly  $2m \times 2m$ , depending on zoom level. Equal processing of the images however is difficult due to them being scaled differently. One reason for this is that many images have different zoom levels due to technical issues. The second reason is that some plants grew rather high and therefore appear much larger in the images.

As mentioned above, the images cover a large time span, i.e., from April to August 2018 in case of the annotated images. Hence, the plants are captured during their complete life cycle, including the different phenological stages they go through, like flowering and senescence.

Occlusion is one of the biggest challenges in the dataset, as it is very dominant in almost every image and makes an accurate prediction of the cover percentages very difficult. The occlusion is caused by the plants overlapping each other and growing in multiple layers. However, as we will mostly focus on the visible parts of the plants, tackling the non-visible parts is beyond the scope of this paper. A small selection from the images of a camera of a single EcoUnit can be seen in Figure 1. Each of the images has an original resolution of 2688x1520 px.

As already discussed in Section 2, we are not able to split up the images into patches and train on these subimages, as we only have the cover annotations for the full image. Therefore, during training we always have to process the complete images. This circumstance, in conjunction with the rather high resolution of the images, the similarity of the plants and massive occlusion, makes this a tremendously hard task.

### 3.2 Annotations

As already mentioned above, the annotations for the images are cover percentages of each plant species, i.e., the percent of ground covered by each species, disregarding occlusion. The cover percentages have been estimated by a botanist using both images of each EcoUnit, if a second image was available. As perfect

estimation is impossible, the estimates have been quantized into classes of a modified Schmidt-scale [34](0, 0.5, 1, 3, 5, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90 and 100 percent). While such a quantization is very common for cover estimation in botanical research [34,30], it introduces label noise and can, in conjunction with possible estimation errors, potentially impair the training and evaluation process of machine learning models. In addition to the cover percentages, we also estimated vegetation percentages, specifying the percentage of ground covered by plants in general, which we use as auxiliary target value.

While both images of each EcoUnit have been used for estimating a single value, the distribution of plants should approximately be the same for both images. Therefore, we increase the size of our dataset by using one annotation for both images, which leads us to 682 image-annotation pairs.

## 4 Approach

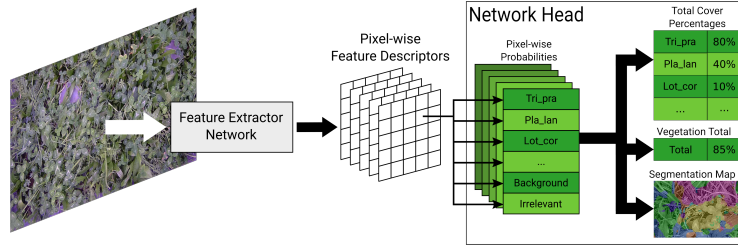
Due to the necessity of using the complete image for the training process, we require a setting, in which it is feasible to process the complete image efficiently without introducing too strong limitations on hyperparameters like the batch size. The most important part of such a setting is the image resolution. As it is hard to train models on very high resolutions due to GPU memory limitations, we chose to process the images at a resolution of 672x336 px, which is several times larger than other common input image resolutions for neural network architectures, like e.g. ResNet [17] training on the ImageNet dataset [38] with a resolution of 224x224 px. To make the results confirmable, we aim to create a segmentation map during prediction that designates, which plant is located at each position in the image. This segmentation map has to be learned implicitly by predicting the cover percentages. Due to the plants being only very small in comparison to the full image, this segmentation map also has to have a high resolution to show the predicted plants as exactly as possible.

The usage of standard classification networks, like ResNet [17] or Inception [41,40], is not possible in this case, as the resolution of the output feature maps is too coarse for an accurate segmentation map. Additionally, these networks and most segmentation networks with a higher output resolution, like Dilated ResNet [49], have large receptive fields. Thus, they produce feature maps that include information from large parts of the image, most of which is irrelevant to the class at a specific point. This leads to largely inaccurate segmentation maps.

We thus require a network, which can process the images at a high resolution, while only aggregating information from a relatively small, local area without compressing the features spatially to preserve as much local information as possible. Our proposed network is described in the following.

### 4.1 General Network Structure

The basic structure of our network is shown in Figure 5. We do a logical separation of the network into two parts: backbone and network head, similar to



**Fig. 5.** The basic structure of the network. It consist of a feature extractor network as backbone, which aggregates information from the input image in a high resolution, and a network head, which performs the cover percentage calculation and generates the segmentation map.

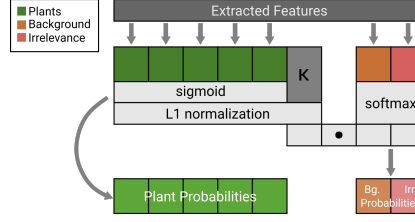
**Table 1.** A detailed view of the network architecture. We use the following abbreviations: k - kernel size, s - stride, d - dilation rate

Layers				Output Shape
Conv k:5x5, s:2x2				336x168x128
Conv k:5x5, s:2x2				168x84x256
9xResidual Bottleneck				168x84x256
Conv k:1x1	Conv k:3x3	Conv k:3x3 Conv k:3x3 Conv k:3x3	Conv k:3x3 Conv k:3x3, d:3,3 Conv k:3x3	168x84x512
Conv k:1x1				168x84x128

Mask R-CNN [16]. The backbone, a feature extractor network, extracts the local information from the image approximately pixel-wise and thus generates a high-resolution feature map, which can then be used by the network head for the cover calculation and generation of the segmentation map. In the network head the pixel-wise probabilities for each plant are calculated, which are then aggregated to calculate the total cover percentage of the complete image. The maxima of the intermediate probabilities are used for generating the segmentation map.

## 4.2 Feature Extractor Network

Feature extractor network initially applies two downscaling operations with 2-strided convolutions, bringing the feature maps to a resolution of 25% of the original image, which is kept until the end of the network. The downscaling layers are followed by nine residual bottleneck blocks as defined in the original ResNet paper [17]. To aggregate information quickly over multiple scales, an inception block, similar to the ones introduced in the papers by Szegedy et al. [41,40] is used. The inception block consists of four branches with different combinations of convolutions, resulting in four different receptive field sizes: 1x1, 3x3, 7x7 and 11x11. In Table 1 the network architecture is shown in detail.



**Fig. 6.** The calculation in the network head. We use a sigmoid function to determine the plant probabilities and a softmax function for the background and irrelevance probabilities. To bring these into a relationship with each other, we use a hyperparameter  $\kappa$ , L1-normalization and a multiplication, denoted with  $\cdot$ .

### 4.3 Network Head & Calculation Model

In the network head we try to calculate the cover percentages as exact as possible. For this, we first introduce two additional classes to the ones already described in Section 3: the background and the irrelevance class. While very similar at the first glance, these two classes differ significantly in meaning. The background class represents every part of the image that is not a plant, but still relevant to cover percentage calculation. The most obvious example for this is the bare soil visible in the images. This class will be abbreviated with *bg* in the following. The irrelevance class, denoted with *irr* in the following, represents all image parts that are not a plant but also not relevant for the cover calculation. Here, the most obvious example are the walls of the EcoUnits, which are visible in many images. The aim of differentiating between these two classes is to separate unwanted objects from the actual plantable area of the EcoUnits and therefore enable the network to work on images without manual removal of such objects, which can be very laborious. If not handled in any way, such objects like the walls of the EcoUnits in our dataset can strongly distort the calculation of cover percentages. For the latter, we require the pixel-wise probabilities of each plant being at the corresponding location in the image as well as the probabilities for both the location being background that is still relevant for the cover percentages, and the location being irrelevant for estimating cover percentages. The calculation scheme is shown in Figure 6.

The extracted features from the backbone are processed by a  $1 \times 1$  convolution to create the classification features for each plant as well as background and irrelevance. As due the occlusion multiple plants can be detected at the same location, we do not consider their probabilities to be mutually exclusive. Hence, we use a sigmoid function to calculate the probability for each plant appearing at this location or not. However, a softmax activation is applied to the classification features for background and irrelevance, as they are mutually exclusive. We also introduce a hyperparameter  $\kappa$ , which we use within the L1-normalization of the probabilities for each plant, and an additional multiplication for the normalized  $\kappa$  to relate the appearance probabilities to those for background and irrelevance, as



they depend on each other. The detailed equations for the complete calculation process are explained in the following.

While the plants already have separate classes, for our formalization we introduce the abstract biomass class, abbreviated with *bio*, which simply represents the areas containing plants. For the introduced classes the following holds:

$$A_{total} = A_{bio} + A_{bg} + A_{irr} , \quad (1)$$

where  $A$  represents the area covered by a certain class. For improved readability we also define the area relevant for cover calculation as

$$A_{rel} = A_{bio} + A_{bg} = A_{total} - A_{irr} \quad (2)$$

As mentioned above we consider the classes of the plants, denoted with  $C^{plants}$ , to be not mutually exclusive due to occlusion enabling the possibility of multiple plants at the same location. However, the classes *bio*, *bg* and *irr* are mutually exclusive. We will refer to these as area classes and denote them with  $C^{area}$ .

Based on this formulation we describe our approach with the following equations. Here, we select a probabilistic approach, as we can only estimate the probabilities of a pixel containing a certain plant. With this, the following equation can be used to calculate the cover percentages for each plant:

$$cover_p = \frac{A_p}{A_{rel}} = \frac{\sum_{\forall x} \sum_{\forall y} P(C_{x,y}^{plants} = p)}{\sum_{\forall x} \sum_{\forall y} 1 - P(C_{x,y}^{area} = irr)} , \quad (3)$$

with  $p$  being the class of a plant, whereas  $x$  and  $y$  determine a certain location in the image.  $C_{x,y}$  is the predicted class at location  $(x, y)$  and  $P(C_{x,y} = c)$  is the probability of class  $c$  being located at the indicated position.

As mentioned before, we also use the vegetation percentages for training to create an auxiliary output. The vegetation percentage represents how much of the relevant area is covered with plants. This additional output helps for determining the area actually relevant for calculation. It can be calculated as follows:

$$vegetation = \frac{A_{bio}}{A_{rel}} = \frac{\sum_{\forall x} \sum_{\forall y} 1 - P(C_{x,y}^{area} = bg) - P(C_{x,y}^{area} = irr)}{\sum_{\forall x} \sum_{\forall y} 1 - P(C_{x,y}^{area} = irr)} . \quad (4)$$

The notation is analogous to Equation 3.

While the probabilities for each plant as well as for background and irrelevance can be predicted, we are still missing a last piece for the construction of the network head: the calculation of the biomass class *bio*. We mentioned above that this class is abstract. This means it cannot be predicted independently, as it is mostly dependent on the prediction of plants in an area. We solve this by introducing the hyperparameter  $\kappa$  as mentioned above, which represents a threshold at which we consider a location to contain a plant (in contrast to background and

irrelevance). We concatenate this value with the plant probabilities  $P(C_{x,y}^{plants})$  to form a vector  $v_{x,y}$ . We normalize this vector using L1-normalization, which can then be interpreted as the dominance of each plant with the most dominant plant having the highest value. As the values of this normalized vector sum up to 1, they can also be treated as probabilities. The value at the original position of  $\kappa$ , which basically represents the probability for the absence of all plants, is higher, if no plant is dominant. Hence, we can define:

$$P(C_{x,y}^{area} = bio) = 1 - \left( \frac{v_{x,y}}{\|v_{x,y}\|_1} \right)_{\kappa} \quad (5)$$

where  $(\cdot)_{\kappa}$  designates the original position of the value  $\kappa$  in the vector. The value  $1 - P(C_{x,y}^{area} = bio)$  can then be multiplied with the background and irrelevance probabilities to generate the correct probabilities for these values. This results in the probabilities of the area classes summing up to one:

$$1 = P(C_{x,y}^{area} = bio) + P(C_{x,y}^{area} = bg) + P(C_{x,y}^{area} = irr). \quad (6)$$

Based on these equations we can construct our network head, which is able to accurately represent the calculation of plant cover in our images.

To generate the segmentation map, we use the maximum values of sigmoidal probabilities of the plant classes together with the ones for background and irrelevance. As these values only have 25% of the original resolution, they are upsampled using bicubic interpolation, resulting in a segmentation map that has the original image resolution.

## 5 Experiments

In the following, we will show our experimental setup, then explain the error measures we used and afterwards will go over the numerical results followed by evaluation of the segmentation maps.

### 5.1 Setup

During our experiments we used an image resolution of 672x336 px and a batch size of 16. We trained the network for 300 epochs using the Adam [22] optimizer with a learning rate of 0.01, decreasing by a factor of 0.1 at epoch 100, 200 and 250. As loss we used the MAE both for the cover percentage prediction as well as for the vegetation prediction weighted equally. Furthermore, we used L2 regularization with factor of 0.0001. The activation functions in the backbone were ReLU functions and we used reflective padding instead of zero padding, as this produces fewer artifacts at the border of the image. During training the introduced hyperparameter  $\kappa$  was set to 0.001. For data augmentation we used horizontal flipping, small rotations in the range of -20° to 20°, coarse dropout, and positional translations in the range of -20 to 20 pixels. We trained the model using the Tensorflow framework [1] with Keras [8] using mixed precision. For a

**Table 2.** The mean cover percentages used for scaling in Equation 8 during evaluation.

Tri_pra	Pla_lan	Med_lup	Cen_jac	Ach_mil	Lot_cor	Sco_aut	Grasses	Dead Litter
33.3%	11.5%	11.7%	13.4%	3.4%	6.4%	3.8%	10.3%	14.1%

**Table 3.** The mean values and standard deviations of the absolute errors and scaled absolute errors.

Plants	Tri_pra	Pla_lan	Med_lup	Cen_jac	Ach_mil
MAE	9.88 ( $\pm$ 10.41)	5.81 ( $\pm$ 5.19)	7.36 ( $\pm$ 6.13)	5.53 ( $\pm$ 5.04)	2.15 ( $\pm$ 2.62)
MSAE	0.30 ( $\pm$ 0.31)	0.50 ( $\pm$ 0.45)	0.63 ( $\pm$ 0.52)	0.41 ( $\pm$ 0.38)	0.63 ( $\pm$ 0.77)
Plants	Lot_cor	Sco_aut	Grasses	Dead_Litter	
MAE	3.20 ( $\pm$ 3.75)	2.31 ( $\pm$ 3.44)	4.49 ( $\pm$ 6.51)	7.23 ( $\pm$ 9.01)	
MSAE	0.50 ( $\pm$ 0.59)	0.61 ( $\pm$ 0.91)	0.44 ( $\pm$ 0.63)	0.51 ( $\pm$ 0.64)	

fair evaluation, we divided the images into training and validation parts based on the EcoUnits. We use 12-fold cross validation, such that each cross validation split consists of 22 EcoUnits for training and 2 for testing. While the cover percentages are not equally distributed over the EcoUnits, this should only have little effect on the results of the cross validation.

## 5.2 Error Measures

To evaluate the numerical results of our approach, we will take a look at two different error measures. The first one is the mean absolute error (MAE), which is defined as follows:

$$MAE(t, p) = \frac{1}{n} \sum_{i=1}^n |t_i - p_i|, \quad (7)$$

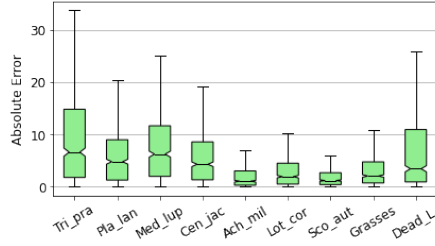
where  $t$  and  $p$  are the true and predicted cover values, respectively. As the mean absolute error can be misleading when comparing the goodness of the predictions for imbalanced classes, we also propose a scaled version of the MAE: the mean scaled absolute error (MSAE), which is defined as follows:

$$MSAE(t, p) = \frac{1}{n} \sum_{i=1}^n \frac{|t_i - p_i|}{m_i}. \quad (8)$$

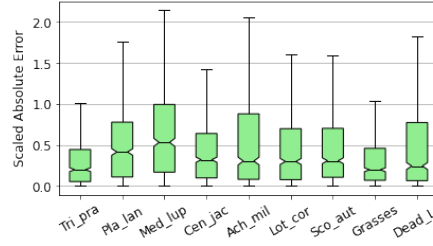
The absolute error values for each class are scaled by a value  $m_i$ , which is the mean cover percentage averaged over the different annotations within the respective class in the dataset. This error will provide a better opportunity for comparing the predictions between the classes. The values that have been used for scaling can be found in Table 2.

## 5.3 Experimental Results

**Cover Predictions.** Our model achieves an overall MAE of 5.3% and an MSAE of 0.50. The detailed results for each species are shown in Table 3 as well as in



**Fig. 7.** An overview over the MAE of the plant cover prediction in the dataset.

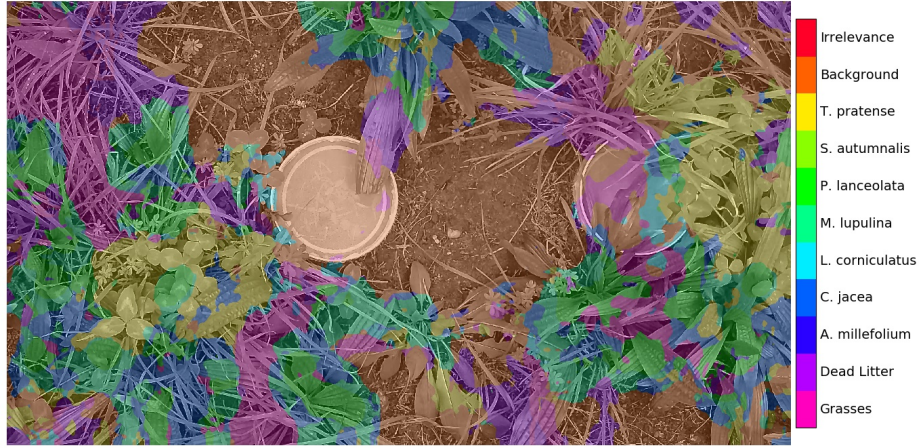


**Fig. 8.** An overview over the MSAE of the plant cover prediction in the dataset.

Figure 7 and Figure 8. With respect to the MAE, we can see that the error of *tri\_pra* appears to be the highest, while the error of the less abundant plants (*ach\_mil*, *lot\_cor*, *sco\_aut*) appears to be much lower. However, as mentioned above, the distribution of the MAE mostly reflects the distribution of the plants in the whole dataset, as the errors for the more abundant plants are expected to be higher. Therefore, to compare the goodness of the results between plants, we take a look at the MSAE depicted in Figure 8, where we can see that *tri\_pra* actually has the lowest relative error compared to the other plants, partially caused by the comparably large amounts of training data for this class. The most problematic plants appear to be *ach\_mil*, *sco\_aut* and *med\_lup*, with MSAE error values of 0.63, 0.61 and 0.63, respectively. For *ach\_mil* the rather high error rate might result from multiple circumstances. The plant is very rare in the dataset, small in comparison to many of the other plants in the dataset and also has a complex leaf structure, most of which might get lost using smaller resolutions. The large error for *med\_lup* might be caused by its similarity to *tri\_pra*, which is very dominant in the dataset. Therefore, the network possibly predicts *Trifolium* instead of *Medicago* on many occasions, causing larger errors. The same might be the case for *sco\_aut* and *pla\_lan* or *cen\_jac*, especially since *sco\_aut* is one of the least abundant plants in the dataset making a correct recognition difficult.

To put these results into perspective, we also provide the results using a constant predictor, which always predicts the mean of the cover percentages of the training dataset, and the results using a standard U-Net [36] as feature extractor. These achieved an MAE of 9.88% and MSAE of 0.84, and an MAE of 5.54% and MSAE of 0.52 for the constant predictor and the U-Net respectively. We can see that our proposed network outperforms the constant predictor by a large margin and also slightly improves the accuracy of a U-Net, despite having less than 10% of the number of parameters compared to the U-Net (3 million vs. 34 million). More details can be found in the supplementary material.

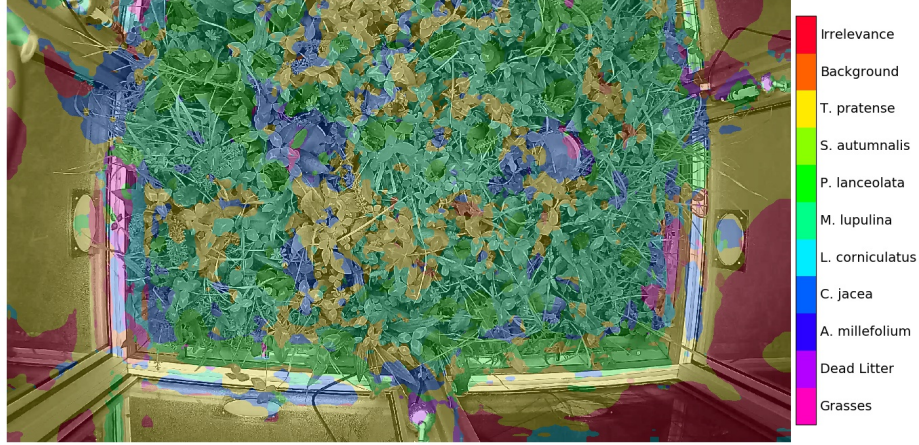
**Segmentations.** To evaluate the result of our network, we also take a look at the results of the segmentation. The first image, shown in Figure 9, is one with a comparably high zoom level. There we can see that *tri\_pra* is detected correctly



**Fig. 9.** Segmentation results for an image with a high zoom level from the validation set. We can see that grasses, *P. lanceolata*, *T. pratense* and background area are segmented correctly in many cases.

in the areas on the left and right sides of the image, while the segmentations are not perfect. *pla.lan* has been segmented well in many cases, especially on the right side of the image. On the left we can see that it is also segmented correctly, even though it is partially covered by grass. Therefore, the approach appears to be robust to minor occlusions to some extent. Despite these results, the segmentation is still mostly incorrect in the top center of the image. Grasses are also detected correctly in most regions of the image, whereas above the aforementioned instances of *pla.lan* they are not segmented at all, which is mostly caused by the low resolution of the segmentation map. This low resolution also appears to impair the segmentation results in many other occasions and we would like to tackle this problem in the future.

The second segmentation image is shown in Figure 10. Here, the zoom level is lower than in the image before, which results in the segmentations getting increasingly inaccurate. We can see that the network correctly captured the presence of most plant species. Notably, the approximate regions of *med.lup* and *tri.pra* are marked correctly. However, the detailed segmentation results are not very accurate. It also appears that some parts of the wall are wrongly recognized as *tri.pra*, while other parts are correctly marked as irrelevant for cover calculation. The segmentations with a U-Net feature extractor can be found in the supplementary materials. All in all, the segmentations appear to be correct for the more prominent plants in the dataset shown in the images with high zoom level and at least partially correct in the images without zoom. Therefore, the segmentation maps can be used to explain and confirm the plant cover predictions for some plants from the dataset.



**Fig. 10.** Segmentation results for a zoomed out image from the validation set. While the network captures the signals of many plants correctly, the segmentations are rather inaccurate leading to a large number of wrongly segmented plant species.

## 6 Conclusions & Future Work

We have shown that our approach is capable of predicting cover percentages of different plant species while generating a high-resolution segmentation map. Learning is done without any additional information other than the original cover annotations. Although not perfect, the segmentation map can already be used to explain the results of the cover prediction for the more prevalent plants in the dataset. Many original images have a very high resolution and are currently downscaled due to computational constraints. Making our approach applicable to images of higher resolution could be one improvement. This would also increase the resolution of the segmentation map, resulting in much finer segmentations. The recognition of the less abundant plants, but also of very similar plants like *T. pratense* and *M. lupulina*, might be improved by applying transfer learning techniques. For example, we could pretrain the network on the iNaturalist datasets [43], since they contain a large number of plant species. Heavy occlusions are still a big challenge in our dataset, making predictions of correct plants and their abundances very hard. While there are already some approaches for segmenting occluded regions in a supervised setting [26,35,50], it is a completely unexplored topic for weakly-supervised semantic segmentation.

## Acknowledgements

Matthias Körschens thanks the Carl Zeiss Foundation for the financial support. In addition, we would like to thank Mirco Migliavacca for additional comments on the manuscript.

## References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). pp. 265–283 (2016)
2. Aggemyr, E., Cousins, S.A.: Landscape structure and land use history influence changes in island plant composition after 100 years. *Journal of Biogeography* **39**(9), 1645–1656 (2012)
3. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218. IEEE (2019)
4. Barré, P., Stöver, B.C., Müller, K.F., Steinhage, V.: Leafnet: A computer vision system for automatic plant species identification. *Ecological Informatics* **40**, 50–56 (2017)
5. Bernhardt-Römermann, M., Baeten, L., Craven, D., De Frenne, P., Hédli, R., Lenoir, J., Bert, D., Brunet, J., Chudomelová, M., Decocq, G., et al.: Drivers of temporal changes in temperate forest plant diversity vary across spatial scales. *Global change biology* **21**(10), 3726–3737 (2015)
6. Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S.M., Botta-Dukát, Z., Chytrý, M., Field, R., Jansen, F., et al.: Global trait–environment relationships of plant communities. *Nature Ecology & Evolution* **2**(12), 1906–1917 (2018)
7. Bucher, S.F., König, P., Menzel, A., Migliavacca, M., Ewald, J., Römermann, C.: Traits and climate are associated with first flowering day in herbaceous species along elevational gradients. *Ecology and Evolution* **8**(2), 1147–1158 (2018)
8. Chollet, F., et al.: Keras. <https://keras.io> (2015)
9. Cleland, E.E., Allen, J.M., Crimmins, T.M., Dunne, J.A., Pau, S., Travers, S.E., Zavaleta, E.S., Wolkovich, E.M.: Phenological tracking enables positive species responses to climate change. *Ecology* **93**(8), 1765–1771 (2012)
10. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1635–1643. IEEE (2015)
11. Eisenhauer, N., Türke, M.: From climate chambers to biodiversity chambers. *Frontiers in Ecology and the Environment* **16**(3), 136–137 (2018)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International journal of computer vision* **88**(2), 303–338 (2010)
13. Fitter, A., Fitter, R.: Rapid changes in flowering time in british plants. *Science* **296**(5573), 1689–1691 (2002)
14. Gerstner, K., Dormann, C.F., Stein, A., Manceur, A.M., Seppelt, R.: Editor’s choice: Review: Effects of land use on plant diversity—a global meta-analysis. *Journal of Applied Ecology* **51**(6), 1690–1700 (2014)
15. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* **235**, 228–235 (2017)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969. IEEE (2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778. IEEE (2016)

18. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023. IEEE (2018)
19. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 532–546. Springer (2018)
20. Kattenborn, T., Eichel, J., Wiser, S., Burrows, L., Fassnacht, F.E., Schmidtlein, S.: Convolutional neural networks accurately predict cover fractions of plant species and communities in unmanned aerial vehicle imagery. *Remote Sensing in Ecology and Conservation* (2020)
21. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885. IEEE (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2015)
23. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision. pp. 695–711. Springer (2016)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 25, pp. 1097–1105. Curran Associates, Inc. (2012), <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
25. Lee, S.H., Chan, C.S., Wilkin, P., Remagnino, P.: Deep-plant: Plant identification with convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). pp. 452–456. IEEE (2015)
26. Li, K., Malik, J.: Amodal instance segmentation. In: European Conference on Computer Vision. pp. 677–693. Springer (2016)
27. Liu, H., Mi, Z., Lin, L., Wang, Y., Zhang, Z., Zhang, F., Wang, H., Liu, L., Zhu, B., Cao, G., et al.: Shifting plant species composition in response to climate change stabilizes grassland primary production. *Proceedings of the National Academy of Sciences* **115**(16), 4051–4056 (2018)
28. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1774–1783. IEEE (2019)
29. Lloret, F., Peñuelas, J., Prieto, P., Llorens, L., Estiarte, M.: Plant community changes induced by experimental climate change: seedling and adult species composition. *Perspectives in Plant Ecology, Evolution and Systematics* **11**(1), 53–63 (2009)
30. Van der Maarel, E., Franklin, J.: *Vegetation ecology*. John Wiley & Sons (2012)
31. Menzel, A., Sparks, T.H., Estrella, N., Koch, E., Aasa, A., Ahas, R., Alm-Kübler, K., Bissolli, P., Braslavská, O., Briede, A., et al.: European phenological response to climate change matches the warming pattern. *Global change biology* **12**(10), 1969–1976 (2006)
32. Miller-Rushing, A.J., Primack, R.B.: Global warming and flowering times in thoreau’s concord: a community perspective. *Ecology* **89**(2), 332–341 (2008)



33. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1796–1804. IEEE (2015)
34. Pfadenhauer, J.: Vegetationsökologie - ein Skriptum. IHW-Verlag, Eching, 2. verbesserte und erweiterte auflage edn. (1997)
35. Purkait, P., Zach, C., Reid, I.: Seeing behind things: Extending semantic segmentation to occluded regions. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1998–2005. IEEE (2019)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
37. Rosenzweig, C., Casassa, G., Karoly, D.J., Imeson, A., Liu, C., Menzel, A., Rawlins, S., Root, T.L., Seguin, B., Tryjanowski, P., et al.: Assessment of observed changes and responses in natural and managed systems. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change pp. 79–131 (2007)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
39. Souza, L., Zelikova, T.J., Sanders, N.J.: Bottom-up and top-down effects on plant communities: nutrients limit productivity, but insects determine diversity and composition. Oikos **125**(4), 566–575 (2016)
40. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826. IEEE (2016)
42. Türke, M., Feldmann, R., Fürst, B., Hartmann, H., Herrmann, M., Klotz, S., Mathias, G., Meldau, S., Ottenbreit, M., Reth, S., et al.: Multitrophische biodiversitätsmanipulation unter kontrollierten umweltbedingungen im idiv ecotron. In: Lysimetertagung. pp. 107–114 (2017)
43. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778. IEEE (2018)
44. Verheyen, K., De Frenne, P., Baeten, L., Waller, D.M., Hédli, R., Perring, M.P., Blondeel, H., Brunet, J., Chudomelová, M., Decocq, G., et al.: Combining biodiversity resurveys across regions to advance global change research. BioScience **67**(1), 73–83 (2017)
45. Wäldchen, J., Mäder, P.: Flora incognita—wie künstliche intelligenz die pflanzenbestimmung revolutioniert: Botanik. Biologie in unserer Zeit **49**(2), 99–101 (2019)
46. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284. IEEE (2020)
47. Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., Shen, C.: From open set to closed set: Counting objects by spatial divide-and-conquer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8362–8371. IEEE (2019)

- 48. Yalcin, H., Razavi, S.: Plant classification using convolutional neural networks. In: 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics). pp. 1–5. IEEE (2016)
- 49. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480. IEEE (2017)
- 50. Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., Loy, C.C.: Self-supervised scene de-occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3784–3792. IEEE (2020)