M. Körschens, P. Bodesheim, C. Römermann, S. F. Bucher, M. Migliavacca, J. Ulrich and J. Denzler:
Weakly Supervised Segmentation Pretraining for Plant Cover Prediction.
DAGM German Conference on Pattern Recognition (GCPR). 2021.

Weakly Supervised Segmentation Pretraining for Plant Cover Prediction

Matthias Körschens[1 [0000−0002−0755−2006]],
Paul Bodesheim[1 [0000−0002−3564−6528]],
Christine Römermann[1 [0000−0003−3471−0951]],
Solveig Franziska Bucher[1 [0000−0002−2303−4583]],
Mirco Migliavacca[2 [0000−0003−3546−8407]],
Josephine Ulrich[1 [0000−0002−5873−8804]], and
Joachim Denzler[1 [0000−0002−3193−3300]]

[1] Friedrich Schiller University Jena, Jena, Germany
{matthias.koerschens,paul.bodesheim,christine.roemermann,
solveig.franziska.bucher,josephine.ulrich,joachim.denzler}@uni-jena.de
[2] Max Planck Institute for Biogeochemistry, Department Biogeochemical Integration, Jena, Germany
mmiglia@bgc-jena.mpg.de

Abstract. Automated plant cover prediction can be a valuable tool for botanists, as plant cover estimations are a laborious and recurring task in environmental research. Upon examination of the images usually encompassed in this task, it becomes apparent that the task is ill-posed and successful training on such images alone without external data is nearly impossible. While a previous approach includes pretraining on a domain-related dataset containing plants in natural settings, we argue that regular classification training on such data is insufficient. To solve this problem, we propose a novel pretraining pipeline utilizing weakly supervised object localization on images with only class annotations to generate segmentation maps that can be exploited for a second pretraining step. We utilize different pooling methods during classification pretraining, and evaluate and compare their effects on the plant cover prediction. For this evaluation, we focus primarily on the visible parts of the plants. To this end, contrary to previous works, we created a small dataset containing segmentations of plant cover images to be able to evaluate the benefit of our method numerically. We find that our segmentation pretraining approach outperforms classification pretraining and especially aids in the recognition of less prevalent plants in the plant cover dataset.

Keywords: Plant Cover Prediction · Biodiversity Monitoring · Plant Segmentation · Computer Vision · Deep Learning · Transfer Learning · Neural Network Pretraining · Weakly Supervised Learning.

1 Introduction

Analyzing the impact of environmental changes on plant communities is an essential part of botanical research. This way, we can find the causes and effects

of such changes and ways to counteract them. A prominent example of an environmental change investigated this way is climate change [30,22,23]. Other environmental aspects can be monitored like this as well, such as land-use [9,1] and insect abundance [33]. One possibility to monitor the effects of such changes on plants is to monitor the species diversity, specifically the species composition of plant communities. This is commonly done by the biologists directly in the field by estimating the so-called plant cover, which is defined as the percentage of soil covered by each plant species, disregarding any occlusion. Performing this task in an automated way based on automatically collected imagery would reduce the massive workload introduced by this recurring and laborious task, and enable an objective analysis of the data in high temporal resolution.

However, developing a correctly working system to perform plant cover prediction (PCP) is a difficult task due to multiple reasons. Firstly, the plant cover estimates are usually noisy due to human error and subjective estimations. Secondly, the plant cover is usually heavily imbalanced by nature, as plants always grow in strongly differing ratios in a natural environment. Thirdly, in addition to this, PCP is an ill-posed problem. This is primarily due to the fact that in the plant cover estimations used as annotation, occlusion is ignored. This makes it near impossible to, for example, train a convolutional neural network (CNN) well on this data alone, as, contrary to human vision, CNNs usually cannot inherently deal with occlusion. Therefore, the network often learns arbitrary features in the images to reduce the error during training in any way possible, which are mostly not the true features the network should use for a correct prediction. To counter this problem, Körschens *et al.* [17], who proposed a first solution for the task of plant cover prediction, suggested utilizing segmentation maps generated in a weakly supervised way by the network for visual inspection of the network's prediction. The segmentation maps make it possible to monitor what the network learned and which areas it uses to generate a prediction, increasing transparency of the system and its gathered knowledge. However, Körschens *et al.* could not solve the problem entirely in the paper, and their results also showed that, when the system is only trained on the plant cover data, the predictions are strongly biased towards the more prevalent plants in the dataset. We argue that these problems can be further alleviated by strong usage of pretraining on domain-similar datasets containing isolated plants, as with such datasets, we can directly control the data balance and have a better influence on the features the network learns. Using such a domain-similar dataset was also recently investigated by Körschens *et al.* [18], who demonstrated the correctness of this assumption on several standard network architectures.

Pretraining, for example the one in [18], is usually done by utilizing a backbone CNN serving as feature extractor, followed by global average pooling and a fully-connected classification layer. We argue that this training method results in the network focusing on the most prevalent features in the pretraining dataset, which might not be optimal since these features are not necessarily contained in the images of the target dataset. An example of this would be the blossom of the plants, which are usually the most discriminative features in iso-

lated plant images but are comparably rare in simple vegetation images for PCP. To solve this problem, we suggest that encouraging the network to focus more on the entire plant instead of only the most discriminative parts by training on segmentation data would be beneficial. However, to the best of our knowledge, there are no comprehensive plant segmentation datasets publicly available. Therefore, we propose a system, which generates weakly supervised segmentations using a classification-trained network. The generated segmentations can then be used for segmentation-pretraining. To this end, we investigate the class activation mapping (CAM) method [48], which is often used in tasks like weakly supervised object localization [48,32,45,46,5,41] and weakly supervised semantic segmentation [15,12,43,34,34,37,40]. While most of these methods use global average pooling (GAP) as basis for their classification network, we found that pooling methods like global max pooling (GMP) or global log-sum-exp pooling (GLSEP), which is an approximation of the former, in parts generate better localizations which result in better segmentations. While similar functions have been investigated before for classification [44], and in another setting in weakly supervised object localization [28], to the best of our knowledge we are the first to apply in as feature aggregation method in conjunction with CAMs.

Furthermore, PCP can be viewed as a task effectively consisting of two parts: analysis of the visual and the occluded parts. Solving occlusion is heavily dependent on a good automated analysis of the visual part, as we can only complete the partially visible plants correctly if our analysis of the species in the visible parts is correct as well. Therefore, we will primarily focus on the correctness of the analysis of the visible parts. Analyzing the visible parts can be done, for example, by investigating segmentation maps as proposed by Körschens *et al.* [17]. However, they merely relied on visual inspection of the segmentations, as no ground-truth segmentations were available for quantitative evaluation. To enable the latter, we manually annotated several images from the plant cover dataset and can therefore also evaluate our approach quantitatively.

Hence, our contributions are the following: We introduce a novel pretraining pipeline, which converts existing classification data of a plant-domain dataset into segmentation data usable for pretraining on the plant cover task. Moreover, we investigate and compare multiple pooling approaches and their differences for generating the abovementioned segmentations. Lastly, we evaluate the segmentations of the final part of our system on the plant cover dataset quantitatively by utilizing a small set of manually annotated plant segmentations.

2 Related Work

2.1 Weakly Supervised Object Localization (WSOL)

Weakly supervised object localization is an established field in Computer Vision research. While there are different kinds of approaches, the most recent ones are based on the method of *Class Activation Mapping* of Zhou *et al.* [48]. In their paper, the authors propose to utilize the classification weights learned by a classification layer at the end of the network to generate a map containing class

activations at each position of the last feature map, also called class activation map (CAM). To this end, the pooling layer is removed, and the fully connected layer is converted into a 1 by 1 convolutional layer. The generated CAM can then be thresholded and weakly supervised bounding boxes or segmentations generated. Multiple methods based on this approach tackle the problem by utilizing occluding data augmentation, e.g., by dropping parts of the images [32,5], or cutting and pasting parts of other images [41]. The aim of such augmentations is to prevent the network from relying too much on the most discriminative features and hence distribute the activations in the CAM more equally over the complete objects. In other approaches, this is done in more sophisticated ways, for example, by using adversarial erasing [45]. Choe *et al.* [4], however, showed recently that methods in this direction primarily gained performance improvements by indirectly tuning hyperparameters on the test set, resulting in almost no effective gain in performance on WSOL benchmarks in the last years. Nevertheless, recently, there have also been other approaches, which tackle the problem differently, for example, by modifying the way the CAMs are generated [26] or generating alternative maps for localizing objects [47].

We also base our method on CAMs. However, in contrast to the methods mentioned above, we also investigate changes to the base method by exchanging the pooling layer used during training. Specifically, utilizing global max pooling, or its approximation global log-sum-exp pooling, can potentially yield more benefits during the weakly supervised segmentation. This is because these methods do not depend on an averaging operation, potentially inhibiting good localization caused by dilution of activations during training.

2.2 Plant Analysis

The continuous developments of convolutional neural networks (CNNs) have also encouraged the development of automated plant analysis methods. These reach from simple plant species identification [39,2,19,10] over the detection of ripe fruit [7] and counting of agricultural plants [24,38] to the prediction of plant diseases [3]. However, the number of works concerned with plant cover determination is still small. The first work in this area was proposed by Kattenborn *et al.* [14], who analyzed the plant cover of several woody species, herbs and shrubs via UAV imagery. In their work, they utilized delineations in the images as training data for their custom CNN, which is not comparable with the data analyzed in this work. Nevertheless, previous works on plant cover analysis of herbaceous plants, specifically on the InsectArmageddon dataset by [36], were published by Körschens *et al.* [17,18]. They did several analyses with a custom network [17], as well as multiple established network architectures with different pretraining methods [18]. As the base of their approach, they model the problem as a weakly supervised segmentation approach, where pixel-wise probabilities for each plant are calculated, which are then aggregated into the final cover percentages. While we also utilize this basic approach, we go more into depth regarding the findings in [18]. Körschens *et al.* found that pretraining on a related dataset, albeit

Fig. 1. Example images from the pretraining dataset. The plant species shown are from left to right: *Achillea millefolium* (Yarrow), *Centaurea jacea* (brown knapweed), *Plantago lanceolata* (ribwort plantain), *Trifolium pratense* (red clover), *Scorzoneroides autumnalis* (autumn hawkbit) and *Grasses*, which are not differentiated into different species.

comparably small, is advantageous when tackling the PCP problem. We, however, argue that the efficiency of pretraining could be massively improved when training on segmentation data. For regular classification training with global average pooling, the network primarily focuses on the most discriminative regions [48]. With segmentation data, however, the network is encouraged to focus on the full extent of the object instead of the few most discriminative parts. For this reason, we investigate a similar approach to [18] with a strongly different pretraining process, in which we initially generate weakly supervised segmentations instead of using classification data directly. Moreover, we also include a numerical evaluation of the prediction quality of the visible plants in the images.

3 Datasets

In our experiment we utilize two separate datasets: one for pretraining and one for the actual plant cover training.

3.1 Pretraining Dataset

The pretraining dataset we use in our experiments contains species-specific randomly selected images from the Global Biodiversity Information Facility[3] (GBIF) [8]. The dataset encompasses images of 8 different plant species in natural settings, which match the ones from the plant cover dataset explained in Section 3.2

The plant species in the datasets and their respective abbreviations used in parts of this paper are *Achillea millefolium* (Ach_mil), *Centaurea jacea* (Cen_jac), *Lotus corniculatus* (Lot_cor), *Medicago lupulina* (Med_lup), *Plantago lanceolata* (Pla_lan), *Scorzoneroides autumnalis* (Sco_aut) and *Trifolium pratense* (Tri_pra).

The pretraining dataset comprises 6000 training images and 1200 validation images, which are evenly distributed across the classes, making the dataset balanced. Example images from the dataset are shown in Figure 1.

[3] http://gbif.org

Fig. 2. Example images from the InsectArmageddon [17,36] dataset. The growth process of the plants, as well as other changes over the time, like the flowering process, are captured in the images.

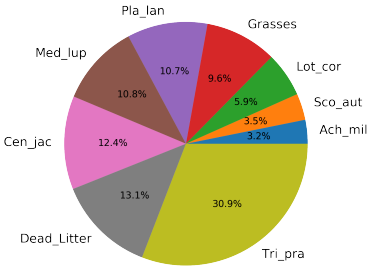

Fig. 3. The distribution of plant cover percentages over the different species from the InsectArmageddon dataset. *Trifolium pratense* takes almost a third of the dataset while *Achillea millefolium* is the least abundant plant with only 3.2% of the total cover. Figure taken from [18].

3.2 Plant Cover Dataset

We utilize the same dataset introduced by Ulrich *et al.* in [36] and Körschens *et al.* in [17]. As the latter refer to it as InsectArmageddon dataset due to its origin in the eponymous project[4], we will also refer to it this way. The dataset contains images from nine different plant species collected in enclosed boxes: so-called EcoUnits. Two cameras collected images in each of the 24 utilized EcoUnits over multiple months from above with a frequency of one image per day, hence also capturing the growth process of the plants. However, due to the laboriousness of the annotation of the images, only weekly annotations are available, leading to 682 images with annotations. Examples of the plants and their state of growth at different points in time are shown in Figure 2.

The nine different plant species in the dataset are the same as introduced in section 3.1 with the addition of *Dead Litter*. The latter was introduced to designate dead plant matter, which is usually indistinguishable regarding the

[4] https://www.idiv.de/en/research/platforms_and_networks/idiv_ecotron/experiments/insect_armageddon.html

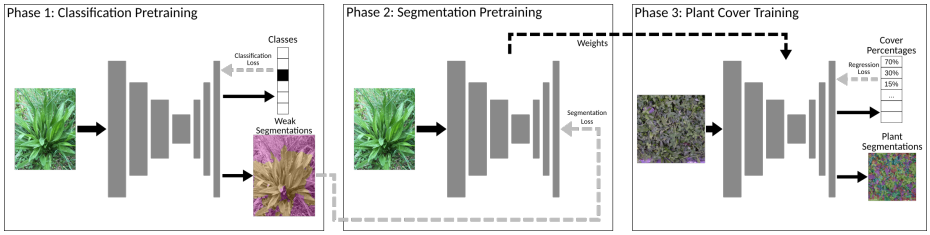

Fig. 4. Our proposed WSOL pretraining pipeline. The first network is trained on classification data and generates segmentation maps for the class in question based on the CAM. The second network is trained on this segmentation data and the trained weights are then utilized in the network used in plant cover training.

original plant species. As seen in Figure 3, the plants are in a heavily imbalanced long-tailed distribution, with *Trifolium pratense* spearheading the distribution and *Achillea millefolium* trailing it.

The plant cover annotations themselves are quantized into the so-called Schmidt-scale [27], so that the values can be estimated better. The scale contains the percentages 0, 0.5, 1, 3, 5, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90 and 100. For details on the data- and image-collection process, we refer to [36,6,35,17].

4 Method

In this section we will explain the general workflow of our method, followed by the pooling methods we use and the network head utilized during plant cover prediction. For better explainability, in the following we will view the network as consisting of three different parts: backbone, pooling layer and head. The backbone represents the feature extractor part of the network, and the head the task specific layer(s), which, in case of classification, is usually a fully connected layer that directly follows a pooling layer.

4.1 Pipeline

Our proposed pipeline can be divided into three distinct phases and an overview is shown in Figure 4. The first phase consists of a simple classification training on the pretraining dataset. The network used during this training has the shape of a typical classification network: a backbone, followed by a pooling operation and a fully connected layer with softmax activation. Hence, a standard classification loss like categorical cross-entropy can be applied. After the training, we remove the pooling layer of the network and convert the fully-connected layer into a convolutional layer as shown in [48] to generate class activation maps (CAMs). We isolate the single CAM belonging to the class annotation of the image and apply a threshold to generate a discrete segmentation map. The threshold is, as described in [48], a value relative to the maximum activation, e.g., 0.2.

In the second phase of our method, we utilize the segmentations generated in the first phase to train another network consisting of only a backbone and a segmentation head, i.e., a simple pixel-wise classifier with sigmoid activation. For training this network, a segmentation loss, e.g., a dice loss [25], can be applied. Upon this network being trained, we can use the weights to initialize the network from phase three for transfer learning.

In the third phase of our approach, we use the network initialized with the segmentation weights to fine-tune on the PCP task. This can be done using a simple regression loss, e.g., the mean absolute error. Finally, the network can be used to generate plant cover predictions in addition to segmentation maps for the plants in the image.

4.2 Pooling Methods

As mentioned above, in our experiments, we will investigate three different pooling methods: global average pooling (GAP), global max pooling (GMP), and global log-sum-exp pooling (GLSEP). GAP is the pooling method usually used for classification training and, therefore, also in the CAM method. However, as shown in [48], networks trained in such a way focus primarily on the most discriminative features in the images. Moreover, as the averaging operation encourages the distribution of higher activations over greater areas in the images, the CAMs dilute and are not optimal for good localization, resulting in worse segmentations. This is not the case for GMP, however, Zhou *et al.* [48] argue that GMP is, in contrast, more prone to focusing on single points with high activations in the images. Therefore, we investigate GLSEP, as log-sum-exp is an approximation of the maximum function and, due to its sum-part, does not only focus on a single point in the image. Hence, GLSEP can also be viewed as a parameter-free compromise between GAP and GMP. While variations of such a pooling method have been investigated before [44,28], to the best of our knowledge, none have been investigated for WSOL in conjunction with CAMs.

4.3 Plant Cover Prediction Head

To take into account the relatively complex calculations done during cover prediction, we utilize the already established plant cover prediction head from [17], with the modification introduced in [18]. To summarize, with this network head the plant cover prediction is viewed as a pixel-wise classification problem, where the classes consist of the plant classes in the dataset in addition to a background class and an irrelevance class. The background class serves as indicator of regions relevant for the plant cover calculation, while not containing any relevant plants (e.g., the soil in the images or weeds not monitored in the experiment). The irrelevance class indicates regions, which are not relevant for the calculation (e.g., the walls of the EcoUnits). Due to possible occlusion between the plants, the plant classes are modeled as not mutually exclusive, while they are mutually exclusive with background and irrelevance. For the details on this approach, we refer to [17,18].

5 Experiments

5.1 Evaluation and Metrics

We evaluate the our method after the last step: the plant cover training. For the evaluation, we utilize the metrics introduced by Körschens *et al.* [17], i.e., the mean absolute error (MAE) and the mean scaled absolute error (MSAE). The latter was considered to enable a fairer comparison of two imbalanced plant species and is defined as the mean of the species-wise absolute errors divided by their respective species-wise mean cover percentage. The exact percentages used during calculation can be found in [17]. To evaluate the quality of the segmentations, we utilize the mean Intersection-over-Union (mIoU), which is commonly used to evaluate segmentation and object detection tasks:

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{P_{pred}^{c_i} \cap P_{true}^{c_i}}{P_{pred}^{c_i} \cup P_{true}^{c_i}} \,, \tag{1}$$

with N being the number of classes, $P_{pred}^{c_i}$ the set of pixels predicted as class i, and $P_{true}^{c_i}$ the respective ground-truth counterpart. Since the original plant cover image dataset [36,17] did not contain any segmentations, we annotated 14 images containing large numbers of plants ourselves and used them for evaluation. It should be noted that these 14 images had no plant cover annotations, and hence are not seen during training. Several example images with our annotations can be found in the supplementary material.

5.2 Plant Cover Prediction

As mentioned above, our training is conducted in three separate phases, and we explain the setups in the following. In all three phases, we utilize the same networks architecture, which is a ResNet50 [11] with a Feature Pyramid Network (FPN) using a depth of 512 and the extraction layer P2, as shown in [20]. We choose this architecture because Körschens *et al.* [18] also investigated this one, among others. While in [18], a DenseNet121 [13] performed best, in our experiments we were not able to successfully train a DenseNet in the same setting and hence choose the ResNet50, as it achieved comparable performance in [18] when using an FPN.

Classification Pretraining. For classification pretraining, we utilize the standard cross-entropy loss and the Adam optimizer [16] with a learning rate of 1e-4, an L2 regularization of 1e-6, and a batch size of 12. The ResNet50 (without FPN) has been initialized with ImageNet weights [31]. We do not train for a fixed number of epochs, but monitor the accuracy on the validation set and reduce the learning rate by a factor of 0.1, if there were no improvements for four epochs. We repeat this process until there have been no improvements for six epochs, after which we end the training. Data augmentation is done by random rotations, random cropping, and random horizontal flipping. During the augmentation process, the images are resized such that the smaller image

Table 1. The results of our experiments with segmentation pretraining (SPT) in comparison to regular classification pretraining, considering different pooling methods. We evaluate the MAE, MSAE, total mIoU and the mIoU for plants only (without background and irrelevance classes). Best results are marked in bold font.

Pooling	SPT	MAE	MSAE	mIoU	mIoU (plants)
GAP	✗	5.23%	0.501	0.148	0.161
	✓	5.23%	0.501	**0.171**	**0.179**
GMP	✗	5.23%	0.501	0.144	0.157
	✓	**5.17%**	**0.494**	0.165	0.176
GLSEP	✗	5.24%	0.504	0.156	0.162
	✓	5.23%	0.503	0.161	0.174

dimension has a size of 512 pixels, then the images are cropped to a size of 448×448 pxused for further processing.

Segmentation Pretraining. The setup for our segmentation pretraining is mostly the same as in the first phase. However, we utilize a dice loss during this training process and monitor the mIoU instead of the accuracy. It should be noted that we initialize the network again with ImageNet weights and do not use the weights from the first pretraining to have a fairer comparison between training using segmentations and training using class information only.

Plant Cover Training. During the plant cover training, we use the same setup as in [18], i.e., image sizes of 1536×768 px, a batch size of 1, Adam optimizer with a learning rate of 1e-5, a training duration of 40 epochs, and a simple horizontal flipping augmentation. For the loss function, we utilize the MAE and we run our experiments in a 12-fold cross-validation ensuring that images from the same EcoUnit are in the same subset, as done in [17,18].

Results. In our experiments, we compare the effect of training using the weakly supervised segmentations with training using only class labels. The latter corresponds to using the weights from the first phase of our method directly in the third phase, and in case of GAP, this is equivalent to the method shown in [18]. The results of these experiments are summarized in Table 1.

We notice that regarding MAE and MSAE, the results using segmentation pretraining always perform at least as good as with standard classification pretraining or even better, albeit often only by small amounts. Moreover, we can see that the top results for MAE and MSAE do not coincide with the best segmentations, as the top error values have been achieved using GMP segmentation pretraining (MAE of 5.17% and MSAE of 0.494), while the best segmentations, measured by mIoU, are achieved when using segmentation pretraining with GAP.

5.3 Detailed Analysis of Segmentations

To have a more thorough insight into the effect of the segmentation pretraining (SPT), we also investigate the different IoU values for each plant species,

Table 2. Detailed results for the composition of the mIoU values from Table 1. Abbreviations used: SPT=Segmentation pretraining, DL=Dead Litter, BG=Background (relevant for calculation), IRR=Background (irrelevant for calculation), PO=Value for plants only (excludes BG and IRR); Top results per species are marked in bold font.

Pooling	SPT	Ach_mil	Cen_jac	Grasses	Lot_cor	Med_lup	Pla_lan	Sco_aut
GAP	✗	0.043	0.101	0.442	0.015	0.082	0.151	**0.001**
		±0.020	±0.028	±0.042	±0.007	±0.014	±0.009	±0.001
GAP	✓	**0.099**	**0.147**	0.438	0.034	0.100	0.152	0.000
		±0.033	±0.007	±0.021	±0.014	±0.010	±0.017	±0.001
GMP	✗	0.024	0.097	0.435	0.010	0.062	0.151	**0.001**
		±0.020	±0.021	±0.024	±0.003	±0.011	±0.016	±0.001
GMP	✓	0.095	0.135	0.429	0.035	0.110	0.159	**0.001**
		±0.028	±0.010	±0.034	±0.016	±0.017	±0.017	±0.001
GLSEP	✗	0.024	0.102	**0.447**	0.019	0.079	0.155	0.001
		±0.020	±0.017	±0.019	±0.007	±0.011	±0.011	±0.000
GLSEP	✓	0.064	0.141	0.411	**0.043**	**0.112**	0.167	0.000
		±0.032	±0.017	±0.034	±0.011	±0.016	±0.015	±0.000

Pooling	SPT	Tri_pra	DL	BG	IRR	Total	PO
GAP	✗	0.528	0.083	0.122	0.062	0.148	0.161
		±0.028	±0.008	±0.009	±0.049	±0.011	±0.011
GAP	✓	**0.556**	**0.088**	**0.144**	0.124	**0.171**	**0.179**
		±0.015	±0.009	±0.011	±0.035	±0.007	±0.006
GMP	✗	0.550	0.086	0.125	0.043	0.144	0.157
		±0.013	±0.006	±0.010	±0.031	±0.004	±0.005
GMP	✓	0.543	0.077	0.125	0.109	0.165	0.176
		±0.016	±0.008	±0.006	±0.087	±0.009	±0.008
GLSEP	✗	0.551	0.082	0.122	**0.138**	0.156	0.162
		±0.016	±0.010	±0.006	±0.097	±0.013	±0.006
GLSEP	✓	0.539	0.086	0.133	0.076	0.161	0.174
		±0.013	±0.006	±0.011	±0.055	±0.009	±0.008

shown in Table 2. It should be noted that due to the small size of many plants, precisely pinpointing their locations with CNNs is a difficult task, leading to relatively small IoU values overall. We see that the IoU values for the less abundant plants in the dataset, especially *Achillea millefolium*, *Centaurea jacea* and *Lotus corniculatus*, increase massively when applying SPT compared to only classification pretraining. Depending on the pooling strategy, the IoU increases by up to 250% for *Achillea millefolium* and *Lotus corniculatus*, and up to 150% for *Centaurea jacea*. This indicates that more relevant features for these species are included during SPT, confirming our intuition. In contrast to this, we also note that the IoU for *Grasses* and in parts for *Trifolium pratense* decreases. For the former, the reason might be that grasses are hard to segment automatically due to their thinness, leading to worse pretraining for these classes during SPT.

To counter this problem, it might be possible to utilize a network with a higher output resolution, e.g., a deeper FPN layer. The decreasing IoU for *Trifolium pratense* can be attributed to the balancing effect in the predictions introduced by the SPT, as a more balanced dataset usually results in worse performance for the more dominant classes. Regarding the varying pooling strategies used during the classification pretraining, we notice that they perform differently, depending on the plant species. We attribute this to the structure of the different plant species. This means that thinner and smaller objects are more easily recognized after training with GLSEP, as this presumably generates features that are more focused on smaller areas. However, plants with large leaf areas are more easily recognizable by networks using the likely more unfocused features generated by networks with GAP due to the averaging operation. Finally, it should be noted that the low IoU for *Scorzoneroides autumnalis* is caused by the small abundance of this species in the segmentation dataset. Hence, no conclusion can be drawn for this plant at this point. Multiple example segmentations using our method can be found in the supplementary material.

To summarize, the segmentation pretraining proved to be superior to using only the standard classification pretraining. It consistently improves the MAE and MSAE over classification pretraining by small amounts and especially improves the quality of the segmentations in general, with more significant improvements for the less abundant plants in the dataset.

6 Conclusions and Future Work

In this work, we proposed a novel pretraining pipeline for plant cover prediction by generating segmentation maps via a weakly supervised localization approach and using these maps for an additional segmentation pretraining. In general, we demonstrated superior performance of our approach compared to only standard classification pretraining for identifying the plants in the visible parts of the image, with more significant improvements for the less abundant plants in the dataset. More specifically, we noticed that the recognition of the less abundant plants improved the most, while the detection of the most prevalent plants, i.e., grasses and *Trifolium pratense*, decreases slightly. This effect is likely caused by the increased training set balance that the segmentation pretraining introduces. We also observed that the investigated pooling strategies perform differently depending on the plant species, which is likely caused by the varying structures of each plant that are handled differently by the individual aggregation methods during training of the network.

Our approach offers multiple directions for further improvements. First, the quality of the segmentation maps generated after the classification pretraining in the first phase could be improved, for example, by applying further data augmentation approaches commonly used in weakly supervised object localization [32,5,41]. Second, the segmentation pretraining could be altered by using different kinds of augmentations, losses, or also different kinds of segmentation networks, e.g., a U-Net [29]. Third, as segmentation maps are generated in the

training process, these maps could also be utilized for applying an amodal segmentation approach [42,21] to better deal with occlusions in the plant cover images. Lastly, the proposed approach might also potentially be applicable for fine-grained classification, as well as segmentation tasks, in general.

Acknowledgement

Matthias Körschens thanks the Carl Zeiss Foundation for the financial support. We would also like to thank Alban Gebler and the iDiv for providing the data for our investigations.

References

1. 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)
2. 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)
3. Chen, J., Chen, J., Zhang, D., Sun, Y., Nanehkaran, Y.: Using deep transfer learning for image-based plant disease identification. Computers and Electronics in Agriculture **173**, 105393 (2020)
4. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3133–3142 (2020)
5. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2219–2228 (2019)
6. Eisenhauer, N., Türke, M.: From climate chambers to biodiversity chambers. Frontiers in Ecology and the Environment **16**(3), 136–137 (2018)
7. Ganesh, P., Volle, K., Burks, T., Mehta, S.: Deep orange: Mask r-cnn based orange detection and segmentation. IFAC-PapersOnLine **52**(30), 70–75 (2019)
8. GBIF.org: Gbif occurrence downloads (13 May 2020), https://doi.org/10.15468/dl.xg9y85,https://doi.org/10.15468/dl.zgbmn2,https://doi.org/10.15468/dl.cm6hqj,https://doi.org/10.15468/dl.fez33g,https://doi.org/10.15468/dl.f8pqjw,https://doi.org/10.15468/dl.qbmyb2,https://doi.org/10.15468/dl.fc2hqk,https://doi.org/10.15468/dl.sq5d6f
9. 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)
10. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
11. 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 (2016)

12. 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 (2018)
13. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
14. 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)
15. Kim, B., Han, S., Kim, J.: Discriminative region suppression for weakly-supervised semantic segmentation. arXiv preprint arXiv:2103.07246 (2021)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Körschens, M., Bodesheim, P., Römermann, C., Bucher, S.F., Ulrich, J., Denzler, J.: Towards confirmable automated plant cover determination. In: European Conference on Computer Vision. pp. 312–329. Springer (2020)
18. Körschens, M., Bodesheim, P., Römermann, C., Bucher, S.F., Ulrich, J., Denzler, J.: Automatic plant cover estimation with convolutional neural networks. arXiv preprint arXiv:2106.11154 (2021)
19. 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)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
21. Ling, H., Acuna, D., Kreis, K., Kim, S.W., Fidler, S.: Variational amodal object completion. Advances in Neural Information Processing Systems **33** (2020)
22. 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)
23. 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)
24. Lu, H., Cao, Z., Xiao, Y., Zhuang, B., Shen, C.: Tasselnet: counting maize tassels in the wild via local counts regression network. Plant methods **13**(1), 79 (2017)
25. Milletari, F., Navab, N., Ahmadi, S., V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571
26. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
27. Pfadenhauer, J.: Vegetationsökologie - ein Skriptum. IHW-Verlag, Eching, 2. verbesserte und erweiterte auflage edn. (1997)
28. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1713–1721 (2015)

29. 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)
30. 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)
31. 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)
32. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE international conference on computer vision (ICCV). pp. 3544–3553. IEEE (2017)
33. 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)
34. Stammes, E., Runia, T.F., Hofmann, M., Ghafoorian, M.: Find it if you can: End-to-end adversarial erasing for weakly-supervised semantic segmentation. arXiv preprint arXiv:2011.04626 (2020)
35. 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)
36. Ulrich, J., Bucher, S.F., Eisenhauer, N., Schmidt, A., Türke, M., Gebler, A., Barry, K., Lange, M., Römermann, C.: Invertebrate decline leads to shifts in plant species abundance and phenology. Frontiers in plant science **11**, 1410 (2020)
37. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1568–1576 (2017)
38. Xiong, H., Cao, Z., Lu, H., Madec, S., Liu, L., Shen, C.: Tasselnetv2: in-field counting of wheat spikes with context-augmented local regression networks. Plant Methods **15**(1), 150 (2019)
39. 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)
40. Yao, Q., Gong, X.: Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access **8**, 14413–14423 (2020)
41. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019)
42. Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., Loy, C.C.: Self-supervised scene deocclusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3784–3792 (2020)
43. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12765–12772 (2020)
44. Zhang, B., Zhao, Q., Feng, W., Lyu, S.: Alphamex: A smarter global pooling method for convolutional neural networks. Neurocomputing **321**, 36–48 (2018)

45. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1325–1334 (2018)
46. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: Proceedings of the European conference on computer vision (ECCV). pp. 597–613 (2018)
47. Zhang, X., Wei, Y., Yang, Y., Wu, F.: Rethinking localization map: Towards accurate object perception with self-enhancement maps. arXiv preprint arXiv:2006.05220 (2020)
48. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)