M. Körschens, S. F. Bucher, P. Bodesheim, J. Ulrich, J. Denzler and C. Römermann:
Determining the Community Composition of Herbaceous Species from Images using Convolutional Neural Networks.
Ecological Informatics. 2024.

Determining the Community Composition of Herbaceous Species from Images using Convolutional Neural Networks

Matthias Körschens[a,b,c,*], Solveig Franziska Bucher[a,c,d], Paul Bodesheim[b], Josephine Ulrich[a,c], Joachim Denzler[b,c,d], Christine Römermann[a,c,d]

[a] *Plant Biodiversity Group, Institute of Ecology and Evolution with Botanical Garden and Herbarium Haussknecht, Friedrich Schiller University, Philosophenweg 16, D-07743, Jena, Germany*

[b] *Computer Vision Group, Institute of Computer Science, Friedrich Schiller University, Ernst-Abbe-Platz 2, D-07743, Jena, Germany*

[c] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, D-04103, Leipzig, Germany*

[d] *Michael Stifel Center Jena, Leutragraben 1, D-07743, Jena, Germany*

Abstract

Global change has a detrimental impact on the environment and changes biodiversity patterns, which can be observed, among others, via analyzing changes in the composition of plant communities. Typically, vegetation relevées are done manually, which is time-consuming, laborious, and subjective. Applying an automatic system for such an analysis that can also identify co-occurring species would be beneficial as it is fast, effortless to use, and consistent. Here, we introduce such a system based on Convolutional Neural Networks for automatically predicting the species-wise plant cover. The system is trained on freely available image data of herbaceous plant species from web sources and plant cover estimates done by experts. With a novel extension of our original approach, the system can even be applied directly to vegetation images without requiring such cover estimates. Our extended approach, not utilizing dedicated training data, performs similarly to humans concerning the relative species abundances in the vegetation relevées. When trained on dedicated training annotations, it reflects the original estimates more closely than (independent) human experts, who manually analyzed the same sites. Our method is, with little adaptation, usable in novel domains and could be used to analyze plant community dynamics and responses of different plant species to environmental changes.

Keywords: Plant Biodiversity, Plant Cover, Deep Learning, Convolutional Neural Networks, Semantic Segmentation, Artificial Intelligence

*Corresponding author
Email addresses: matthias.koerschens@uni-jena.de (Matthias Körschens),

1. Introduction

The estimation of plant cover is an essential part of plant-ecological research, as the thereby investigated plant species composition allows assessing the effect of, for example, land use [14, 19] and insect abundance [60, 64] on ecosystems. Moreover, climate change is one of the most important factors influencing the plant community composition [53, 38, 41] and thus also subject of a large number of projects [38, 14, 60, 6]. Typically, vegetation is monitored only once per year, though we do know that it changes in the course of the season and across years. To capture these seasonal variations in species composition is, however, time consuming and costly. Especially in the context of experimental approaches, information on higher temporal resolution would be advantageous (see also [64]).

Collecting images capturing the vegetation can be automated with dedicated camera setups. The collected images can then be analyzed automatically using state-of-the-art computer vision and machine learning methods. In combination with such methods, the camera setups have the potential to generate high-quality data on plant communities. Furthermore, this high-quality data can also be generated in a very high temporal frequency, allowing for investigations with high temporal resolution. The latter is usually infeasible with manual analysis due to being too laborious. Hence, manually generated datasets often only have a small number of data points, while the high frequency of automated methods allows for a much better analysis of seasonal variation. Apart from the aforementioned advantages, automated methods also alleviate the workload of researchers regarding the estimation and generate consistent output over time while alleviating biases and potential errors introduced by human estimation. Lastly, with automated methods, expert knowledge, usually necessary for manual estimation, is not required, and such systems can be used by non-experts as well.

Machine learning methods, and especially deep learning approaches, have become prevalent tools in a large number of different disciplines over the last years. In the area of computer vision, they represent the state of the art in tasks like image classification [12, 39, 40] image instance segmentation [17], and object detection [7]. Also in the field of plant image analysis they are applied in many different scenarios, like plant species classification [59], plant disease detection [63] and agricultural applications [22, 45, 1]. Especially the usage of convolutional neural networks (CNNs) for image analysis has gotten ubiquitous, enabling automatic object identification and resulting in the automation of a large number of tedious tasks in biological areas, like animal species identification [5, 51] and plant species identification [2, 25, 34, 49]. A major reason for their prevalent usage is the strong abundance of image data, generated not only by omnipresent smartphones with cameras but also by automated camera systems continuously collecting new data. All this data can either be used to satisfy the rather large training data requirements of CNNs or be automatically analyzed using trained deep learning models.

solveig.franziska.bucher@uni-jena.de (Solveig Franziska Bucher), paul.bodesheim@uni-jena.de (Paul Bodesheim), josephine.ulrich@uni-jena.de (Josephine Ulrich), joachim.denzler@uni-jena.de (Joachim Denzler), christine.roemermann@uni-jena.de (Christine Römermann)

Nevertheless, despite their large potential, CNNs have seen little use in plant ecology research. Up until now, mostly rather simple tasks have been investigated with CNNs in this area, like single-image species identification [2, 25, 34, 49], blossom detection [70] or phenological analysis of homogeneous communities [61, 71, 23]. However, even more complex tasks can be solved with CNNs, for example, the prediction of plant cover from images.

The task of predicting the plant cover via CNNs has been investigated before in several instances. For example, there are two approaches in the area of remote sensing: Kattenborn et al. [24] investigate UAV imagery containing several kinds of shrubs and trees using a custom CNN architecture. They utilize delineations in the image as training data for their segmentation model, which is then used to estimate the cover percentages. In contrast to our investigations in this work, the trees and shrubs are visually relatively easy to discern, and no relevant occlusion is taking place. Du et al. [13] similarly analyze hyperspectral data of wetland communities taken by UAVs, also training with segmentation annotations. The plant communities surveyed were quite homogeneous and also visually easy to discern.

The automated prediction of vegetation cover from images, i.e., the fraction of ground covered by vegetation, has also been investigated in several instances [44, 3, 26, 56, 10]. Existing approaches in this area mostly tackle this problem by simple color analysis, as usually this problem is solvable in most parts by separating the green and non-green parts in the images. In multiple approaches additional differentiations are performed with classical computer vision methods. McCool et al. [44], for example, utilize local binary patterns to differentiate grasses and forbs by texture, Bauer and Strauss [3] use a hand-crafted rule-based system to separate residues, vegetation, stones and shadow and, similarly, King et al. [26] also use a rule-based system to distinguish mosses and their health. Sellers et al. [56] utilize classical (non-deep) machine learning algorithms in conjunction with object-based methods [4] to differentiate bryophytes, forbs, graminoids, shrubs and lichens. Existing methods, hence, mostly differentiate vegetation on a high level, while usually also using delineations as training target, which are not available in our case.

A detailed automatic analysis of species-rich vegetation plots of herbaceous plant species, despite its significant potential, is a much more complex problem. In addition to subtle differences between several plant species, the plants in the vegetation plots often grow in multiple layers, leading to heavy occlusion that is enormously problematic for visual identification. Moreover, if the plants are monitored over larger time spans, they also undergo optical changes induced by the growth and aging processes. Their size, form, and color change over time, making it challenging to consistently identify even the same plant individual in a series of images.

Concerning herbaceous plant species, Körschens et al. [33] and Körschens et al. [30, 31] previously introduced approaches that are the only ones concerned with the automatic analysis of plant cover. These approaches do not utilize manual segmentation annotations to differentiate the plants from one another but train with plant cover estimations from human annotators directly. These estimates are, especially in large and diverse plant communities, cheaper to collect than pixel-wise plant segmentation masks like in the abovementioned approaches.

However, gathering many image annotations is nevertheless a tedious task. Therefore, we supplement the training process with freely available images from web sources. Nowadays, several web platforms, e.g., GBIF [66] and iNaturalist [68] gather and curate observation and image data, which usually includes species information usable as annotations for network training.

Here, we use and build on the approach from Körschens et al. [31], which involves a three-phase training process. In the first phase, aforementioned images from web sources are used to train a classifier to differentiate plant species, which is then used to automatically generate segmentations, i.e., delineations of the plants in the images. In the second phase, these generated delineations are used to train another network to segment plant species in images, thus gathering more comprehensive knowledge about the appearance of the plant species. This second network is used as the starting point of the plant cover training, during which we train the network on the cover estimates provided by an expert. In this work, we propose an extension of the approach from Körschens et al. [31], which eliminates the need to train on plant cover estimates by directly applying the trained segmentation model from phase two to the vegetation images for estimating the plant cover. With this enhancement, no manual data annotations, i.e., no cover estimates from an expert, are required for automated plant cover prediction, enabling easy adaptation of the network to new situations and environments. We will refer to this approach as *zero-shot* plant cover prediction. As in Körschens et al. [30, 31] only images of plant individuals taken in nature have been used as pre-training data (i.e., before training with plant cover estimates), in this work, we also investigate how beneficial the usage of images of preserved herbarium specimens is instead. While the latter have been investigated with deep learning methods before [62], to the best of our knowledge, such data has not yet been utilized for pre-training a neural network for downstream tasks. Herbarium specimens are usually more evenly sized in the images than specimens taken in the wild, and the former are completely visible, resulting in potentially valuable training data. However, due to the drying process, the plants often also change shape and color, making the plants look rather dissimilar to living individuals and two-dimensional. Moreover, the herbarium images are usually showing a side-view of the plants, which can potentially also be detrimental to training, as pictures of vegetation plots are typically taken from above.

We evaluate our approaches with different configurations on the InsectArmageddon dataset [64, 33] using previously applied and novel metrics and methods of evaluation. We also compare the performance of our method to the estimations of human experts to receive an indication of how well a CNN compares to estimates from biologists directly, which has not been done before.

To summarize, we formulate the following research questions:

- How well can CNNs predict the plant species composition in terms of species abundances in grassland communities from vegetation images?
- How well can such an analysis be performed without using plant cover estimations

from humans for training but only exploiting web images with species labels?

- How beneficial is the usage of images of preserved herbarium specimens with their drastically different image and plant structure in comparison to plant images taken in the wild during network pre-training?
- How does the performance of automated plant cover estimation compare to traditional estimates done by humans?

2. Materials & Methods

Our plant cover estimation method is evaluated on the so-called InsectArmageddon dataset, which we describe in detail in section 2.2. As this dataset comprises only a comparably small number of annotated images, we utilize the technique of transfer learning, in which we pre-train the network on another, larger and related dataset first and finally on the target dataset. Usually, the knowledge gained during pre-training can then, in parts, be applied in the final target task, resulting in a better performance in comparison to training only on the latter. Hence, we also utilize pre-training datasets, which we introduce in the following section (section 2.1).

2.1. Pre-Training Data

Similar to our previous works [30, 31], we utilize images of plant species extracted from the Global Biodiversity Information Facility (GBIF) [66] to pre-train our CNN. This pre-training serves as a preparation of the network for the plant species in the plant cover dataset to be analyzed. This way, the network is primed for the typical appearances and characteristics of the species, which not only increases performance in automatic cover estimation, but also enables us to apply the pre-trained model to cover data directly.

We use the eight plant species of the InsectArmageddon datasets described in section 2.2, namely *Achillea millefolium* L., *Lotus corniculatus* L., *Trifolium pratense* L., *Centaurea jacea* L., *Plantago lanceolata* L., *Scorzoneroides autumnalis* (L.) Moench, *Medicago lupulina* L. and a collective class of grasses, here represented by images from the genus *Festuca* Tourn. ex L. to reflect a variety of different possible grass species.

We built two datasets with these classes, one with images taken in a natural environment [65], in the following denoted as *GBIF natural*, and one comprising only preserved herbarium specimens [67], denoted as *GBIF preserved*. In these datasets, each species is represented by 900 images, from which 750 are utilized for training and 150 for validation purposes, resulting in 6000 training and 1200 validation images. Example images for both datasets can be found in Figure 1.

Figure 1: Example images for the GBIF natural dataset (left) and GBIF preserved (right). The former contains images taken in the natural environment with differing sizes and view points, while the latter contains plant specimens from herbaria scanned with mostly homogeneous sizes and view points. The plant species depicted in a certain row and column on the right side are the same as the ones depicted in the same row and column on the left side.

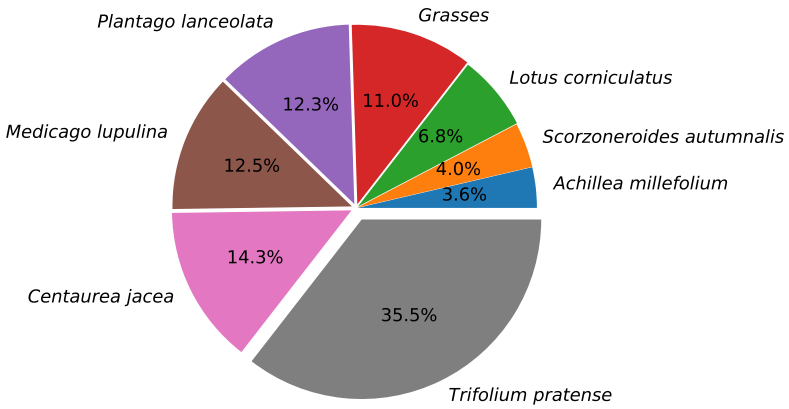

Figure 2: The distribution of plant species abundances in the InsectArmageddon dataset.

2.2. Plant Cover Data

For our experiments we utilize the so-called InsectArmageddon dataset introduced in Ulrich et al. [64], Körschens et al. [33], which was collected during the eponymous project[1] in 2018. The dataset covers imagery and respective plant cover annotations collected weekly over 18 weeks, from April to August. The images were collected by an automated camera system in so-called EcoUnits, which are experimental units with a base area of about 1.25×1.25 meters containing small enclosed ecosystems [55]. As described in [64], in these units, twelve herbaceous plant species were sown, ten of which were visible in the images and monitored throughout the experiment, collecting image data from two cameras per unit in the process. Three of the sown plant species were grasses (*Poaceae/Festucaceae*), which are not differentiable in the images and, therefore, were summarized in a single class of *Grasses*, resulting in eight plant categories in the dataset, seven of which represent species. It should be noted that the soil was not sterilized before or during the experiment. Hence, in rare cases unsown plant species occur in the images, which were removed in several cycles of weeding. The experiment also included several different treatments, leading to a variety of plant communities and reflecting a realistic experimental protocol. For details on the experimental setup, please refer to [64].

Species Abundances. The species abundances within the EcoUnits are heavily imbalanced, as shown in Figure 2. In the InsectArmageddon dataset, the most prevalent species is *T. pratense*, taking up about a third of the total cover in the dataset according to human estimates. The least abundant plant species are *A. millefolium*, *S. autumnalis* and *L. corniculatus*, which together amount to only about 14% of the dataset's cover.

[1] https://www.idiv.de/en/research/platforms-and-networks/idiv-ecotron/experiments/insect-armageddon.html

Figure 3: Example images from the InsectArmageddon dataset showing the change of the observed plants over the months of monitoring.

Images. The images in the dataset show, depending on the camera angle, direction of view and zoom level, up to about two thirds of the base area of the EcoUnits. In the first weeks, primarily bare soil can be seen, which is quickly overgrown after a few weeks. As the plants grow, they also heavily overlap each other, creating large areas with occlusions. The images capture most of the plants' life cycle and the different phenological stages, e.g., their flowering or senescence. Hence, the images cover the plant species in changing sizes, shapes, and colors. Several example images from a single EcoUnit are shown in Figure 3. It should also be noted that the borders of the EcoUnits are visible in many of the images. Therefore, not all of the image is relevant for cover calculation, and some parts should be excluded from the estimation process. Moreover, due to technical reasons, zoom levels can vary from camera to camera and week to week, leading to a rather inconsistent view of the vegetation plots, which had to be accounted for during method development. The viewing angle of the cameras, which was approximately vertical for all cameras, and their direction of view stayed consistent throughout the image collection process.

Annotations. The plant cover was surveyed using the images generated by every single camera of each unit and extracted from Ulrich et al. [64]. Due to being very laborious, this survey has been performed only weekly. The cover values were estimated by a single ecologist on the aforementioned image data alone using a slightly adapted version of the so-called Schmidt-scale [48], i.e., they are estimated in quantized percentages of 0, 0.5, 1, 3, 5, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90 and 100 percent. In the following, we will refer to these annotations as reference estimates or observed cover values. It should be noted that the estimates likely contain noise, which can, for example, be introduced by the subjectivity of the estimator, unclear image parts, and, of course, the quantization itself. As the cover values are independently estimated for each plant species, the sum of these values over all species contained in a single image can exceed 100% as it includes overlap of the plants.

After combining the images with the surveyed plant cover estimates and filtering out unus-

able images, we are left with a dataset of 682 annotated images distributed over all EcoUnits.

2.3. Methods for Automatically Extracting the Plant Cover

In this work, we utilize the base method we proposed in Körschens et al. [31], which is applied with several small modifications in comparison to the original work, including different kinds of data augmentations and additional postprocessing of images. Moreover, we present a novel extension to the system from Körschens et al. [31] to extract plant cover information without training on actual annotated plant cover data.

2.3.1. Base Network

To train the network, we follow the base method proposed in Körschens et al. [31]. This approach, which we will refer to as segmentation pre-training, is motivated by four aspects. First, there is usually only little training data available for plant cover estimation data, as annotating such images is highly laborious. CNNs usually require large amounts of training data to perform well. Therefore, it would be advantageous to be able to utilize additional external training data to improve the training results. Second, transfer learning, the task of training the network on larger datasets first, followed by fine-tuning the network on the target task, has been shown to improve the performance of CNNs drastically [28]. Third, in previous investigations [30] also found that performing pre-training on domain-related image data (e.g., images containing the same plant species as in the target dataset) is more beneficial in the plant cover prediction task than using unrelated data, like ImageNet images [54]. The fourth and last aspect to consider is the amount of image data freely available on the web, which is steadily increasing. Several facilities, e.g., GBIF or iNaturalist, have started collecting and structuring large amounts of occurrence and image data, specifically in biological areas. This kind of data is not only easily accessible but, in large parts, also pre-annotated with taxonomics and therefore also easily usable as training data. Therefore, this amount of well-annotated data is an optimal supplement to the already existing plant cover annotation data.

While we have the option to pre-train our network for classification and apply it on plant cover data afterward, we have shown in Körschens et al. [31] that training a network for classification only often leads to the network focussing on a small number of discriminative parts of the plants instead of all visible parts. This can be especially detrimental in plant cover prediction, where often only the leaves of the plants are visible, while the network focuses on the blossoms of the plants during pre-training, as they are usually the most discriminative parts.

Because of this, the approach we introduced in Körschens et al. [31] trains the network in three distinct phases, as shown in Figure 4. In the first phase, we perform standard classification training, i.e., we train the network to predict the respective plant species when given an input image. The trained network then utilizes the so-called class activation mapping (CAM) approach [73] to generate segmentation maps for each image (here also referred to as "weak segmentations" or "weak segmentation maps"), which usually cover most of the plant, including blossoms and leaves. This segmentation data is then used in

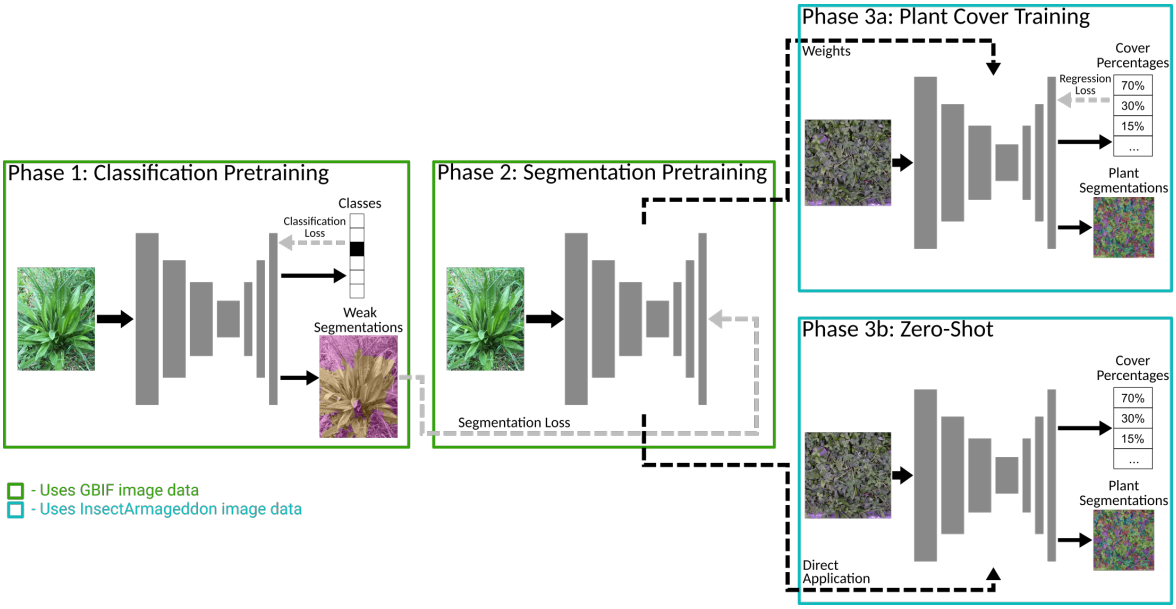

Figure 4: The basic 3-phase processing pipeline for training our network. In the first phase, a classification network is trained using GBIF image data, which applies class activation mapping to generate weak segmentation maps. In the second phase, these maps are used for training a segmentation network on the GBIF image data. In the third phase, the networks weights can be used to either train a plant cover prediction network on the actual vegetation data (like the InsectArmageddon dataset), or the network can be applied directly to vegetation data to generate zero-shot predictions without any training on plant cover annotations. Image adapted from Körschens et al. [31].

the second training phase to train a segmentation network, which, due to the nature of the previously generated pseudo-ground-truth segmentation maps, focuses on the complete plant instead of only single parts. The parameter values of this network, also known as weights, are then used as initialization of our plant cover prediction network in the third training phase, in which we utilize the plant cover annotations to train the network for regression of the cover values, as shown in Phase 3a in Figure 4. This approach is described more in detail in section 2.3.2. Alternatively, we can also apply the trained segmentation network from Phase 2 to the vegetation images directly to determine plant cover predictions without additional training, as shown in Phase 3b in Figure 4. This approach is referred to as our zero-shot approach and is explained more in detail in section 2.3.3.

2.3.2. The Cover-Trained Approach: Plant Cover Prediction with Dedicated Training Data

The network consists of two parts: a feature extractor and a network head. To be able to predict species cover values for our investigations with the neural network, the network head is dedicated to the computation of these values. To this end, we utilize the calculation model from Körschens et al. [30]. The plant cover is calculated by determining the occurrence probability for each plant species per pixel and then aggregating these into a numerical cover vector. Based on presence or absence of plant species, the calculation model also

determines probabilities for background areas like bare soil, as well as areas irrelevant for cover calculation. This kind of calculation of the plant cover values also has the advantage of being easily interpretable. I.e., the predicted pixel-wise classifications (the most probable species per pixel) can be viewed as segmentations and therefore analyzed visually by the user to confirm the correctness of the predictions.

The approach from Körschens et al. [30] we use here can be seen as a probabilistic approach, as the network predicts a probability of each plant species being contained in each pixel and aggregates these over the complete image.

Training a network with dedicated plant cover data has the advantage that the CNN can get more information about the underlying data and, therefore, can perform better than without dedicated training data. However, a potential drawback is that some underlying biases in the dataset or from the annotator could potentially also be introduced into the network's predictions, in addition to the laboriousness of gathering the annotations.

2.3.3. The Zero-Shot Approach: Plant Cover Prediction without Training Data

To avoid a laborious collection process of possibly biased training data, we also investigate a novel zero-shot plant cover prediction approach, i.e., plant cover prediction without the network having been trained on annotated cover data. To do this, we utilize the segmentation network trained in phase 2 of the abovementioned training process. In this phase, the network learns to segment the complete plant and determine its species, which is a necessity in plant cover prediction. Hence, we can directly apply this network to plant cover data, assign the class of each pixel to be the most probable class as predicted by the network, and then aggregate the prediction similar to before, i.e., average the pixel-wise predictions over the entire image. This has the advantage of not requiring additional annotations and being applicable on images of all resolutions, as CNN inference is usually much cheaper compared to network training. However, this approach also has certain caveats, which are discussed in section 2.3.4 and section 2.3.5.

2.3.4. Applying a Segmentation Network to Plant Cover Images

A direct application of the pre-trained segmentation network yields several advantages but poses challenges as well. An example of this is high-resolution processing. Training on high-resolution images without the possibility to train on image parts only is extremely expensive but still necessary for plant cover prediction training. Therefore, the images usually have to be scaled down to a smaller resolution than the original one, leading to missing details. When using the network only to predict the cover estimates from images of vegetation plots, without dedicated training beforehand, we can separate the image into smaller patches and predict on these sequentially. This makes it possible to process the vegetation images independent of their resolution without losing details. In the following, we will separate the original full-resolution image in full resolution into patches of 512×512 px in two interlaced grids as shown in Figure 5. We chose this patch size as we found it performed best in comparison with similar alternatives (e.g., 256×256 px). The reason for this is likely that 512×512 px is the most similar image size to the one used during the pre-training

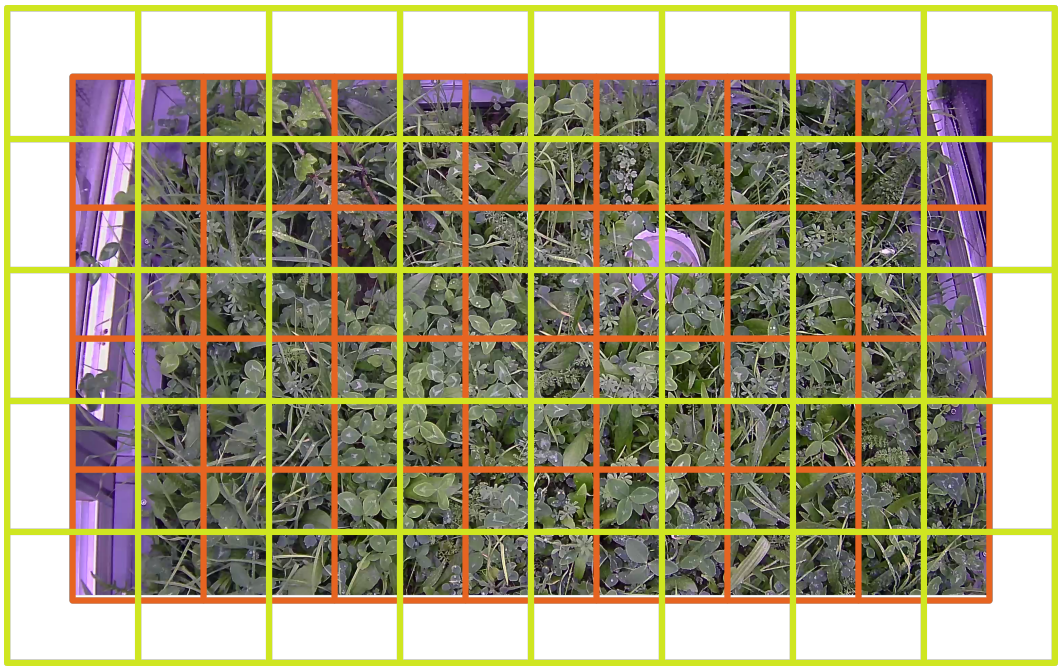

Figure 5: An example of an interlaced prediction grid. We split the image into evenly-sized patches, predict on these separately, and fuse the result. To reduce border effects, we do the same with a second grid which is offset by half a patch size and average the results over both grids. Different colors are used in this figure for a clear differentiation of the two prediction grids.

(448 × 448 px). The patches are processed independently from one another by the network and the output is put together with respect to the grid after prediction. The predictions of both grids are averaged, which mitigates prediction irregularities at the grid cell borders. It should be noted that the image can also be processed in a fully convolutional way (see [42]) without requiring patches with similar performance. However, the application of patches leaves the memory footprint constant, allowing processing on machines with suboptimal hardware, in contrast to fully convolutional processing.

Another challenge is the domain shift. The plants in the vegetation plots will not necessarily look equal to the images in the pre-training dataset, possibly differing in size, color, and shape. Such a shift should ideally be accounted for. We do this by performing test-time augmentation, i.e., predicting on several modified instances of the original image and averaging over these instances. During the prediction, we use the network to predict pixel-wise probabilities on horizontally and vertically flipped vegetation images, as well as images scaled to half the original resolution, and average the results over these. This process leads to better prediction results, especially for different plant sizes.

Another problem when applying zero-shot plant cover estimation is the prediction of areas irrelevant for cover calculation. As in some cases, parts of the images cannot contain plants, e.g., the walls of the EcoUnits in the InsectArmageddon dataset (see section 2.2), for a correct calculation, these should be excluded. Similarly, another challenge is the prediction of empty areas relevant for plant cover prediction, like bare soil.

2.3.5. Determination of Background and Irrelevant Areas

The images contain background areas in which plants could grow. Hence, they are relevant for cover calculation (e.g., bare soil). Other areas like the walls of EcoUnits are irrelevant. As both kinds of areas can take a considerable part of the image, they should be included in the calculation to prevent implicitly erroneous computations. When training with annotated plant cover directly, these areas can automatically be learned by the model and are therefore also included in the network's calculation [30]. However, including these areas in the automatic cover estimations is more difficult without dedicated plant cover annotations as pre-training data for such areas are not available. Hence, alternative solutions have to be found. Our proposed solution denoting areas irrelevant for cover calculation is to receive the annotations externally, i.e., from a source other than our pre-trained network. One possibility is providing an existing delineated mask of relevant areas directly, which can be very straightforward if the camera viewpoints and zoom do not change. Alternatively, especially for more inconsistent camera views, such a mask could be provided by a simple dedicated binary segmentation network, e.g., a U-Net [52] with a small set of annotated image data. It should be noted that, as irrelevant and relevant areas are usually very coarse, annotating a few images is usually not very laborious. As the InsectArmageddon dataset represents such an inconsistent scenario, we utilize the latter method for discerning relevant and irrelevant areas. Regarding background prediction, the solution is more complex as a respective annotation would require segmenting the soil while leaving out small and large plants from the foreground, which would be comparably time-consuming and therefore not desirable. Due

to its complexity, in this paper we ignore the issue of the missing background prediction and evaluate the method using the prediction of arbitrary plant species in background locations. It should be noted that the issue of background estimation can even be neglected in many cases, since usually most of the soil is overgrown by plants. This makes the influence of background pixels in the cover calculation rather small, while the main difficulty lies in the differentiation of the plants growing on top.

2.4. Network Setup for Investigations

For extracting plant species cover information for our investigations, we use the same setup as described in Körschens et al. [31] with slight differences. During all three phases, we use a ResNet50 [18], which is initialized with ImageNet [54] weights from Keras [9] before phase 1 and phase 2, as well as the AdamW optimizer [27, 43]. During each phase, we use the ResNet in conjunction with a Feature Pyramid Network (FPN) [36] to increase the network output resolution. Phase-specific parameters are listed in the following. The algorithm is implemented in Python [69] using the PyTorch framework [47].

The evaluations are done in a 12-fold cross validation manner. As in the InsectArmageddon dataset there are 24 EcoUnits, for each cross validation split we select two of them to validate on, and use the rest for training, if applicable. Results below are averaged over the 12 validation splits. Hence, for every cross validation split, we train on about 92% of the data (625 images on average), while testing on the remaining 8% (57 images on average). It should also be noted that, while the dataset consists of images in a time-series, we train and predict on images independently, ignoring the temporal aspect and relationships between the images.

Phase 1. During the first phase, we use an FPN with layer P2 and 256 features, a learning rate of 10^{-4} and a weight decay also of 10^{-4}. Moreover, we utilize Global Log-Sum-Exp Pooling (GLSEP) [29], as we found in previous work [29] that training with this pooling method improves segmentation when one applies the network in weakly supervised object localization (WSOL) [73, 8] tasks in comparison to standard global average pooling [35]. In this phase, we also utilize horizontal flipping as well as random rotation for data augmentation [57]. To improve the generated segmentation results, during the prediction of the segmentation maps, we use horizontal and vertical flipping as test-time augmentations and average over the network predictions for each augmentation. The loss used for training is the standard softmax cross entropy loss [15].

Phase 2. In the second phase, we utilize different configurations for the FPN. To increase the output resolution of the FPN further, we extend the FPN to include the other higher resolution parts of the ResNet50; creating the FPN layers P1 and P0, which we investigate in conjunction with the P2 layer and different numbers of features (P2: 512, P1: 256, P0: 128). We also utilize an additive combination of dice loss and binary cross entropy (BCE). Moreover, we use horizontal flipping and random rotation for data augmentation in this phase.

Phase 3. The third phase happens as described in Körschens et al. [31], with the difference that we utilize the mean scaled absolute error as loss, which we found worked slightly better than the standard mean absolute error. As described in Körschens et al. [31], we construct a cover estimation network by using the pre-trained ResNet50 with FPN from phase 2. Instead of the dedicated segmentation network used during phase 2, we modify the network to reflect the cover calculation model as described in Körschens et al. [30], in which the probabilities for each plant species, as well as background and irrelevance are predicted in a pixel-wise fashion and averaged over the image afterwards.

In this phase the network is trained with the mean scaled absolute error as regression loss, using a batch size of 1, an input image resolution of 1536×768 pixels, and a learning rate of 10^{-5}. Moreover, we utilize only horizontal flipping for data augmentation in this phase.

2.5. Baseline Comparison

To also have a comparison of our method with a simpler approach, we compare our results with the ones from a ResNet50 that was pre-trained on the ImageNet classification dataset [54] with added FPN, later also referred to at *ImageNet baseline model*. As the ImageNet dataset is a normal classification dataset that does not contain the plant species from our dataset, this network cannot be utilized for zero-shot cover prediction. Hence, we only evaluate this network after training it on our dedicated plant cover data.

2.6. Comparison with Human Experts

To evaluate the performance of the complete method, with and without training on plant cover data, we performed a study with several human estimators for comparison. The study comprised six plant ecologists with experience in plant cover estimation. In the study, the biologists estimated the cover values for each plant species in 12 different images from the InsectArmageddon dataset. The cover estimates, similar to the ones done by our CNN, have then been compared against reference cover estimates, which have been done by a single ecologist as described in section 2.2.

2.7. Statistical Analysis

To measure, how well our employed models can predict the species compositions, we utilize several metrics to shed light on different aspects of the models' predictions. For the analysis of the size of the species-wise error, we utilize the mean scaled absolute error with standard deviation as scale value (MSAE_σ), which we define as

$$MSAE_{\boldsymbol{\sigma}}(\boldsymbol{t}, \boldsymbol{p}) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{t_i}{\sigma_i} - \frac{p_i}{\sigma_i} \right| , \tag{1}$$

where $\boldsymbol{t}$ and $\boldsymbol{p}$ are the vectors of true and predicted values, respectively, and $\boldsymbol{\sigma}$ represents the species-wise standard deviations over the complete dataset. Due to the large imbalance in the plant cover values, even in single images, this metric relativizes the disproportions by

relating the values to the respective species-wise standard deviations. In the following, for simplicity, this metric will be referred to as MSAE.

The second metric utilized is the Intersection over Union metric (IoU), also known as Jaccard Distance [21], which we use to quantify the quality of the segmentations generated by the models. This quality tells us how well plant species in the topmost layer in the vegetation images are predicted. The IoU metric is defined as

$$IoU(T, P) = \frac{|T \cap P|}{|T \cup P|} \tag{2}$$

where T and P denote the sets of ground-truth segmentation pixels and predicted segmentation pixels, respectively. It should be noted that the IoU is evaluated on a small subset of 14 images, as done in Körschens et al. [31]. This is because the number of plant individuals per image is extremely high such that few images are already representative of the actual distribution. Additionally, manual delineations of such a large number of plant individuals very time intensive. It should be noted that the segmentations generated by the trained models only contain the class with the highest probability, i.e., the most dominant class or species per pixel, and do not provide an indication of the prediction of other classes in the same location.

The third metric we use we will refer to as DCA-Procrustes-Correlation, DPC in short. As the name suggests, the metric is a combination of a Detrended Correspondence Analysis (DCA) [20], and a Procrustes analysis [16]. To generate the DPC, we first predict the respective cover estimate outputs of our network for each plant species for each image in the dataset. Afterward, we apply a DCA on the matrix of these predicted values, as well as the respective reference estimates and compare these using the Procrustes analysis. This combination of the two analyses allows for a multivariate analysis of the distributions of predicted and observed values. For calculation, we utilize the *decorana* and *protest* functions as implemented in the *vegan* *R* package [46, 50] and report the resulting correlation value from the *protest* function. The advantage of the DPC is that, in contrast to metrics like the MSAE_σ, it quantifies, how well the entire joint cover distribution of all plant species is reflected, instead of considering the species independently.

We utilize these metrics to compare the different model setups and training methods with each other to answer two questions: 1. How well does the zero-shot model without any additional training data compare to the cover-trained model, that uses cover estimates as annotations; and 2. How well does the model perform when pre-trained with herbarium image data in comparison to images taken in nature.

For a better indication of our models' performance in comparison to the reference estimates, we compare the predicted values by our model with these estimates, and a linear model, in which we use the observed values as explanatory and the predicted values as dependent variables (i.e., a model of the form `observed = species * predicted`). To indicate the fit,

we provide the coefficient of determination R^2, and the corresponding p-values to indicate significance.

To get a further indication of the similarity of the model output in comparison with the reference human estimation, we also investigate the Shannon-diversity [58]. This provides insights into how the overall distributions of predictions and observed values differ.

Finally, to find out, how the models compare to human experts, we evaluate the average DPC of the human estimators and compare with the ones resulting from our models' predictions.

3. Results

The results of the evaluation of our models' performance are shown in Table 1. We compare the experimental results of the three different network setups in the zero-shot setting and the setting with training on cover values. Moreover, we also compare the zero-shot and cover-trained method with natural and preserved imagery used as pre-training. We find that the best-performing model for the zero-shot scenario is FPN P1 256 with GBIF natural pre-training, while the best-performing cover-trained model FPN P2 512 with GBIF natural pre-training. These two models will also represent the zero-shot approach and the cover-trained approach in the following analyses, respectively.

3.1. Comparison of Predicted and Observed Values

Figure 6 shows that the plant species composition is predicted well, with R^2 values ranging from 0.506 for the zero-shot model to 0.813 for the cover-trained model. The linear model reveals that both models are prone to underestimating cover values, which is more severe in the zero-shot model. More detailed species-wise plots can be found in the supplementary material S1.

3.2. Diversity

A comparison of the Shannon diversity scores over all sites is shown in Figure 7. While the overall trend is similar for both model types, the model fit is significantly better for the cover-trained model.

3.3. Segmentation Quality

Table 2 shows the prediction performance of the top-layer of plants of both models as measured by species-wise Intersection-over-Union (IoU). Figure 8 also shows several qualitative segmentations generated by the models.

3.4. Comparison with Human Experts

Our zero-shot cover prediction model achieves with a DCA-Procrustes-Correlation (DPC) of 0.55 ($\pm$0.072) a similar score in comparison to the human estimators (DPC of 0.62 ($\pm$0.14)). The cover-trained model, outperforms both alternatives by a large margin with a DPC of 0.77 ($\pm$0.003).

Table 1: The investigation results on the two datasets. "FPN PX N" denotes the usage of a Feature Pyramid Network with layer X and N features. Abbreviations: MSAE - Mean Scaled Absolute Error, IoU (Plants) - the intersection over union averaged over all plant species, DPC - DCA-Procrustes-Correlation. Top results are marked in **bold** font. All results are averaged over three repetitions.

Network Configuration	Pre-Training Dataset	No Cover Training (Zero-Shot)			Cover Training		
		MSAE	IoU (Plants)	DPC	MSAE	IoU (Plants)	DPC
FPN P0 128	ImageNet	-	-	-	0.534	0.128	0.715
FPN P1 256		-	-	-	0.521	0.153	0.750
FPN P2 512		-	-	-	0.510	0.169	0.757
FPN P0 128	GBIF Natural	1.113	**0.163**	0.503	0.534	0.156	0.693
FPN P1 256		**1.087**	0.162	**0.548**	0.507	0.181	0.758
FPN P2 512		1.096	0.157	0.527	**0.499**	**0.195**	**0.771**
FPN P0 128	GBIF Preserved	1.597	0.079	0.362	0.549	0.127	0.689
FPN P1 256		1.614	0.070	0.366	0.523	0.154	0.727
FPN P2 512		1.736	0.066	0.385	0.513	0.168	0.754

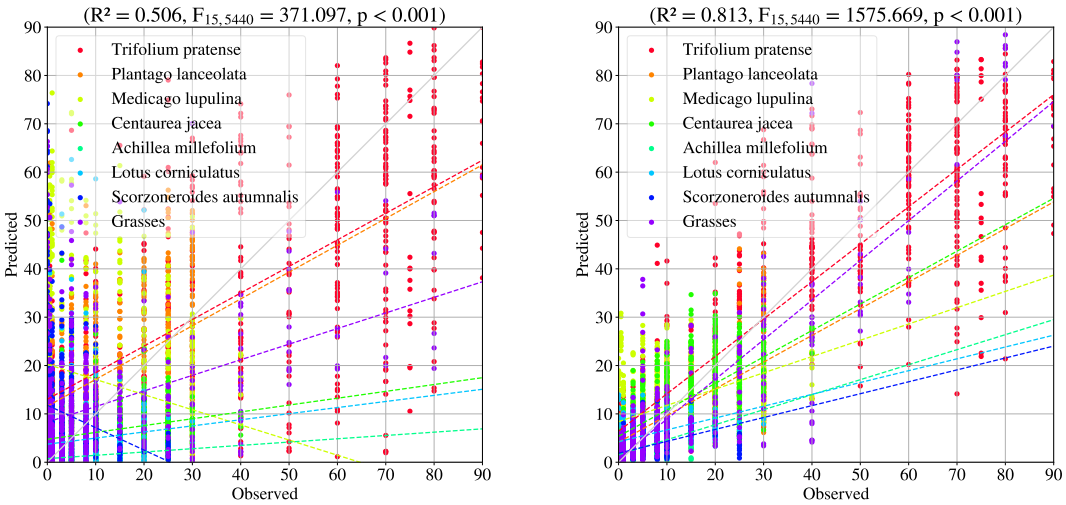

Figure 6: A comparison between the values predicted by our zero-shot model and the observed values on the left, and the respective comparison for the cover-trained estimation model on the right side. Each data point represents an image taken by a camera in a single week in the dataset. The continuous diagonal represents the identity, while the dashed lines represent the result of the linear regression analysis.

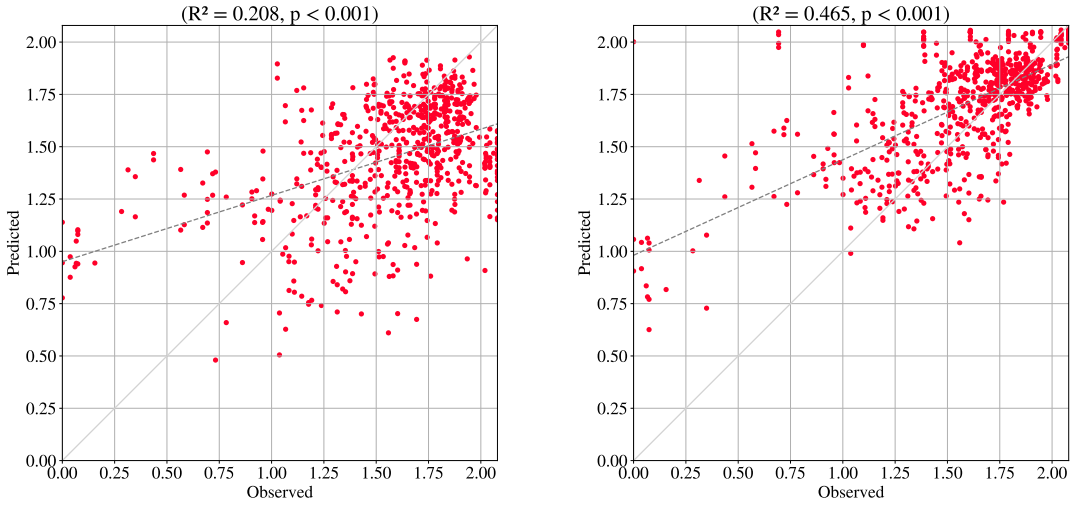

Figure 7: A comparison of the Shannon diversity calculated over the observed values and the predicted values of our zero-shot approach (left) and cover-trained approach (right).

Table 2: A quantitative comparison of Intersection-over-Union (IoU) values for the segmentation quality (i.e., correctness of the prediction of the top layer of plants) for our zero-shot method and the cover-trained method.

	A. millefolium	*C. jacea*	Grasses	*L. corniculatus*	*M. lupulina*
Zero-Shot	0.080	0.029	0.269	0.094	0.129
Cover-Trained	0.003	0.151	0.491	0.010	0.131
	P. lanceolata	*S. autumnalis*	*T. pratense*	**Total**	
Zero-Shot	0.207	0.014	0.471	0.162	
Cover-Trained	0.210	0.000	0.568	0.195	

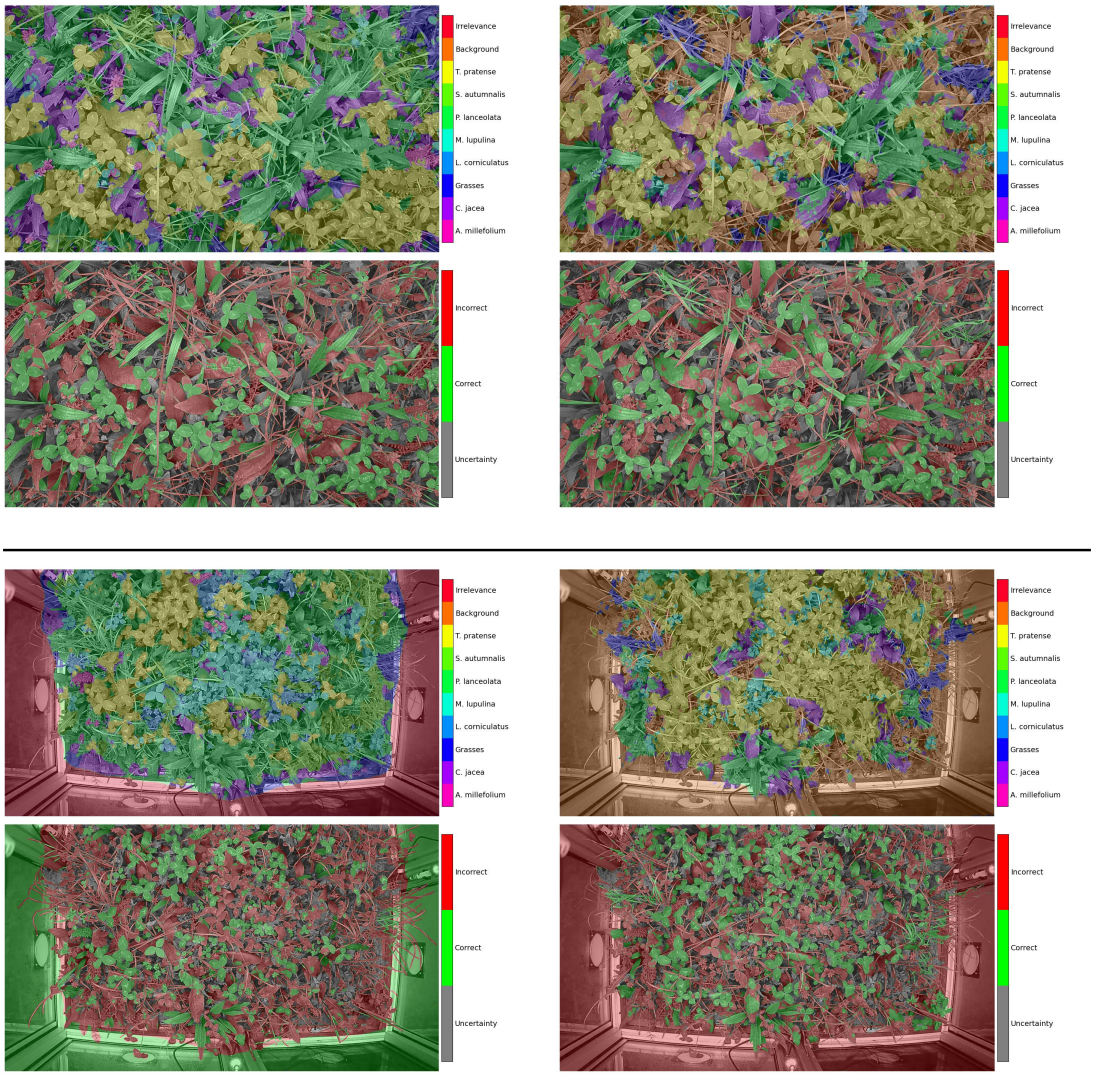

Figure 8: A qualitative comparison of the segmentation quality of the zero-shot models (left) and the cover-trained models (right). Below each segmentation map an error map indicates, in which locations the segmentation has been correct. Uncertainty areas mark locations in which experts were not able to determine a ground truth due to occlusions, empty areas, etc. These areas are exempt from the evaluation process. It is visible that, while the predictions of several less abundant plants are better using the zero-shot model in comparison to its counterpart, it cannot predict several common plants well and also no background (e.g., soil), leading to miscalculation of the cover percentages.

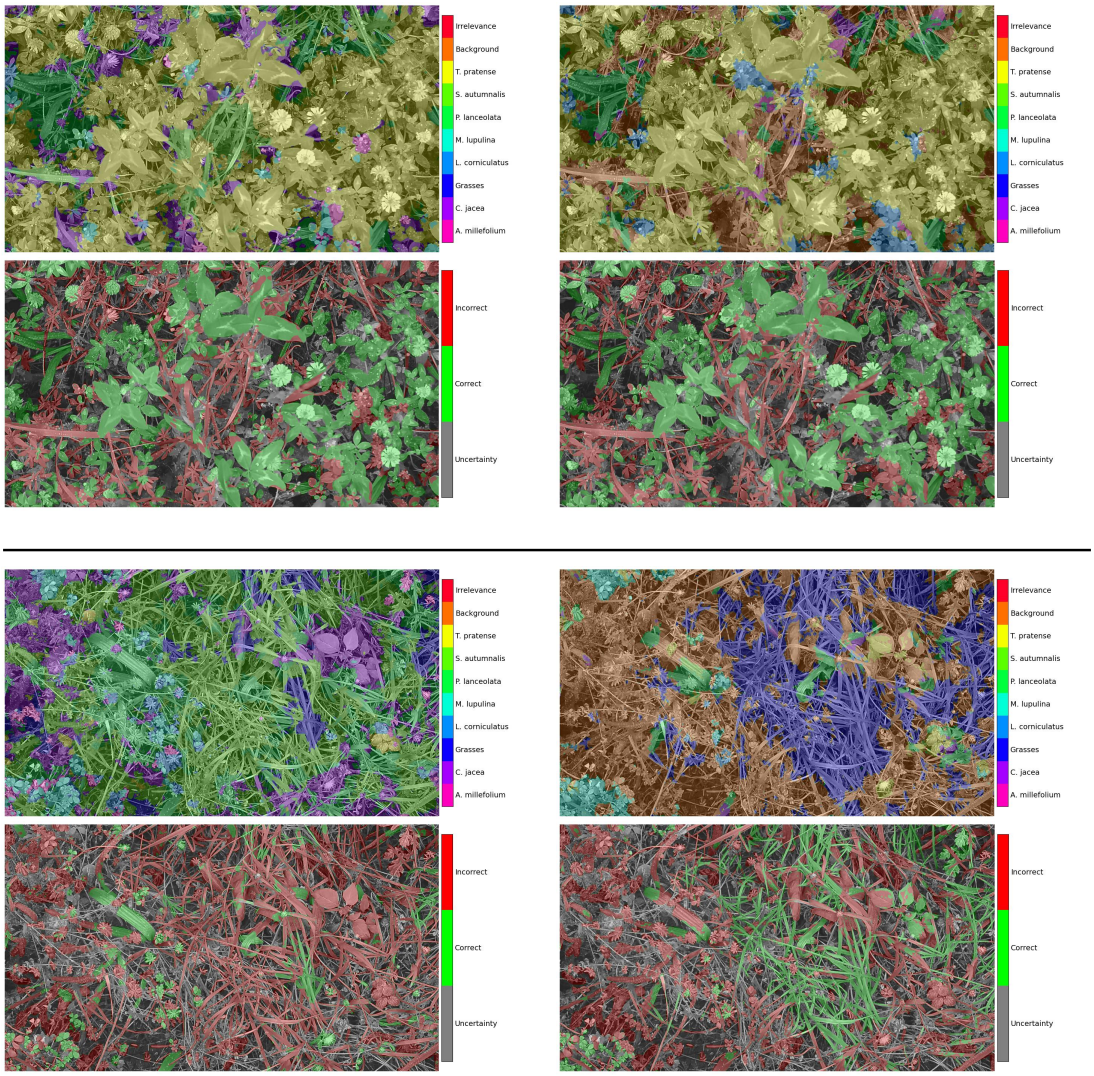

Figure 8: (continued) A qualitative comparison of the segmentation quality of the zero-shot models (left) and the cover-trained models (right). Below each segmentation map an error map indicates, in which locations the segmentation has been correct. Uncertainty areas mark locations in which experts were not able to determine a ground truth due to occlusions, empty areas, etc. These areas are exempt from the evaluation process. It is visible that, while the predictions of several less abundant plants are better using the zero-shot model in comparison to its counterpart, it cannot predict several common plants well and also no background (e.g., soil), leading to miscalculation of the cover percentages.

4. Discussion

General Aspects. The developed methods capture variations in species coverages well. More specifically, natural images used for pre-training seem generally preferable in comparison to images of preserved plant individuals, as well as ImageNet pre-training. This is shown in Table 1, in which, for almost every metric and setup, the method using the natural image pre-training outperforms the preserved-image counterpart. The reason for this discrepancy is likely that the domain difference between images of dried specimens and images of vegetation plots is too big for herbarium imagery to be useful in cover estimation. Moreover, the herbarium imagery often contains plants in the same sizes and perspectives, while images taken in nature are much more diverse, capturing a much larger set of variations useful in cover prediction. Images of herbarium specimens are often collected in a very high resolution; therefore, it might be possible to develop more specialized methods that can fruitfully utilize herbarium image data for pre-training CNNs. However, from our results, we conclude that, when trained on in our 'naive' way, the domain differences are too high to apply herbarium data for pre-training beneficially.

Furthermore, as shown in Table 1, while the top-performing ImageNet pre-trained model outperforms several of our cover-trained models, the best cover-trained model with natural GBIF images as pre-training still performs better than all ImageNet baseline models. We can also see that the best-performing cover-trained model outperforms the best-performing zero-shot model in all metrics. However, while the former reduces the MSAE to about 50% compared to the best zero-shot model, the values for IoU and DPC are only slightly worse for the zero-shot approach in comparison to its fully-trained counterpart. This indicates that the zero-shot model cannot precisely predict the exact reference cover estimates but can still predict the top layer of plants and the relative species distribution reasonably well.

Comparison of Predicted and Observed Values. The model outcomes comparing predicted vs. observed values suggest that, in general, the zero-shot prediction appears to be most reliable for *Trifolium pratense*, *Plantago lanceolata* and *Medicago lupulina*, which are likely the easiest to recognize in the vegetation images due to their morphology. Moreover, for several of the species, we see very high predicted cover values for nearly and also completely unobserved plant species in certain images, with a strong negative influence on the general linear trend. These strong mispredictions can be explained by the zero-shot model's inability to predict bare soil. In the case of bare soil, the model predicts arbitrary classes instead, leading to large errors in these cases. Hence, using the model in areas containing no or only little background areas such as bare soil likely leads to better cover predictions.

As the cover-trained model can predict background areas like bare soil, the relationship between the predicted and observed values is much better for all species, as shown in Figure 6 (right). The species with the best linear fits are *Medicago lupulina*, *Lotus corniculatus* and grasses, followed by *Trifolium pratense* and *Plantago lanceolata*, most of which belong to the most prevalent plants in the dataset. Due to this, the network also has seen many instances of these plants in the training data, which leads to a better prediction. In contrast, the three worst-predicted species are *Achillea millefolium*, *Scorzoneroides autumnalis* and *Centaurea*

jacea, the former two of which are the species with the smallest abundances in the dataset, likely leading to the opposite effect. For *Centaurea jacea*, the reason for the mispredictions is likely that the fine leaf texture cannot be captured by the network due to a too low resolution.

When observing the general trend of the predictions in Figure 6, in both plots one can see that the regression lines lie below the 1:1 line, indicating a systematic underestimation of plant cover estimates. This underestimation is likely caused by the heavy occlusion in the dataset and the network's current inability to deal with it.

When looking at the Shannon diversity in Figure 7, similarly, the cover-trained model produces better values compared to its zero-shot counterpart, while, for the latter, there is also a large number of values lying on the diagonal, indicating a good match. Moreover, it is visible for both models that, as there are many outliers below the diagonal, the models often overestimate the diversity. This likely happens because the models predict several additional species in images with only a few observed species, possibly indicating more species in the image than initially estimated in the reference estimates. It should also be noted that, as the reference estimates have been done on images as well, they also underlie the problem of occlusion, possibly leading to inaccuracies in the model's evaluation.

Segmentation Quality. The numerical results for the segmentation are similar for the zero-shot approach and the cover-trained approach, as also shown in Table 2. However, while the total segmentation quality averaged over all plant species is similar, both models have different recognition rates, depending on the species investigated. For the cover-trained model, the plants recognized best are usually the most prevalent species in the InsectArmageddon dataset. The zero-shot model, in contrast, demonstrates better recognition rates for some of the rarer species in the dataset, like *Lotus corniculatus* and *Scorzoneroides autumnalis*. This is also visible in the qualitative results shown in Figure 8. The reason for that is likely that the zero-shot model is provided with balanced training data in the GBIF dataset and, hence, does not focus on some plants more than others. In contrast, the cover-trained model is trained on the heavily imbalanced InsectArmageddon dataset and, thus, focuses more on the more abundant species therein. Overall, we can conclude that both model types can predict the plants in the top layer similarly well, while focusing on different species. However, the top layer does not necessarily represent the distribution of the plant community accurately, as the partially occluded plants can still have a big effect on the total cover. The occluded plant parts are not taken into account when analysing the top layer only. The high MSAE in conjunction with the comparably good IoU value shown in Table 1 suggest that the problems for the zero-shot model do not lie in the prediction of the top layer of plants but elsewhere, like the prediction of the lower layers or the prediction of the most abundant plants. As mentioned above, the zero-shot model, in its current state, is unable to predict background (e.g., bare soil) and instead predicts arbitrary plant species, also visible in Figure 8, leading to drastically wrong cover predictions in these instances. Another problem with the zero-shot model is likely the occlusion in the images. While the cover-trained model can implicitly gather knowledge about occlusion when training on the annotated cover data,

the zero-shot model has no such knowledge available, leading to worse results in images with heavy occlusions. As the DPC represents the relative relations of the cover values between each other, we can also conclude that the DPC for the zero-shot model is highest for images in which the distribution of the top layer of plants approximately matches the overall cover distribution.

Comparison with Human Estimators. In our comparison study we can see a similar performance of the human estimators with the zero-shot model and a more significant gap between these values and the results from the cover-trained model. The significant difference in DPC for the human estimators, as well as the comparably large standard deviation, indicates strongly varying cover values even in manual estimation, demonstrating the difficulty of plant cover estimation from images themselves. Moreover, the study shows that, despite its drawbacks, the zero-shot cover estimation approach can still potentially be used to replace human estimation, primarily due to its consistency and freedom of bias. Consistency is an especially important factor, as we have seen in our study by means of the standard deviation that human estimators can differ greatly during manual estimation. These inconsistencies can lead to potentially large estimation errors that can be mitigated when applying our method. When plant cover annotations are provided, the model more closely reflects the reference estimates used during training than the human and the zero-shot estimates, demonstrating that dedicated cover annotations can boost the estimation quality, if sophisticated cover estimates are provided during training. In contrast, the zero-shot model reasonably well reflects the species composition that would be predicted by other experts.

Summary & Usability Notes. While the cover predictions of the zero-shot model are not numerically accurate concerning the reference cover estimates, with the DPC values calculated over the whole dataset, we have shown that the relative distribution of the estimated cover values is similar to human performance. However, as this analysis includes images with background areas currently unpredictable by the network (e.g., soil and litter), the estimates likely improve in images without such areas. Moreover, as the model in its current state does not perform any occlusion handling but merely predicts the top layer of plants, the model's estimates will also be better in images with only one or few layers of plants, i.e., images with little occlusion. Lastly, as irrelevant areas should ideally be excluded for a good estimate, these areas should be provided by either a separate prediction model, as done here, or provided directly by a user. This additional effort, however, is removed if the images are already taken in a way that no such areas appear. Therefore, with several caveats, such a zero-shot model for cover estimation offers an easily adaptable method for vegetation surveys without requiring dedicated cover annotations. Therefore, this extension of the original method from Körschens et al. [31] offers a good alternative in situation where training data is not available. In comparison with existing approaches like Kattenborn et al. [24] and Du et al. [13], our zero-shot approach works similarly to theirs, as the top-layer of plants is segmented to predict the cover. However, our approach does not require any manually made segmentation annotations, which drastically reduces the amount of work necessary to utilize our cover prediction approach. Moreover, also in contrast to Kattenborn

et al. [24] and Du et al. [13], our approach can work with detailed herbaceous plant species instead of mostly homogeneous UAV image data.

When adding dedicated cover annotations, the results can be improved even further compared to our zero-shot approach, as shown in the previous experiments. With the training data, the network significantly improves on the most dominant plants in the dataset while possibly gathering some information about occlusion relationships.

While also similar to existing vegetation cover prediction approaches [44, 3, 26, 56, 10] on the first glance, our methods cannot only differentiate between vegetation and non-vegetation, or merely analyse the vegetation composition on a high level, but can do so on species-level, especially without any image delineations and only estimates as training data, or even without training data at all. It does not require any hand-crafted rules for differentiation, in contrast to [3, 26] and is able to differentiate with a much higher taxonomic granularity than [44, 56].

Our approaches can be applied in different scenarios. An example is an application to images taken by dedicated camera setups, as done in this work. Camera setups ensure that the images are collected the same way and, therefore, aid the approaches in analyzing the images in a consistent manner. However, our methods can also be applied to images taken by mobile devices directly in the field. While, in this scenario, a consistent image collection setup is not ensured, our methods can nevertheless drastically improve the collection of vegetation data.

Generally, our approaches can be utilized to analyze species communities, as we have shown that the overall species composition is well represented by the predictions of our model. In addition to this, our approaches might also be utilized to analyze the species-wise cover. However, in this case there often are larger deviations between predicted and observed cover estimates, as the different species can not always be predicted equally well by the model. The results can vary due to morphology, size and visibility of the respective plants and the prediction quality is there strongly dependent on the species.

5. Conclusion

We presented two approaches for automatic plant cover prediction: a cover-trained and a zero-shot approach. The latter produces predictions that are comparable with ones produced by human estimators, while the former produces cover estimates that are closer to the reference estimates. Hence, both approaches can be of value to ecologists. As per our study, the zero-shot approach can be used to retrieve data on plant communities, even when no dedicated cover annotations are available, while the cover-trained approach can be utilized to tune the outputs of the model to be more similar to a set of reference estimates.

Moreover, we compared the usage of plant images taken in their natural environment versus using dried herbarium specimens during pre-training and found that natural images perform far better in a large number of metrics in comparison to the herbarium specimens. We

conclude that the domain difference of dried specimens to images of vegetation plots is likely far too large for such imagery to be useful in cover estimation and therefore recommend using natural imagery in future endeavours.

We have seen that, while our approach produces comparable or better results to human estimation concerning the relative values (DPC), the concrete estimates, especially the zero-shot model, underestimate cover. As the reason for this is likely the huge amount of occlusion in the image, we will tackle the problem of occlusion as a next step. A possibility to do this is with means of data augmentation, e.g., Cutout [11], Inverted Cutout [32] or methods utilized in amodal segmentation [72, 37].

In addition to that, to improve the performance of the zero-shot model on plots containing bare soil and other background-like areas, the approach could be extended with another small model or an anomaly detection approach to recognize such areas.

Furthermore, as the zero-shot approach and the cover-trained approach both have advantages and disadvantages, they could also be combined into a single approach, to mitigate each of their disadvantages. In such a combined approach, for example, the predictions could be performed by each model individually and then weighted afterwards, depending on the strengths and weaknesses of each model.

Lastly, the approach for predicting plant cover can also be extended to predict the phenology of the plants in the images. Due to pre-training, the network already has knowledge of the blossoms of the plants in addition to the plant leaves, and the extension of the detection of senescing or flowering plants is therefore likely possible with only little adaptation of the method.

Acknowledgements

Matthias Körschens thanks the Carl Zeiss Foundation for the financial support. We acknowledge funding from the German Research Foundation (DFG) via the German Centre for Integrative Biodiversity research (iDiv) Halle-Jena-Leipzig (FZT 118) for the support of the FlexPool project PhenEye (09159751). We thank Alban Gebler for enabling the image collection process in the iDiv EcoTron. We also thank Gabriel Walther, Markus Bernhardt-Römermann, Robert Rauschkolb, Carolin Plos, Marco Patrzek and Solveig Franziska Bucher for participating in our comparison study. The Insect Armageddon dataset was collected in the framework of the iDiv Insect Armageddon project in the iDiv Ecotron financially supported through the German Research Foundation DFG-FOR 5000.

Competing Interests

The authors declare no conflict of interest.

Data Availability

The code for the developed system will be made available on github and the InsectArmageddon image dataset will be provided via the iDiv data repository.

References

[1] Altalak, M., Ammad uddin, M., Alajmi, A., and Rizg, A. (2022). Smart agriculture applications using deep learning technologies: A survey. *Applied Sciences*, 12(12):5919.

[2] Bambil, D., Pistori, H., Bao, F., Weber, V., Alves, F. M., Gonçalves, E. G., de Alencar Figueiredo, L. F., Abreu, U. G., Arruda, R., and Bortolotto, I. M. (2020). Plant species identification using color learning resources, shape, texture, through machine learning and artificial neural networks. *Environment Systems and Decisions*, 40(4):480–484.

[3] Bauer, T. and Strauss, P. (2014). A rule-based image analysis approach for calculating residues and vegetation cover under field conditions. *Catena*, 113:363–369.

[4] Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 65(1):2–16.

[5] Bodesheim, P., Blunk, J., Körschens, M., Brust, C.-A., Käding, C., and Denzler, J. (2022). Pre-trained models are not enough: active and lifelong learning is important for long-term visual monitoring of mammals in biodiversity research—individual identification and attribute prediction with image features from deep neural networks and decoupled decision models applied to elephants and great apes. *Mammalian Biology*, pages 1–23.

[6] Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., Botta-Dukát, Z., Chytrỳ, M., Field, R., Jansen, F., et al. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*, 2(12):1906–1917.

[7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

[8] Choe, J., Oh, S. J., Lee, S., Chun, S., Akata, Z., and Shim, H. (2020). Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3133–3142.

[9] Chollet, F. et al. (2015). Keras. https://keras.io.

[10] Coy, A., Rankine, D., Taylor, M., Nielsen, D. C., and Cohen, J. (2016). Increasing the accuracy and automation of fractional vegetation cover estimation from digital photographs. *Remote Sensing*, 8(7):474.

[11] DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[13] Du, B., Mao, D., Wang, Z., Qiu, Z., Yan, H., Feng, K., and Zhang, Z. (2021). Mapping wetland plant communities using unmanned aerial vehicle hyperspectral imagery by comparing object/pixel-based classifications combining multiple machine-learning algorithms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8249–8258.

[14] Gerstner, K., Dormann, C. F., Stein, A., Manceur, A. M., and Seppelt, R. (2014). Editor's choice: Review: Effects of land use on plant diversity–a global meta-analysis. *Journal of Applied Ecology*, 51(6):1690–1700.

[15] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[16] Grey, D. (1981). Multivariate analysis, by kv mardia, jt kent and jm bibby. pp 522.£ 14· 60. 1979. isbn 0 12 471252 5 (academic press). *The Mathematical Gazette*, 65(431):75–76.

[17] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969.

[18] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[19] Helm, J., Dutoit, T., Saatkamp, A., Bucher, S. F., Leiterer, M., and Römermann, C. (2019). Recovery of mediterranean steppe vegetation after cultivation: Legacy effects on plant composition, soil properties and functional traits. *Applied Vegetation Science*, 22(1):71–84.

[20] Hill, M. O. and Gauch, H. G. (1980). Detrended correspondence analysis: an improved ordination technique. In *Classification and ordination*, pages 47–58. Springer.

[21] Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

[22] Kamilaris, A. and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90.

[23] Katal, N., Rzanny, M., Mäder, P., and Wäldchen, J. (2022). Deep learning in plant phenological research: A systematic literature review. *Frontiers in Plant Science*, 13.

[24] Kattenborn, T., Eichel, J., Wiser, S., Burrows, L., Fassnacht, F. E., and Schmidtlein, S. (2020). Convolutional neural networks accurately predict cover fractions of plant species and communities in unmanned aerial vehicle imagery. *Remote Sensing in Ecology and Conservation*, 6(4):472–486.

[25] Kaur, S. and Kaur, P. (2019). Plant species identification based on plant leaf using computer vision and machine learning techniques. *Journal of Multimedia Information System*, 6(2):49–60.

[26] King, D. H., Wasley, J., Ashcroft, M. B., Ryan-Colton, E., Lucieer, A., Chisholm, L. A., and Robinson, S. A. (2020). Semi-automated analysis of digital photographs for monitoring east antarctic vegetation. *Frontiers in plant science*, 11:766.

[27] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[28] Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671.

[29] Körschens, M., Bodesheim, P., and Denzler, J. (2022). Beyond global average pooling: Alternative feature aggregations for weakly supervised localization. In *VISIGRAPP*.

[30] Körschens, M., Bodesheim, P., Römermann, C., Bucher, S. F., Migliavacca, M., Ulrich, J., and Denzler, J. (2021a). Automatic plant cover estimation with convolutional neural networks. In *Computer Science for Biodiversity Workshop (CS4Biodiversity), INFORMATIK 2021*, pages 499–516.

[31] Körschens, M., Bodesheim, P., Römermann, C., Bucher, S. F., Migliavacca, M., Ulrich, J., and Denzler, J. (2021b). Weakly supervised segmentation pretraining for plant cover prediction. In *DAGM German Conference on Pattern Recognition*, pages 589–603. Springer.

[32] Körschens, M., Bodesheim, P., and Denzler, J. (2022). Occlusion-robustness of convolutional neural networks via inverted cutout. In *International Conference on Pattern Recognition (ICPR)*.

[33] Körschens, M., Bodesheim, P., Römermann, C., Bucher, S. F., Ulrich, J., and Denzler, J. (2020). Towards confirmable automated plant cover determination. In *ECCV Workshop on Computer Vision Problems in Plant Phenotyping (CVPPP)*.

[34] Lasseck, M. (2017). Image-based plant species identification with deep convolutional neural networks. In *CLEF (Working Notes)*.

[35] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[36] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125.

[37] Ling, H., Acuna, D., Kreis, K., Kim, S. W., and Fidler, S. (2020). Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33.

[38] Liu, H., Mi, Z., Lin, L., Wang, Y., Zhang, Z., Zhang, F., Wang, H., Liu, L., Zhu, B., Cao, G., et al. (2018). Shifting plant species composition in response to climate change stabilizes grassland primary production. *Proceedings of the National Academy of Sciences*, 115(16):4051–4056.

[39] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international*

conference on computer vision, pages 10012–10022.

[40] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

[41] Lloret, F., Peñuelas, J., Prieto, P., Llorens, L., and Estiarte, M. (2009). Plant community changes induced by experimental climate change: seedling and adult species composition. *Perspectives in Plant Ecology, Evolution and Systematics*, 11(1):53–63.

[42] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

[43] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[44] McCool, C., Beattie, J., Milford, M., Bakker, J. D., Moore, J. L., and Firn, J. (2018). Automating analysis of vegetation with computer vision: Cover estimates and classification. *Ecology and evolution*, 8(12):6005–6015.

[45] Ojo, M. O. and Zahid, A. (2022). Deep learning in controlled environment agriculture: A review of recent advancements, challenges and prospects. *Sensors*, 22(20):7965.

[46] Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2019). *vegan: Community Ecology Package*. R package version 2.5-6.

[47] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[48] Pfadenhauer, J. (1997). *Vegetationsökologie - ein Skriptum*. IHW-Verlag, Eching, 2. verbesserte und erweiterte auflage edition.

[49] Quoc Bao, T., Tan Kiet, N. T., Quoc Dinh, T., and Hiep, H. X. (2020). Plant species identification from leaf patterns using histogram of oriented gradients feature space and convolution neural networks. *Journal of Information and Telecommunication*, 4(2):140–150.

[50] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[51] Ravoor, P. C. and Sudarshan, T. S. B. (2020). Deep learning methods for multi-species animal re-identification and tracking - a survey. *Comput. Sci. Rev.*, 38:100289.

[52] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer.

[53] Rosenzweig, C., Casassa, G., Karoly, D. J., et al. (2007). 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*, pages 79–131.

[54] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[55] Schmidt, A., Hines, J., Türke, M., Buscot, F., Schädler, M., Weigelt, A., Gebler, A., Klotz, S., Liu, T., Reth, S., Roy, J., Trogisch, S., Wirth, C., and Eisenhauer, N. (2021). The idiv ecotron - a flexible research platform for multitrophic biodiversity research.

[56] Sellers, H. L., Vargas Zesati, S. A., Elmendorf, S. C., Locher, A., Oberbauer, S. F., Tweedie, C. E., Witharana, C., and Hollister, R. D. (2023). Can plot-level photographs accurately estimate tundra vegetation cover in northern alaska? *Remote Sensing*, 15(8):1972.

[57] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning.

Journal of big data, 6(1):1–48.

[58] Smith, R. L., Smith, T. M., Hickman, G. C., and Hickman, S. M. (1998). *Elements of ecology*. Number 577 S6E5 1998. Benjamin Cummings Menlo Parie, CA.

[59] Sobha, P. M. and Thomas, P. A. (2019). Deep learning for plant species classification survey. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–6. IEEE.

[60] Souza, L., Zelikova, T. J., and Sanders, N. J. (2016). Bottom–up and top–down effects on plant communities: nutrients limit productivity, but insects determine diversity and composition. *Oikos*, 125(4):566–575.

[61] Taylor, S. D. and Browning, D. M. (2022). Classification of daily crop phenology in phenocams using deep learning and hidden markov models. *Remote Sensing*, 14(2):286.

[62] Triki, A., Bouaziz, B., Gaikwad, J., and Mahdi, W. (2021). A deep learning-based approach for segmenting and counting reproductive organs from digitized herbarium specimen images using refined mask scoring r-cnn. In *TACC*.

[63] Tugrul, B., Elfatimi, E., and Eryigit, R. (2022). Convolutional neural networks in detection of plant leaf diseases: A review. *Agriculture*, 12(8):1192.

[64] Ulrich, J., Bucher, S. F., Eisenhauer, N., Schmidt, A., Türke, M., Gebler, A., Barry, K., Lange, M., and Römermann, C. (2020). Invertebrate decline leads to shifts in plant species abundance and phenology. *Frontiers in plant science*, 11:1410.

[65] GBIF.org (2020). Gbif occurrence downloads. 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; Accessed 13 May 2020.

[66] GBIF.org (2022a). Gbif home page. https://www.gbif.org, Accessed 9 June 2022.

[67] GBIF.org (2022b). Gbif occurrence downloads. https://doi.org/10.15468/dl.pj3kmk, https://doi.org/10.15468/dl.bvwuja, https://doi.org/10.15468/dl.zce4rt, https://doi.org/10.15468/dl.z5cc9j, https://doi.org/10.15468/dl.8kzh4u, https://doi.org/10.15468/dl.x9rsx4, https://doi.org/10.15468/dl.a7g778, https://doi.org/10.15468/dl.mje3p5; Accessed 9 June 2022.

[68] iNaturalist.org (9 June 2022). inaturalist home page. https://www.inaturalist.org.

[69] Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

[70] Wang, X. A., Tang, J., and Whitty, M. (2021). Deepphenology: Estimation of apple flower phenology distributions based on deep learning. *Computers and Electronics in Agriculture*, 185:106123.

[71] Yalcin, H. (2018). Phenology recognition using deep learning: Deeppheno. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

[72] Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., and Loy, C. C. (2020). Self-supervised scene de-occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3784–3792.

[73] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.