PlantCAPNet: A Deep Learning System for Image-Based Plant Cover and Phenology Analysis

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

[a] *Plant Biodiversity Group, Institute of Biodiversity, Ecology and Evolution, 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] *Senckenberg Institute for Plant Form and Function (SIP), Philosophenweg 12, D-07743, Jena, Germany*

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

Abstract

Plant community data, like the species composition of the community and the phenology of the occurring species, are paramount for environmental research. Such data can be used to detect species responses to environmental changes, but the collection is very laborious, slow, and prone to human error. These detriments can be counteracted with automatic camera systems in combination with machine learning approaches that are able to extract the vegetation data from collected images in a consistent and fast manner. We introduce PlantCAPNet, an application to automate the analysis of herbaceous plant communities from images by extracting plant cover and phenology, addressing the tedious and biased nature of manual field collection. The system has an easy-to-use web interface with a single image prediction tool, a batch prediction function for image series, and a training interface for users to build novel models. We offer PlantCAPNet with two operational modes: a 'cover-trained' mode for predicting cover and phenology using user-provided labeled data, and a 'zero-shot' mode capable of predicting cover using only web-sourced data, thus lowering the barrier for entry. Our evaluations show that PlantCAPNet performs comparably or better than independent human experts in estimating plant cover. The zero-shot method reflects the reference estimates with a correlation of 0.625, and the cover-trained method with one of 0.790 compared to a correlation of 0.620 from independent experts. Moreover, we show that our system performs reliably for dataset with few species, and the cover prediction is also reliable for the most abundant species in datasets with many species, while the phenology prediction is dependent on the amount of training data. In total, our system offers higher consistency than human experts, and enables the extraction of high-temporal-resolution ecological data, facilitating novel environmental research.

Keywords: Plant Biodiversity, Plant Cover, Deep Learning, Convolutional Neural Networks, Artificial Intelligence, Application, Web-tool

1. Introduction

Collecting data on plant communities is a tedious task, but of paramount importance for many ecological studies. For example, the plant species composition is an important indicator for environmental changes like climate change (Rosenzweig et al., 2007; Liu et al., 2018; Lloret et al., 2009), but also responds to changes in insect abundance (Souza et al., 2016; Ulrich et al., 2020) and land use (Gerstner et al., 2014; Helm et al., 2019). Beyond the species composition, also their phenology responds to these changes in the environment (Root et al., 2003; Rosenzweig et al., 2007; Liu et al., 2018; Ulrich et al., 2020; Souza et al., 2016; Gerstner et al., 2014; Helm et al., 2019).

Community data is usually collected manually in the field by ecologists. However, this manual collection introduces several issues. The potentially strongest limitation is that experts require a lot of time to assess the community, and consequentially can only collect such data on few plots and in a low temporal resolution. Moreover, human estimation can be very inaccurate and biased, potentially changing the estimates due to influences like observer bias, leading to reproducibility issues. However, with automatic image collection systems combined with automated intelligent analysis methods, we can reduce the observer bias and provide detailed data across the season, therewith offering larger and more concise datasets.

While several freely available applications already exist for plant species determination, like Flora Incognita (Mäder et al., 2021) or Pl@ntNet (Affouard et al., 2017), they focus only on single plant individuals. In Flora Incognita, for example, the user can take several images of the same plant individual to improve species identification. However, as many ecological research projects focus on plant communities, such applications facilitate species identification, but not estimation of species covers.

While the automated estimation of plant cover from images has been investigated, existing scientific approaches are not suited for determining species-specific cover in the complex, dense herbaceous communities that are central to many ecological studies. For instance, some methods use CNNs on UAV imagery to assess vegetation (Kattenborn et al., 2020; Du et al., 2021), but these typically analyze visually distinct subjects like trees or shrubs where plants are relatively easy to discern and occlusion is not a major issue. Other ground-level approaches are often limited to simple color analysis, separating green from non-green parts of an image, or use classical computer vision and machine learning to differentiate between broad functional groups like grasses, forbs, or mosses (McCool et al., 2018; Bauer and Strauss, 2014; King et al., 2020; Sellers et al., 2023; Coy et al., 2016). Finally, while a recent work by Picard-Krashevski et al. (2025) also employs deep learning to predict the plant cover of herbaceous and shrub species by performing a patch-wise classification for each image, the setup is comparably simple with

*Corresponding author
Email addresses: matthias.koerschens@uni-jena.de (Matthias Körschens), solveig.franziska.bucher@uni-jena.de (Solveig Franziska Bucher), paul.bodesheim@uni-jena.de (Paul Bodesheim), joachim.denzler@uni-jena.de (Joachim Denzler), christine.roemermann@uni-jena.de (Christine Römermann)

few plant species and occlusion is not considered. Critically, many of these methods rely on pixel- or patch-level delineations for training, which are extremely labor-intensive to produce and not available in our case. Consequently, a significant gap persists: there is no existing, accessible solution that can automatically and accurately estimate the cover of individual species within dense, overlapping plant communities directly from images without requiring manual annotations.

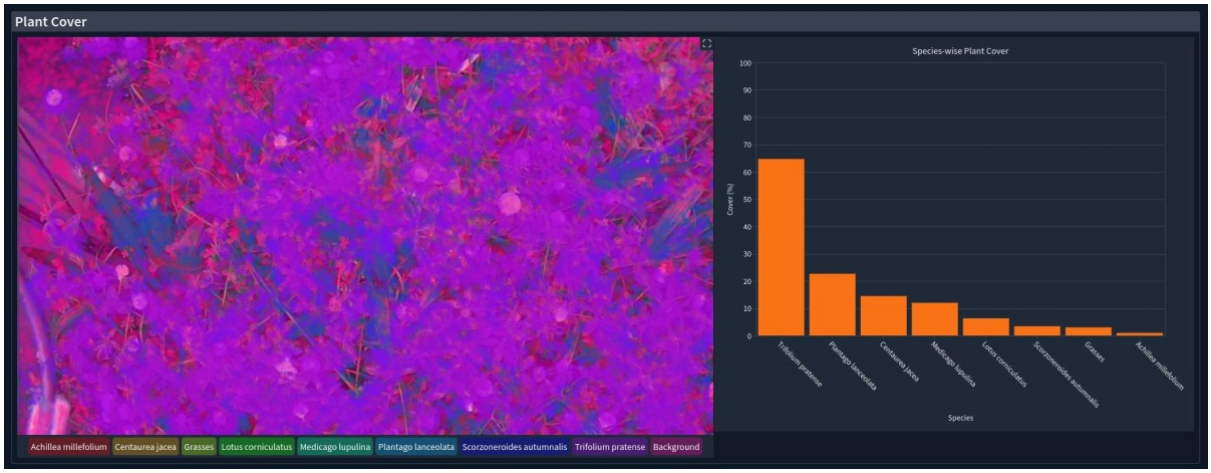
We introduce PlantCAPNet as an easy-to-use application enabling the automatic extraction of the species composition and phenology of herbaceous plant communities from images. The system comprises an interface of three parts. The first part is for single image predictions and can be used to check the correctness of the underlying model's predictions. The second interface is for batch-processing a series of images in a fast and convenient way, which enables inference for very long image sequences, and offers a download of the prediction results in form of CSV files. Finally, the third interface allows for configuring model training, providing default parameters and generating a downloadable script. When executed, the script trains a new model on custom data such that a user can easily create custom models on their own data with which they can obtain the most accurate predictions.

The methods used in PlantCAPNet are based on our previous works (Körschens et al., 2024, 2023b). Thus, PlantCAPNet features two modes: a cover-trained mode and a zero-shot mode. The former allows for training and inference of the species-wise plant cover and phenology, if a fully-labeled vegetation dataset is available. The zero-shot mode, in contrast, allows for training cover-prediction-only models solely based on web-sourced training data, and, thus, does not require a labeled vegetation dataset.

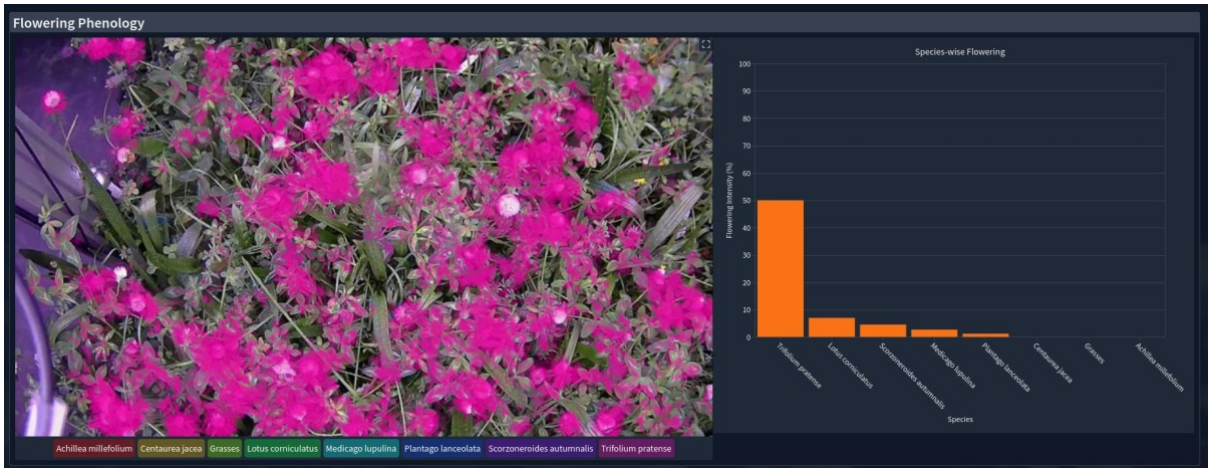
Additionally, PlantCAPNet features three major practical extensions to the methods introduced in previous works. The first extension is the usage of two recent network architectures, namely the EfficientNetV2 (Tan and Le, 2021) and ConvNeXt (Liu et al., 2022) families of architectures, thereby replacing the ResNet50 (He et al., 2016) used in previous works and providing a considerable performance boost. Second, we employ model ensembles, which enable the aggregation of information of several equally or differently trained models to make predictions more robust. Finally, we also introduce temporal aggregation of the predictions, with which is becomes possible to use temporally close images (like images from the previous and following week of the image to predict) to make the predictions even more robust and accurate.

With our novel application it becomes straightforward to train models on series of images of vegetation relevées and perform predictions using such models to extract consistent data in a high temporal resolution. To be able to use or train our system, no coding is required, resulting in simple adaptations to custom datasets.

In the following, we provide an introduction to the system, including a detailed overview over its components and present the typical workflow when working with it. Additionally, we also show that it performs comparable or better than independent experts, and present important aspects we found that are paramount for good prediction performance when applying the system to custom data.



(a) Cover predictions



(b) Flowering phenology predictions

Figure 1: An example of the single prediction interface showing heatmaps and bar plots for the detected plants, their flowers and senescent areas.

2. PlantCAPNet

In this section, we first introduce the web interface and its components, followed by an explanation of the backend and the general usage workflow, and conclude details and requirements for setting up the application.

2.1. Web Interface

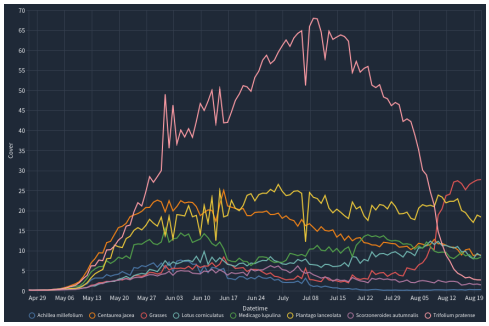
The web interface to our system comprises three interfaces: two for prediction, and one for setting up a training workflow to create new models.

2.1.1. Single Prediction

The single prediction interface allows running the system on individual images in order to obtain visualizations of the models' detections and for the user to confirm the correctness of the prediction results. For the cover and phenology prediction modes, the interface shows a heatmap for each species. These heatmaps show the model's detections, i.e., the pixel-wise likelihood for each species according to the model. In addition to each of these maps, the model's cover and phenology percentage predictions are summarized as bar

filenames	year	month	day	hour	minute	Achillea millefolium	Centaurea jacea
2018-04-28T1800140200.tif	2018	4	28	18	0	0.187592551112175	0.14013828337192
2018-04-29T1800140200.tif	2018	4	29	18	0	0.2077176719903946	0.16426610946655
2018-04-30T1800140200.tif	2018	4	30	18	0	0.2203075885772705	0.17341983318328
2018-05-01T1800140200.tif	2018	5	1	18	0	0.2130817323923111	0.16901150345802
2018-05-02T1800140200.tif	2018	5	2	18	0	0.2995823919773102	0.26698684692382
2018-05-03T1800140200.tif	2018	5	3	18	0	0.3403624892234802	0.39743533730506
2018-05-04T1800140200.tif	2018	5	4	18	0	0.3715320825576782	0.50859290361404
2018-05-05T1800140200.tif	2018	5	5	18	0	0.43959879875183105	0.73458838462829
2018-05-06T1800140200.tif	2018	5	6	18	0	0.47872886061668396	0.9983422160148
2018-05-07T1800140200.tif	2018	5	7	18	0	0.6012144684791565	1.36870694160461
2018-05-08T1800140200.tif	2018	5	8	18	0	0.6390894055366516	1.73099908199316
2018-05-09T1800140200.tif	2018	5	9	18	0	0.9451730251312256	2.28262972831726

(a) An exemplary output table for the batched prediction interface showing variations in cover for all species over time.



(b) An exemplary plot for the batched prediction interface. Each line represents the development of the plant cover of a single species over time.

Figure 2: Exemplary outputs of the batched cover prediction interface.

plot for a better overview of the total model predictions. An example of this interface is shown in Figure 1.

2.1.2. Batched Prediction

To speed up analysis, we introduce a batch prediction interface that processes an entire series of images at once. The images can be handled independently or as a time series by interpreting dates and times from the filenames. This temporal information is used in two ways. First, for aggregation, where images over a set time unit (like a day) are averaged to produce more robust predictions by reducing noise from lighting changes or movement. This can be done over minutes, days, or years. Second, temporal smoothing uses neighboring data points over weeks or months to further refine the final predictions.

After processing, results are presented in tables (Figure 2a) and line plots (Figure 2b), and can be downloaded as a CSV file for further ecological analysis.

2.1.3. The Training Interface

The provided training interface facilitates fine-grained configuration across the entire model training pipeline, encompassing pre-training, cover and phenology training, and ensemble generation.

Pre-training Configuration. Users can independently configure parameters for both classification and segmentation pre-training stages (see Körschens et al. (2024)). Common adjustable parameters include the learning rate schedule, dataset selection, choice of loss function, and the underlying network architecture.

Plant Cover Training Configuration. For the final plant cover training phase, the interface permits the selection of task-specific hyperparameters and target datasets. Additionally, parameters controlling the methodology for subsequent plant cover and phenology calculations can be precisely defined.

Ensemble Methods. Integrated support for ensemble methods allows users to configure, train and automatically construct model ensembles to improve generalization and robustness. Ensembles can be formed using various strategies, such as: (i) combining models based on heterogeneous architectures, (ii) aggregating model checkpoints trained with different training epochs, or (iii) averaging predictions from multiple independent training runs with identical settings (repetitions).

Execution Environment. To accommodate diverse computational needs, the interface provides options for both local execution and distributed training on a SLURM-managed cluster.

After configuring, the corresponding code can either be copied or downloaded into a script file that can be executed via the backend framework.

2.2. The Backend Framework

The backend framework, the code-base for system inference and training, can be operated independently of the interface. It builds upon Körschens et al. (2024, 2023b) but includes additional features. The backend supports pre-training phases for classification and segmentation, followed by a dedicated cover prediction training phase. Alternatively, the trained segmentation model can function directly as a zero-shot cover predictor if cover training data is absent. Thus, the training pipeline is usable with specific plant cover and phenology data, or even without any annotated data. The GBIF-Downloader detailed in section 3 can retrieve species images from GBIF[1] to train or pre-train a zero-shot or cover-trained model.

Beyond these core capabilities, the system supports the construction of model ensembles using techniques like model averaging (Goodfellow et al., 2016). This approach involves training multiple models independently (e.g., with different initializations, epochs or architectures) and averaging their predictions, potentially improving accuracy and reliability by leveraging diverse model errors and reducing variance (Goodfellow et al., 2016). Additionally, the system incorporates training speed optimizations for high-resolution images, such as Monte-Carlo Cropping (MCC) (Körschens et al., 2023a).

Segmentation pre-training incorporates the Inverted Cutout (IC) augmentation (Körschens et al., 2022) to enhance robustness against partial occlusion. As plants often overlap in community analysis, causing significant occlusion, methods addressing this can improve predictions.

Furthermore, the system currently supports two model architectures: ConvNeXt (Liu et al., 2022) and EfficientNetV2 (Tan and Le, 2021). They are available via PyTorch Torchvision (Paszke, 2019), using model weights from ImageNet (Russakovsky et al., 2015) pre-training.

As mentioned, the configuration of the training can be done via the web system, but the training needs to be done in the backend directly to have the highest amount of control and to not interfere with the running web system. As the backend can be operated independently, it can be deployed without the interface on dedicated computing resources like HPC clusters. Users can configure runs through the web interface, download the configuration as a bash script, and execute it on remote servers.

2.3. Workflow

The general workflow using our application is shown in Figure 3, with the usage of the web interface in Figure 3a and the training of a custom model in Figure 3b. In the training workflow, the user can first utilize our provided GBIF-Downloader (section 3) to obtain

[1] https://gbif.org

a pre-training dataset comprising the species of interest to supplement training. Then, they can utilize the training web interface to configure the training procedure, generating a complete training script therewith. In case of training a complete cover-trained model, annotated cover- and phenology training data has to be prepared before the training can start. In the case of zero-shot models, as no annotated data is used, no such preparation is required before the training, and only the pre-training data is used.

After the training process, the newly generated model has to be integrated into the web interface, after which any of the two prediction interfaces can be used. The single prediction interface generates visualizations on the localizations of each species in a single image, while the batch prediction interface can be used to obtain predictions for an entire series of images at once. After the batch prediction, the results can be downloaded for further processing and the prediction can be repeated for more images. Notably, the prediction process only takes a few seconds per image when using a GPU.

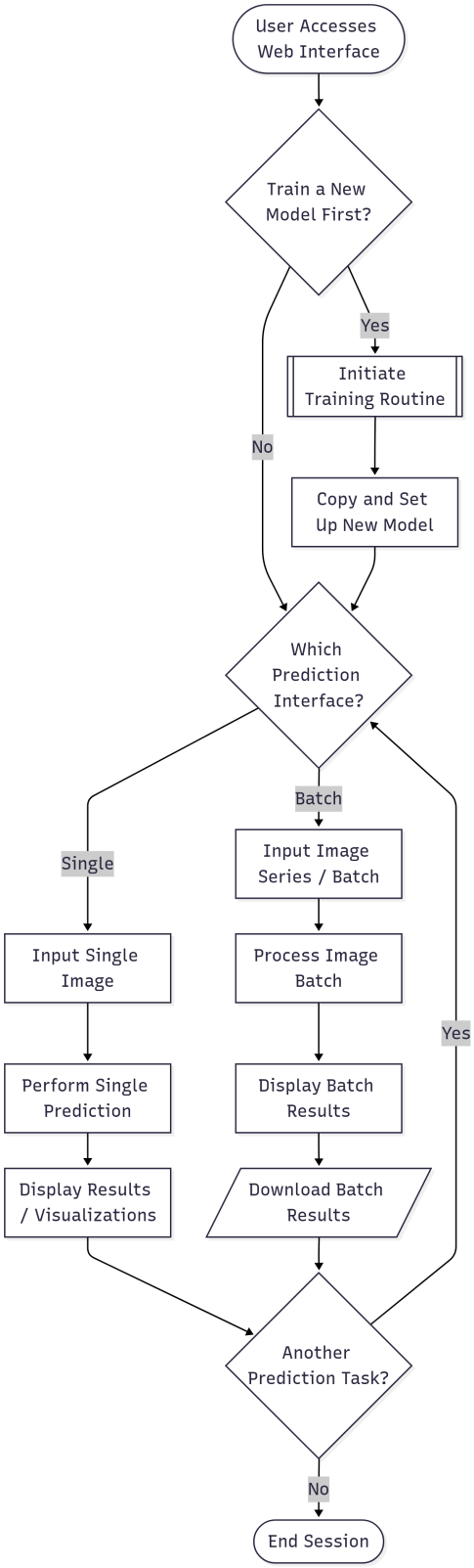
2.4. Setup

This section outlines the necessary prerequisites and recommended hardware for installing and running the application.

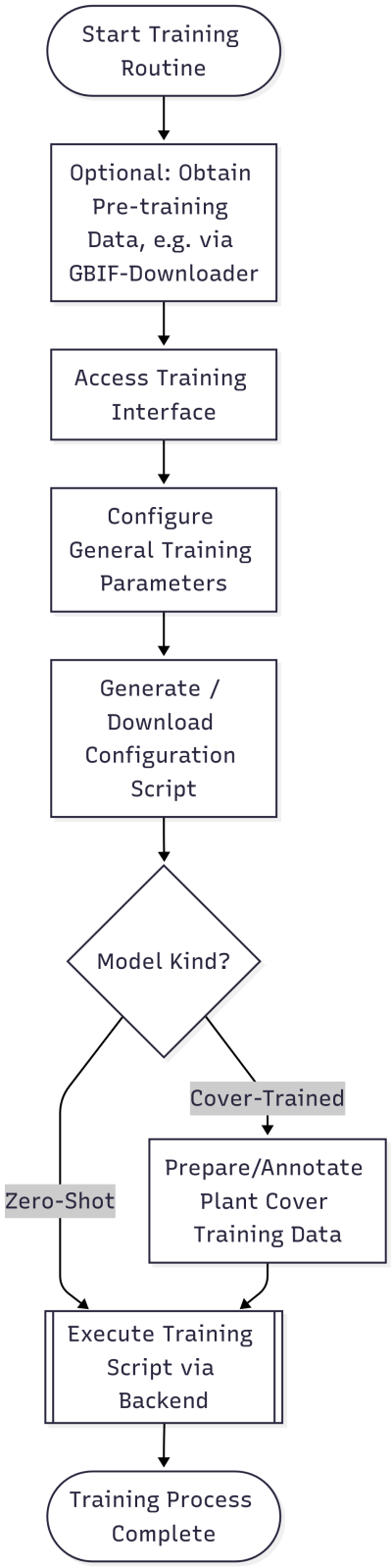
Software Prerequisites. A Linux distribution is recommended as the operating system; development and testing were performed primarily on Ubuntu 22.04 LTS and 24.04 LTS. The application necessitates Python version 3.10 or newer. We recommend managing dependencies using Anaconda or a similar virtual environment manager. The core functionality relies on PyTorch (Paszke, 2019) for deep learning model training and inference, complemented by Torchvision for standard computer vision models. Furthermore, the interactive web interface is built using the Gradio (Abid et al., 2019) library. Setup scripts are provided to facilitate the installation of these and other dependencies; please refer to the accompanying `README` file or documentation for detailed instructions.

Hardware Recommendations. Hardware requirements vary based on the intended use (inference or training). For inference tasks, a CUDA-compatible NVIDIA GPU possessing at least 8 GB of VRAM (e.g., NVIDIA GeForce RTX 2080 or equivalent) is recommended to achieve reasonable performance. Model training, being more computationally intensive, benefits significantly from a CUDA-compatible NVIDIA GPU with 12 GB of VRAM or more (e.g., NVIDIA GeForce RTX 3080, RTX 4070 Ti, or data center-grade GPUs like A100). Regardless of the primary task, a minimum of 64 GB of system memory (RAM) is recommended for the host machine, particularly when performing model training.

Client Access. It should be noted that the software prerequisites and hardware recommendations detailed above apply specifically to the server component responsible for hosting the backend logic and executing model computations. The web interface itself is designed to be lightweight and can be accessed and utilized from any standard web browser on any operating system, provided there is network connectivity allowing communication with the server.



(a) The main process of using the web interface.



(b) The process of using the web interface, GBIF-Downloader and backend for training.

Figure 3: The principal process for using the PlantCAPNet system in UML flowchart notation.

3. The GBIF-Downloader

We also provide a tool for image retrieval from the Global Biodiversity Information Facility (GBIF) platform (GBIF.org, 2025). Its primary function is to assemble image datasets for specified taxonomic groups, initially focused on plant species observed in situ or preserved as herbarium specimens. However, the tool's capabilities extend beyond botany, enabling dataset creation for various taxonomic ranks (e.g., genus, family) and kingdoms (e.g., Animalia), encompassing the diverse range of image data hosted within the GBIF database. This flexibility makes it a valuable resource for constructing comprehensive image datasets tailored to the requirements of training computer vision algorithms, particularly when pre-existing, curated datasets for specific taxa or imaging conditions are unavailable.

4. Evaluation

To provide an overview over the system's capabilities, we present results evaluated on the InsectArmageddon dataset (Ulrich et al., 2020; Körschens et al., 2020, 2024) and a novel dataset collected in the botanical garden of Jena, Germany called BotGardJena21.

4.1. Metrics

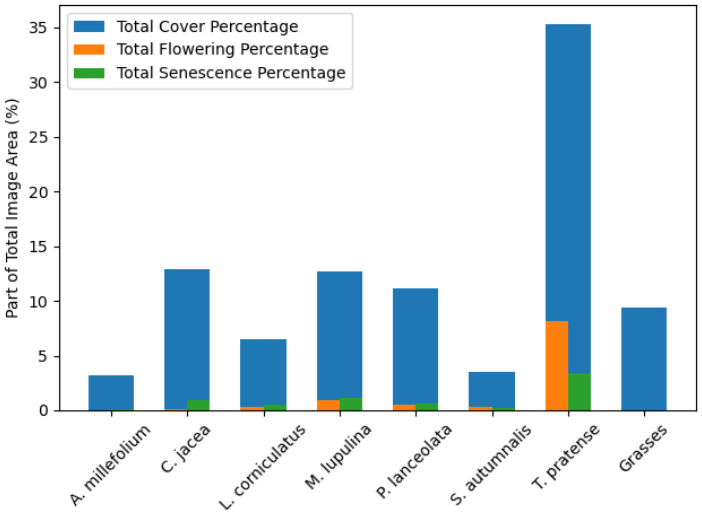
We present numerical results based on the Pearson-correlation of the predicted values and the original annotations, and the DCA-Procrustes-Correlation (DPC), as defined in Körschens et al. (2024). The Pearson-correlation is used to evaluate how well the species-wise cover and phenology predictions relate to the ones observed by an ecologist. It is calculated independently between each predicted and target value for each species, followed by averaging. The DPC indicates, how well the entire plant cover distribution over all species is reflected in the predictions compared to the original estimates. The metric is a combination of the Detrended Correspondence Analysis (DCA) (Hill and Gauch, 1980), and a Procrustes analysis (Grey, 1981). It is calculated by transforming both the predicted and target outputs with a DCA, followed by a comparison using a Procrustes test, resulting in a correlation-like value ranging from 0 to 1, with 1 being the best.

4.2. InsectArmageddon Dataset

In this section, we introduce the InsectArmageddon dataset, followed by the evaluation of our methods on it.

4.2.1. Dataset

The InsectArmageddon dataset comprises eight herbaceous plant species from central Europe in a strongly imbalanced species distribution, as shown in Figure 4a. In the figure, we also see the *flowering* and *senescence* percentages as fractions of the total plant cover. We see that the amounts of phenological training data are considerably smaller than the already small and imbalanced cover data, typical for plant community data. Example images for the *flowering* and *senescence* stages are shown in Figure 4b and Figure 4c, and example images for the species in the dataset can be found in Figure 5. While the original dataset comprises 682 weekly images with complete annotations, we employed Label Interpolation (Körschens et al., 2023a) to weakly label the unlabeled daily images. This increased the size of the dataset to approximately 4900 images. We



(a) Plant cover and phenology percentages in the Insect-Armageddon dataset calculated over the entire dataset

(b) Flowering plants

(c) Senescent plants

Figure 4: An overview over the InsectArmageddon dataset.

maintain the evaluation protocol from Körschens et al. (2024) and evaluate in a 12-fold cross-validation. For a comparison with human experts, we compare the results of our methods with the estimates from the study with six human experts from Körschens et al. (2024), in which they independently estimated the plant cover for a selection of images. The DPC is calculated for the estimates for each expert independently and then averaged for comparison with our provided method. For more details on the InsectArmageddon dataset, we would like to refer to the aforementioned publications.

4.2.2. Setup

For both the cover-trained model, and the zero-shot model, we utilize ensembles. During the classification phase (phase 1 in Körschens et al. (2024)), we use an EfficientNetV2 for both setups. For the cover-trained model, we create an ensemble by averaging predictions from a total of 9 models. Specifically, we average three models trained for 10 epochs, three models trained for 20 epochs, and three models trained for 40 epochs. Moreover, for our zero-shot model, we construct a different ensemble comprising 15 models. For this ensemble, we average the predictions from five models each of the ConvNeXt Tiny, ConvNeXt Base, and ConvNeXt Large architectures, which were trained in the same way. When we employ temporal smoothing, we use an exponential kernel with base of 0.8. I.e., we average the cover percentages by weighting the week to evaluate with a factor of $0.8^0 = 1$, the week before and after with a factor of $0.8^1 = 0.8$, the second weeks before

(a) *Achillea millefolium*

(b) *Centaurea jacea*

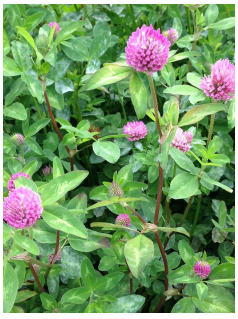
(c) *Grasses*

(d) *Lotus corniculatus*

(e) *Medicago lupulina*

(f) *Plantago lanceolata*

(g) *Scorzoneroides autumnalis*



(h) *Trifolium pratense*

Figure 5: Example images of the species contained in the InsectArmageddon dataset.

Table 1: Comparison of DCA-Procrustes-Correlation (DPC) values for different methods/individuals for all species.

Method/Individual	DPC Mean ± Std Dev
Human Experts	0.620 ± 0.140
Cover-Trained Model	0.790 ± 0.004
Zero-Shot Model	0.625 ± 0.023
Zero-Shot Model (Temporal)	0.677 ± 0.011

with a factor of $0.8^2 = 0.64$ et cetera.

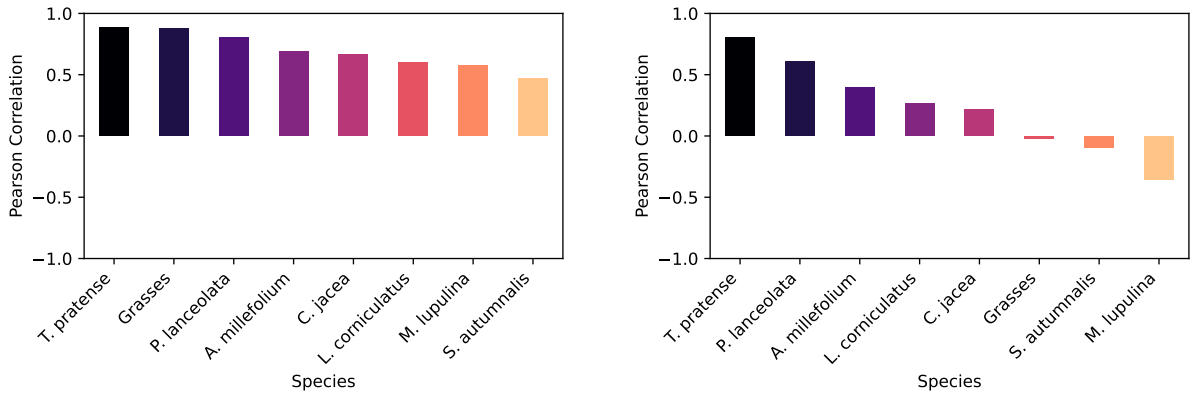
4.2.3. Results

In Table 1, we present the DPC results for our methods and the expert ecologists. The human experts achieved a mean DPC of 0.620±0.140, indicating relatively consistent performance but with considerable inter-expert variability initially. The cover-trained model demonstrates the highest performance overall with a DPC of 0.790±0.004, showcasing very low variance. Furthermore, the standard zero-shot model yields a DPC of 0.625±0.023, comparable to the human experts' mean. When employing the zero-shot model and smoothing the species-wise cover values over time using an exponential kernel, its prediction results in a DPC of 0.677±0.011. Notably, all models exhibit considerably lower standard deviations compared to the human experts, indicating higher consistency of the prediction models.

In Figure 6 we present the Pearson-correlations of the predictions with the original expert estimates. For the cover-trained model, the correlations are generally high for all the species, peaking at 0.9 for *Trifolium pratense* and *Grasses* (not distinguished into different species in this dataset), with *Scorzoneroides autumnalis*, the least abundant species in the dataset, having the lowest correlation with about 0.4. For the zero-shot model, the higher correlations are primarily focused on the *T. pratense* and *Plantago lanceolata*, two of the most dominant species in the dataset. *S. autumnalis* and *Medicago lupulina* show slightly negative correlations, in parts due to small size in the images (*S. autumnalis*), and in parts possibly due to considerable similarity to other species (*T. pratense* vs. *M. lupulina*). Notably, the species with the highest correlations are visually very distinct, leading to a reliable prediction, while small, thin or visually similar species can be problematic.

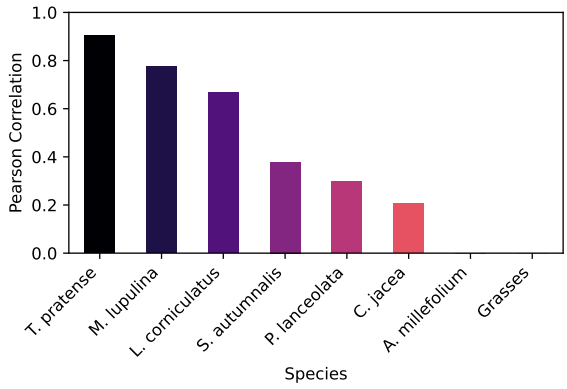
Regarding the cover-trained model's ability to predict *flowering* phenology (Figure 6c), we see that the more obvious flowers of *T. pratense*, *M. lupulina* and *Lotus corniculatus* are recognized reliably, while other, more inconspicuous blossoms, like the one of *P. lanceolata*, can be more problematic. Furthermore, the detection of blossoms of a certain plant species is also relying on the correct identification of the species of the respective plant, which can potentially result in detrimental results for *flowering* or *senescence*, if the species identification is not working well due to low resolution or similar issues. On average, the cover-trained model achieves a *flowering* correlation of 0.529±0.014.

Concerning the *senescence*-prediction ability shown in Figure 6d, we see that most correlations are high, besides the ones for species with little to no phenological training data. The average *senescence* correlation achieved by the model is 0.454±0.009.

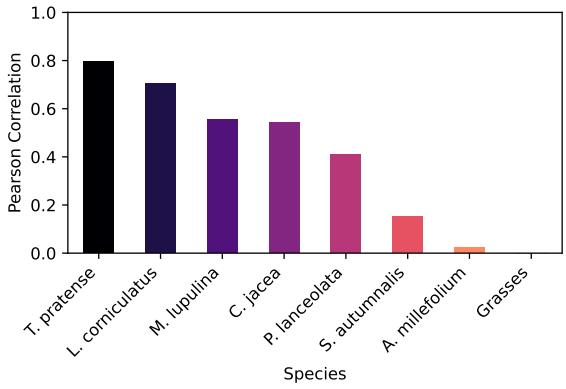


(a) Plant cover correlation for the cover-trained model

(b) Plant cover correlation for the zero-shot model



(c) Flowering correlation for the cover-trained model



(d) Senescence correlation for the cover-trained model

Figure 6: Correlations of model predictions with the reference estimates of the annotating expert.

Overall, for cover and phenology prediction with the cover-trained model, the correlations considerably depend on the amount of cover and phenology data available. *T. pratense*, the most dominant species in the dataset, shows the best correlations with the reference estimates, due to its prevalence in the dataset. Species with considerably less training data in phenology or cover perform drastically worse. Regarding the zero-shot model, clearly differentiable and abundant species show the best performance. This effect is shown in Figure 7. We see that for all cover and phenology prediction, the amount of data plays a considerable role in the prediction accuracy, and a clear trend is visible.

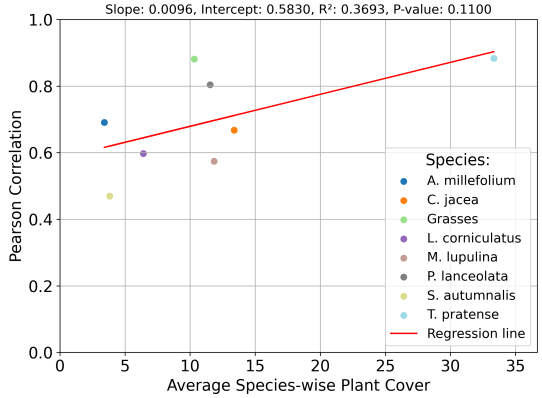
Again, it should be noted that the negative correlations for the zero-shot model can be caused by considerable similarities of the plant leaves, leading to the model mistaking a species for another. These similarities are, unfortunately, only hard to solve without any annotated cover data.

4.3. The BotGardJena21 (BGJ21) Dataset

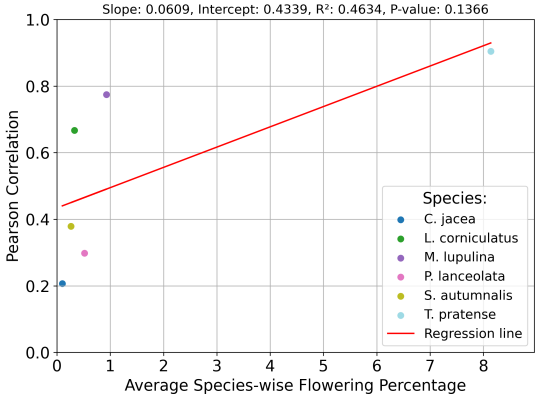
To provide an indication of the performance of our system in considerably more complex plant communities, we introduce a novel dataset, which we also use for evaluation.

4.3.1. Dataset

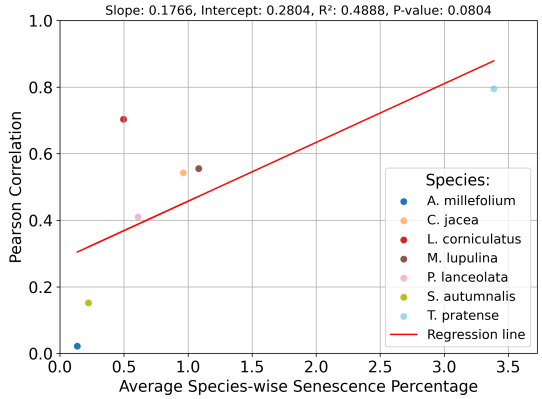
The BotGardJena21 (BGJ21) dataset was collected in 2021 at the botanical garden of Jena, Germany. It consists of hourly images (≈8 AM to 8 PM) of two distinct vegetation plots, captured by two overhead cameras in a height of about 2 m from March



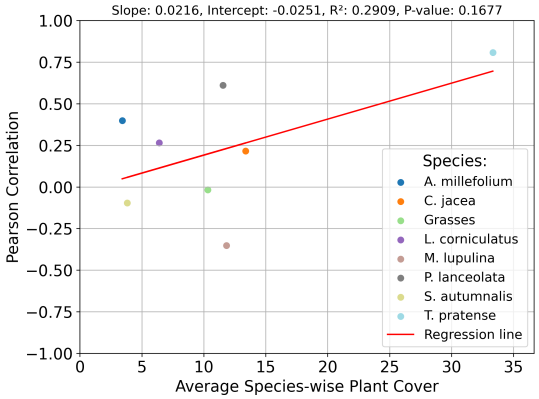
(a) Accuracy of cover predictions of the cover-trained model vs. species-wise cover



(b) Accuracy of flowering predictions of the cover-trained model vs. amount of species-wise flowering data



(c) Accuracy of senescence predictions of the cover-trained model vs. amount of species-wise senescence data



(d) Accuracy of cover predictions of the zero-shot model vs. species-wise cover

Figure 7: The relationship between the accuracy of cover and phenology predictions of the cover-trained and zero-shot model and amount of species-wise cover and phenology data in the dataset. The cover and phenology percentages are calculated as averages over the entire dataset.

(a) Plot 1: June 15, 2021

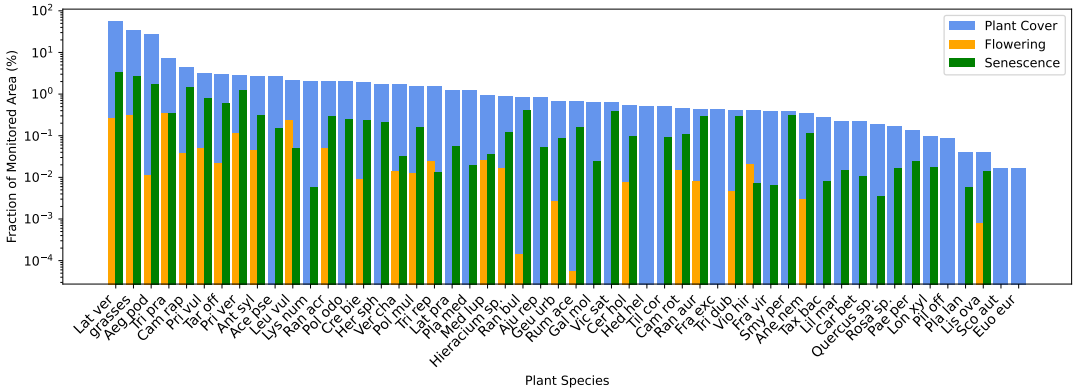
(b) Plot 2: September 15, 2021

Figure 8: Example images from the BGJ21 dataset showing the two monitored plots at different times of the year.

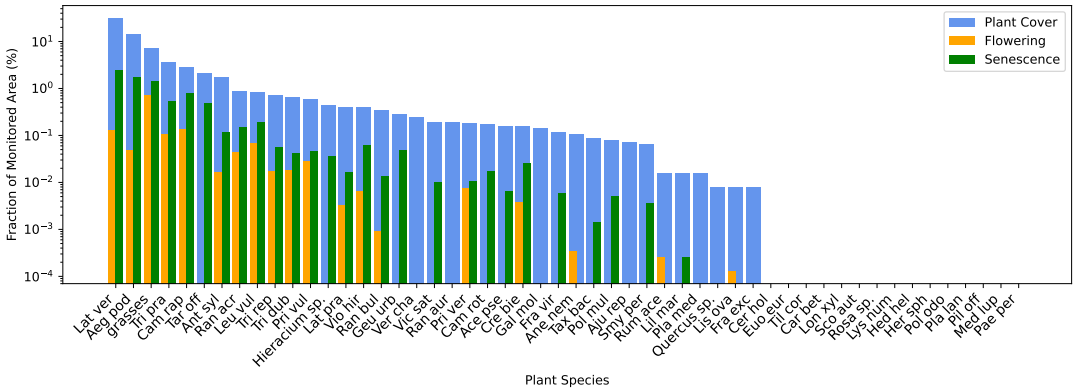
to November. Two example images are shown in Figure 8. The dataset documents 52 classes, comprising 51 herbaceous central European plant species and a single combined "Grasses" class, and is accompanied by weekly vegetation surveys that provide reference data for plant cover and phenology. The species distribution is highly imbalanced, dominated by a few frequent species like *Lathyrus vernus* and Grasses, with a long tail of many rare species, as shown in the following section. The two monitored plots also exhibit notable differences in their species composition, which presents a challenge for model training and validation.

Annotations. We obtained and compare two annotation methods of the images: field-based and image-based observations. Field-based surveys were conducted on-site by an ecologist who could interact with the vegetation, providing highly accurate but potentially biased data that includes occluded plants not visible to the camera. In contrast, image-based annotations were performed by an expert viewing only the camera footage, aligning the data source with what a machine learning model would see but at the cost of missing hidden information. This discrepancy led to significant differences. For example, field-based surveys identified 52 species, while image-based surveys found only 37. Furthermore, image-based annotations tended to overestimate the cover of dominant, visible species and underestimate species like grasses or those hidden in lower vegetation layers. The species cover and phenology distributions for the two annotation types are shown in Figure 9. The most abundant species in this dataset are *Aegopodium podagraria* (Aeg pod), *Campanula rapunculoides* (Cam rap), *Lathyrus vernus* (Lat ver), *Trifolium pratense* (Tri pra) and the collective class of *Grasses*. Notably, the amount of phenological information, i.e., *flowering* and *senescence* information is extremely low, in addition to the considerably skewed cover distributions. Few species surpass a *flowering* rate of 1% of the total plant cover, while *senescence* is generally more frequent in the dataset.

While this dataset also contains only weekly annotations, as before, we employed Label Interpolation (Körschens et al., 2023a) to get daily labels, which we use for every hourly image for each day. Therefore, this dataset has approximately $12 \times 30 \times 9 = 3240$ training images for each of the two sites.



(a) Field-based annotation: Number of observations per species and phenological stage (log scale).



(b) Image-based annotation: Number of observations per species and phenological stage (log scale).

Figure 9: Comparison of phenology annotation distributions for field-based and image-based surveys in the BGJ21 dataset.

Dataset Limitations. The dataset has several limitations that pose challenges for analysis. These include a relatively low resolution per centimeter, a significant five-week gap in data collection during the peak flowering season, and heavy occlusion of vegetation by fallen leaves in autumn, which complicates species identification and cover estimation. The lack of data during the key flowering period is particularly detrimental to training models for phenology prediction.

4.3.2. Setup

We train our network using images of 2560 x 1920 pixels. To manage this resolution, we employ Monte-Carlo-Cropping (MCC) (Körschens et al., 2023a) to create eight smaller patches of 448 x 448 pixels from each image for processing. To achieve robust and stable final predictions, we construct an ensemble of 12 sub-models by training three models for 6, 10, 15, and 20 epochs each.

Since the dataset provides hourly images, we leverage this rich temporal data. We aggregate the predictions of the hourly images over an entire day by averaging them. We also optionally apply temporal smoothing (TS) to these daily aggregated predictions to further improve temporal consistency. For this, we employ the same exponential kernel as described in section 4.2.2.

Our evaluation is performed using a cross-validation strategy. We train the model on the

data from one plot and validate it on the other, then average the results to account for the differing species compositions between the two sites. For phenology evaluation, we evaluate the mean correlation across all species, setting the correlation to 0 for species with no phenology data to ensure a fair average.

4.3.3. Results

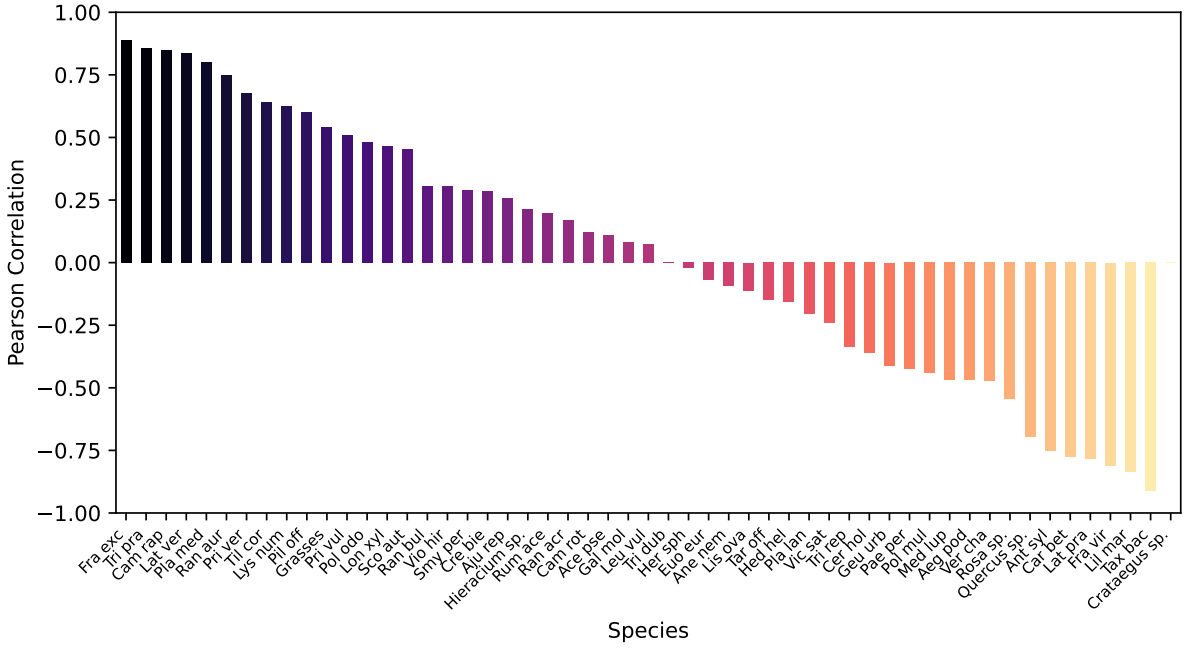
The results of our experiments with this dataset are shown in Table 2. We notice that the cover-trained models have a similarly high DPC with 0.90 for the image-based annotations, and 0.94 for the field-based ones. Also the zero-shot methods perform similar on both kinds of annotations, with values of about 0.8. The prediction of *flowering* phenology was very unreliable for both annotation kinds, as the amount of phenological training data in the dataset was very small. Regarding the *senescence* prediction, the model predictions are considerably more accurate for the field-based model, but not for the image-based one. The reason for this is likely that the senescent plants are usually in the lower layers of the vegetation plot and therefore usually not visible in the images. The field-based model learned the prediction of the phenology data from other features in the image, introducing a substantial bias in the *senescence* prediction. The image-based model, however, does not have access to this information and therefore only has little phenological data to train or predict, resulting in low accuracy.

Finally, the application of temporal smoothing appears to be beneficial in most instances, raising the performance by up to several percent points. Only in some cases, like the application in the zero-shot scenario of the image-based annotations it appears to be detrimental. A possible reason for this is that the weekly predictions are very noisy, requiring a differently-sized aggregation window to be effective.

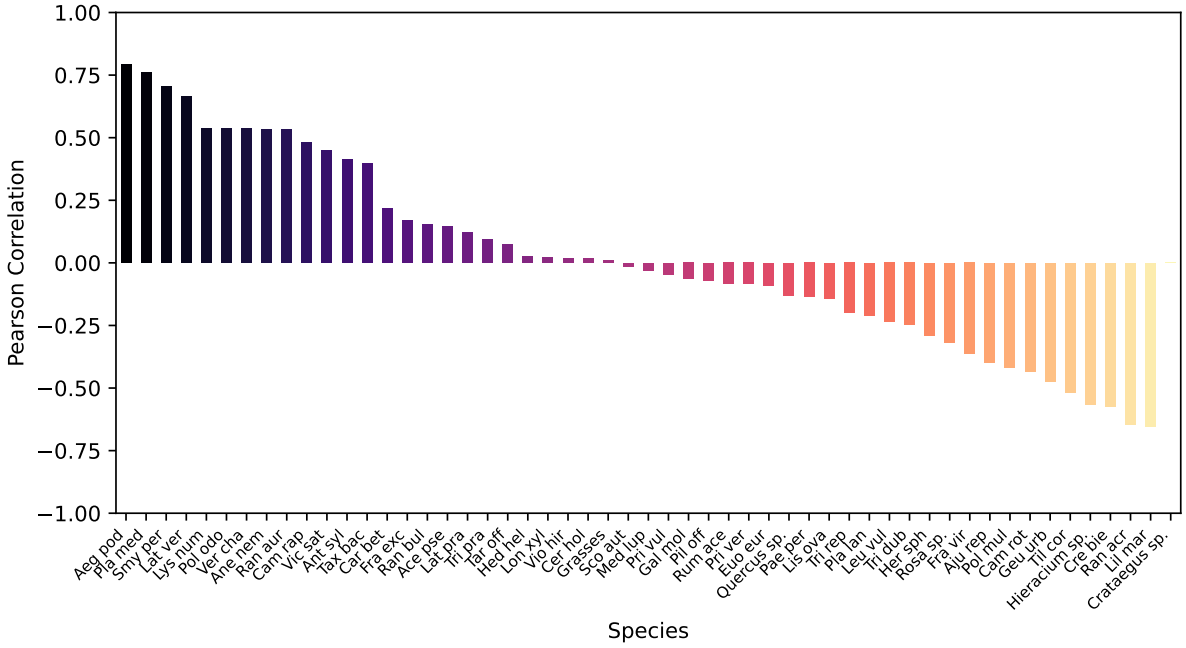
An indication of the species-wise correlations for the cover prediction is shown in Figure 10. Again, in most methods we notice that the most dominant species in the datasets are among the ones with the highest correlations. Among the most dominant species, only the *Grasses* appear to be a considerable challenge to estimate from images due to their thinness.

Table 2: The results for plant cover and phenology prediction trained on the BGJ21 field- and image-based estimates, and the zero-shot predictions evaluated on the each of these estimates. TS is abbreviated for temporal smoothing.

Base of Annotation	Approach	TS	DPC	Correlation *flowering*	Correlation *senescence*
Field	Cover-Trained	✗	0.936 ±0.009	0.036 ±0.012	0.537 ±0.019
		✓	0.962 ±0.004	0.045 ±0.013	0.550 ±0.016
	Zero-Shot	✗	0.793 ±0.036	—	—
		✓	0.859 ±0.007	—	—
Image	Cover-Trained	✗	0.901 ±0.002	-0.004 ±0.016	0.042 ±0.028
		✓	0.894 ±0.006	-0.001 ±0.018	0.044 ±0.024
	Zero-Shot	✗	0.805 ±0.030	—	—
		✓	0.653 ±0.010	—	—

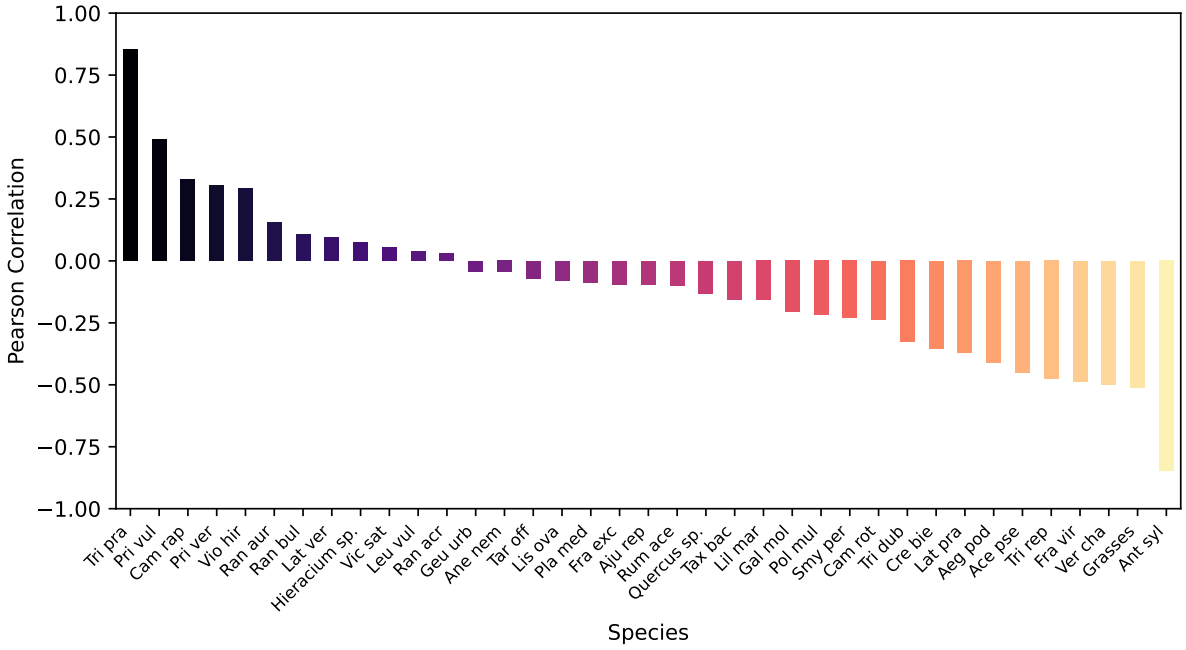


(a) Field-based cover-trained

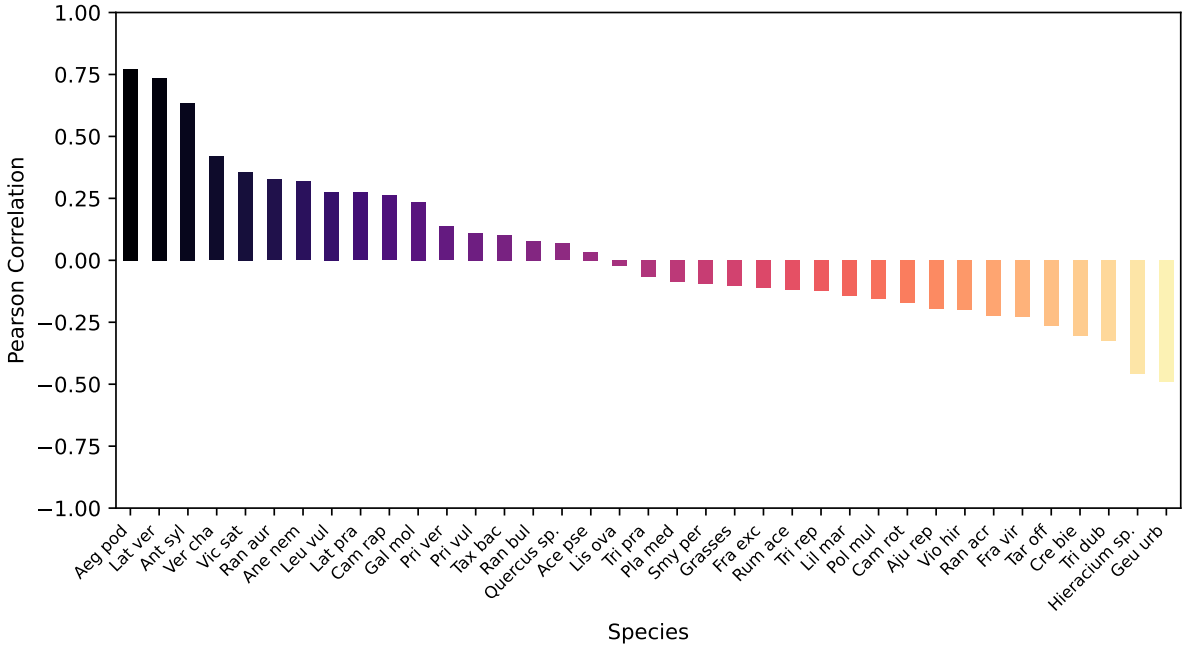


(b) Field-based zero-shot

Figure 10: Correlation plots for plant cover predictions on the BGJ21 dataset. Each panel shows the results for a different annotation source and prediction approach.



(c) Image-based cover-trained



(d) Image-based zero-shot

Figure 10: Correlation plots for plant cover predictions on the BGJ21 dataset. Each panel shows the results for a different annotation source and prediction approach.

5. Discussion

As shown in our experiment, users can achieve the most accurate cover and phenology predictions by providing their own annotated data, especially when sufficient training data for the species and their phenological stages is available. Additionally, PlantCAPNet's zero-shot capability allows for data extraction even without any user-provided vegetation training annotations. This significantly lowers the barrier to entry and enables its application across a wider range of ecological scenarios and novel research studies.

Unlike existing applications such as Flora Incognita (Mäder et al., 2021) or Pl@ntNet (Affouard et al., 2017), which typically analyze single plant specimens often using multiple detailed images, PlantCAPNet processes entire plant communities within single images. This approach drastically accelerates the analysis of community-level vegetation dynamics. While top-down imagery inherently involves potential occlusion of plant features, a challenge mitigated by the detailed, multi-image approach of single-specimen apps, analyzing each individual within a community in such detail is often temporally prohibitive. Therefore, PlantCAPNet is optimized for rapid community assessment, whereas tools like Flora Incognita are better suited for in-depth analysis of individual plants.

With our easy-to-use system many novel use-cases become possible for plant ecologists. Since the system can automatically analyze large numbers of images in a short amount of time, it is not necessary to collect sizable vegetation data manually anymore, lifting a considerable limitation of data collection. Therefore, future studies can investigate many plots at the same time by merely setting up a stationary camera collecting images in regular intervals. With such an application, the scale of ecological experiments can be extended considerably. Similarly, cameras can be stationed in remote locations or even in severely underrepresented regions in ecological research, significantly reducing travel times for researchers.

6. Usability Notes & Limitations

PlantCAPNet offers significant deployment flexibility; it can be hosted on a central computation server and accessed remotely from any network-connected device. It features two distinct prediction interfaces: one for visually verifying results and the spatial localization of detected species, and another for efficient batch processing of large image series, including capabilities for temporal smoothing. This design ensures both transparency and practical usability.

To clarify the optimal use cases and inherent limitations of our system beyond our evaluation results, we will discuss the key factors influencing its performance.

Our investigations revealed that system performance is critically dependent on data characteristics. Its effectiveness is constrained by a trade-off between image resolution and species count. While optimal results are achieved with high-resolution images (e.g.,$\approx 3000 \times 1500$ px) and a limited number of species (5-20), the requirements become more stringent as complexity increases. For a small number of visually distinct plant species, a lower resolution might be sufficient. However, as more species are added, the likelihood of visual similarity between them grows, making subtle distinguishing features critical for identification. A significant practical limitation is that these subtle features can often only be captured in very high-resolution images, raising the barrier for data

acquisition. Furthermore, the system is not well-suited for identifying species with low abundance. These species are difficult to identify reliably, and attempting to include them can negatively impact prediction accuracy for more dominant species. Consequently, the approach is best applied by focusing on a subset of the 10 to 20 most prominent species in an area.

Another significant constraint is the need for comprehensive data. The system's accuracy is severely impaired by data imbalances or a lack of representation across all relevant species and their different phenological stages. The annotation strategy also introduces potential limitations. Field-based annotations, where experts work directly in the field, might introduce biases by including information not visible in the images, which can improve accuracy for similar sites but harm generalizability. Conversely, image-based annotations foster evidence-based models that predict only what is visible. Our zero-shot approach is designed to mitigate this concern as it is inherently evidence-based. Finally, the reliability of predictions from a single image is a notable limitation. We found that predictions become considerably more robust and accurate when results from several images, such as those taken from different angles or at different times, are aggregated by averaging. This suggests that for critical applications, relying on a single viewpoint may be insufficient to ensure robust and accurate results.

7. Conclusion

We introduce PlantCAPNet, a novel application designed for ecological research to automatically extract species-specific plant cover and phenology from images of herbaceous plant communities. This tool simplifies the application of CNN architectures to user-collected image data and facilitates the training of custom models.

As shown in our evaluation, it enables the extraction of plant cover data and phenological data, especially on dominant species in the community. Moreover, we showed that our cover prediction methods perform well in real-world scenarios, but the phenology prediction is very dependent on the amount of available training data. Hence, depending on the amount of training data available, our approach can generate high-quality predictions for images of plant communities. We also showed that the zero-shot approach not requiring any training data can be a helpful asset, predicting most dominant plant species in the images well. Hence, it can be utilized in situations, where no training data is available. Finally, on our second dataset, we showed that both annotations done in the field, but also ones directly from images can be used. The former foster more biases for data that can not be seen in the images, potentially leading to more accurate predictions. The latter lead to more evidence-based models, which, however, leads to potentially worse predictions.

In summary, PlantCAPNet provides an automated system for extracting high-quality ecological data at high temporal resolutions. By significantly reducing the required effort, it facilitates novel, more fine-grained ecological studies than were previously feasible.

8. Future Work

Future work could focus on several key areas. Model architecture can be improved with more sophisticated loss functions, such as approximations of the non-differentiable DPC

metric similar to Wasserstein-GANs (Arjovsky et al., 2017), and by adapting amodal segmentation techniques (Zhan et al., 2020; Ling et al., 2020; Liu et al., 2024) to better handle occlusion.

Data collection can be enhanced by using dynamic platforms like UAVs (Sun et al., 2021) and by incorporating advanced sensors like light-field (Lippmann, 1908; Bergen and Adelson, 1991; Taugourdeau et al., 2022) or hyperspectral cameras (Du et al., 2021; Rogers et al., 2024; Li et al., 2024).

The pre-training and zero-shot pipeline could be strengthened by using superior CAM methods (Selvaraju et al., 2017; Wang et al., 2020; Ramaswamy et al., 2020) and by extending zero-shot prediction to phenology by combining models like SAM (Kirillov et al., 2023) with Grounding DINO (Liu et al., 2025; Ren et al., 2024).

The methods could also be applied to new data types, such as digitized herbarium specimens (Hussein et al., 2022) and remote sensing imagery (Isaienkov et al., 2020; Hızal et al., 2024; Hnatushenko and Honcharov, 2024; Kattenborn et al., 2020; Du et al., 2021). Finally, the analysis could be deepened by extending temporal analysis to the pixel-level (Moskolaï et al., 2021) and employing fine-grained classification techniques (Wei et al., 2021), like part-based models (Korsch et al., 2021), for more accurate species identification.

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).

Conflict of Interest

The authors have no conflict of interest to declare.

Data Availability

PlantCAPNet and its sub-applications are freely available on github under `https://github.com/Atlas8008/PlantCAPNet`.

Author Contributions

Matthias Körschens, Christine Römermann, Joachim Denzler, Paul Bodesheim and Solveig Franziska Bucher acquired funding, conceived the ideas, designed methodology. Joachim Denzler and Christine Römermann administrated the project and led the supervision. Matthias Körschens conceptualized and developed the system, performed the evaluation and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used the Google Gemini 2.5 Pro model in order to improve readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., and Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. arXiv preprint arXiv:1906.02569.

Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., and Joly, A. (2017). Pl@ntnet app in the era of deep learning. In ICLR: International Conference on Learning Representations.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In International conference on machine learning, pages 214–223. PMLR.

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.

Bergen, J. R. and Adelson, E. H. (1991). The plenoptic function and the elements of early vision. Computational models of visual processing, 1(8):3.

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.

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.

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.

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

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.

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.

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.

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

Hızal, C., Gülsu, G., Akgün, H., Kulavuz, B., Bakırman, T., Aydın, A., and Bayram, B. (2024). Forest semantic segmentation based on deep learning using sentinel-2 images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48:229–236.

Hnatushenko, V. and Honcharov, O. (2024). Land cover mapping with sentinel-2 imagery using deep learning semantic segmentation models.

Hussein, B. R., Malik, O. A., Ong, W.-H., and Slik, J. W. F. (2022). Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. Ecological Informatics, 69:101641.

Isaienkov, K., Yushchuk, M., Khramtsov, V., and Seliverstov, O. (2020). Deep learning for regular change detection in ukrainian forest ecosystem with sentinel-2. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:364–376.

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.

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.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026.

Korsch, D., Bodesheim, P., and Denzler, J. (2021). End-to-end learning of fisher vector encodings for part features in fine-grained recognition. In DAGM German Conference on Pattern Recognition, pages 142–158. Springer.

Körschens, M., Bucher, S. F., Bodesheim, P., Ulrich, J., Denzler, J., and Römermann, C. (2024). Determining the community composition of herbaceous species from images using convolutional neural networks. Ecological Informatics, page 102516.

Körschens, M., Bucher, S. F., Römermann, C., and Denzler, J. (2023a). Improving data efficiency for plant cover prediction with label interpolation and monte-carlo cropping. In DAGM German Conference on Pattern Recognition, pages 321–334. Springer.

Körschens, M., Bucher, S. F., Römermann, C., and Denzler, J. (2023b). Unified automatic plant cover and phenology prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 685–693.

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

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).

Körschens, M., Ulrich, J., and Gebler, A. (2024). Insectarmageddon image dataset (version 1.0) [dataset]. iDiv Data Repository. https://doi.org/10.25829/idiv.3542-wqwf56.

Li, H., Tang, X., Cui, L., Zhai, X., Wang, J., Zhao, X., Li, J., Lei, Y., Wang, J., Wang, R., et al. (2024). Estimating aboveground biomass of wetland plant communities from hyperspectral data based on fractional-order derivatives and machine learning. Remote Sensing, 16(16):3011.

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

Lippmann, G. (1908). Epreuves reversibles donnant la sensation du relief. J. Phys. Theor. Appl., 7(1):821–825.

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.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2025). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pages 38–55. Springer.

Liu, Z., Liu, Q., Chang, C., Zhang, J., Pakhomov, D., Zheng, H., Lin, Z., Cohen-Or, D., and Fu, C.-W. (2024). Object-level scene deocclusion. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11.

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.

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.

Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H. C., Deggelmann, A., and Wäldchen, J. (2021). The flora incognita app–interactive plant species identification. Methods in Ecology and Evolution, 12(7):1335–1342.

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.

Moskolaï, W. R., Abdou, W., Dipanda, A., and Kolyang (2021). Application of deep learning architectures for satellite image time series prediction: A review. Remote. Sens., 13:4822.

Paszke, A. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

Picard-Krashevski, C., Germain, M., and Laliberté, E. (2025). Multilabel classification of peatland plant species from high-resolution drone images. Ecological Informatics, page 103366.

Ramaswamy, H. G. et al. (2020). Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pages 983–991.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159.

Rogers, J. A., Robertson, K. M., Hawbaker, T. J., and Sousa, D. J. (2024). Classifying plant communities in the north american coastal plain with prisma spaceborne hyperspectral imagery and the spectral mixture residual. Journal of Geophysical Research: Biogeosciences, 129(9):e2024JG008217.

Root, T. L., Price, J. T., Hall, K. R., Schneider, S. H., Rosenzweig, C., and Pounds, J. A. (2003). Fingerprints of global warming on wild animals and plants. Nature, 421(6918):57–60.

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.

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.

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.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626.

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.

Sun, Z., Wang, X., Wang, Z., Yang, L., Xie, Y., and Huang, Y. (2021). Uavs as remote sensing platforms in plant ecology: review of applications and challenges. Journal of Plant Ecology, 14(6):1003–1023.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In International conference on machine learning, pages 10096–10106. PMLR.

Taugourdeau, S., Dionisi, M., Lascoste, M., Lesnoff, M., Capron, J. M., Borne, F., Borianne, P., and Julien, L. (2022). A first attempt to combine nirs and plenoptic cameras for the assessment of grasslands functional diversity and species composition. Agriculture, 12(5):704.

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.

GBIF.org (2025). Gbif home page. https://www.gbif.org.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS), pages 24–25.

Wei, X.-S., Song, Y.-Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., and Belongie, S. (2021). Fine-grained image analysis with deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(12):8927–8948.

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.