Utilising DINOv2 for Domain Adaptation in Vegetation Plot Analysis

Stephen Foy[1,2], Simon McLoughlin[2]

[1]Atlantic Technological University, Galway, Ireland
[2]Technological University Dublin, Dublin, Ireland

Abstract

The Plant CLEF 2024 initiative seeks to improve ecological research through the analysis of vegetation plot inventory photographs, facilitating standardised sampling, biodiversity assessment, long-term monitoring, and extensive remote surveys. This challenge was structured as a multi-label classification task, targeting the prediction of all plant species visible in high-resolution images of vegetation plots. Team Atlantic focused on developing a domain adaptation detection pipeline to bridge the gap between training and test data, adopting a comprehensive methodology. This included a sliding-window strategy for systematically identifying and labelling regions of interest. Probabilistic analysis was used to identify the highest scoring plant species. Optimal results were achieved by integrating these pipelines with the Segmentation and Annotation Method (SAM) to minimise false positives. This paper provides an overview of the challenge, details the methods and insights employed by the researchers, and presents a detailed analysis of the critical findings that led to achieving the highest rank in the challenge. The researchers code used is available at https://github.com/stevefoy/AtlanticAnalytica

Keywords

Imaging, Deep learning, Plant detection, Domain adaptation, Vision transformer, Machine vision

1. Introduction

Quadrat sampling is used by botanists to analyse the types of plant species within a small area of land. The process consisted of using a square or rectangular plot of a known area-known as a quadrat and manually counting the types and abundance of individual species within that area. The process allowed for accurate comparisons and analysis of plant communities across different locations and time periods. The data can be used to assess for biodiversity, ecological health, habitat mapping, invasive species monitoring, climate change studies, restoration projects, agricultural uses and research. Botanists typically use quadrat sampling methods to analyse small plots of land to identify plant species. In addition, they can quantify abundance using various metrics, including biomass, which involves measuring the weight of plant material either in fresh or dry form. Other factors, such as the area occupied by each species and additional qualification factors, are also assessed to provide a comprehensive understanding of the distribution of species and abundance [1].

Despite the extensive application of quadrat sampling, traditional research has focused primarily on evaluating single-species distributional aggregation, leaving multi-species aggregation patterns largely unexamined [2]. Liao et al. have addressed this gap by exploring species distribution across various quadrat sizes, employing advanced statistical models to infer both single- and multi-species aggregation patterns. Their work provides a comprehensive framework for understanding spatial interactions within plant communities, enabling more accurate assessments of biodiversity and species co-existence dynamics [3]. A review by Wäldchen examines the increasing interest in automating plant species identification through computer vision techniques, addressing the limitations of traditional manual identification methods. The review presents a detailed analysis of the workflow involved in automated plant species identification, focussing on critical stages such as image acquisition, pre-processing, feature

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ stephen.foy@atu.ie (S. Foy); simon.d.mcloughlin@tudublin.ie (S. McLoughlin)
🌐 https://stevefoy.github.io/ (S. Foy)
iD 0009-0001-9736-2264 (S. Foy); 0000-0002-9846-4617 (S. McLoughlin)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

extraction, and classification. It highlights the potential of digital techniques to improve accuracy and efficiency in the conservation of biodiversity, representing a substantial advancement over traditional manual methods [4].

Wäldchen's research also highlighted the effectiveness of convolutional neural networks (CNN) like AlexNet, VGGNet, and ResNet, emphasising their robust feature extraction capabilities and high classification accuracy. More recently, strategies have increasingly employed deep learning models for vegetation-type classification, with a shift in focus towards vision foundation models. Mathilde et al. demonstrated that self-supervised Vision Transformers (ViT), deep learning models introduced by Google researchers, are highly effective in generating feature representations for large-scale image classification. This suggests that these foundation models can produce robust and discriminate representations, although they also come with significant computational requirements [5].

Recent research challenges, such as PlantCLEF 2023, utilised datasets comprising a total of 4 million images across 80,000 plant species. Many of the highest-performing outcomes demonstrated that deep learning models are the most effective methods to achieve high accuracy on large datasets [6, 7, 8].

The motivation for Plant CLEF 2024 was to improve ecological studies through vegetation plot inventory photographs, allowing standardised sampling, biodiversity assessment, long-term monitoring, and large-scale remote survey [9]. The challenge initiative aimed to incorporate artificial intelligence to increase the efficiency of specialists, thus broadening the scope and comprehensiveness of ecological research [1, 9].

The inherent complexity of the challenge arises from the domain shift between the training data (source domain) and the test data (target domain). The training images featured single species from various perspectives, while the test images were multi-species and captured from a top-down perspective. The following sections detail the procedures and observations for the training and test datasets. These observations guided the development of the strategy for the challenge submissions. The task was structured as a multi-label classification challenge, requiring the prediction of all plant species present in high-resolution images of vegetation plots.

2. Dataset and Evaluation Metric

Vegetation plot inventories play a crucial role in ecological research, providing standardised methods for sampling, assessing biodiversity, conducting long-term observations, and facilitating extensive remote surveys [10]. In the challenge test set for PlantCLEF 2024, a total of 1,694 images were captured, generally at a resolution of 3065x3065 pixels. The images were typically taken over a 50x50 cm vegetation plot, with the camera positioned parallel to the ground for a top-view perspective. The capture angles were noted to be more or less perpendicular to the ground. Each image could potentially contain multiple species simultaneously. The image capture protocol involved using wooden frames or measuring tape to delimit the plot. In addition, the quality of the images varied depending on the weather, resulting in more or less pronounced shadows and blurry areas.

The primary evaluation metric for the challenge is the F1 score, which seeks a balance between recall and precision to avoid overprediction and underprediction of plant species. These inventories generate critical data that support ecosystem management, biodiversity preservation, and informed environmental policy making. The dataset contains approximately 1.4 million images, extended with trusted labels aggregated from the GBIF platform to enhance the representation of less illustrated species.

2.1. Task Description

The PlantCLEF 2024 challenge was a multi-label classification task aimed at predicting the visible plant species in a set of images. The organisers noted that a single plot generally did not contain dozens and dozens of species simultaneously. Participants were allowed three submissions per day per team, over a period of nine weeks, from 21 March to 24 May.

2.2. Classifiers Analysis

The CLEF Plant organisers fine-tuned two versions of a Vision Transformer architecture known as DINOv2. DINO (DIstillation with NO labels) is a self-supervised learning method designed to train vision transformers (ViTs) without the need for labelled data. DINOv2 is an improved version of the original DINO model, offering improvements in various aspects of self-supervised learning for computer vision tasks [11]. DINOv2 can generate robust and meaningful image representations that can be used for various downstream tasks such as image classification, segmentation, and object detection. These models serve as a foundational base upon which more specific applications can be built through fine-tuning.

In this challenge case, the downstream task was fine-tuned for classification. Two fine-tuned models based on the DINOv2 architecture were used, each with an input resolution of 518x518 pixels. A classifier head was used for both models, consisting of the number of classes in the dataset and a softmax layer, which classified 7,806 species. The first model had the original backbone network frozen, while the second model underwent additional training with the unfrozen backbone network.

The rationale for this training approach was based on the principle that freezing the backbone network (the main body of the model before the classifier head) during fine-tuning is a common technique to prevent overfitting and reduce computational load. Conversely, allowing the backbone to be unfrozen (i.e., allowing its weights to be updated during training) can lead to better adaptation to the new dataset but also risks overfitting, especially if the new dataset is not large [12]. The two DINOv2 models, developed by the competition team, were shared along with the inference code on the Zenodo repository [10]. The models were fine-tuned using the timm library (version 0.9.16) on PyTorch (version 2.2.1+cu121) and incorporated the Exponential Moving Average (EMA) technique to enhance performance metrics. The timm library provided a standardised version of DINOv2 for the challenge.

2.3. Pretrained Models Dataset

These models were showcased during the PlantCLEF 2024 challenge, as detailed on the official pages of ImageCLEF and Hugging Face. The models leverage a substantial dataset pre-trained on approximately 1.4 million images, encompassing 7,806 vascular plant species from the PlantNet collaborative platform for southwestern Europe [13].

Due to the varying sizes of images in the training dataset, a specific set of transformations was applied using the T.Compose function, as reflected in the pre-trained model weights. Transformations included resizing the images to 518x518 pixels, center-cropping to a size of 518x518 pixels, followed by normalisation. Normalisation was performed using mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], which are standard for ImageNet normalisation and standardisation using RGB input images. The dataset was methodically divided into three subsets for model training purposes.

2.4. Challenge Evaluation Method

The primary evaluation metric for the challenge is the F1 score, which seeks a balance between recall and precision to avoid overprediction and underprediction of plant species. The metric used for the primary evaluation is the macro-averaged F1 score per sample, denoted as $F1_{\text{macro, sample}}$, which is calculated by averaging the individual F1 scores of each vegetation plot, formulated as:

$$F1_{\text{macro, sample}} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

where N is the number of samples, and $F1_i$ is the F1 score of the i-th sample. For additional insights, two other variations of the F1 score were also recorded: the Macro F1 Score Averaged Per Species ($F1_{\text{macro, species}}$), which is less relevant for this competition as not all species are evaluated, and the

Micro F1 Score ($F1_{\text{micro}}$) [14], known to be sensitive to data imbalances, and is calculated as:

$$F1_{\text{micro}} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

where precision and recall are aggregated across all predictions. The sklearn framework had an implementation of f1_score function.

3. Methodology

In this section, we outline the methodology behind the submission to the challenge, detailing the various techniques and strategies employed to analyse and process the data. The figure 1 illustrates the results of Team Atlantic over the duration of the competition. The competition period spanned from 21 March 2024 to the deadline on 24 May 2024. It should be noted that not all strategies produced consistently improving results, as evidenced in the plot.

Initial development focused on integrating the shared DINOv2 weights into a classifier pipeline. Subsequently, a sliding-window approach was employed on the challenge data to systematically identify optimal regions for classification. This strategy captured the coordinates of the bounding box and a list of plant species along with their respective classification scores, enabling precise identification and labelling of regions of interest. Probabilistic analysis was then used to identify the regions with the highest classification score. To further optimise object selection and minimise false positives, a segmentation algorithm was incorporated to exclude regions containing rocks and soil.

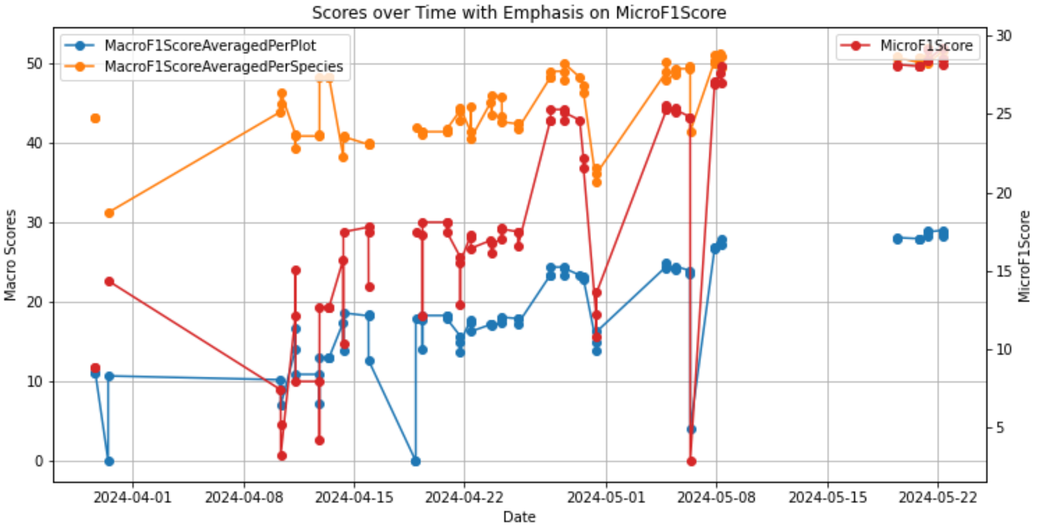

Figure 1: Atlantic scores by week on key indicators

3.1. Region Proposal Pipeline

The objective of implementing a sliding window approach was to address multiple challenges associated with plant detection and feature extraction using DINOv2. Many popular object detection algorithms, such as YOLO (You Only Look Once) and its variants, use a grid and anchor boxes to achieve real-time performance [15, 16]. The selection of the grid dimensions and anchor box sizes is usually derived from clustering of the object sizes in the target domain. However, a sliding-window approach was more suitable with DINOv2. This method allows for the generation of region proposals that better match the varying scales and contextual nuances present in the training and deployment domains. The inference process used a sliding-window technique, capturing the top five classes for each region. Earlier deep learning object detection models like OverFeat [17] used similar sliding windows to generate region proposals. Sliding windows have been shown to be an effective approach, but are not computationally efficient to handle complex image structures and achieve robust detection performance [18].

The DINOv2 model required an input resolution of 518x518 pixels and normalisation to ImageNet standards. Adjustable parameters for the pipeline included window size, stride step, and border offset in pixels, which were necessary to optimise the starting window position to effectively capture regions of interest. Most of the images had wooden frames around the borders, which were excluded by applying a simple border offset of 150 pixels. An example of different windows can be seen in 2. All of these results of the region proposals were stored in a database format as comma-separated values (CSV) files for further analysis. Each line in the file includes the image name, the bounding-box coordinates, and the top five classes along with their associated probability scores. The bounding boxes are defined using the coordinate system (X_1, Y_1, X_2, Y_2), where (X_1, Y_1) represents the top-left corner and (X_2, Y_2) represents the bottom-right corner of the bounding box.

The selection of the window size and the step size was based on a combination of observations of the size of plants in the training data and the test data. Using the following three scales and sliding windows, steps were considered to adequately capture the spatial representation. This approach and its implications are discussed further in the conclusion section.

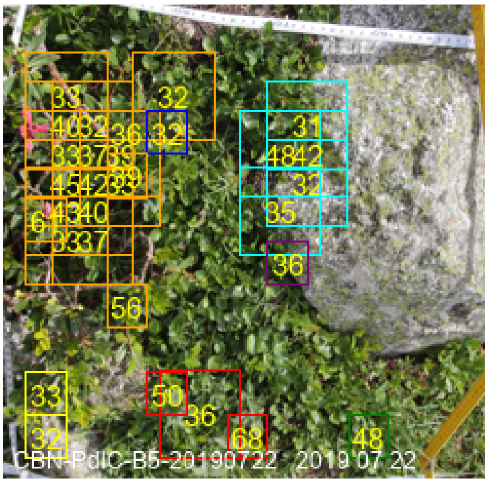

Figure 2: Example debug image of regions, colored by Class ID, with class scores shown in yellow.

- **Set 1: 518 Window, 172 Stride**: This set was generated by cropping the original images to 224 pixels and upsampling them to 518 pixels, using a sliding window with a stride of 172 pixels and a 150-pixel border offset. The generation time on a single RTX 4080 was approximately 12 hours.
- **Set 2: 172 Window, 172 Stride**: This set was generated by cropping the original images to half the size of 518 pixels (resulting in a raw window size of 259 pixels). During inference, these cropped images were upsampled to 518 pixels, with a sliding window of 172 pixels and a 150-pixel border offset. The generation time on a single RTX 4080 was approximately 18 hours.
- **Set 3: 224 Window, 112 Stride**: This set was generated by cropping the original images to 224 pixels and upsampling them to 518 pixels, using a sliding window with a stride of 112 pixels. Additionally, each crop was rotated by +90 and -90 degrees, and these results were also saved. No padding was used in the border. The generation time on a single RTX 4080 was over 24 hours.

3.2. Top Species Selection

The application of sliding windows resulted in three sets of data, which required further filtering to identify the highest values per image and to format the data according to the Hugging Face PlantCLEF competition requirements. This process involved identifying all unique species in each challenge image.

Initially, a fixed probability threshold of 30 was established, which yielded moderate success in selecting the top plant species in the challenge images. However, adopting a statistical approach to achieve a dynamic threshold using the 90th percentile threshold proved to be superior [19]. Adjusting this parameter up to the 98th percentile during the initial trials did not result in further improvements.

Additionally worth noting, was the high number of detection's in some plots and low detection rates in others, the number of classes per plot was capped between 8 to 12 plant species for submissions in many of the resultant runs. Generally, this basic strategy yielded results in the following ranges: Macro F1 score averaged per plot: ∼7 to ∼19, macro F1 score averaged per species: ∼30 to ∼48, and Micro F1 score: ∼16 to ∼18.

3.3. Additional Optimisations

Upon analysing some trials using a segmentation algorithm discussed in the next section, it was observed that similar masks were generated. Initially, it was assumed that there was an issue with the generation of segmentation masks, but further exploration revealed that the images had a specific naming convention. These groups of images helped with debugging detection's and increased the overall scores when incorporated into the highest species selection on plots.

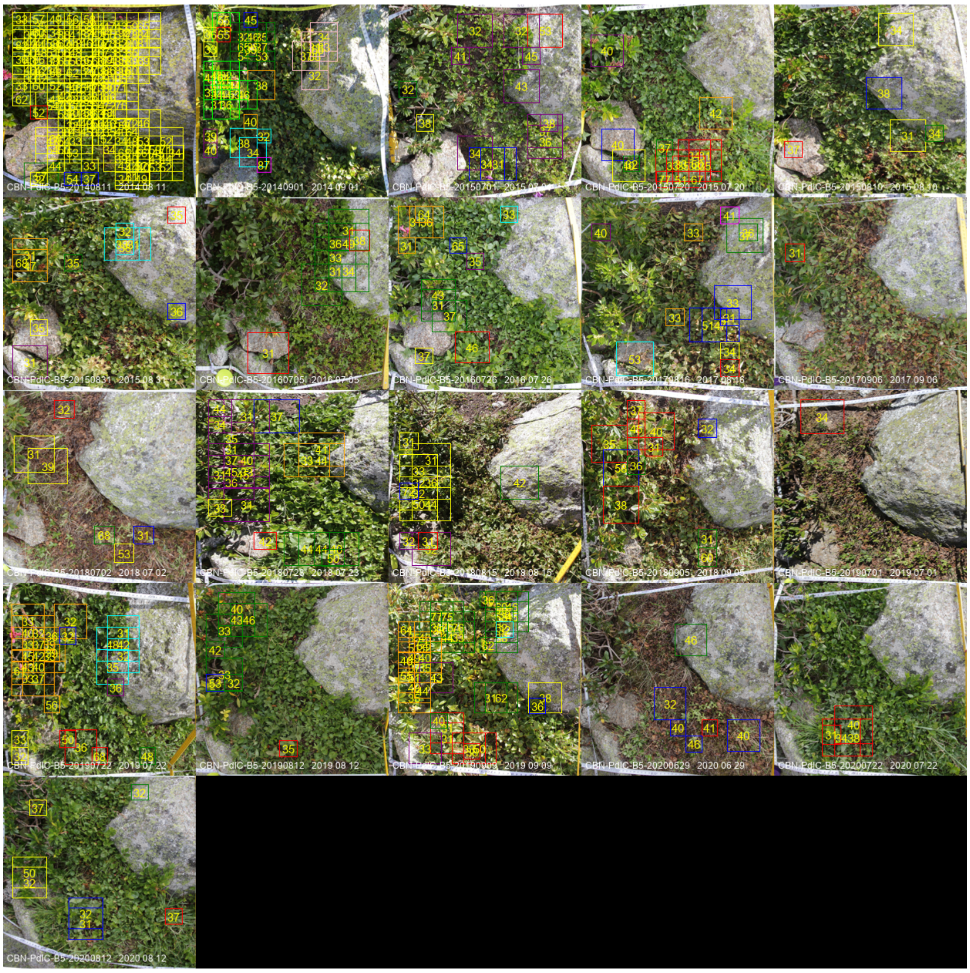
CBN-PdlC-B5-20140811 2014 08 11
CBN-PdlC-B5-20140901 2014 09 01
CBN-PdlC-B5-20150701 2015 07 01
CBN-PdlC-B5-20150720 2015 07 20
CBN-PdlC-B5-20150810 2015 08 10
CBN-PdlC-B5-20150831 2015 08 31
CBN-PdlC-B5-20160705 2016 07 05
CBN-PdlC-B5-20160726 2016 07 26
CBN-PdlC-B5-20170816 2017 08 16
CBN-PdlC-B5-20170906 2017 09 06
CBN-PdlC-B5-20180702 2018 07 02
CBN-PdlC-B5-20180723 2018 07 23
CBN-PdlC-B5-20180815 2018 08 15
CBN-PdlC-B5-20180905 2018 09 05
CBN-PdlC-B5-20190701 2019 07 01
CBN-PdlC-B5-20190722 2019 07 22
CBN-PdlC-B5-20190812 2019 08 12
CBN-PdlC-B5-20190909 2019 09 09
CBN-PdlC-B5-20200629 2020 06 29
CBN-PdlC-B5-20200722 2020 07 22
CBN-PdlC-B5-20200812 2020 08 12

Figure 3: Images from the group CBN-PdlC-B5_21, 21 image plots at different dates

The filename was divided into components using the hyphen (-) as the delimiter. For example, the test image filename "OPTMix-0598-P1-149-20231205" was divided into a list as "OPTMix", "0598", "P1", "149", and "20231205". The substrings at index 0, 1, and 2 were used to create file groups, as these images were approximately captured over the same plot but at different times of the year. This method enabled the application of statistical analysis on a plot to determine the potential plant species present in these

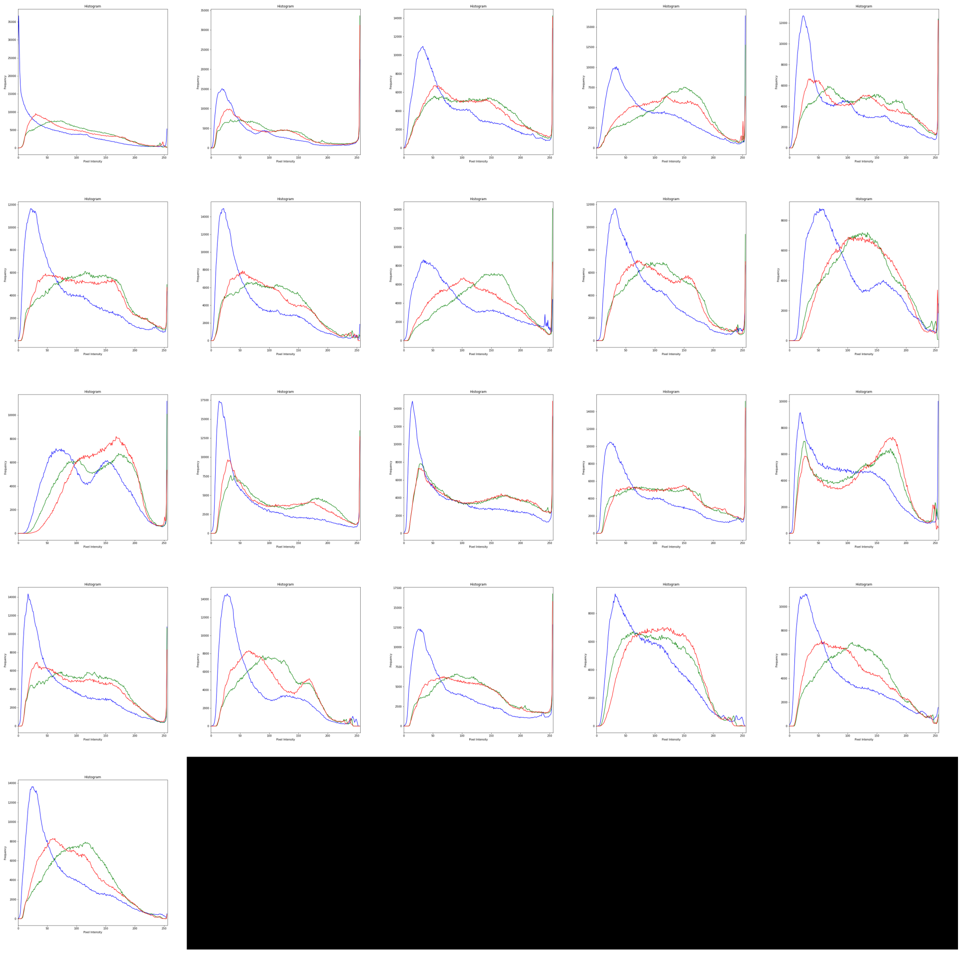

Figure 4: Histograms of images from the group CBN-PdlC-B5_21, 21 image plots at different dates

areas. However, despite the expectation that similar plots would contain similar plant species, the figure 3 illustrates the limitations of the trained model on the detection of similar classes throughout the image group. There are substantial differences between the histograms, as seen in figure4, indicating significant variability in the image characteristics. Generally, this strategy yielded results in the following ranges: Macro F1 score averaged per plot: ∼22 to ∼24, macro F1 score averaged per species: ∼46 to ∼50, and Micro F1 score: ∼21 to ∼26.

3.4. False Positive Removal

During the debugging process, it was observed that plant classifications were occurring on soil and rocks. Further investigation revealed that the high confidence scores originated from training data with a high percentage of rock or soil in the background. This led to the implementation of a trial using HSV (Hue, Saturation, Value) values. If a percentage of pixels within the bounding box region of interest exceeded a specified threshold, the region proposal was ignored. The primary objective was to remove false positives; however, while this method showed some performance improvement, the inherent issues of selecting appropriate HSV values ranges across all images proved to be too challenging. The selection of appropriate HSV ranges for green vegetation was based on a combination of empirical observation with the GIMP (GNU Image Manipulation Program) software and established literature on colour segmentation. The work of Yang et al. examines the use of the HSV colour space to isolate green plants by applying thresholds [20].

An alternative approach was to use deep learning for segmentation of the images which can automatically identify and separate objects based on their common colors, textures, and other common features. The Segment Anything Model (SAM) algorithm is a computer vision technique designed to automatically segment objects within an image by generating precise masks [21]. Due to the extensive size of the SAM code base, the decision was made to utilise the pre-existing SAM algorithm to produce masks for each test image, rather than coding it from scratch. There are numerous alternative CNNs for segmentation, but SAM is robust across various datasets without requiring model retraining [22, 23].

The inference time for the SAM model on a single NVIDIA RTX 4080 GPU required approximately 45 hours to process 1,586 out of a total of 1,695 challenge images. Despite the fact that the process was inadvertently halted early, the generated data was deemed sufficient to test the concept. A total of 2.53 GB of data were produced, and each image resulted in a corresponding folder of data. Each folder contained multiple masks for segmented objects. Note that these masks do not have classifications per se for each mask.

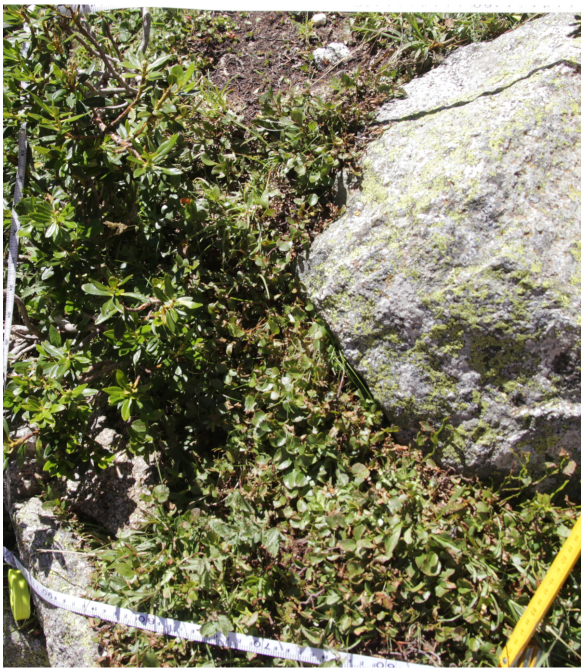

Figure 5: CBN-PdlC-B5-20180815.jpg Raw

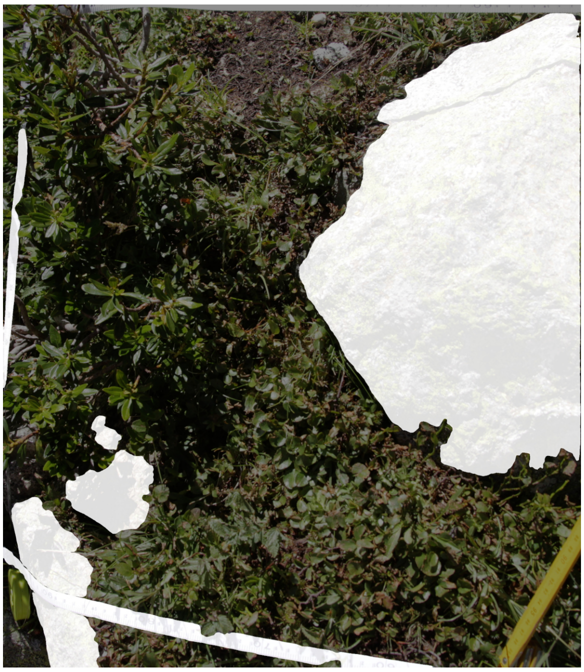

Figure 6: CBN-PdlC-B5-20180815.jpg SAM objects

A single mask was created for each image folder, containing all large rocks and soil areas. This mask was generated using a simple classification strategy for soil and rocks. An example of a challenge image can be seen in the figure 5 and this same image has an overlay of the mask in figure 6 where white pixel are objects. The strategy involved using the generated masks and the original image to identify objects larger than 100 pixels with an aspect ratio of one to one and either a grey texture or a brown soil-like texture. This resultant mask was used to reject plants on rocks and single plants in large areas. If a bounding box from the sliding-window algorithm had a codable parameter of percentage intersection with this mask, these objects were directly rejected.

The DINOv2 model failed to detect similar plant species in different plots under varying lighting conditions, resulting in false negative (FN) detection issues. The following score ranges were observed using all the combined techniques mentioned. The codable parameters for the maximum number of plants per plot and the percentage intersection threshold were tested, respectively, with the values 8 to 14 and 50 to 90 percent, resulting in ∼26 to 29.62 for the Macro F1 score averaged per plot, ∼48 to 51.2 for Macro F1 score averaged per species, and ∼28 to 30.01 for Micro F1 score.

4. Summary and Conclusions

The methodologies and strategies implemented in this study significantly improved the precision of the detection of multi-plant species on a plot. Integrating the DINOv2 model with a sliding-window approach and probabilistic analysis yielded promising results in identifying the top species. Furthermore, employing the Segmentation and Annotation Method (SAM) to reduce false positives from rocks and soil resulted in a notable increase of 6 percentage points bringing the Micro F1 score to a value slightly exceeding 30. Additionally, a regional proposal strategy using SAM would be a perceptible next step, and if computational resources allow, fine-tuning the DINOv2 model would be a considerable direction to further enhance detection performance.

Future work should explore the implementation of region proposal strategies using SAM or an alternative segmentation algorithm to refine detection regions and exclude irrelevant or non-target regions. Furthermore, given sufficient computational resources, fine-tuning the DINOv2 model with an augmentation strategy could provide even more accurate and robust results. Training data could be optimised to focus on plant characteristics. These advances have great potential to improve detection accuracy.

Acknowledgments

We extend our gratitude to the developers of the CLEF2024 team for their pre-trained DINOv2 model. This aspect of the project was particularly interesting, as it transformed the project into a potential competition using only a limited compute, leveraging a state-of-the-art vision foundation model.

References

[1] H. Goëau, V. Espitalier, P. Bonnet, A. Joly, Overview of PlantCLEF 2024: Multi-Species Plant Identification in Vegetation Plot Images, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[2] F. He, K. J. Gaston, Estimating Species Abundance from Occurrence., The American Naturalist 156 (2000) 553–559. URL: https://www.journals.uchicago.edu/doi/10.1086/303403. doi:10.1086/303403, publisher: The University of Chicago Press.

[3] Z. Liao, J. Zhou, T.-J. Shen, Y. Chen, Inferring single- and multi-species distributional aggregation using quadrat sampling, Ecological Indicators 156 (2023) 111085. URL: https://www.sciencedirect.com/science/article/pii/S1470160X2301227X. doi:10.1016/j.ecolind.2023.111085.

[4] J. Wäldchen, P. Mäder, Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review, Archives of Computational Methods in Engineering 25 (2018) 507–543. URL: http://link.springer.com/10.1007/s11831-016-9206-z. doi:10.1007/s11831-016-9206-z.

[5] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin, Emerging Properties in Self-Supervised Vision Transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9630–9640. URL: https://ieeexplore.ieee.org/document/9709990. doi:10.1109/ICCV48922.2021.00951, iSSN: 2380-7504.

[6] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, Y. Li, X.-S. Wei, A Deep Learning based Solution to FungiCLEF2023, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR, Thessaloniki, Greece, 2023, pp. 2051–2059. URL: https://ceur-ws.org/Vol-3497/#paper-173, iSSN: 1613-0073.

[7] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2023: Image-based Plant Identification at Global Scale, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR, Thessaloniki, Greece, 2023, pp. 1972–1981. URL: https://ceur-ws.org/Vol-3497/#paper-167, iSSN: 1613-0073.

[8] M. Xu, S. Yoon, C. Wu, J. Baek, D. S. Park, PlantCLEF2023: A Bigger Training Dataset Contributes More than Advanced Pretraining Methods for Plant Identification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR, Thessaloniki, Greece, 2023, pp. 2168–2180. URL: https://ceur-ws.org/Vol-3497/#paper-183, iSSN: 1613-0073.

[9] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. \Šulc, M. Hrúz, M. Servajean, J. Matas, others, Overview of lifeclef 2024: Challenges on Species Distribution Prediction and Identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[10] H. Goëau, J.-C. Lombardo, A. Affouard, V. Espitalier, P. Bonnet, A. Joly, PlantCLEF 2024 pretrained models on the flora of the south western Europe based on a subset of Pl@ntNet collaborative images and a ViT base patch 14 dinoV2 (2024). URL: https://zenodo.org/records/10848263, publisher: Zenodo.

[11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning Robust Visual Features without Supervision, 2024. URL: http://arxiv.org/abs/2304.07193. doi:10.48550/arXiv.2304.07193, arXiv:2304.07193 [cs].

[12] R. Bao, Y. Sun, Y. Gao, J. Wang, Q. Yang, H. Chen, Z.-H. Mao, Y. Ye, A Survey of Heterogeneous Transfer Learning, 2023. URL: http://arxiv.org/abs/2310.08459. doi:10.48550/arXiv.2310.08459, arXiv:2310.08459 [cs].

[13] PlantNet, PlantNet https://plantnet.org/en/, 2019. URL: https://plantnet.org/en/, publication Title: Pl@ntNet.

[14] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. Google-Books-ID: Np9SDQAAQBAJ.

[15] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv:2004.10934 [cs, eess] (2020). URL: http://arxiv.org/abs/2004.10934, arXiv: 2004.10934.

[16] J. Redmon, A. Farhadi, YOLO9000: Better, Faster, Stronger, arXiv:1612.08242 [cs] (2016). URL: http://arxiv.org/abs/1612.08242, arXiv: 1612.08242.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, 2014. URL: http://arxiv.org/abs/1312.6229. doi:10.48550/arXiv.1312.6229, arXiv:1312.6229 [cs].

[18] A. Lavin, S. Gray, Fast Algorithms for Convolutional Neural Networks, arXiv:1509.09308 [cs] (2015). URL: http://arxiv.org/abs/1509.09308.

[19] N. Developers, numpy.percentile, 2024. URL: https://numpy.org/doc/stable/reference/generated/numpy.percentile.html.

[20] W. Yang, X. Zhao, W. Sile, J. Zhang, J. Feng, Greenness identification based on HSV decision tree, Information Processing in Agriculture 2 (2015). doi:10.1016/j.inpa.2015.07.003.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment Anything, 2023. URL: http://arxiv.org/abs/2304.02643. doi:10.48550/arXiv.2304.02643, arXiv:2304.02643 [cs].

[22] H. Cuevas-Velasquez, A.-J. Gallego, R. B. Fisher, Segmentation and 3D reconstruction of rose plants from stereoscopic images, Computers and Electronics in Agriculture 171 (2020) 105296. URL: https://www.sciencedirect.com/science/article/pii/S0168169919323919. doi:10.1016/j.compag.2020.105296.

[23] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, 2018. URL: http://arxiv.org/abs/1703.06870. doi:10.48550/arXiv.1703.06870, arXiv:1703.06870 [cs].