Zero-Shot Segmentation through Prototype-Guidance for Multi-Label Plant Species Identification

Luciano Araujo Dourado Filho[1,*], Almir Moreira da Silva Neto[1], Rodrigo Pereira David[2] and Rodrigo Tripodi Calumby[1]

[1]*ADAM - Advanced Data Analysis and Management, University of Feira de Santana (UEFS), Feira de Santana, Brazil*
[2]*National Institute of Metrology, Technology and Quality (Inmetro)*

Abstract

This paper presents an approach developed to address the PlantClef 2025 challenge, which consists of a fine-grained multi-label species identification, over high-resolution images. Our solution focused on employing class prototypes obtained from the training dataset as a proxy guidance for training a segmentation Vision Transformer (ViT) on the test set images. To obtain these representations, the proposed method extracts features from training dataset images and create clusters, by applying K-Means, with K equals to the number of classes in the dataset. The segmentation model is a customized narrow ViT, built by replacing the patch embedding layer with a frozen DinoV2, pre-trained on the training dataset for individual species classification. This model is trained to reconstruct the class prototypes of the training dataset from the test dataset images. We then use this model to obtain attention scores that enable to identify and localize areas of interest and consequently guide the classification process. The proposed approach enabled a domain-adaptation from multi-class identification with individual species, into multi-label classification from high-resolution vegetation plots. Our method achieved fifth place in the PlantCLEF 2025 challenge on the private leaderboard, with an F1 score of 0.33331. Besides that, in absolute terms our method scored 0.03 lower than the top-performing submission, suggesting that it may achieved competitive performance in the benchmark task. Our code is available at https://github.com/ADAM-UEFS/PlantCLEF2025.

Keywords

vision transformer, multispecies classification, plant identification, fine-grained classification, prototype learning, clustering

1. Introduction

The automated identification of plant species from images of complex vegetation plots presents a significant challenge in ecological research and biodiversity monitoring. Initiatives like the PlantCLEF aim to advance this field by providing benchmark datasets and tasks, fostering innovation in multi-label classification from high-resolution plot imagery [1, 2, 3, 4]. A primary challenge in this task is the notable domain shift between training data, which mostly consist of images of individual plants, and test data, comprising dense, multi-species vegetation plots captured under varied environmental conditions [3, 4, 5, 6]. Therefore, it demands robust methods that can effectively generalize and differentiate target species within cluttered natural scenes [7].

Vision Transformers (ViTs), particularly those pre-trained using self-supervised learning, have demonstrated remarkable effectiveness in generating discriminative visual features for various tasks, including large-scale plant identification [8, 9]. In the broad context of image classification, models such as DINOv2 [10] has demonstrated to offer powerful backbones for feature extraction. Specifically in The PlantCLEF 2025, organizers supported participants by providing ViT models pre-trained on

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*]Corresponding author.
lucianoadfilho@gmail.com (L. A. D. Filho); almirneto338@gmail.com (A. M. d. S. Neto); rpdavid@inmetro.gov.br (R. P. David); rtcalumby@uefs.br (R. T. Calumby)
https://github.com/FalsoMoralista (L. A. D. Filho); https://github.com/almirneeto99/ (A. M. d. S. Neto); https://github.com/rpdavid78/ (R. P. David); https://www.rtcalumby.com.br (R. T. Calumby)
0000-0002-0507-2201 (L. A. D. Filho); 0009-0008-5042-5556 (A. M. d. S. Neto); 0000-0001-9218-8191 (R. P. David); 0000-0001-8515-265X (R. T. Calumby)

relevant flora, considering a DINOv2-based architectures [11]. Our proposed method goes beyond the usual direct end-to-end classification of the entire plot images (which can be confounded by background noise and species overlap) by focusing on a preliminary segmentation-driven filtering step. We propose a “decrease-and-conquer” strategy which relies in a dedicated ViT that first identifies plant-relevant regions, thereby simplifying the subsequent classification task for specialized heuristics. This ViT leverages its attention mechanisms as a proxy for spatial segmentation, aiming to isolate pertinent areas from irrelevant background elements.

A crucial aspect of our methodology is the training of this segmentation ViT without direct pixel-level supervision. To overcome the absence of segmentation labels, we introduce a novel proxy task as our core contribution. We trained a narrow, customized ViT to reconstruct a predetermined representation of the *training dataset*—specifically, class prototypes derived from k-Means clustering on DINOv2 embeddings of the original training set images. Uniquely, this ViT learns this reconstruction objective using, as its input, feature embeddings extracted by a separate, frozen DINOv2 model from *test images*. Our hypothesis is that for the ViT to successfully map features from unseen test plots back to these known training set prototypes, it must learn to generate attention scores that highlight plant-relevant patches while attenuating signals from the background (as illustrated in Figure 3).

This paper provides a detailed exposition of this proxy task-driven ViT technique. The generation of the target class prototypes, the specific architecture of the custom ViT, and the operational steps of its unique training regimen as detailed in Section 3. Furthermore, we describe how the learned attention maps are utilized within our inference pipeline to select regions of interest prior to applying classification heuristics for final species identification. The experimental settings, key hyperparameters, and the results achieved in the PlantCLEF 2025 challenge are presented and discussed, offering insights into the effectiveness of this approach for multi-label classification in complex ecological imagery.

2. Task Description

The PlantCLEF 2025 challenge addresses the intricate problem of multi-label species identification within high-resolution images of vegetation plots. The core objective is to predict all plant species present in a given plot image, a fundamental task to ecological research, biodiversity assessment, and long-term environmental monitoring.

The primary difficulty of the challenge stems from a significant domain shift between the provided datasets. A comprehensive training dataset comprising roughly 1.4 million images representing 7,806 plant species. These training images typically feature single plant species from various perspectives. In contrast, the test dataset consists of 2,105 high-resolution plot images from Mediterranean and Pyrenean regions. These test images depict complex, multi-species scenes, often captured from a top-down perspective within delimited sampling areas (quadrats), and under varying environmental conditions that can introduce shadows or blurry areas.

Two pre-trained ViT models based on the DINOv2 architecture were provided along with the dataset. These models were not only pre-trained on a massive dataset of images but were also specifically fine-tuned on the PlantCLEF 2024 training data which includes isolated plant samples, considering multiple organs and perspectives. The fine-tuned models available were: a) one where only the final classification layer was trained; b) another where all model layers were unfrozen and fine-tuned for the task. These models offer a powerful, ready-to-use foundation resources avoiding new investigations to demand the extensive computational resources needed to train such large-scale models from scratch. The classification effectiveness was evaluated using the macro F1 score averaged per plot, a metric designed to balance precision and recall for each individual plot images, with results tracked on a public leaderboard. The final score is computed as described in Equation 1, where N is the number of transects (plots) , T^i is the number of quadrats in transect i, $\frac{1}{T_i} \sum_{j=1}^{T_i} F1^j$ is the macro-averaged F1-score per sample of transect i and $F1^j$ is the F1 score for test image j.

$$Score = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{T_i}\sum_{j=1}^{T_i}F_1^j) \tag{1}$$

For each test image j the F1 score is described as follows in Equation 2.

$$Macro\ F_1^j = \frac{2 \times Precision_j \times Recall_j}{Precision_j + Recall_j} \tag{2}$$

The *Precision* and *Recall*, for each test image j is represented in Equation 3 and 4, respectively.

$$Precision = \frac{TP_j}{TP_j + FP_j} \tag{3}$$

$$Recall = \frac{TP_j}{TP_j + FN_j} \tag{4}$$

Where TP_j is the number of species correctly predicted, FP_j is the number of plant species incorrectly predicted and FN_j number of plant species missed.

3. Proposed Method and Experiments

The proposed zero-shot learning approach focuses on the segmentation of plants from irrelevant background elements within test images. In order to achieve that, we trained a dedicated segmentation model, whose primary function is to identify plant regions. We theorize that a preliminary segmentation step could allow for more precise application of classification heuristics, over pertinent regions in the image, in a "decrease-and-conquer" fashion.

For the segmentation task, we employed a ViT due to its capability to model contextual relationships between image patches (tokens), according to an objective. We propose that given the appropriate objective, these relationships could be guided towards enabling the model to differentiate plant structures from other objects. In other words, we anticipate that the ViT's attention scores could provide a mechanism for selecting relevant regions from test images.

The caveat of our approach is that in order for this model to learn semantically meaningful (patch-wise) relationships, a reasonable objective has to be given, which we initially do not have. To tackle such limitation, a proxy task was designed consisting of reconstructing an approximate representation of the training dataset. We used the training dataset to obtain features that represented each of its individual species generating cluster prototypes, which we employ as reconstruction objective for our ViT. We considered that in order to successfully reconstruct the training dataset representation from test-set images, the model would have to learn patch-wise relationships (attention scores) that enabled

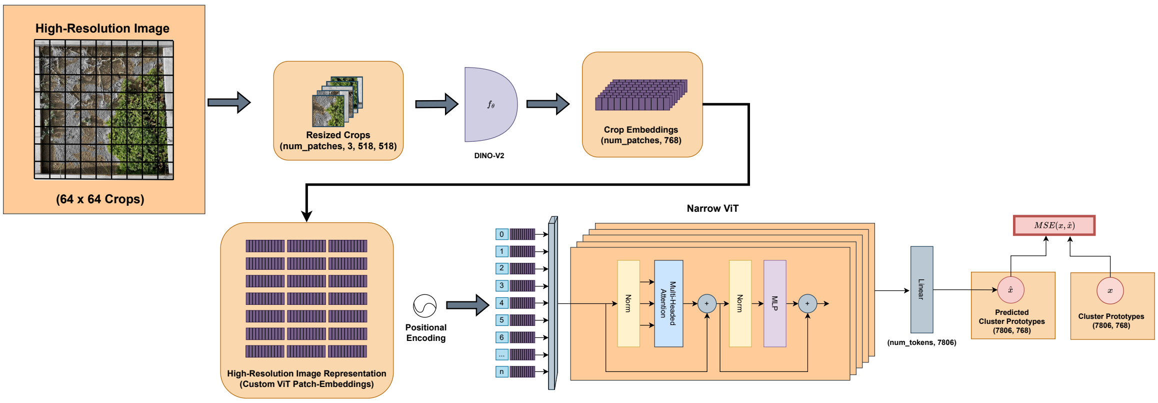

Figure 1: Training process diagram.

the maximization of this objective. As a consequence, our assumption is that the model would learn to assign low attention scores to irrelevant regions (background noise) and high scores to patches containing plants (as illustrated in Figure 1).

3.1. Clustering as proxy task

To obtain the target representation that guides the ViT's optimization, we applied K-Means clustering to the image embeddings of the training dataset. These embeddings were extracted using a separate, identical instance of the pretrained DINOv2 model, from which the classification head has been removed. Following embedding extraction, K-Means was performed with K set to the number of classes (7806 species). The resulting cluster centroids, or 'class prototypes', constitute the predetermined representation of the training dataset that the narrow ViT is addressed to reconstruct. Figure 1 illustrates the complete training process wherein the ViT learns this reconstruction.

Figure 1 details the operational steps for training our narrow ViT on this proxy reconstruction objective. The process begins with resizing each image from the test dataset into a predefined $H \times W$ resolution, where H and W denote image height and width. Following that, each high-resolution image is segmented into a grid of non-overlapping 64×64 pixel patches (crops). These patches are subsequently resized to 518×518 pixels using bicubic interpolation to meet the input specifications of the pre-trained DINOv2 model, which serves as the patch feature extractor. This procedure yields $num\_patches = \frac{H \times W}{64^2}$ tokens per image.

These resized 518×518 pixel crops from the test images are then processed by the aforementioned frozen DINOv2 model to extract the corresponding (768-dimensional) feature embeddings. Positional encodings are subsequently integrated with these patch embeddings for the ViT processing. This final sequence of augmented patch embeddings serves as the input to the narrow ViT. The ViT architecture is a ViT-Base (ViT-B) variant configured with 6 transformer blocks, 12 attention heads per block, an embedding dimension of 768, and a linear layer. This model is optimized to reconstruct the predetermined target matrix of class prototypes (dimensions 7806×768) from these input embeddings.

3.2. Inference

Following the training procedure, an optimized Vision Transformer (ViT) is employed to generate attention maps for the test images. These scores facilitate the removal of background and other irrelevant elements from images, prior to the classification stage. To accomplish this, an image is first processed through the ViT. The attention scores from all transformer blocks are then aggregated by averaging them across both blocks and attention heads. This averaged result is subsequently normalized to create a final attention map. This attention map guides the selection of relevant image regions by filtering tokens: only tokens whose normalized attention scores exceed a predefined threshold t are retained. After this filtering step, which discards potentially irrelevant elements, the process advances to the species identification pipeline. For this identification task, we implemented two distinct procedures which are detailed in the following.

As an initial step, the first heuristic comprises resizing each relevant crop to conform to the input dimensions required by the pretrained DINOv2 model. Following this preprocessing, each resized crop serves as input to the pretrained DINOv2 (with classifier) to obtain the classification scores for each crop. From that, the predicted species was determined by selecting the most confident prediction that surpassed a predefined threshold: *prob*. In instances where no prediction met this criterion, the species associated with the highest overall probability score was selected by default. The final list of predicted species for a given quadrat was then compiled by aggregating the unique species IDs from all individual crop-level predictions within that quadrat.

Also following the application of a threshold (t) to the normalized attention map, the second heuristic consisted of constructing a composite image as follows. For each patch with an attention score above t, an image was assembled as a $K \times K$ grid of neighbouring patches, using the relevant patch as the central element. This newly assembled image was then resized via bicubic interpolation to match the input

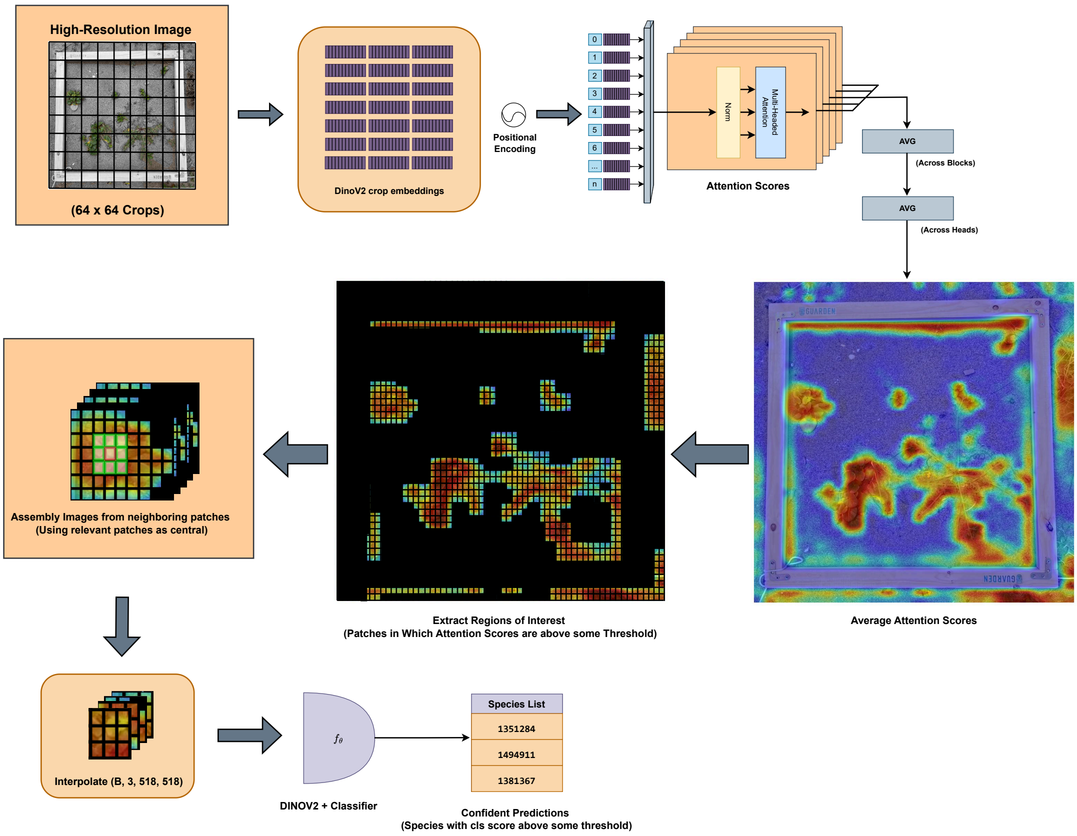

Figure 2: Inference stage diagram.

dimensions of the pretrained DINOv2 model and processed for classification. Analogous to the first heuristic, this approach assumes the classifier can inherently filter irrelevant or outlier regions by prioritizing predictions that achieve a confidence score above a specified threshold (*prob*). Figure 2 illustrates this procedure, which was the approach that yielded the best performance of the proposed method. During this inference stage, we experimented over the classification parameters depicted in Table 1 to optimize predictive performance.

3.3. Experimental Settings

Our custom ViT implementation is based on the I-JEPA architecture [12], in which the `[cls]` token is disregarded. We evaluated two input image resolutions: 2048 × 2048 pixels and 3072 × 2048 pixels. When processed into a grid of non-overlapping 64 × 64 pixel crops, these resolutions correspond to 1024 and 1536 patches (tokens) per image, respectively.

Model optimization was performed for up to 30 epochs using the AdamW optimizer. This included a cosine learning rate decay schedule [12], typically applied over 10 to 20 epochs of the training period, and a weight decay schedule that linearly increased from 0.04 to 0.4. A consistent learning rate profile was adopted across all experiments, defined by: a starting learning rate (LR_{start}) of 5.0×10^{-6}, an effective

Table 1
Classification parameters values

Classification Parameter	Range
Attention threshold (t)	[0.5, 0.6, 0.7]
Probability threshold (prob)	[0.3, 0.45 ,0.5, 0.6, 0.7, 0.8]
Grid Size (K)	[5, 9]

peak learning rate ($LR_{\text{effective}}$) of 1.0×10^{-3}, and a final learning rate (LR_{final}) of 1.0×10^{-6}. Early stopping was implemented, triggered by observed degradation in the quality of attention maps and the onset of model overfitting. Ablation studies, extended to 100 epochs, confirmed that training for 30 epochs with this early stopping criterion was sufficient to achieve optimal performance.

All experiments were conducted on a system equipped with two NVIDIA A100-SXM 80GB GPUs. A per-GPU batch size of 128 was employed, achieved through gradient accumulation. Our observations indicated that this VRAM capacity constrained the maximum number of tokens processed simultaneously to approximately 2048 for the given model configuration.

4. Results

Table 2 summarizes the results all the configurations evaluated. Configurations 1-4 used our first classification strategy, which consisted of performing direct patch-wise classification over resized crops.

As demonstrated in Table 2, the patch-wise classification strategy yielded the lowest performance among all evaluated. Specifically, after applying an attention threshold of $t = 0.6$ and classifying the resulting patches without any probability constraints (Configuration 1), the model achieved an F_1-score of 0.02114 on the private set. While a marginal improvement to an F_1-score of 0.09511 was obtained by introducing a probability threshold of $prob = 0.5$ (Submission 3), the overall strategy demonstrated a low performance. We attribute these suboptimal results to the low-resolution nature of the individual patches, which likely misses the necessary contextual information required for the model to identify a complete plant specimen in a effective way.

In contrast, the alternative approach, adopting the strategy of assembling a contextual grid of adjacent patches around each relevant token (as depicted in Figure 2) yielded a remarkable performance improvement. Using the exact same model weights, this strategy increased the F_1-score from 0.09511 to 0.33131. This outcome corroborates our hypothesis that individual, low-resolution patches lack sufficient contextual information for effective classification. Subsequent experiments with a smaller grid size ($K = 5$ in submissions 15 and 16) provided further evidence, demonstrating that reducing the amount of contextual information by using fewer adjacent patches led to deterioration in the classification performance. More information regarding checkpoint epoch selection is described in Section 4.1

The best results were obtained with configuration 14, where the segmentation model was trained with an input resolution of 3072 × 2048. This submission achieved F1= 0.33331 on the private score, which represented an improvement of approximately 0.5% over our first submission with this strategy (with an input resolution of 2048 × 2048 pixels). This findings indicated that increasing the input resolution may not be worthy, specially considering the computational and memory overhead associated to token processing in the ViT.

In summary this decrease-and-conquer strategy enabled the proposed method to reach the 5th place on the private leaderboard of PlantCLEF 2025. Moreover, our best configuration achieved a F_1-score of 0.33331, in contrast to the best performing proposal in the leaderboard, which achieved 0.36479. In absolute terms, our strategy presented roughly a 0.03 difference in contrast to the first place, which demonstrates its high competitiveness. We believe that with further refinements, this strategy holds the potential to achieve even higher performance in future iterations.

4.1. Training Dynamics and Model Selection

Figure 3 shows a qualitative example of a well-formed attention map from an early training stage, where the model correctly learns to assign low attention scores (in blue) to irrelevant regions, such as the quadrat frame, and higher scores (in red) to regions containing plants.

Regarding to model training, a counter-intuitive phenomenon was observed: a clear degradation in the quality of the attention maps coincided with the convergence the loss function (Figure 4).

For the 2048 × 2048 input resolution, for example, the model performed as intended, up to approximately Epoch 15. Beyond this point, a clear reversal in behavior was noted: attention scores for relevant

Table 2

Performance metrics of the evaluated 16 configurations with varying hyperparameter configurations. The configurations are grouped by classification strategy. The absence of parameter K indicates the use of the direct patch-wise classification strategy (Heuristic 1). Public Score results are reported on 11% of the data, whereas the Private Score is obtained from the remaining 89%.

Configuration	Input Resolution	Checkpoint Epoch	t	prob	K	Public Score	Private Score
			Patch-wise Classification (Heuristic 1)				
1	2048 × 2048	15	0.6	-	-	0.02709	0.02114
2	2048 × 2048	15	0.6	0.15	-	0.08554	0.07328
3	2048 × 2048	15	0.6	0.5	-	0.11809	0.09511
4	2048 × 2048	15	0.6	0.7	-	0.13248	0.09367
			Grid Assembly Classification (Heuristic 2)				
5	2048 × 2048	15	0.6	0.5	9	0.32606	0.33131
6	2048 × 2048	15	0.6	0.8	9	0.26701	0.32461
7	3072 × 2048	10	0.6	0.6	9	0.33069	0.33007
8	3072 × 2048	10	0.6	0.7	9	0.31004	0.32181
9	3072 × 2048	10	0.6	0.5	9	**0.34427**	0.33183
10	3072 × 2048	10	0.6	0.45	9	**0.34427**	0.33183
11	3072 × 2048	10	0.6	0.3	9	**0.34427**	0.33183
12	3072 × 2048	10	0.5	0.5	9	0.34060	0.33151
13	3072 × 2048	10	0.5	0.7	9	0.31153	0.32595
14	3072 × 2048	**10**	**0.7**	**0.5**	**9**	0.33867	**0.33331**
15	3072 × 2048	10	0.6	0.60	5	0.30671	0.28580
16	3072 × 2048	10	0.6	0.8	5	0.29442	0.28908

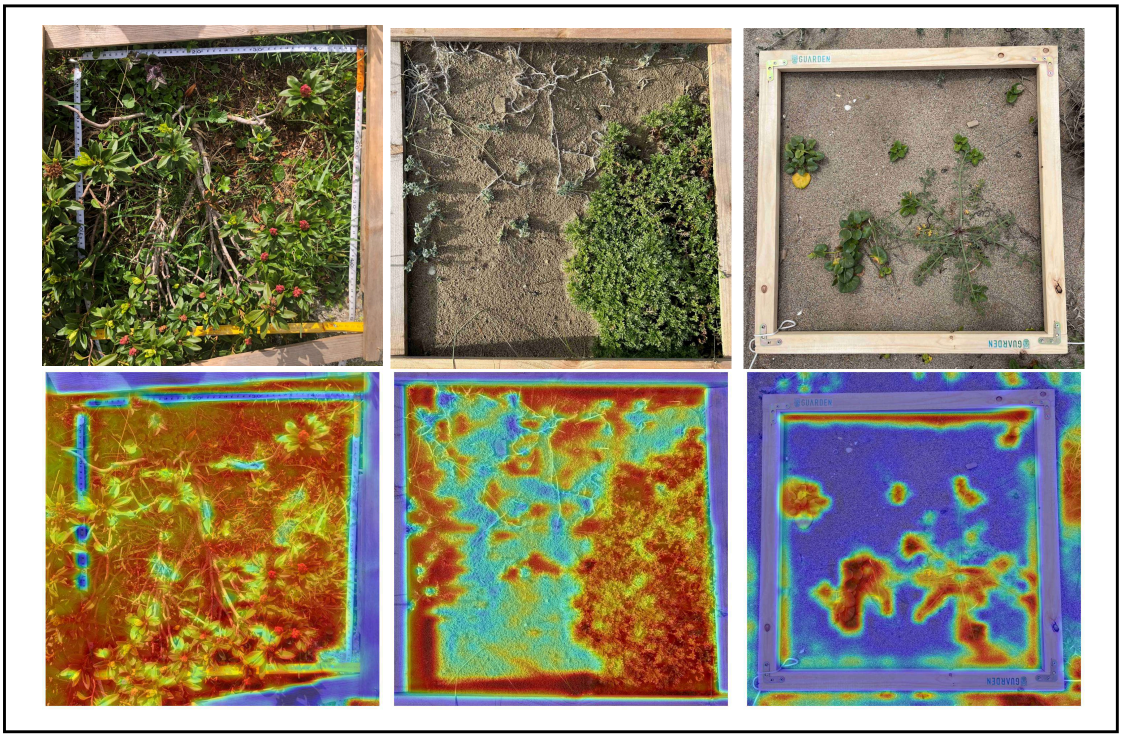

Figure 3: Qualitative results from the segmentation ViT during an early training epoch. The model correctly assigns high attention scores (red) to plant-containing regions and low scores (blue) to background elements, including the quadrat frame.

plant areas began to decrease, while scores for background elements started to intensify. An analysis of the training loss curve (Figure 5) revealed that this deterioration coincided precisely with the onset of loss convergence. The same behavior was observed for the 3072 × 2048 resolution, although the degradation began earlier, after Epoch 10. This analysis suggested that later checkpoints could represent overfitted models with less meaningful semantic attention relationships. Because of that, we selected the checkpoints from Epoch 15 (for the 2048 × 2048 model) and Epoch 10 (for the 3072 × 2048 model) for all inference tasks presented in Table 2.

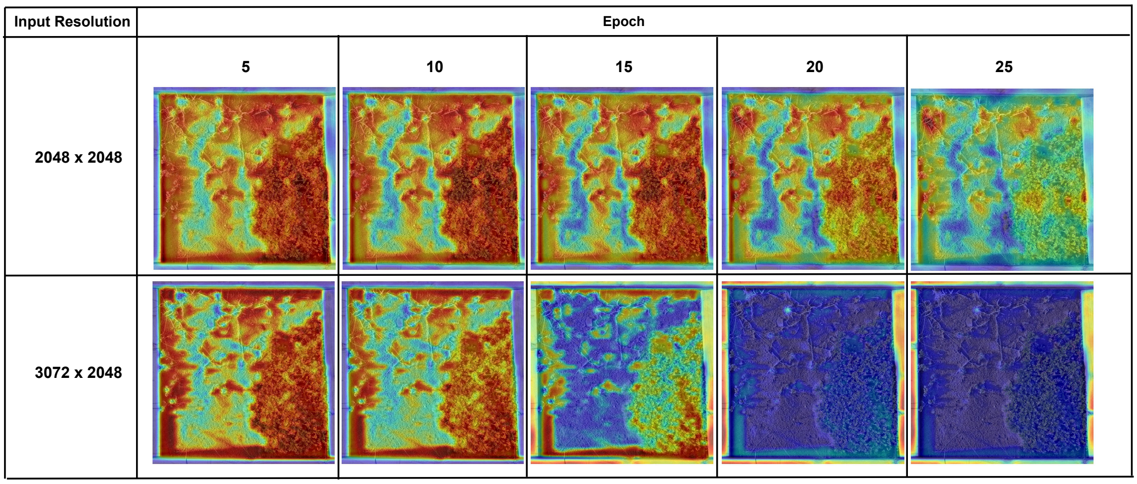

Figure 4: Illustration of attention map degradation over several training epochs. While initial epochs exhibit focused attention on plant regions, later epochs show a diffusion of attention and an incorrect intensification on background areas.

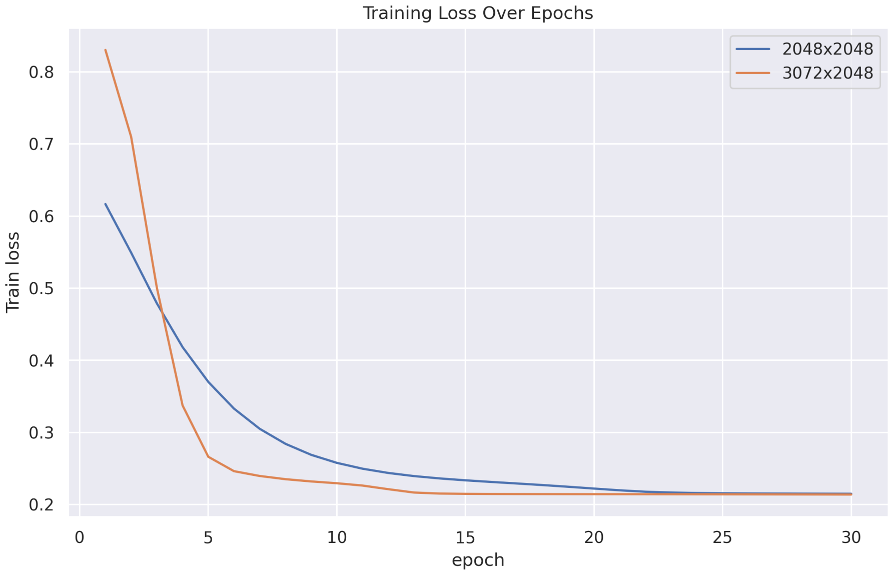

Figure 5: Training loss curve over epochs. The onset of loss convergence (plateau) aligns with the observed degradation in attention map quality, suggesting the model began to overfit to the reconstruction task.

In other words, we observed that although the model initially learns correct semantical relations, focusing its attention on plant features to reconstruct the class prototypes, as training progresses, it collapses by focusing on semantically irrelevant features, such as the quadrat frame. Our hypothesis is that this model collapse is somewhat expected, specially considering the objective of reconstructing a constant target matrix.

A possible conclusion is due a divergence between the optimization of an explicit reconstruction objective and the implicit goal of semantic segmentation, which without the appropriate regularization leads to the trivial solution of focusing on semantically irrelevant features. As a matter of fact, we anticipated that the objective of reconstructing a constant target matrix could lead to the straightforward trivial solution of the model ignoring the inputs and producing the outputs that perfectly reconstructs the target matrix. We believe that this does not occurs in an early stage of training, due to the prior information added by replacing the original patch-embedding layer with the pre-trained DINOv2 model. Despite of that, the use of earlier training checkpoints is essential, as an explicit regularization was not provided.

5. Conclusion

In this paper, we introduced a decrease-and-conquer strategy for the PlantCLEF 2025 challenge, where a Vision Transformer (ViT) was trained on a proxy task to perform segmentation previous to plant species identification. This strategy demonstrated competitive effectiveness, securing a 5th place on the private leaderboard of the challenge, with a final F_1-score of 0.33331. Our experiments confirmed that providing local context by assembling adjacent patches for classification is a crucial and highly effective strategy, significantly outperforming direct patch-wise classification. Given the narrow performance gap to top-ranking methods in the leaderboard, we are confident the proposed approach represents a promising direction for state-of-the-art multi-label classification in complex ecological imagery through domain-adaptation with further refinements holding the potential to achieve superior results.

Acknowledgments

This work was partially supported by UEFS-AUXPPG 2023/2024/2025, CAPES-PROAP 2023/2024/2025, CAPES grant 88887.159255/2025-00 and 88887.594676/2020-00 and UEFS FINAPESQ (grant 047/2023).

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4o in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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, CEUR-WS.org, 2024.

[2] 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, et al., 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.

[3] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[4] G. Martellucci, H. Goëau, P. Bonnet, F. Vinatier, A. Joly, Overview of PlantCLEF 2025: Multi-species plant identification in vegetation quadrat images, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[5] S. Chulif, H. A. Ishrat, Y. L. Chang, S. H. Lee, Patch-wise inference using pre-trained vision transformers: Neuon submission to plantclef 2024, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024. CEUR-WS.org/Vol-3740/paper-192.pdf.

[6] S. Foy, S. McLoughlin, Utilising dinov2 for domain adaptation in vegetation plot analysis, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024. CEUR-WS.org/Vol-3740/paper-196.pdf.

[7] R. Bao, Y. Sun, Y. Gao, J. Wang, Q. Yang, H. Chen, Z.-H. Mao, Y. Ye, A survey of heterogeneous transfer learning, arXiv preprint arXiv:2310.08459 (2023). URL: http://arxiv.org/abs/2310.08459. doi:10.48550/arXiv.2310.08459, arXiv:2310.08459 [cs].

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9650–9660.

[9] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (cnn) in vegetation remote sensing, ISPRS journal of photogrammetry and remote sensing 173 (2021) 24–49.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. V. 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. Jégou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[11] 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, https://doi.org/10.5281/zenodo.10848263, 2024. doi:10.5281/zenodo.10848263.

[12] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15619–15629.