

# Tile-Based ViT Inference with Visual-Cluster Priors for Zero-Shot Multi-Species Plant Identification

Murilo Gustineli<sup>1,\*</sup>, Anthony Miyaguchi<sup>1,\*</sup>, Adrian Cheung<sup>1</sup> and Divyansh Khattak<sup>1</sup>

<sup>1</sup>*Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332*

## Abstract

We describe DS@GT’s second-place solution to the PlantCLEF 2025 challenge on multi-species plant identification in vegetation quadrat images. Our pipeline combines (i) a fine-tuned Vision Transformer ViTD2PC24All for patch-level inference, (ii) a  $4 \times 4$  tiling strategy that aligns patch size with the network’s  $518 \times 518$  receptive field, and (iii) domain-prior adaptation through PaCMAP + K-Means visual clustering and geolocation filtering. Tile predictions are aggregated by majority vote and re-weighted with cluster-specific Bayesian priors, yielding a macro-averaged F1 of 0.348 (private leaderboard) while requiring no additional training. All code, configuration files, and reproducibility scripts are publicly available at [github.com/dsgt-arc/plantclef-2025](https://github.com/dsgt-arc/plantclef-2025).

## Keywords

Computer Vision, Vision Transformers, Information Retrieval, Transfer Learning, CEUR-WS

## 1. Introduction

The PlantCLEF task [1] within the LifeCLEF lab [2] asks competitors to identify all plant species present in high-resolution ( $\approx 2000\text{px}$ ) images of  $50 \times 50\text{cm}$  vegetative quadrat frames placed on the ground to determine a specific area for sampling plant species. Recent advancements in deep learning and collaborative data platforms have created unprecedented opportunities to automate species identification across diverse ecological contexts. The main challenges of the task lie in the domain-shift between **single-label training images** and **multi-label test images**, an extreme class imbalance with over 800 species in the test set and 7,806 in the training set, the size of the training set with 1.4 million images totaling 281GB, and the high intra-class variation across growth stages, organ types, and environmental contexts. To address these challenges, we adopt a transfer-learning strategy built on the ViTD2PC24All backbone—a Vision Transformer that was first self-supervised with DINOv2 and later fine-tuned by the PlantCLEF 2025 organizers on the 2024 single-label dataset. In this work, we investigate how far a publicly released, PlantCLEF-fine-tuned ViT can go on the 2025 multi-label task using zero-shot inference enhanced by tile-based classification, geospatial filtering, and visual-cluster Bayesian priors.

## 2. Related Work

There were a total of seven teams participating in the PlantCLEF 2024 challenge [3], in which three of those shared their solutions as working note papers. Most participants leveraged the fine-tuned ViT provided by the organizers, as training models from scratch using the 1.4M single-label images in the training set poses significant computational challenges. The three main methods used were: (1) Tiling-based inference with false positive reduction (best approach) [4]; (2) Embedding extraction and dimensionality reduction for classification [5]; and (3) Multi-label classification with composite training images [6].

---

*CLEF 2025: Conference and Labs of the Evaluation Forum, September 9-12, 2025, Madrid, Spain*

\*Corresponding author.

✉ murilogustineli@gatech.edu (M. Gustineli); acmiyaguchi@gatech.edu (A. Miyaguchi);acheung@gatech.edu (A. Cheung); dkhattak6@gatech.edu (D. Khattak)

🌐 <https://murilogustineli.com> (M. Gustineli)

>ID 0009-0003-9818-496X (M. Gustineli); 0000-0002-9165-8718 (A. Miyaguchi); 0009-0006-8650-4550 (A. Cheung)

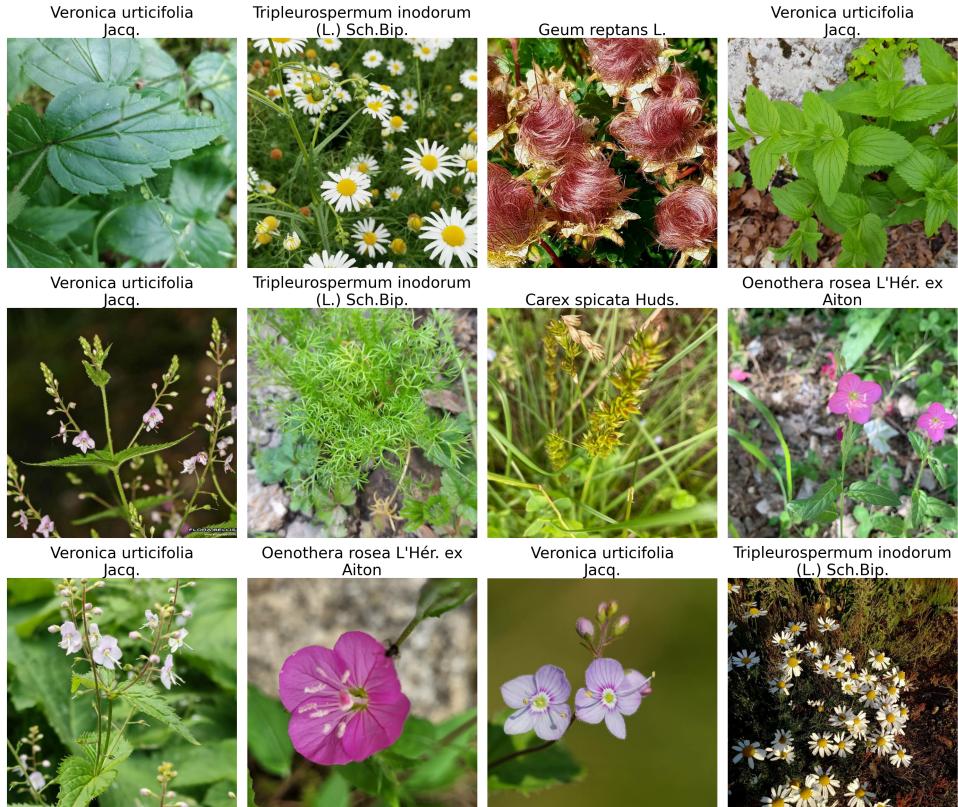


© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

### 3. Datasets and Models

#### 3.1. Training dataset

The training dataset is a subset of the Pl@ntNet [7] training data, composed of single-label plant species, focusing on southwestern Europe (Figure 1). As supplied by the organizers, the dataset comprises 7,806 plant species in 1.4 million images, totaling 281GB (Table 1). The high-resolution images have 800 pixels on their longest side, allowing the use of classification models that can handle large resolution inputs and facilitating the prediction of small plants in large vegetative plots. The images were organized into subfolders by class (i.e., species) and split into predefined train/validation/test sets to facilitate the training of classification models.



**Figure 1:** Single-label training images displaying the following six species: *Veronica urticifolia* Jacq., *Tripleurospermum inodorum* (L.) Sch.Bip., *Geum reptans* L., *Carex spicata* Huds., *Oenothera rosea* L'Hér. ex Aiton, *Lamium bifidum* Cirillo.

**Table 1**

PlantCLEF 20204 [3] dataset overview. The dataset was split into **train/validation/test** sets to facilitate the training of classification models on the individual plant species. Note that the single-label test set used for model training differs from the challenge test set, which contains large multi-label images.

| Datasets | Images    | Observations | Species | Genera |
|----------|-----------|--------------|---------|--------|
| All      | 1,408,033 | 1,151,904    | 7,806   | 1,446  |
| Train    | 1,308,899 | 1,052,927    | 7,806   | 1,446  |
| Val      | 51,194    | 51,045       | 6,670   | 1,415  |
| Test     | 47,940    | 47,932       | 5,912   | 1,375  |

### 3.2. Test dataset

The test set features image quadrats of many floristic environments, emphasizing Pyrenean and Mediterranean flora. All datasets are curated by experts and include a total of 2,105 high-resolution quadrat images (Figure 2). The shooting protocols can differ considerably depending on the context, with variations such as using wooden frames or measuring tape to outline the plot, or capturing images from angles that may not be perfectly perpendicular to the ground due to the site's slope. Furthermore, image quality can fluctuate based on weather conditions, leading to factors like pronounced shadows, blurred areas, and other visual inconsistencies.



**Figure 2:** Subset of twelve test set images showcasing the significant domain shift between different quadrats.

### 3.3. Fine-tuned models

The ViTD2 and ViTD2PC24 models are Vision Transformers (ViTs) pretrained using the DINOv2 Self-Supervised Learning (SSL) approach on the LVD-142M dataset, which contains 142 million images [8]. These models were fine-tuned on the PlantCLEF 2024 dataset to address plant species identification [3]. The original ViTD2 model serves as the backbone, pretrained with DINOv2 without the classifier head, and was not further fine-tuned on PlantCLEF data. This model is mainly used for extracting general image embeddings. The ViTD2PC24 models, however, build on top of the backbone with additional supervised training tailored for plant classification.

To simplify their naming, ViTD2PC24OC refers to the version where only the classifier head was fine-tuned, while ViTD2PC24All refers to the model where both the backbone and classifier head were fine-tuned. The models were made available to participants to facilitate their experiments, particularly those with limited computational resources, and played a key role in developing solutions for the PlantCLEF 2024 challenge. We exclusively utilized the **ViTD2PC24All** model, as it was more effective in extracting richer embedding representations and achieving higher classification scores as compared to its counterpart.

### 3.4. Evaluation metric

The task is evaluated using the *Macro-Averaged F1 score per sample*, which provides a balance between precision and recall for multi-label classification. We reproduce the formula for completeness (Eq. 1–2). The goal is to predict the presence of one or more plant species in high-resolution quadrat images. This evaluation metric takes the average of the F1 scores computed individually for each vegetation plot. The F1 score for each quadrat image is calculated as the harmonic mean of precision and recall:

$$F1_j = \frac{2 \cdot \text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j} \quad (1)$$

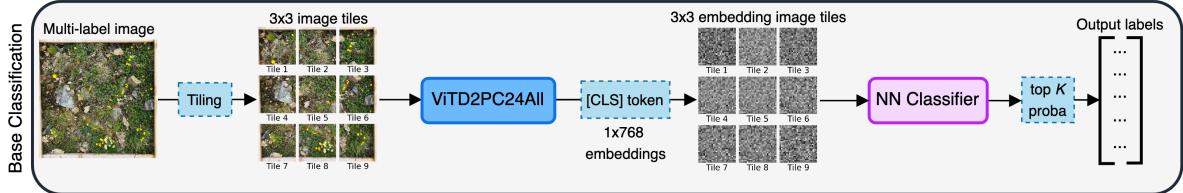
Where  $\text{Precision}_j = \frac{TP_j}{TP_j + FP_j}$  and  $\text{Recall}_j = \frac{TP_j}{TP_j + FN_j}$  with  $TP_j$ ,  $FP_j$ , and  $FN_j$  denoting true positive, false positive, and false negatives for image  $j$ . To ensure fairness across ecological regions (transects), macro-averaging is applied in a two-step process:

1. F1 scores are averaged across all quadrat images within each transect.
2. These per-transect averages are then averaged across all transects to yield the final score:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^N \left( \frac{1}{T_i} \sum_{j=1}^{T_i} F1^j \right) \quad (2)$$

Where  $N$  is the number of transects,  $T_i$  is the number of quadrats in transect  $i$ , and  $F1^j$  is the F1 score of image  $j$ .

## 4. Methodology



**Figure 3:** Overview of our proposed transfer learning method. We perform a tiling approach on the test set, classify each tile using the ViTD2PC24All model, and aggregate the results by selecting the top-K species based on their frequency count.

Our approach leverages the embedding space learned by the ViTD2PC24All model as a generalized feature representation of images, which is used for classification (Figure 3). ViTD2PC24All learns robust feature representations by processing images as sequences of fixed-size patch tokens with an additional [CLS] token for classification tasks [9]. These tokens serve as low-dimensional representations of the image patches, similar to words in a phrase for language models. The main challenge lies in overcoming the domain shift between single-label training images and multi-label test images. We perform a tiling approach, dividing each test image into a grid of  $N \times N$  tiles. Our code is available at [github.com/dsgt-arc/plantclef-2025](https://github.com/dsgt-arc/plantclef-2025).

### 4.1. Tiled inference

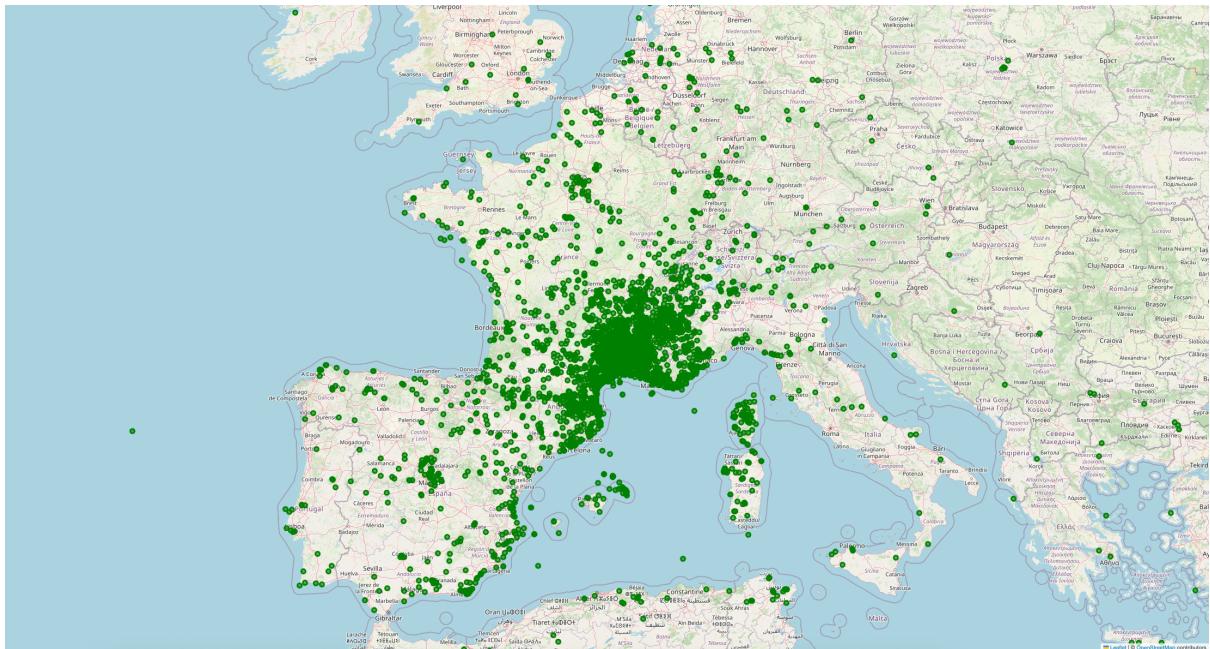
We use the fine-tuned ViTD2PC24All model as a baseline classifier. To bridge the gap between single-label training images and multi-label test images, we perform a tiling-based classification strategy on the multi-label test images. We begin by leveraging the fine-tuned classification head at inference time, where each high-resolution test image is partitioned into a fixed-size grid of non-overlapping tiles (e.g.,

$3 \times 3$  or  $4 \times 4$ ). Each tile is independently classified using the fine-tuned ViT model. This method enables localized prediction within the image and helps with the mismatch between the global multi-label test images and the local single-label learning context of the model. To produce image-level predictions, we aggregate tile-level predictions across the image and rank species based on their frequency of occurrence among the top-K predictions per tile. The most frequent predicted species are selected as the final image-level labels.

We empirically determined the optimal grid size to be  $4 \times 4$  tiles. That aligns with the input resolution of the fine-tuned model ViTD2PC24All, which is based on the `timm/vit_large_patch14_dinov2.1vd142m` architecture and expects images resized to  $518 \times 518$ . A typical high-resolution quadrat image has a width of approximately 2000 pixels, partitioning in  $4 \times 4$  tiles yields sub-images of roughly 500 pixels per side, closely matching the model’s expected input size. Using smaller or larger grid sizes leads to image downscaling or upscaling during preprocessing, resulting in degradation of feature quality and decreasing classification performance.

## 4.2. Geolocation filtering

To address the domain-shift problem between training and test data – where the training data has 7,806 species and the test data has roughly 800 species from Southwestern Europe – we used geolocation metadata from the training images to narrow down likely species candidates. We defined a reference point in Southern France (44°N, 4°E) and computed the squared Euclidean distance between this point and each species observation. For each species, we selected the closest known geotagged observation. We then filtered species whose nearest observation falls within the geographic boundaries of countries relevant to the test set (France, Spain, Italy, and Switzerland), as shown in Figure 4. This geospatial filtering reduced the search space from thousands of global species to a plausible subset of 4,981 species, improving prediction relevance and mitigating the long-tailed class distribution (Table 3).



**Figure 4:** Geolocation of plant species based on their latitude and longitude metadata. Using a reference point (latitude=44, longitude=4) located in the Southwestern European region, we computed the squared Euclidean distance between each plant species geolocation and the reference point, rank observations per species by proximity to the reference point, and selected the closest point (rank = 1) for each species.

### 4.3. Visual-Cluster Bayesian prior adaptation

To address the domain shift and class imbalance between training and test sets, we introduce a strategy to **prioritize likely species** in the test set. We grouped test images by their corresponding region identifiers, which are present in the quadrat\_id field. These identifiers represent the origin of the vegetation plots and were used to cluster images based on their location. We defined 13 regions based on the test set quadrat\_id naming format and assigned each image to its respective region (Table 2).

**Table 2**

Grouping of the 13 regions based on the test set quadrat\_id naming patterns, including the number of images per region and their corresponding dominant K-Means cluster derived from visual embedding similarity.

| Region         | Image Count | K-Means Cluster |
|----------------|-------------|-----------------|
| CBN-PdIC       | 816         | 2               |
| CBN-Pla        | 628         | 3               |
| GUARDEN-CBNMed | 165         | 1               |
| RNNB           | 141         | 1               |
| LISAH-BOU      | 82          | 1               |
| OPTMix         | 78          | 1               |
| LISAH-BVD      | 76          | 1               |
| GUARDEN-AMB    | 36          | 1               |
| LISAH-PEC      | 35          | 1               |
| CBN-can        | 30          | 2               |
| LISAH-JAS      | 15          | 1               |
| CBN-Pyr        | 2           | 1               |
| 2024-CEV3      | 1           | 1               |

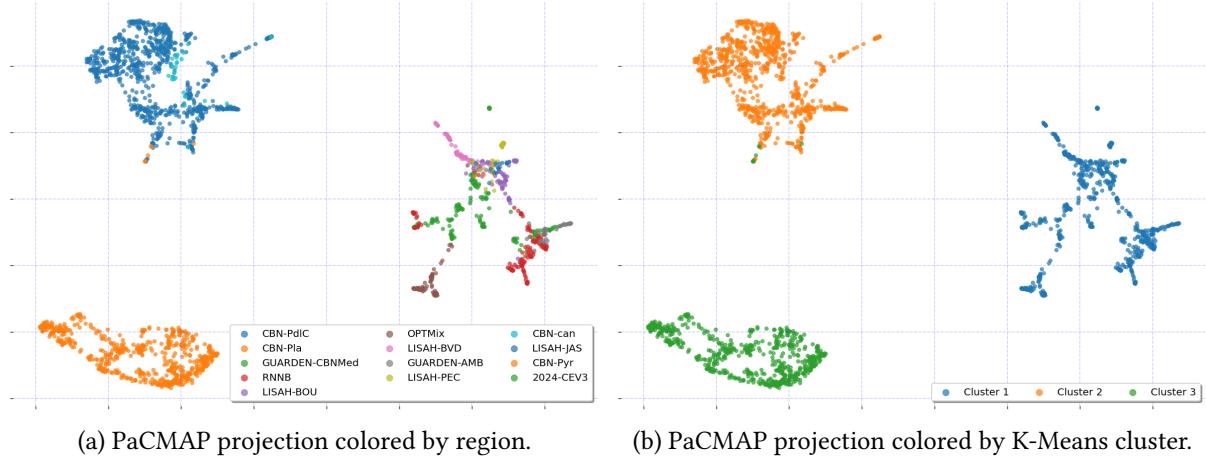
We utilized the ViTD2PC24All model to extract the [CLS] token embeddings of the test set images and projected the embeddings into two dimensions using PaCMAp [10] to visually explore their structure. (Figure 5a). By coloring points based on their region labels, we observed that quadrats from the same region tend to cluster together, revealing three well-defined clusters. This suggests that certain geographic or ecological similarities—such as altitude or vegetation type—may drive visual similarity among regions, which can be leveraged to improve classification performance under domain shift.

After visualizing the PaCMAp embeddings, we applied K-Means clustering to group the quadrats into three unsupervised clusters based on visual similarity. We then assigned each region to its dominant cluster by identifying where the majority of its images were grouped (Figure 5b). This provided a meaningful stratification of the test set, allowing us to model regional variation in species composition and incorporate cluster-specific priors for downstream classification. The final region distribution in the test set is summarized in Table 2.

We hypothesize that the three dominant K-Means clusters represent different altitude levels where the test set images were taken:

- **Cluster 1: “Coastal and Salt-Tolerant Plants”** – Salt-tolerant and drought-resistant, coastal dunes, salt marshes, and sandy habitats.
- **Custer 2: “Alpine and Sub-alpine Specialists”** – Hardy, low-growing plants adapted to cold, high-altitude environments (alpine meadows and rocky slopes).
- **Cluster 3: “Alpine Grasses and Ferns”** – Resilient grasses and ferns, this cluster thrives in alpine grasslands and sub-alpine zones, often in rocky or well-drained soils.

To assign the descriptive labels in the bullet list, we first identified the most frequent species within each visual cluster by averaging the per-image class-probability vectors. The top species for every cluster were then given to ChatGPT, which returned concise ecological summaries that we adopted as cluster names.



**Figure 5:** PaCMAP projections of the test set [CLS] token embeddings. 5a Two-dimensional projection of test image embeddings using PaCMAP, with each point colored by its region label (13 regions total). The plot reveals that quadrats from the same region tend to cluster together, indicating strong visual similarity and potential ecological coherence within regions. 5b PaCMAP projection colored by the dominant K-Means cluster assigned to each region. After clustering the embeddings into three groups, we assigned each region to the cluster where most of its quadrats reside. This unsupervised stratification highlights coherent visual groupings that align with ecological or geographic distinctions.

We subsequently incorporated region-specific Bayesian priors into the tile-based inference pipeline. The PaCMAP + K-Means step yields, for every cluster  $c$ , an empirical prior distribution  $P(y|c)$  obtained by averaging the model’s predicted probability vectors across all images in that cluster. During inference, we re-weight each tile’s class probabilities by this prior, increasing the bias towards species that are visually and geographically likely for that cluster. This approach helped narrow down the candidate species space for each test image and improve robustness to underrepresented classes. This is particularly important given the shift from single-label training images to multi-label plot images in the test set.

## 5. Results

We evaluated our approaches on the hidden test set provided on the Kaggle competition leaderboard. Table 3 presents an ablation study comparing different variants of our classification pipeline. The naive baseline—selecting the top-K most frequent species in the training data—achieved negligible performance across both public and private leaderboard splits. Introducing the fine-tuned ViT-based classifier without tiling improved results only marginally, highlighting the difficulty of processing high-resolution vegetation plots holistically. Tiling test images into a  $4 \times 4$  grid (matching the input resolution of the fine-tuned ViTD2PC24All model) led to a substantial performance gain. Specifically, selecting the top-9 predictions per tile yielded a private leaderboard F1 score of 0.3442, representing a strong baseline for multi-label classification using patch-level aggregation.

To further mitigate the domain shift and long-tailed class distribution challenges, we incorporated two complementary strategies: **(1)** cluster-based priors derived from PaCMAP+K-Means embeddings, and **(2)** spatial filtering using geolocation priors. Applying cluster-specific Bayesian reweighting improved the private leaderboard score to 0.3483. Alternatively, geolocation-based filtering—removing species unlikely to occur near the test region—resulted in a private score of 0.3449 and the highest public leaderboard score of 0.3160. These findings demonstrate that both spatially-aware inference and prior reweighting provide valuable regularization, yielding competitive performance without modifying the underlying model architecture.

**Table 3**Ablation study of our different approaches.  $4 \times 4$  is the tiling size.

| Method              | Top-K Predictions | Tiles      | Private (%)    | Public (%)     |
|---------------------|-------------------|------------|----------------|----------------|
| Naive baseline      | top-5             | -          | 0.00422        | 0.00736        |
| Naive baseline      | top-10            | -          | 0.00776        | 0.00466        |
| Naive baseline      | top-25            | -          | 0.00571        | 0.00440        |
| ViT                 | top-20            | -          | 0.00633        | 0.01157        |
| ViT                 | top-20            | 4x4        | 0.26313        | 0.25239        |
| ViT                 | top-12            | 4x4        | 0.32667        | 0.30203        |
| ViT                 | top-10            | 4x4        | 0.33926        | 0.30906        |
| ViT                 | top-9             | 4x4        | 0.34420        | 0.30810        |
| <b>ViT + GEO</b>    | <b>top-10</b>     | <b>4x4</b> | <b>0.34489</b> | <b>0.31600</b> |
| <b>ViT + PRIORS</b> | <b>top-9</b>      | <b>4x4</b> | <b>0.34834</b> | <b>0.29293</b> |

## 6. Discussion

Our ablation study shows that simply forwarding the full-resolution quadrat image through the fine-tuned ViT barely surpasses a frequency-based baseline ( $F1 \approx 0.006$ ; Table 3). Once the image is tiled into  $4 \times 4$  sub-images whose side length roughly matches the 518px receptive field of ViTD2PC24All, macro-F1 jumps by two orders of magnitude (0.34). This finding echoes recent work on high-resolution ViTs, where tile- or window-based inference is consistently reported as the most reliable way to preserve fine-grained cues without exceeding GPU memory limits [11, 12].

Adding visual-cluster Bayesian priors yields a further +0.004 improvement. By averaging the model’s own probability vectors inside PaCMAP + K-Means clusters, we obtain an empirical prior  $P(y|c)$  that captures region-specific floristic bias; re-weighting tile probabilities with this prior is related to the “context-conditioned” re-ranking used in training-free zero-shot pipelines [13] and to Bayesian reweighting strategies explored for low-shot recognition [14, 15]. The alternative geolocation filter achieves the best public-leaderboard score (0.316) but only matches the prior-adapted model privately. This suggests that purely spatial heuristics over-prune plausible long-tail species that remain detectable when appearance cues and cluster priors are combined.

### 6.1. Limitations and future work

While our training-free pipeline demonstrates that tile-based ViT inference plus cluster-aware Bayesian priors can reach competitive accuracy, several limitations remain that shape our next research steps. First, the backbone we rely on is already fine-tuned on single-label PlantCLEF 2024 data, so our “zero-shot” claim holds only for the 2025 task; extending this strategy to domains that lack such a pre-fine-tuned model remains an open challenge. Second, non-overlapping square tiles risk bisecting plants at tile boundaries; sliding-window inference [16], learned token merging [17], or adaptive receptive-field [18] methods such as ViT-AR [19] could recover boundary context without prohibitive compute. Finally, a lightweight round of self-training on high-confidence tile pseudo-labels, or ensembling with CNN backbones that capture texture cues absent in ViTs [20], could raise the current 0.348 macro-F1 ceiling while keeping compute modest.

## 7. Conclusion

We presented a fully training-free pipeline that combines tile-based ViT inference, geolocation filtering, and visual-cluster Bayesian priors to tackle the PlantCLEF 2025 multi-label plant identification challenge. Starting from a publicly released, PlantCLEF-fine-tuned ViT, our method boosts macro-F1 from 0.006 to 0.348 on the private leaderboard—good for second place—without updating a single model weight. The study confirms three take-aways: (1) matching the inference tile scale to the ViT’s receptive field

is critical for high-resolution plant imagery; (2) unsupervised visual clustering provides a cheap yet powerful prior that complements spatial heuristics; and (3) zero-training adaptation is competitive when domain-specific compute or labels are scarce. All code and artifacts are open-sourced to support follow-up research on even more challenging biodiversity datasets.

## Acknowledgements

We thank the Data Science at Georgia Tech (DS@GT) CLEF competition group for their support. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA [21].

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools for writing this working note paper.

## References

- [1] 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.
- [2] 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 (CLEF), Springer, 2025.
- [3] 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.
- [4] S. Foy, S. McLoughlin, Utilising dinov2 for domain adaptation in vegetation plot analysis, in: Conference and Labs of the Evaluation Forum, 2024.
- [5] M. Gustineli, A. Miyaguchi, I. Stalter, Multi-label plant species classification with self-supervised vision transformers, arXiv preprint arXiv:2407.06298 (2024).
- [6] 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: Conference and Labs of the Evaluation Forum, 2024.
- [7] H. Goëau, P. Bonnet, A. Joly, V. Bakić, J. Barbe, I. Yahiaoui, S. Selmi, J. Carré, D. Barthélémy, N. Boujemaa, et al., Pl@ntnet mobile app, in: Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 423–424.
- [8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
- [9] L. Wu, W. Zhang, T. Jiang, W. Yang, X. Jin, W. Zeng, [cls] token is all you need for zero-shot semantic segmentation, arXiv preprint arXiv:2304.06212 (2023).
- [10] Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik, Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization, Journal of Machine Learning Research 22 (2021) 1–73. URL: <http://jmlr.org/papers/v22/20-1061.html>.
- [11] V. Leroy, J. Revaud, T. Lucas, P. Weinzaepfel, Win-win: Training high-resolution vision transformers from two windows, arXiv preprint arXiv:2310.00632 (2023).

- [12] Z. Li, S. F. Bhat, P. Wonka, Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10016–10025.
- [13] B. An, S. Zhu, M.-A. Panaiteescu-Liess, C. K. Mummadri, F. Huang, Perceptionclip: Visual classification by inferring and conditioning on contexts, arXiv preprint arXiv:2308.01313 (2023).
- [14] Y. Miao, Y. Lei, F. Zhou, Z. Deng, Bayesian exploration of pre-trained models for low-shot image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23849–23859.
- [15] Z. Ji, X. Chai, Y. Yu, Z. Zhang, Reweighting and information-guidance networks for few-shot learning, Neurocomputing 423 (2021) 13–23.
- [16] A. Dede, H. Nunoo-Mensah, E. T. Tchao, A. S. Agbemenu, P. E. Adjei, F. A. Acheampong, J. J. Kponyo, Deep learning for efficient high-resolution image processing: A systematic review, Intelligent Systems with Applications (2025) 200505.
- [17] Y. Niu, Z. Song, Q. Luo, G. Chen, M. Ma, F. Li, Atmformer: An adaptive token merging vision transformer for remote sensing image scene classification, Remote Sensing 17 (2025) 660.
- [18] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, J. Gall, Adaptive token sampling for efficient vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 396–414.
- [19] Q. Fan, Q. You, X. Han, Y. Liu, Y. Tao, H. Huang, R. He, H. Yang, Vitar: Vision transformer with any resolution, arXiv preprint arXiv:2403.18361 (2024).
- [20] W. Hussain, M. F. Mushtaq, M. Shahroz, U. Akram, E. S. Ghith, M. Tlija, T.-h. Kim, I. Ashraf, Ensemble genetic and cnn model-based image classification by enhancing hyperparameter tuning, Scientific Reports 15 (2025) 1003.
- [21] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: <http://www.pace.gatech.edu>.