Multi-label Plant Species Classification with Vision Transformers at PlantCLEF 2025

2nd place
PlantCLEF2025 @ LifeCLEF & CVPR-FGVC

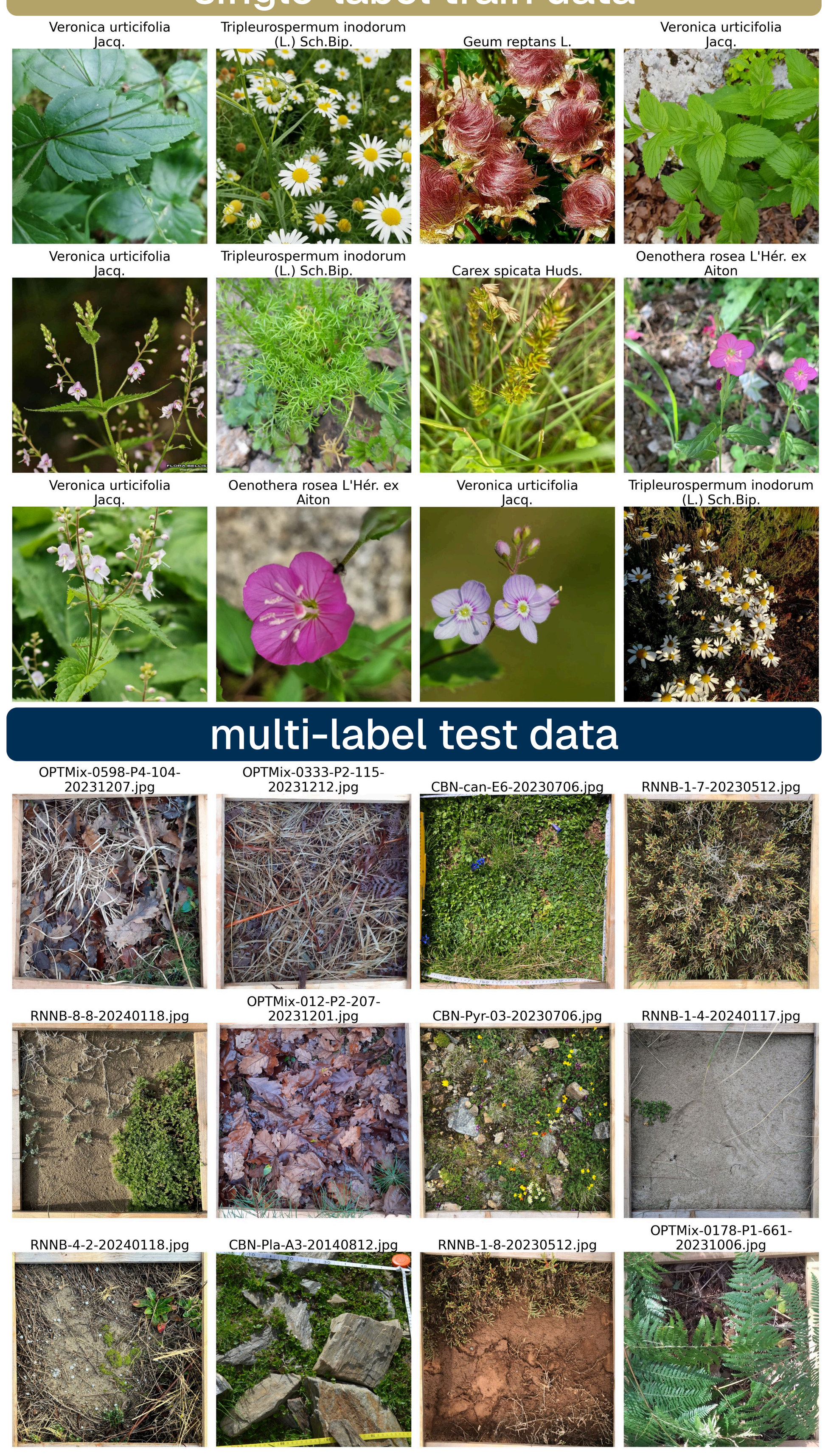
Team DS@GT-ARC: Murilo Gustineli, Anthony Miyaguchi, Adrian Cheung, Divyansh Khattak

kaggle GitHub

Domain Shift

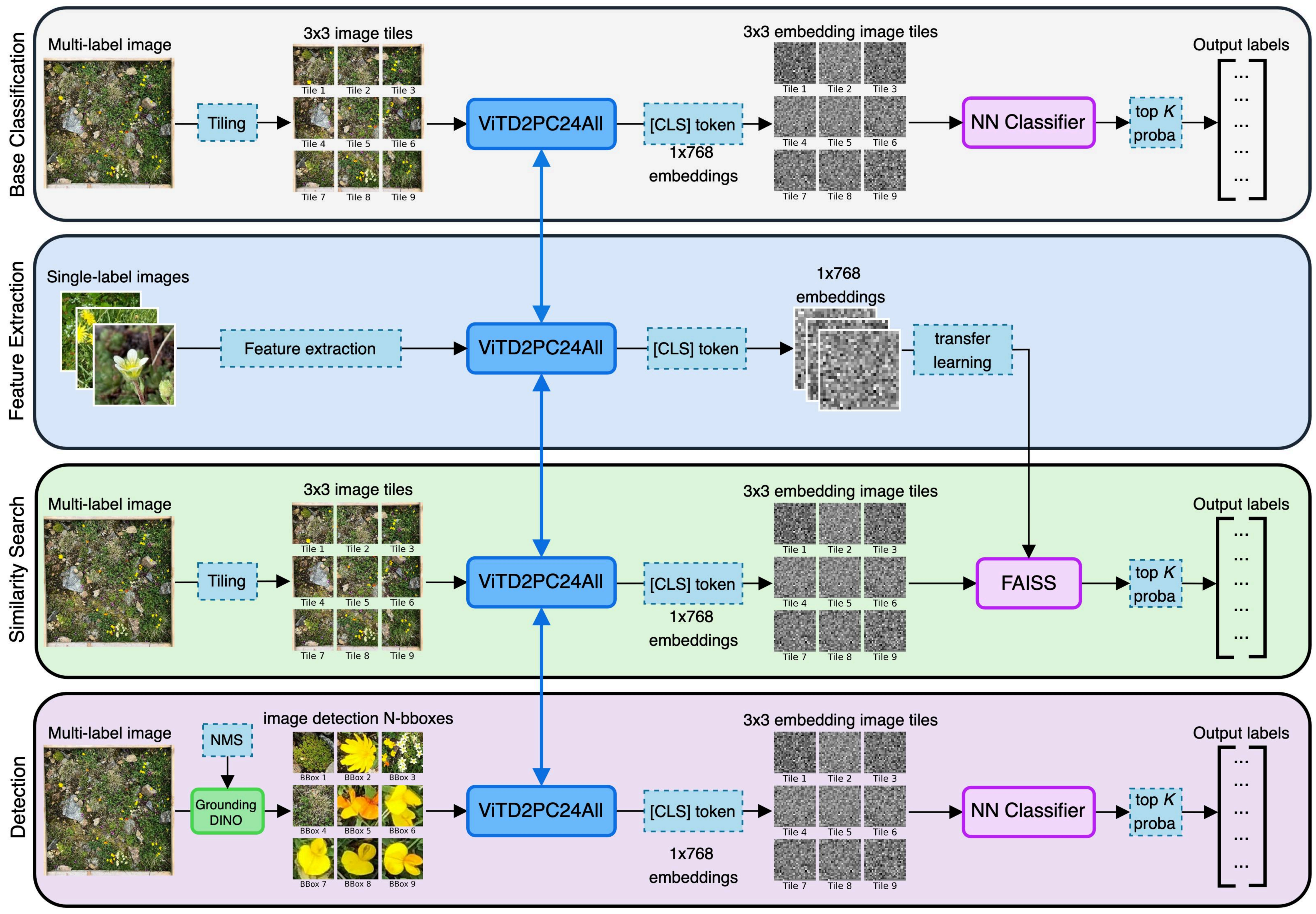
single-label train data

multi-label test data



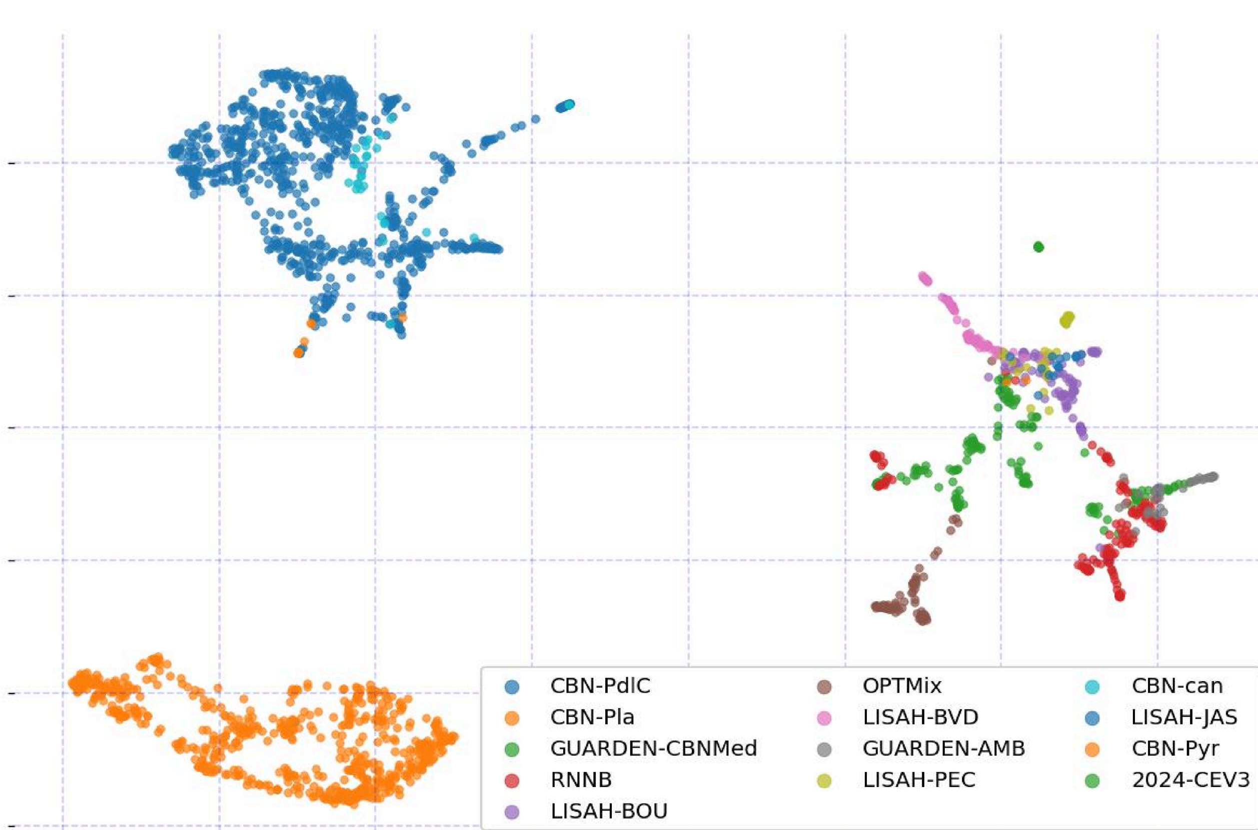
Pipeline

We leverage the embedding space learned by the **ViTD2PC24All** model as a generalized feature representation of images, which are used for classification

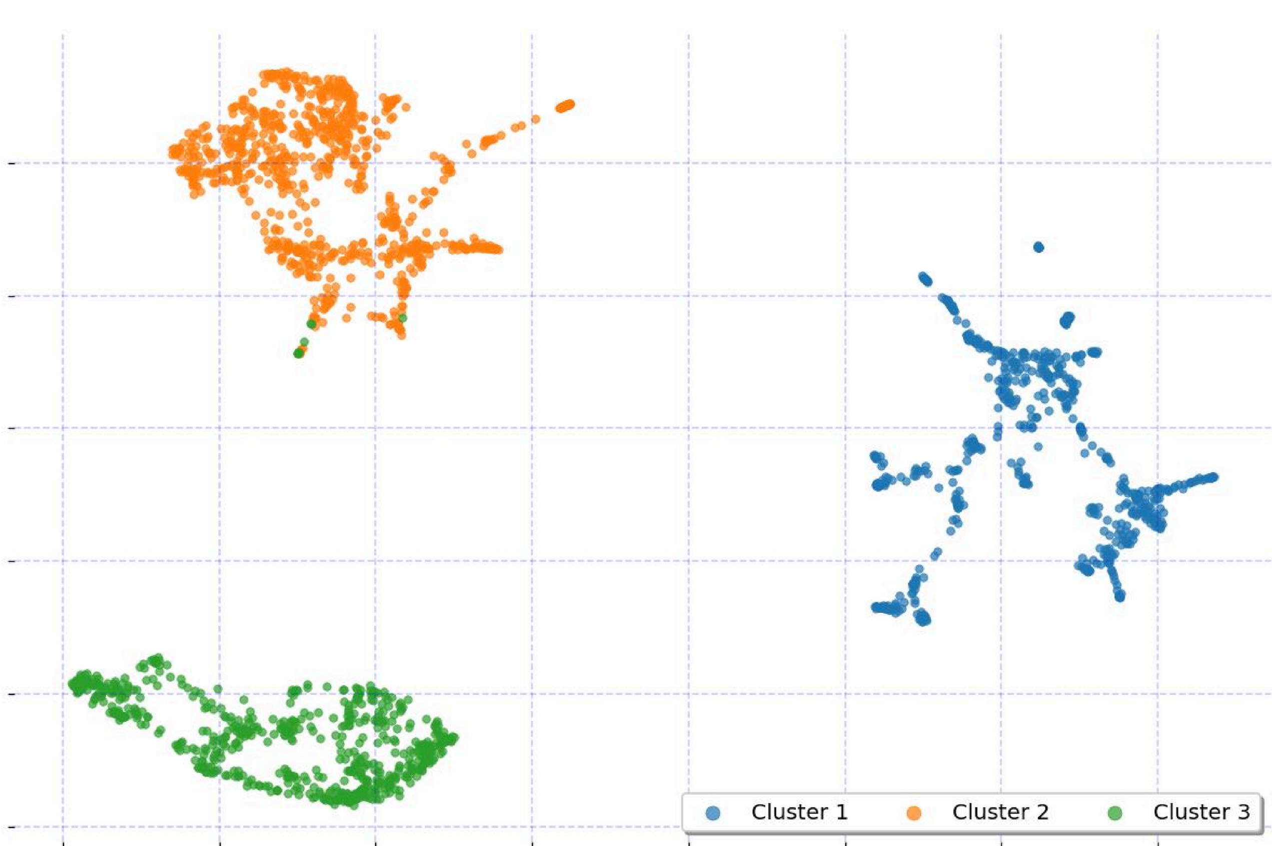


Prior Adaptation

PaCMAP embedding projection colored by region



PaCMAP embedding projection colored by KMeans cluster



Cluster-wise Prior Estimation

$$\pi^{(k)} = \frac{1}{|\mathcal{X}_k|} \sum_{x_i \in \mathcal{X}_k} f(x_i)$$

where:

- $\pi^{(k)} \in \mathbb{R}^C$ is the average softmax probability vector for cluster k
- $f(x_i)$ is the softmax output from the fine-tuned DINOv2 model
- $\mathcal{X}_k = \{x_i | c(i) = k\}$ is the set of tiles in cluster k

Final Prediction with Cluster-Aware Prior

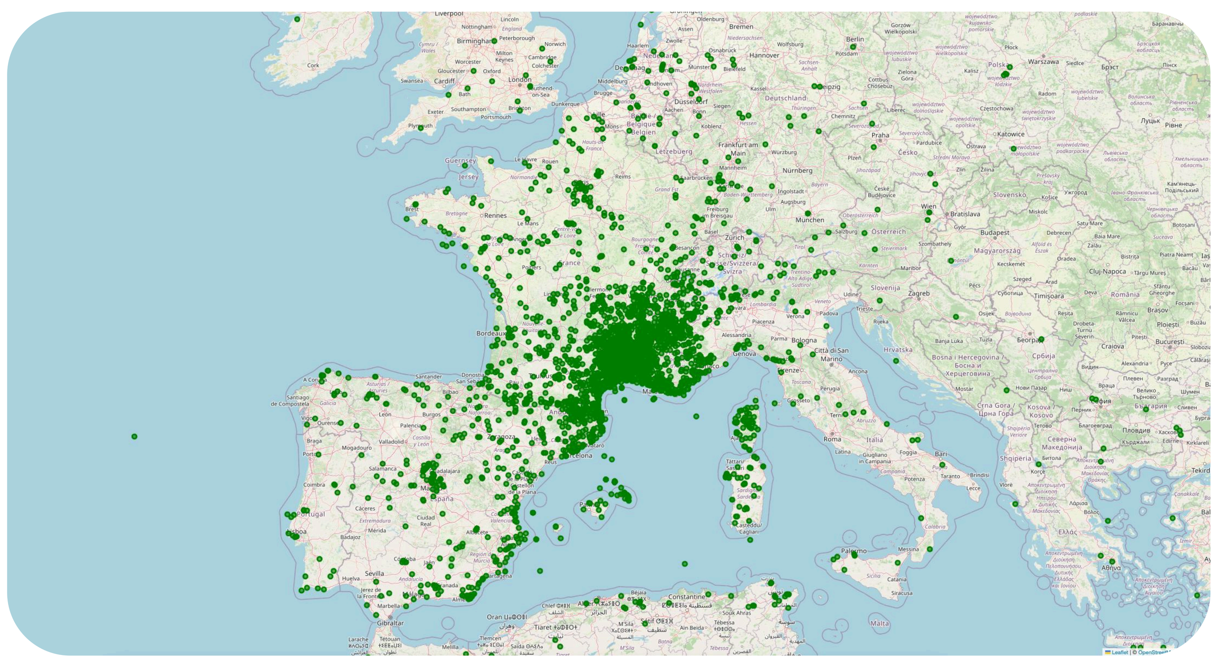
$$\hat{y}_i = \frac{f(x_i) \odot \pi^{(c(i))}}{\sum_{j=1}^{C} \left[f(x_i) \odot \pi^{(c(i))} \right]_j}$$

where:

- $\odot$ denotes element-wise multiplication
- $c(i)$ is the cluster assignment for tile x_i
- $\hat{y}_i \in \mathbb{R}^C$ is the final prediction vector for tile x_i

- **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

Geolocation Filtering



Let each plant observation be defined by a geolocation:

$$x_i = (\phi_i, \lambda_i)$$

where:

- ϕ_i is the latitude of observation i
- λ_i is the longitude of observation i

Let the reference point in Southwestern Europe be:

$$r = (\phi_i, \lambda_i) = (44,\ 4)$$

We compute the **squared Euclidean distance** from each observation to the reference point:

$$d_i = (\phi_i - \phi_0)^2 + (\lambda_i - \lambda_0)^2$$

For each species $s \in S$, let I_s be the set of indices of observations belonging to species s. Then, define:

$$i_s^* = \text{argmin}_{i \in I_s} d_i$$

That is, i_s^* is the observation of species in s that is closest to the reference point

Main Takeaways

Classification method	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
Baseline, top-20, no tiling	0.00633	0.01157
Baseline, top-20, 4x4	0.26313	0.25239
Baseline, top-9, 4x4	0.34420	0.30810
Baseline, top-9, 4x4, PRIOR	**0.34834**	**0.29293**
Baseline, top-10, 4x4	0.33926	0.30906
Baseline, top-10, 4x4, GEOLOCATION	**0.34489**	**0.31600**
Baseline, top-12, 4x4	0.32667	0.30203
Baseline, top-10, Grounding DINO + NMS	0.21083	0.23913

Our Approach

- **Patch-wise inference** bridges the gap between single-label training data and multi-label plot images by tiling test images into a 4x4 grid
- Each tile is independently classified using a ViT fine-tuned on PlantCLEF data, and predictions are aggregated to produce image-level labels
- We empirically matched tile size to the model resolution, reducing input distortion and improving accuracy

Best Solutions

- **Baseline + tiling** significantly boosts performance compared to naive frequency based methods
- **Cluster-based priors** derived from PaCMAP + KMeans embeddings yield the highest private score (0.3483)
- **Geolocation filtering** using spatial metadata yields the best public leaderboard score (0.3160), effectively reducing the long-tailed label space

Future Work

- **Hierarchical taxonomic structure** (genus/family) into the model to improve generalization → fine-tune CLIP model
- Leverage **external metadata** (altitude, soil type) to refine spatial priors
- Explore **contrastive learning** or **multi-view training** to better align single-plant training with multi-label test plots
- Investigate **ensemble methods** combining prior-reweighted, geolocation-filtered, and object-detection-based classifiers