Optimizing a Vision Transformer with Ecological Context for Multi-Label Plant Species Identification*

Nelly Semenova[1,*]

[1]Moscow Pedagogical State University (MPGU University), 1/1 Malaya Pirogovskaya St., Moscow, 119435, Russian Federation

Abstract

This paper presents an ecology-oriented post-processing pipeline that optimizes a pre-trained Vision Transformer (DINOv2) for the PlantCLEF 2025 challenge. The task requires predicting the complete species list in vegetation-plot images (0.25 m², 2-12 MP), whereas the model was trained almost exclusively on single-plant images, which results in a pronounced domain shift. The pipeline comprises (i) multi-scale tiling with test-time augmentation, (ii) artifact down-weighting via zero-shot segmentation, and (iii) ecological correction of prediction scores using Global Biodiversity Information Facility (GBIF) occurrence statistics, seasonal windows and niche similarity derived from Ecological Indicator Values for Europe (EIVE). Without retraining the Vision Transformer on any task-specific data, the public F1-score rises from 21.84 to 38.13 (+16.3 pp) and the private score rises from 20.12 to 33.45, ranking 1st of 38 teams (public leaderboard) and 4th (private). Three consecutive ecological filters contribute approximately +4 pp to this improvement. These results show that ecology-aware post-processing is a reproducible alternative to costly model retraining for multi-species identification.

Keywords

vision transformer, vegetation classification, multi-species identification, ecological niche, test-time augmentation, species co-occurrence

1. Introduction

Automatic identification of *all* plant species in a vegetation-plot image is a challenging multi-label classification task. In the PlantCLEF 2025 challenge participants must predict the complete species list for every top-down photograph of a vegetation plot ($\approx 0.5 \times 0.5$ m, ≈ 0.25 m²) [1, 2].

While mobile apps such as Pl@ntNet and iNaturalist solve the single-label problem (recognizing one plant per close-up image) detecting dozens of overlapping taxa in a single scene is far harder [3]. A pronounced domain shift complicates the task: The training set comprises single-plant close-ups, whereas the test set contains high-resolution (2–12 MP) multi-species plots. Patch-wise processing is currently regarded as the most effective approach. Moreover, the heavy tailed distribution of the European flora (*Gini coefficient* = 0.88; *skewness* = 6.46; *kurtosis* = 55.4, and 45.7% of species have ≤100 records in 2019–2023, whereas the top 1% of species account for 25.6% of all records [4]) means that many rare species are represented by only a few images, which leads to systematic false negatives and positives.

Automating full-species inventories is critical for ecological studies, yet manual annotation is time-consuming and expertise-intensive. The self-supervised Vision Transformer DINOv2 [3] supplies strong generic embeddings, but its accuracy improves markedly when ecological context is exploited. This paper proposes an ecology-oriented post-processing pipeline that enhances a pre-trained ViT without any fine-tuning on new data. The pipeline combines multi-scale tiling with test-time augmentation (TTA), artifact removal via zero-shot segmentation, ecological priors: GBIF range filtering, seasonal phenology windows, Ellenberg indicator scaling from EIVE [5], and species co-occurrence.

The pipeline increases the F1-score from 21.84 to 38.13 on the public leaderboard (public score) and from 20.12 to 33.45 on the private leaderboard (team Webmaking), ranking 1st/38 and 4th, respectively, with fewer than 200 GPU-hours and no additional training (Fig. 1 illustrates the end-to-end inference

CLEF 2025: Conference and Labs of the Evaluation Forum, September 09–12, 2025, Madrid, Spain

*Corresponding author.

nelli.semenova@mail.ru (N. Semenova)

0000-0002-0190-8382 (N. Semenova)

pipeline). These results confirm that lightweight ecology-aware post-processing can substantially boost pre-trained ViTs for large-scale multi-label plant-species identification, and remaining limitations are discussed.

2. Data and Baseline Model

2.1. Pre-trained Vision Transformer

The baseline rests on the checkpoint `dinov2_patch14_reg4_onlyclassifier_then_all` distributed by the organisers on Kaggle [6]. Its backbone is *DINOv2ViT-B*/14 with four learnable register tokens, originally pre-trained on 142 M images [7]. A two-stage supervised phase followed: first, a linear head was fitted on the 1.4 M single-plant pictures of PlantCLEF 2024; afterward the entire network was fine-tuned on the same data. The final classifier therefore produces logits for 7806 European taxa—the exact label set of the 2025 challenge.

All experiments keep these weights frozen. The model expects an input of 518 × 518 px, which coincides with the crop size used in both tiling schemes (Section 3.1). Each window yields a 7806-element logit vector; applying the sigmoid transforms it into class-wise confidences. Only the five highest scores are retained per window to minimize I/O without information loss for later aggregation. No further fine-tuning, domain adaptation or ensembling at the feature level is attempted; every subsequent improvement described in Sections 3–4 operates solely on the fixed predictions of this Vision Transformer.

2.2. Competition data

The official test set comprises 2105 JPEG images taken vertically above vegetation plots of roughly 0.25 m^2. Native resolutions range between 2 MP and 12 MP; the organizers distributed the files exactly as recorded in the field. Each file name encodes a persistent plot identifier followed by the acquisition date in the form *YYYYMMDD* [8]. These two tokens enable later aggregation of images that depict the same location in different years, yet the images themselves are completely unannotated: no geographic coordinates, species labels, masks or bounding boxes accompany the pictures. Every method must therefore infer the full multi-species content from a single high-resolution frame without spatial supervision.

Alongside the test set the organizers provided two additional resources: a 1.4-million image single-plant collection covering 7806 European species, and an archive of 212,782 unlabelled pseudo-quadrat views derived from LUCAS Cover photographs. Neither resource is used in the present study. All experiments are conducted with the competition's pre-trained Vision Transformer, and no further fine-tuning or self-supervised adaptation is performed.

2.3. External ecological context

To supply the ecological background that the image files themselves lack, two public data sets are linked to the baseline predictions. Daily and monthly sighting counts for every target species were extracted from the European section of GBIF [4]; later allow the method to judge whether a taxon suggested by the network is seasonally plausible for the date embedded in the file name. In addition, the five numerical Ellenberg indicator values compiled in the EIVE project (light, temperature, moisture, soil reaction and nitrogen supply) are available for most of the species list. These figures are converted into a quantitative measure of potential niche overlap and will be used to assess the ecological compatibility of the candidate labels produced by the Vision Transformer. A coarse spatial prior derived from a five-kilometre raster of GBIF occurrences was also evaluated but provided no measurable benefit, so seasonal frequency and niche overlap constitute the only contextual signals carried forward into the methodological section that follows.

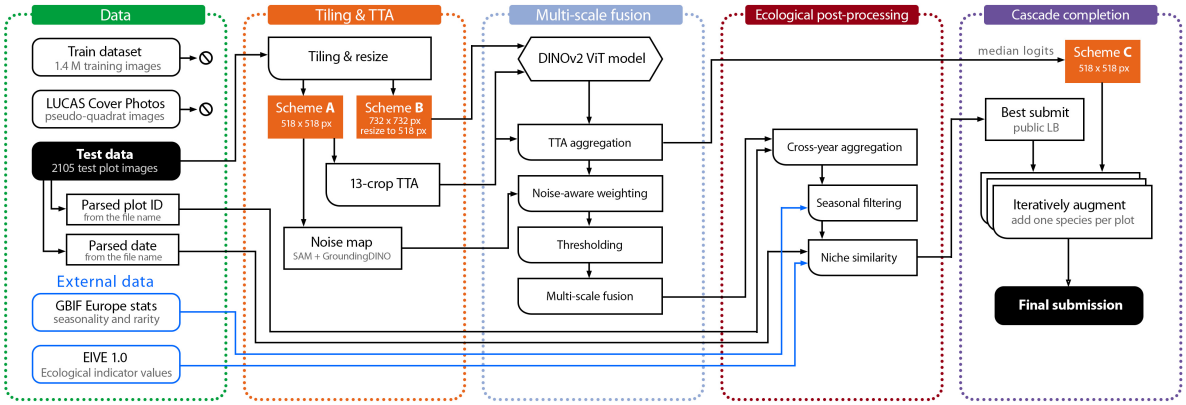

Figure 1: End-to-end inference pipeline for the PlantCLEF 2025 solution. Test images and ecological priors (GBIF statistics, EIVE indicators) enter the system, are tiled into two window sizes with 13-crop TTA, and passed through a DINOv2 Vision Transformer. Window logits are aggregated, noise-weighted, thresholded and fused across scales. The resulting species set is then refined by three ecological checks (cross-year duplication, seasonal plausibility and niche overlap) before a final cascade adds any extra high-confidence taxa from an auxiliary median-logit stream.

2.4. Image pre-processing and tiling

Each test image is analyzed at its original resolution. A 150-pixel margin on every side is simply skipped during the sliding-window pass, so that no crop ever includes the wooden frame, ruler, or color card lying along the borders. Two window sizes are employed, forming a multi-scale (double-scale) tiling strategy.

The fine-scale pass (Scheme A) sweeps the interior with 518×518 px squares taken every 172 px; across the 2105 plots this produces 303,558 fragments—on average 144 per image (median 154, minimum 4, maximum 272). The coarse pass (Scheme B) uses 732×732 px windows on the same grid; each fragment is down-scaled so that its longer side equals 518 px, yielding roughly one third as many tiles and capturing larger leaves or inflorescences that may be split across fine crops.

For every tile the frozen DINOv2 ViT returns class confidences; the five highest scores and their species IDs are stored. Tile-wise probabilities are summed per scale, and the 18 most confident species are retained. Empirical tuning shows that a single-scale configuration is optimal when taxa below 0.20–0.30 confidence are discarded; with Scheme A alone, a 0.26 threshold followed by limiting the output to eight labels per plot yields 20.12/21.84 F1-score (private / public score).

The baseline used throughout the paper combines both scales: Scheme A filtered at 0.30 and Scheme B at 0.26; their probabilities are summed and the eight highest species are submitted. This multi-scale ensemble attains 29.15 private score and 31.30 public score, whereas percentile cut-offs or different limits on the number of submitted labels proved consistently weaker.

3. Methodology

3.1. Visual inference pipeline

Every test image is processed at two spatial scales. The fine-scale stream (Scheme A) slides 518×518 px crops across the usable area with a stride of 172 px, yielding on average 144 windows per plot (median 154, min 4, max 272). The coarse stream (Scheme B) extracts 732×732 px crops on the same grid and rescales each crop so that its longer side equals 518 px, capturing broader plant structures that might be split across fine crops.

Each crop is forwarded to the frozen ViT-B/14 DINO v2 model introduced in Section 2.1. The model returns class confidences, of which only the five highest are stored. A species is retained for a plot if its best score in Scheme A reaches 0.30 or its best score in Scheme B reaches 0.26. The two scale-

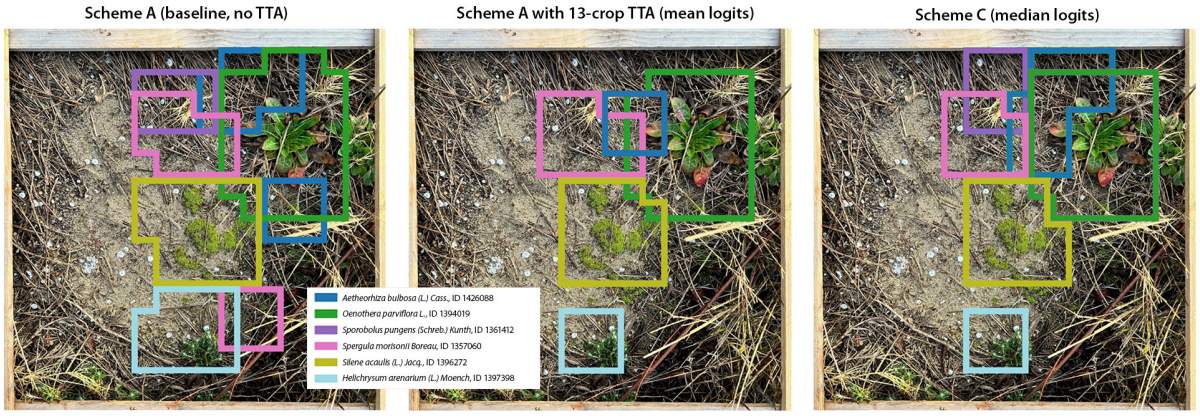

Figure 2: Model predictions using three logit-aggregation strategies: Scheme A (baseline, no TTA), Scheme A + 13-crop TTA (mean logits), and Scheme C (median logits). The colored contours mark merged high-confidence windows for the six taxa shown (only species with confidence above 0.265 are displayed). *The example uses the image RNNB-4-2-20240118.jpg*

specific lists are merged by taking, for every taxon, the single highest confidence observed in either stream; probabilities are not summed. The merged list is truncated to the eight most confident taxa and constitutes the baseline label set that subsequent stages refine.

Alternative aggregation rules were examined. Global percentile cut-offs (80–95%) and hybrid strategies that combined a percentile threshold for one scale with a fixed threshold for the other both lowered the overall public score. The simple per-scale thresholds described above therefore represent the strongest purely visual baseline.

3.2. Thirteen-crop self-ensemble

For each 518 × 518 window generated by Scheme A a fixed 13-crop set was created. The first element is the window itself; the remaining twelve crops are derived from it: four concentric centre crops covering 90%, 80%, 70% and 60%; eight corner crops extracted at 80% and 70% of the shorter side (top-left, top-right, bottom-left, bottom-right for each scale) [9].

Each crop is passed through the evaluation transform, which rescales it to the model's native 518 px input if necessary. The frozen ViT returns a logit vector for every crop; the arithmetic mean of the 13 logit vectors is computed, and only then is a softmax applied to obtain class confidences. A single threshold of 0.265, tuned on the public score, is used to filter these probabilities. The coarse 732 px stream remains unchanged and keeps its 0.30 threshold. After both scales are processed, their outputs are merged exactly as in Section 3.1 and truncated to the eight most confident species.

Alternative reductions were evaluated: the geometric mean of logits, a hand-tuned weighted mean favouring central crops, and the element-wise median of logits. None of these surpassed the simple arithmetic average on either leaderboard, so the arithmetic-logit ensemble is retained as the default; the median variant is revisited later in the cascade described in Section 3.7.

3.3. Multi-scale fusion

After the introduction of the 13-crop self-ensemble, each test image is analyzed twice: once with 518 ×518 px windows enhanced by test-time augmentation and once with larger 732 ×732 px windows that are down-scaled to 518 px. For the fine-scale pass, the logits of the thirteen geometric variants are averaged, passed through a softmax and filtered with a confidence threshold of 0.265. For the coarse pass, the single forward prediction is accepted when its confidence reaches 0.30. The two tiling schemes are then reconciled by a simple rule: for every species the higher of the two confidences is taken; probabilities are neither summed nor renormalized. The resulting list is sorted and trimmed to the nine most confident taxa, a slight relaxation compared with the eight-label limit used in the baseline.

This "highest-confidence" fusion exploits the complementary strengths of the two spatial resolutions. The TTA-augmented fine windows are sensitive to small or partially occluded plants, whereas the coarse windows stabilize predictions on larger leaves or inflorescences that may be split across finer crops. Alternative fusion strategies—probability summation, geometric or weighted averaging, as well as percentile cut-offs—were systematically evaluated but always yielded a lower public score. With the adopted thresholds of 0.265 for the fine scale and 0.30 for the coarse scale, and with the Top-9 restriction, the purely visual stage already attains 33.50 public score and 29.66 private score, providing a strong basis for the ecological post-processing introduced in the following sections.

3.4. Cross-year plot aggregation

Because every test image name encodes a persistent PlotID, plots imaged more than once can be identified, often in different years. All images that share the same identifier are therefore grouped and treated as a temporal series of the same physical quadrat.

Within such a series, the confidence scores already produced for each image are examined, and the single taxon attaining the highest confidence in any member of the group is selected. This most reliable "anchor" species is then added to the prediction list of every other image of the same plot if it is not already present. Copying exactly one label in this way consistently raises the leaderboard score, whereas propagating two or more labels degrades performance; the aggregation is therefore limited to a single species per series. The operation is applied after the purely visual stage yet before the seasonal and niche-based filters discussed in Section 3.6.2, so that the inherited label can still be removed if subsequent ecological checks deem it implausible.

3.5. Noise-aware weighting with SAM and GroundingDINO

A visual inspection of the test images revealed that many frames contain substantial non-botanical objects (wooden plot frames, rulers, stones, pieces of plastic or metal) that occupy a noticeable fraction of the field of view. Their texture sometimes activates the Vision Transformer and produces false species labels. Simply discarding contaminated windows, however, deprives the classifier of genuine plant information along the borders of those windows. A soft penalty that down-weights, rather than deletes, the evidence coming from noisy regions was therefore chosen.

To locate the unwanted objects in a domain-agnostic manner the full test image is first processed by GroundingDINO [10], using the open-vocabulary prompt *"stone, shell, roulette, plastic, metal, hand, ice, snow, measure, ruler, wood, board, paper"*. For every bounding box returned by the detector, the Segment Anything Model (SAM) [11] produces a pixel-accurate mask. The masks are stored as a thirteen-channel tensor—one channel per prompt—so that the contribution of each query can be inspected visually and disabled if necessary. In the production pipeline the channels are merged by a logical OR into a single binary "noise" map.

When the image is later divided into windows by schemes A and B, the fraction of noise pixels inside a window is denoted $m \in [0, 1]$. The confidence of every species predicted in that window is then rescaled according to

$$p' = p\big(1 - \alpha m\big), \qquad \alpha = 0.35.$$

Thus the penalty grows linearly with the proportion of contamination, yet the window is discarded entirely only when it is fully covered by the mask ($m = 1$). The adjusted confidences are subjected to the same thresholds as before, namely 0.265 for the 518 px windows and 0.30 for the 732 px windows that are subsequently resized to 518 px, and then pass on to the remaining stages of the pipeline.

This linear weighting proved more reliable than both hard rejection and non-linear penalty functions: it consistently reduced false positives originating from frames and rulers while preserving the recall of true plant instances that share the window. The setting $\alpha = 0.35$ offered the best trade-off on the public score; larger values harmed recall, whereas smaller values left many noisy detections untouched.

3.6. Ecological post-processing

Most plant species exhibit pronounced seasonality: even widespread taxa are usually observable only in the months when their vegetative or reproductive organs are visible in photographs. In addition, species that co-occur within a 0.25 m² plot typically share similar requirements for light, moisture, and other abiotic factors. These observations motivate an ecological post-processing filter that complements the purely visual pipeline, discarding predictions that are clearly out of season and removing taxa that are ecologically incompatible with the rest of the plot's label set.

Seasonal consistency is assessed with five year occurrence statistics from GBIF Europe, providing a month-wise plausibility check without the need for detailed spatial range modelling [12]. The same criterion applies to both rare and common species, an advantage when the training data are strongly imbalanced.

Niche compatibility is evaluated with the numerical Ellenberg indicators supplied by EIVE. Each species is represented as a multidimensional Gaussian cloud on the axes of light, temperature, soil moisture, soil reaction, and nitrogen supply; the extent to which two clouds overlap gives a direct measure of their likelihood of co-occurrence. This screening step simultaneously increases recall, by reinstating rare but ecologically plausible taxa, and reduces false positives that arise from visually similar yet ecologically unsuitable species.

3.6.1. Seasonal filtering of candidate species

Every test image encodes its acquisition date in the file name (YYYYMMDD), providing the month of observation [8]. For each of the 7806 target taxa, GBIF Europe statistics were compiled over the past five years (2020-2024), yielding monthly and daily counts of confirmed occurrences. Two variant filters were designed.

The hard seasonal filter excludes a taxon when its European record count for the month of the photograph falls between 1 and 10 inclusive; counts of zero are preserved to avoid removing species that may have been mis-matched in GBIF. This rule produced the most stable improvement on the private leaderboard, and exhibited further gains when species with exactly zero observations were also discarded.

The soft seasonal filter applies the same 1-10 threshold to a three-month window centred on the month of acquisition. Although this broader window achieved a larger increase on the public score, the effect on the private set proved inconsistent, so the hard variant is retained as the default while the soft variant is reserved for sensitivity analysis.

Two alternative designs were abandoned. A day-level filter, which removed species never observed on the exact calendar day, reduced performance because daily statistics are too sparse. Restricting predictions to the thousand most frequent species of the month likewise failed to yield any benefit. Seasonal filtering is applied directly after the visual fusion described in Sections 3.3–3.4 and before the rarity and niche-overlap procedures outlined in Section 3.6.2.

3.6.2. Niche overlap filtering based on EIVE indicators

The multi–scale fusion step (Sec. 3.3) keeps the nine most confident taxa for each image and also records a *reserve list* of the next-best nine candidates, yielding at most 18 species per plot. Ecological consistency among those taxa is assessed with a *Niche Similarity Index* $S \in [0, 1]$ that is computed from the five Ellenberg indicator values supplied by EIVE (soil moisture M, nitrogen N, soil reaction R, light L, temperature T) [5].

Similarity measure For species i and j let μ_i, μ_j be their indicator means and σ_i, σ_j the published standard deviations on every axis where both species are defined. Two complementary notions of overlap are combined:

Let μ_{ik}, μ_{jk} denote the consensus niche positions of species i and j, and let σ_{ik}, σ_{jk} denote the corresponding *niche widths* (rather than statistical standard deviations) on every axis k where both species are defined. The niche–similarity index is computed in three steps:

$$\text{centre overlap:} \quad \Delta = \exp\!\Big(-\tfrac{1}{2}D^2\Big), \qquad D^2 = \sum_k \frac{(\mu_{ik}-\mu_{jk})^2}{\sigma_{ik}^2+\sigma_{jk}^2};$$

$$\text{shape overlap:} \quad \mathrm{BC} = \Big[\prod_k \frac{2\sigma_{ik}\sigma_{jk}}{\sigma_{ik}^2+\sigma_{jk}^2}\Big]^{1/4} \exp\Big[-\tfrac{1}{8}\sum_k \frac{(\mu_{ik}-\mu_{jk})^2}{\sigma_{ik}^2+\sigma_{jk}^2}\Big];$$

$$\boxed{S \;=\; \frac{\Delta + \mathrm{BC}}{2}}.$$

Here Δ converts the squared ***Mahalanobis distance*** D^2 between the two niche centres into a kernel, so that values close to 1 correspond to centres that are virtually coincident, whereas small values indicate large Mahalanobis separation.

The second term BC, the ***Bhattacharyya coefficient***, quantifies how strongly the two diagonal multivariate normal clouds interpenetrate, falling from 1 (perfect overlap) towards 0 as the clouds drift apart or narrow. Both components therefore lie in the interval $[0, 1]$, and their arithmetic mean S inherits the same scale, offering an interpretable measure of ecological similarity.

Δ_{ij} reflects how far the niche centres lie apart, whereas BC_{ij} measures the actual overlap of the Gaussian "clouds". A value of $S_{ij} = 1$ indicates virtually identical ecological conditions; scores near 0 imply almost complete separation. If two species share no common indicator, S_{ij} is left undefined.

Although the underlying ecological distributions are neither strictly normal nor necessarily symmetric, the subsequent similarity calculation treats each niche as a multivariate Gaussian centred at μ_{ik} with variance σ_{ik}^2; this simplification proved adequate for fast large-scale filtering while retaining an interpretable overlap score.

Filtering procedure For every image, the algorithm first considers the set P of taxa that survive the visual and seasonal stages. For each species in this set the niche–similarity index is averaged over its partners in P; whenever this mean similarity falls below 0.015, the species is judged ecologically incompatible with the rest of the community and is discarded. After this pruning step the algorithm turns to the "reserve" list—the next nine candidates that were kept only for ecological checks. For every taxon in the reserve list the average of its similarity scores to the species still present in P is calculated; if that average reaches at least 0.750, the taxon is inserted into the prediction set. Similarities contribute to these averages only on axes where both species have indicator values, so missing data never penalise a comparison. In practice a single pass of removal followed by addition is sufficient; further iterations do not change the composition.

Rationale Neighboring plants on a 0.25 m^2 plot rarely exhibit radically different indicator profiles, whereas taxa with highly similar niches often co-occur even when visual evidence is weak. Combining a distance kernel with the Bhattacharyya coefficient keeps S in the convenient range $[0, 1]$, remains continuous, and captures both the location and the breadth of a niche without discretising the environmental axes. The twin thresholds (0.015 for removal and 0.750 for addition) proved to reduce false positives, particularly when visually plausible but ecologically impossible species were present, yet preserved recall by reinstating rare taxa whose niches closely match those already accepted.

3.7. Cascade completion with the median-logit stream (Scheme C)

In addition to the arithmetic 13-crop stream of Scheme A (Sec. 3.2), a parallel set of predictions was produced in which the logits of the thirteen image crops were aggregated by the coordinate-wise median. This "median-logit" variant, hereafter Scheme C, delivered a top-1 species that differed from

Table 1
Incremental impact of successive modules on validation, Private and public scores (all values are macro-F1 ×100, rounded to two decimals; Δ columns give the change with respect to the previous step).

Step	Modification introduced	Private	Δ	Public	Δ
1	No tiling: single resize of the entire image to 518×518 px	13.52	—	13.74	—
2	Scheme A — single-scale tiling, 518 px windows, max 12 species	20.12	+6.60	21.84	+8.10
3	**Baseline:** Scheme A (no TTA, thr 0.26) + Scheme B (732 px, thr 0.30), max 8	29.15	+9.03	31.30	+9.46
4	Multi-scale fusion: Scheme A with 13-crop TTA (thr 0.265) + Scheme B, max 9	29.66	+0.51	33.50	+2.20
5	Cross-year plot aggregation (add top-1 from earlier years)	29.78	+0.12	33.99	+0.49
6	Noise-aware weighting with SAM masks ($\alpha = 0.35$)	32.42	+2.64	36.60	+2.61
7	Seasonal filtering: soft month window *and* hard filter (1–10 GBIF records)	32.49	+0.07	36.75	+0.15
8	Niche drop: remove species with similarity $S < 0.015$	33.20	+0.71	36.99	+0.24
9	Niche re-add: if $S \geq 0.75$ add one reserve species	**33.77**	+0.57	37.72	+0.73
10	Scheme C: median-logit stream, single extra species	33.45	−0.32	**38.13**	+0.41

the arithmetic stream in roughly 16% of the test images, indicating complementary information that might recover plants overlooked by the main ensemble (Fig. 2).

To inject that information in a controlled manner, a three-stage cascade was applied to the submission obtained after ecological filtering.

First, plots that still contained no labels received the single most confident species of Scheme C.

Second, for any plot whose list remained shorter than nine taxa, one additional species was copied from Scheme C provided the median confidence exceeded 70%.

A final pass repeated the procedure with a softer 66% threshold, again limited to at most one extra label per plot. At each step the copied species were appended only if absent, preserving the original rank order of confidences.

Near the close of the competition a mild over-tuning of the ensemble toward the public score became apparent, complicating the assessment of the cascade's true merit. The procedure improved the public score by only +0.4 pp, yet reduced the private score by −0.3 pp. Consequently, the practical value of Scheme C remains questionable and warrants a separate evaluation on the full test data.

4. Results and Discussion

The solution that ranked 4th on the PlantCLEF 2025 private leaderboard (1st on the public leaderboard) can be decomposed into ten incremental steps, each raising the final score; Table 1 lists all steps and their resulting metrics.

Steps 1–3. Earlier PlantCLEF 2024 papers already explored tiling; experiments identified 518 ×518 px windows with a 172 px stride and a 150 px border margin as the most effective setting [8]. A larger-window (Scheme B) was introduced for multi-scale fusion (Sec. 3.3), and confidence thresholds were tuned separately for each scheme. Smaller tiles or non-overlapping grids yielded lower public scores than the combined Scheme A + Scheme B setup. Randomly positioned windows appeared promising but were not evaluated. Limiting predictions to the eight most confident species per plot further improved performance. After steps 1–3, the baseline reached 31.30 public score.

Step 4. A new prediction set was generated for the same Scheme A windows using 13 test-time augmentations; the logits from all transforms were averaged arithmetically, followed by a softmax (Sec.

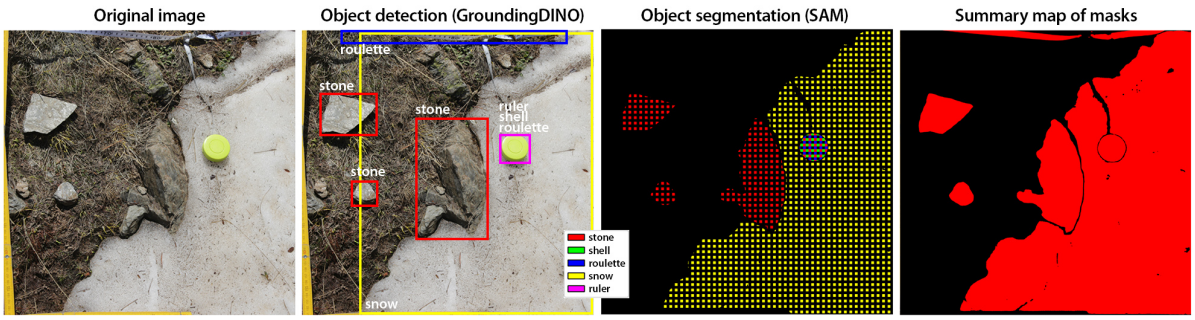

Figure 3: Noise-aware weighting. From left to right: (1) original vegetation-plot image; (2) GroundingDINO detects extraneous objects and returns bounding boxes; (3) SAM refines each detection into pixel-accurate masks; (4) combined binary mask indicating all artifact pixels (red) that trigger confidence down-weighting inside every overlapped window. *The example uses the image CBN-Pla-E4-20200630.jpg*

3.2), and the resulting probabilities were subjected to the same confidence threshold as in the baseline. This arithmetic-logit stream replaced the original Scheme A branch in the ensemble. Experiments further showed that retaining the top-9 species per plot outperformed the previous top-8 limit. The score gained an additional +2.20 points at this stage.

Step 5. Cross-year plot aggregation offered a modest improvement of +0.49 pp: for each plot the top-1 species predicted for the same PlotID in other years was added to the current list (Sec. 3.4). Including the top-2 candidates, however, reduced the score, so the aggregation was limited to a single additional taxon.

Step 6. To mitigate noise-related errors, each image was first processed by GroundingDINO, which detected bounding boxes for objects frequently observed on the test set (stone, shell, roulette, plastic, metal, hand, ice, snow, tape measure, ruler, wood, board, and paper) (Sec. 3.5). Pixel-accurate masks for these detections were then produced by SAM. The most effective strategy was a soft penalty: the confidence of every window was multiplied by $(1 - 0.35 \times \textit{mask coverage})$, where "mask coverage" is the fraction of masked pixels inside the window. The penalty factor 0.35 was found empirically. This noise-aware weighting increased the public score by +2.61 pp; an illustration is provided in Fig. 3.

Step 7. Monthly seasonality was quantified from five years of GBIF Europe occurrence data: for every target species the number of confirmed records was tallied for each calendar month (Sec. 3.6.1). Among several tested rules, only the soft-month filter delivered a gain: a species is discarded when its European record count does not exceed ten in the observation month and in each of the two adjacent months. This constraint raised the public score by +0.15 pp.

Steps 8-9. For every test plot, pairwise ecological similarity was computed among the taxa already predicted. The calculation relied on a composite Niche Similarity Index S that merges the Mahalanobis distance with the Bhattacharyya coefficient (Sec. 3.6.2), using the calibrated Ellenberg indicators from EIVE 1.0. Within each plot the species whose mean S to all other candidates was the lowest (and fell below the empirically fixed threshold of 0.015) was removed. Attempts to remove more than one taxon per plot yielded smaller gains, so the pruning was restricted to a single removal.

After the least compatible taxon had been dropped, the mean S was recomputed between the remaining prediction set and every species in the original top-18 list. Whenever a previously filtered taxon achieved $S > 0.75$, it was reinstated (Sec. 3.6.2). Together, the removal-and-add cycle improved the public score by +0.97 pp.

The procedure yields ecologically interpretable results: manual inspections confirmed that pairs with high S share similar environmental preferences, whereas taxa with low values occupy contrasting niches.

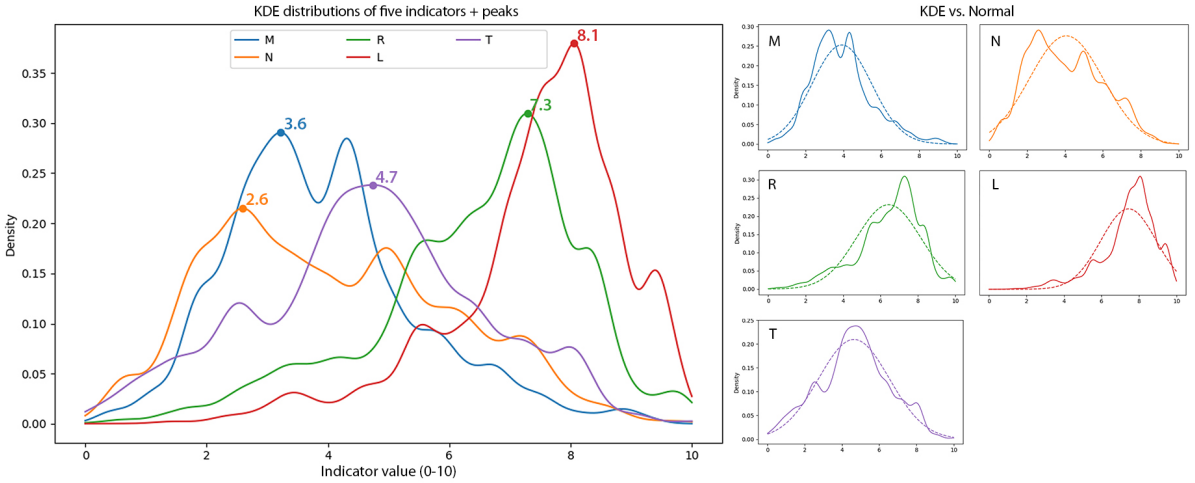

Figure 4: The large panel on the left displays kernel-density estimates (solid lines) for moisture (M, blue), nitrogen (N, orange), reaction (R, green), light (L, red) and temperature (T, violet) with their modal peaks annotated. The small panels on the right juxtapose, for each axis, the empirical KDE (solid) with the normal curve that shares the same mean (μ) and standard deviation (σ) for that indicator (dashed). The fitted parameters are $\mu_M = 3.9,\ \sigma_M = 1.6$; $\mu_N = 4.0,\ \sigma_N = 1.8$; $\mu_R = 7.0,\ \sigma_R = 1.7$; $\mu_L = 7.8,\ \sigma_L = 1.2$; and $\mu_T = 4.6,\ \sigma_T = 1.3$.

However, Fig. 4 reveals that the calibrated Ellenberg distributions deviate from normality and are often asymmetric around the mean. Niche Similarity Index S treats each niche as a symmetric Gaussian; a more faithful model that captures these asymmetries (and estimates indicator spread individually for every taxon) may further refine the similarity metric.

Step 10. Scheme C, which aggregates window logits by the median and is therefore more robust to outliers than Scheme A, frequently ranked a previously filtered taxon as its top-1 candidate. The final cascade iteratively inserted the most confident Scheme C predictions into the submission—up to three passes, stopping once the public score no longer increased. This procedure increased the public score by +0.41 pp, yet lowered the private score, indicating a degree of overfitting to the public leaderboard.

The pipeline achieved a 33.447% private F1-score, securing 4th place in PlantCLEF 2025 [1].

5. Conclusion

The presented approach demonstrates that a combination of a ViT classifier, tiling, test-time augmentation, zero-shot segmentation and multi-step ecological post-processing can substantially improve multi-label plant identification accuracy without additional network training. Nevertheless, several aspects of the methodology require critical reassessment and further development.

Computational costs Test-time augmentation (Sec. 3.2) yields a gain in F1 (Tab. 1) but prolongs inference by ≈ 25 GPU-hours per dataset. In order to accelerate inference, it is necessary to experimentally select two to three augmentations from the available twelve and to determine an averaging scheme for their logits that provides the maximum improvement.

Biological plausibility Cross-year aggregation (copying the top-1 species from previous seasons, Sec. 3.4) adds +0.1 pp, yet may mask local extinctions and shifts in taxonomic diversity observable in long-term vegetation resurvey studies [13].

Ellenberg/EIVE indicators are primarily calibrated for Central Europe; application elsewhere requires local gradient tables or estimation of missing indicator values for species of the new region [14].

The niche approximation by a single Gaussian (Sec. 3.6.2) may inadequately represent multimodal or asymmetric distributions and thus distort the similarity metric S; the *Schoener D* may therefore be considered [15]. Future work should explore more precise metrics based on *kernel density estimation* (KDE), which constructs a continuous density without assuming normality, or *Gaussian mixture models* (GMM), which approximate the distribution by a sum of Gaussians and often provide a more compact representation than KDE at comparable accuracy.

Overfitting and lack of independent validation Local validation was *not performed*, although the single-plant close-up dataset (≈ 1.4 million images) is formally available within the challenge. Because the team was unable to download this volume in time due to technical constraints, hyperparameter tuning relied exclusively on the public score, leading to overfitting and a loss of three positions on the private leaderboard.

Several methods can local estimate model quality, overfitting and hyperparameter selection on unlabelled data. Under PlantCLEF 2025 conditions it would have been reasonable to select the output limiter k and confidence threshold on a small subset (about 50k) of single-plant close-up images using proxy metrics such as F1@k. For automatic optimization of these two hyperparameters in future work, a single-plant validation set is proposed. It has been theoretically shown that, for a narrow distribution of ground-truth label counts, such optimization transfers to multi-label images without bias [16].

Complexity and redundancy of certain components Although the GroundingDINO + SAM combination effectively suppresses artificial objects (Sec. 3.5), it increases memory consumption and inference time; preliminary experiments with simple rectangular masks resulted in a reduction of merely 0.3 pp in F1, indicating that the segmentation step can be simplified.

The cascade scheme with three additional species (Scheme C, Sec. 3.7) increased the public score but reduced the private score, representing further overfitting that could have been avoided through local validation.

Use of external data The PlantCLEF rules permit external data; however, the ecological layer described in this work extends beyond a purely computer-vision task and limits transferability to other regions and taxonomic groups.

Future work is planned to: (i) develop and implement a local validation scheme based solely on label-free proxy metrics; (ii) create code for computing an ecological neighborhood validity metric based on niche intersection (KDE/GMM hypervolumes) and global co-occurrence statistics (GBIF); and (iii) openly publish a repository containing the full ecological post-processing pipeline (https://github.com/nellyse-menova, release planned for Q3 2025). These developments should enhance reproducibility and extend applicability to new regions and datasets.

Acknowledgments

Sincere gratitude is extended to all naturalists, researchers, and experts who contribute data to and support the work of the Global Biodiversity Information Facility (GBIF) worldwide; their tireless efforts and commitment to open science have made the results presented in this paper possible.

Declaration on Generative AI

During the preparation of this work, the following Generative AI tool was employed:

- **ChatGPT o3** (OpenAI, May 2025 model) — *Text Translation* of the paper from Russian to English, *Grammar and spelling check*, and *Improve writing style*.

All AI-generated suggestions were reviewed and edited manually; the authors assume full responsibility for the final content. No Generative AI system was used to create original scientific ideas, analyse data, or draw conclusions.

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, 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] GBIF.Org User, Occurrence download, 2025. URL: https://www.gbif.org/occurrence/download/0003757-250227182430271. doi:10.15468/DL.QT3PHR.

[5] J. Dengler, F. Jansen, O. Chusova, E. Hüllbusch, M. P. Nobis, K. V. Meerbeek, et al., Ecological indicator values for europe (eive) 1.0, Vegetation Classification and Survey 4 (2023) 7–29. doi:10.3897/VCS.98324.

[6] 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://doi.org/10.5281/zenodo.10848263. doi:10.5281/zenodo.10848263.

[7] M. El Oquab, T. Darcet, T. Moutakanni, H. V. Vo, et al., DINOv2: Learning robust visual features without supervision, arXiv:2304.07193, 2023.

[8] S. Foy, S. McLoughlin, Utilising DINOv2 for domain adaptation in vegetation plot analysis (plantclef 2024), in: Working Notes of CLEF 2024, 2024.

[9] L. Picek, M. Šulc, Y. Patel, J. Matas, Plant recognition by ai: Deep neural nets, transformers and knn in deep embeddings, Frontiers in Plant Science 13 (2022) 787527. doi:10.3389/fpls.2022.787527.

[10] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, L. Zhang, Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL: https://arxiv.org/abs/2303.05499. arXiv:2303.05499.

[11] 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: https://arxiv.org/abs/2304.02643. arXiv:2304.02643.

[12] H. C. Wittich, M. Seeland, J. Wäldchen, M. Rzanny, P. Mäder, Recommending plant taxa for supporting on-site species identification, BMC Bioinformatics 19 (2018) 190.

[13] U. Jandt, H. Bruelheide, C. Berg, M. Bernhardt-Römermann, V. Blueml, F. Bode, J. Dengler, M. Diekmann, H. Dierschke, I. Doerfler, U. Döring, S. Dullinger, W. Haerdtle, S. Haider, T. Heinken, P. Horchler, F. Jansen, T. Kudernatsch, G. Kuhn, M. Wulf, Resurveygermany: Vegetation-plot time-series over the past hundred years in germany, Scientific Medical Data 9 (2022) 631. doi:10.1038/s41597-022-01688-6.

[14] L. Leccese, G. Fanelli, V. E. Cambria, M. Massimi, F. Attorre, M. Alfò, S. Aćić, E. Bergmeier, A. Čarni, M. Cuk, R. Custerevska, P. Dimopoulos, P. Hoda, A. Mullaj, U. Šilc, Z. Skvorc, Z. Stancic, Z. Dajic Stevanovic, R. Tzonev, K. Vassilev, L. Malatesta, M. De Sanctis, Estimation of missing ellenberg indicator values for tree species in south-eastern europe: a comparison of methods, Ecological Indicators 160 (2024) 111851. URL: https://www.sciencedirect.com/science/article/pii/S1470160X2400308X. doi:https://doi.org/10.1016/j.ecolind.2024.111851.

[15] D. Warren, R. Glor, M. Turelli, Enmtools: A toolbox for comparative studies of environmental niche models, Ecography 33 (2010) 607 – 611. doi:10.1111/j.1600-0587.2009.06142.x.

[16] N. Xu, C. Qiao, J. Lv, X. Geng, M.-L. Zhang, One positive label is sufficient: Single-positive multi-label learning with label enhancement, 2022. URL: https://arxiv.org/abs/2206.00517. arXiv:2206.00517.