

Beyond Domain Shift: A Three-Tier Diagnostic Framework for Multi-Species Plant Identification in Vegetation Quadrats

LifeCLEF lab at CLEF 2026

Hamza Ahmed Ishrat*, Sue Han Lee

Swinburne University of Technology Sarawak Campus, 93350, Sarawak, Malaysia

Abstract

We present the third submission in our series investigating multi-species plant identification in vegetation quadrat images, completing a systematic examination of preprocessing and inference strategies without backbone modification. Through a comprehensive ablation of tiling granularity across five strategies, we establish that fixed multi-scale tiling centred around ecologically motivated spatial scales outperforms resolution-adaptive alternatives, achieving 36.051% macro-averaged F1 and placing 12th of 105 teams under the team name 'kiwimckiwiman'.

We conduct an exploratory failure mode analysis across a three-tier inference hierarchy: patch, plot, and site, identifying four mechanistically distinct failure modes that attempt to uncover challenges beyond domain adaptations. At the patch level, we distinguish attentional from representational failure and introduce an epistemic-aleatoric uncertainty decomposition for patch reliability weighting. At the plot level, we examine dominant species capture, visual similarity contamination, and low-confidence spatial suppression as structurally distinct behaviours with different implied remedies. At the site level, we demonstrate that ecological implausibility is invisible to visual methods and propose community-level Gaussian Mixture Model filtering over ecological indicator trait spaces as a candidate remedy. These findings motivate the transition from pipeline optimisation toward model-centric approaches as the next phase of this research.

Keywords

multi-species plant identification, vegetation quadrats, failure mode analysis, tiling granularity, computer vision uncertainty, ecological spatial scale, domain adaptation, PlantCLEF

1. Introduction

Automated identification of plant species within vegetation quadrat images, as shown in Figure 1, remains an open problem at the intersection of computer vision and computational ecology. A quadrat, typically a 50 by 50 centimetre plot, serves as a structured sampling unit for inferring broader ecological patterns, with ecologists employing them to quantify species occurrence, estimate percent cover, and characterise vegetation community structure [1, 2, 3]. The method has been foundational to quantitative plant ecology and remains the dominant approach for biodiversity assessment and long-term ecological monitoring [3]. Historically this process has been labour-intensive and reliant on human expertise [4], motivating a growing body of work on automated species identification from field imagery [5].

The PlantCLEF 2024, 2025, and 2026 challenges [6, 7, 8], part of the LifeCLEF series [9] formalise this problem as a computer vision task: given a high-resolution top-down quadrat image, predict all plant species present. The difficulty is substantial; the highest accuracy across the 2024 and 2025 editions did not exceed 37%. The 2026 edition saw first place reach 75%, a jump large enough to confirm that the ceiling belongs to a specific class of approach rather than the problem itself, while second place and below did not break 44%.

Our submission is the third in a series. In 2024 [10] we established that post-hoc Bayesian aggregation over patch-level predictions is competitive with model-centric pipelines, granted no fine-tuning of the model was performed. In 2025 [11] we characterised the aggregation parameter space and identified a

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

✉ hamza.ishrat@yahoo.com (H. A. Ishrat); shlee@swinburne.edu.my (S. H. Lee)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

performance ceiling near 35%. Although the winner of the 2025 challenge had worked on preprocessing by utilising JPEG recompression techniques as a normalisation method [12], this work addresses the remaining preprocessing question, specifically, how tiling granularity affects what the model sees and performs the diagnostic analysis necessary to understand why that ceiling exists.

We achieve 36.051% macro-averaged F1 through tiling strategy alone, with no backbone modification and no change to the aggregation pipeline from 2025. This places us 12th of 105 teams. The score is not the primary contribution of this work. Rather, we use the competition setting as a controlled empirical platform to conduct a systematic exploratory analysis of where and why aggregation-based inference fails in this specific problem domain. Our aim is to reframe multi-species quadrat identification not as a competition challenge to be optimised incrementally, but as a structured research problem with identifiable failure modes that constrain the design space for future solutions.



Figure 1: Examples of vegetation plots from the PlantCLEF challenge. Taken from our 2024 paper [10]

2. Dataset

The dataset is unchanged from PlantCLEF 2025 [7]. We refer to [10, 11] for full details and summarise only what is relevant to this work.

2.1. Training set

The training set is a subset of the Pl@ntNet corpus [13], consisting of 1,408,033 single-species images across 7,806 taxa predominantly found in South Western Europe. Class sizes are highly imbalanced, ranging from fewer than 10 to over 500 images per species. Images span multiple organ types and capture conditions reflecting the natural visual variability the model must handle at deployment.

2.2. Test set

The test set consists of 2,105 high-resolution vegetation plot images (roughly 1000 to 4000 pixels per side) taken from above by experts across Pyrenean and Mediterranean ecological contexts. Each plot may contain between 1 and many co-occurring species, captured under varying domain conditions including seasonal change, occlusion by rocks and measuring tools, motion blur, and phenological variation.

3. Methodology

We use the same pretrained backbone as our prior submissions [10, 11]: the `vit_base_patch14_reg4_dinov2_lvd142m_pc24_only_classifier_then_all` DINOv2-based vision transformer provided by the challenge organisers [14]. No training is performed; all

modifications are confined to preprocessing and inference-stage aggregation, which follows our 2025 pipeline without modification [11].

3.1. Tiling

Tiling has emerged as the dominant inference strategy for high-resolution quadrat imagery across PlantCLEF submissions [10, 11, 15, 16, 12, 17, 18], decomposing each plot into smaller patches that are independently classified before aggregation.

This approach serves two purposes: it reduces the multi-species classification problem to a pseudo single-species task under the assumption that sufficiently small patches are likely dominated by one taxon, and it brings the effective input resolution within the operational range of standard vision transformers without requiring architectural modification.

Our central hypothesis is that patch granularity governs the real-world spatial scale each patch represents, and that this ecological spatial scale is the primary determinant of performance rather than pixel-level information preservation alone. We systematically evaluate five tiling strategies to test this hypothesis.

3.1.1. Fixed n

Each plot is resized using Lanczos interpolation to $\sqrt{n} \times 518$ pixels per side, then divided into n non-overlapping patches of 518×518 pixels, conforming directly to the model input size. We evaluate square values of n from 9 to 64.

The choice of n determines the real-world area each patch represents: given that quadrats cover a fixed physical area of 50×50 cm, a plot divided into $n = 25$ patches assigns each patch approximately 10 cm^2 of physical space regardless of native image resolution. This framing reinterprets tiling not as a sampling trade-off between detail and context, but as a selection of ecological spatial scale.

3.1.2. Dynamic n

As an alternative, we explored a resolution-adaptive tiling strategy in which n is derived from the native image dimensions, centring patch size around the model input of 518×518 . Three variants are evaluated:

$$n_{\text{ceil}} = \left\lceil \frac{\max(w, h)}{518} \right\rceil \quad n_{\text{floor}} = \left\lfloor \frac{\max(w, h)}{518} \right\rfloor \quad n_{\text{round}} = \text{round}\left(\frac{\max(w, h)}{518}\right) \quad (1)$$

The motivation is to minimise rescaling distortion by adapting n to the native resolution of each plot. However, this approach sacrifices real-world spatial consistency; two plots of different native resolutions but identical physical dimensions receive different values of n , such that corresponding patches represent different physical areas. We include these variants to directly compare information preservation against spatial scale consistency.

3.1.3. Border cropping

Several prior submissions reported gains from cropping plot borders before tiling, removing wooden frames, rulers, and other measurement artifacts. Foy & McLoughlin [17] applied a fixed 150-pixel border removal; Herasimchyk et al. [16] evaluated a range of 5–15%, finding that 5% generalised better to the private test set despite 10% performing better on the public set; Semenova [18] applied similar cropping in conjunction with their sliding window scheme. We evaluate border removal at 1%, 2.5%, 5%, and 10% across fixed $n \in \{16, 25, 36, 49\}$.

3.1.4. Multi-scale tiling

Tiling granularity varies substantially across prior submissions. Gustineli et al. [15, 19] used 3×3 and 4×4 grids respectively; Herasimchyk et al. [16] combined 4×4 and 5×5 grids, finding non-overlapping tiles sufficient. Espitalier [12] combined grids from the full image through to 6×6 , totalling 91 patches, noting diminishing returns beyond this. Rather than committing to a single granularity, we evaluate combinations of fixed n values processed simultaneously, allowing the aggregation pipeline to integrate predictions across multiple spatial scales. Combinations evaluated are reported in Table 6.

3.1.5. Sliding window

Following approaches in the 2024 and 2025 challenges [17, 18], we also evaluate overlapping patch extraction via a sliding window. After scaling to $\sqrt{n} \times 518$ per side, patches of 518×518 are sampled with a stride of 259 pixel (half the patch size) ensuring that the boundary region between adjacent non-overlapping patches is captured as a contiguous patch. We evaluate this scheme at $n \in \{16, 25, 36\}$ and in combination with dynamic n derivations.

3.1.6. JPEG recompression

Following Espitalier [12], who demonstrated that tuning JPEG compression quality and YCbCr chroma subsampling acts as a domain regularisation step: harmonising the compression statistics of test plots with those of the training data, we apply recompression at 2 modes that were found to increase model performance: quality 94, subsampling factor of 4:1:1 and quality 85, subsampling factor of 4:2:2, prior to tiling. This is evaluated in combination with fixed n to isolate the contribution of recompression from tiling granularity.

3.2. Aggregation

Post-hoc aggregation follows our 2025 submission [11] without modification, employing variance-weighted Bayesian aggregation with fixed hyperparameters $k = 5000$ and $z = 1.6$ established through ablation in prior work. Species selection uses z-score thresholding to identify statistically significant outliers in the aggregated distribution, in contrast to the fixed confidence thresholds or top- n selection strategies employed by most submissions [12, 16, 18]. Aggregation parameters are held fixed across all tiling experiments to ensure that observed score differences are attributable to tiling strategy alone.

This design choice has a direct bearing on the preprocessing results reported in Section 4. Approaches such as border cropping and JPEG recompression produced gains in prior submissions that employed naive aggregation (summing probabilities or selecting top- n predictions) where removing noisy or artefactual patches before aggregation reduces false positives directly. In our pipeline, patches are not treated as equal contributors; each patch is weighted by its estimated reliability through the joint epistemic and aleatoric uncertainty decomposition described in Section 5.1. A patch containing a wooden frame or inorganic content produces a diffuse, low-confidence distribution that receives low weight in aggregation regardless of whether it is explicitly removed.

4. Results

We would like to note that, for all intents and purposes, all scores reported and discussed are from the private leaderboard, evaluated on approximately 89% of the test data.

4.1. Tiling

4.1.1. Fixed n

Table 1 reports private scores across fixed tiling configurations. Dividing each plot into 25 patches yields the highest individual score at 35.672%, surpassing our previous best of 35.128% achieved with a

combined 64+16 multi-scale scheme. Performance degrades at both extremes: $n = 64$ requires upscaling most plots to approximately 4144×4144 pixels, nearly twice the native resolution, introducing blur and noise amplification, while $n = 9$ downscales the image to approximately 1554×1554 pixels, increasing the likelihood of multiple species co-occurring within a single tile and reducing discriminability as well as compressing information.

Table 1

Private scores for fixed n tiling. Bold results exceed the prior best of 35.128% [11].

n	Private score
64	33.396
49	34.109
36	34.557
25	35.672
16	35.441
9	31.849

Table 2

Native resolution distribution of test plots.

Image size (x)	Count
$x < 1500 \times 1500$	10
$1500 \times 1500 < x < 2000 \times 2000$	92
$2000 \times 2000 < x < 2500 \times 2500$	206
$2500 \times 2500 < x < 3000 \times 3000$	873
$3000 \times 3000 < x < 3500 \times 3500$	869
$3500 \times 3500 < x$	55
Total	2105

4.1.2. Dynamic n

Contrary to expectations, dynamic tiling consistently underperforms fixed tiling despite better preserving native image resolution. As shown in Table 3, the floor scheme concentrates most plots at $n = 25$, which accounts for its relatively strong performance (Table 4). However, all dynamic variants fall below fixed $n = 25$ and $n = 16$.

Table 3

Plot counts per n value under each dynamic tiling scheme.

n_patches	Dynamic ceiling	Dynamic floor	Dynamic round
64	24	0	5
49	454	24	130
36	1231	465	1167
25	284	1220	581
16	97	284	162
9	15	97	59
4	0	15	1

Dynamic tiling consistently underperforms fixed tiling despite better preserving native image resolution. The failure is interpretable through the ecological spatial scale argument: fixed $n = 25$ rescales all plots to ≈ 2590 pixels per side, ensuring each patch represents $\approx 10 \text{ cm}^2$ of physical quadrat area regardless of native resolution. Dynamic tiling assigns different values of n to plots of different

Table 4

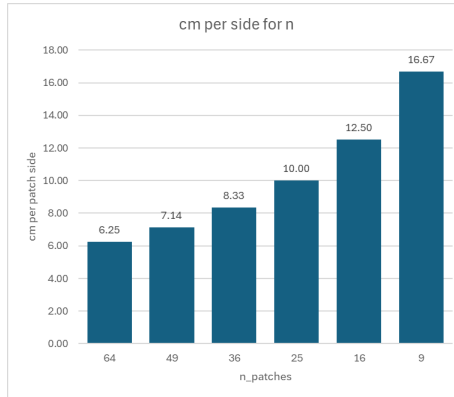
Private scores for dynamic tiling schemes.

Scheme	Private score
Ceiling	35.137
Floor	35.192
Round	34.655

resolutions, producing patches that represent different physical areas despite the quadrats covering identical ground; a structural inconsistency that fixed tiling avoids.

The strong performance of $n = 25$ admits two complementary explanations that cannot be fully disentangled. First, 2590 pixels per side closely matches the native resolution of most test plots as shown in Table 2, minimising rescaling distortion. Second, $n = 25$ corresponds to $\approx 10 \text{ cm}^2$ per patch; an ecologically motivated scale at which single-species dominance within a patch is plausible. The two explanations are not independent: consistent field collection methodology produces images within a predictable resolution range, such that the resolution-optimal and ecologically-motivated scales naturally coincide. The dynamic floor scheme, which concentrates most plots at $n = 25$, still underperforms fixed $n = 25$ (35.192 vs 35.672), providing partial evidence that cross-plot scale consistency contributes beyond resolution fidelity alone.

We acknowledge this interpretation is circumstantial; controlled experiments varying resolution independently of physical scale would be required to confirm it. As shown in Figure 2, the $n \in \{16, 25, 36\}$ range occupies the region where patch scale is fine enough for morphological discrimination but coarse enough to limit within-patch species overlap.

**Figure 2:** Physical scale (cm per patch side) as a function of n , for a $50 \times 50 \text{ cm}$ quadrat.

4.1.3. Border cropping

Border cropping consistently degrades performance across all tested configurations as shown in Table 5, in contrast to prior work reporting gains at 5–10% removal [17, 16, 18]. Even 1% removal reduces scores in nearly all cases, with the sole exception of $n = 36$ at 2.5% (34.715 vs 34.557), which remains unexplained.

As noted in Section 3.2, our uncertainty-weighted pipeline naturally down-weights inorganic patches, making explicit border removal redundant and the discarded signal costly.

4.1.4. Multi-scale tiling

Combining multiple values of n simultaneously allows the aggregation pipeline to integrate predictions across different physical scales, consistently outperforming any single-scale configuration. Table 6

Table 5

Private scores for border cropping at varying removal percentages.

n	Border crop %				
	0	1	2.5	5	10
49	34.109	33.403	33.058	31.439	30.881
36	34.557	33.836	34.715	32.745	32.704
25	35.672	35.458	33.960	33.055	33.029
16	35.441	35.338	34.536	34.536	33.699

reports results for all evaluated combinations. The combination 36_25_16_9 achieves the highest score at 36.000%, the first result to exceed 36% and one of the best in our submission series.

Table 6

Private scores for multi-scale tiling combinations.

Patch levels	Total patches	Private score
49_36_25_16_9	135	35.649
49_36_25_16	126	35.492
49_36_25	110	35.124
36_25_16_9	86	36.000
36_25_16	77	35.403
36_25	61	35.172
25_16_9	50	35.842
25_16	41	35.714
ceil_floor	see Table 7	35.993

Performance is highest when combinations are centred around $n \in \{36, 25, 16\}$, consistent with these values occupying the ecologically meaningful scale range identified in the fixed n experiments. Including $n = 49$ degrades performance relative to starting from $n = 36$, suggesting that patches smaller than ≈ 7 cm per side introduce more noise than signal. The contribution of $n = 9$ is positive when combined with $n \in \{36, 25, 16\}$ but not when $n = 49$ is also present, indicating that coarse patches provide useful global context only when the complementary fine-scale signal is sufficiently strong.

Interestingly, the ceil_floor tiling strategy that merged the ceiling and floor values of n obtained dynamically, performed just as well as 36_25_16_9 despite only containing 2 spatial scales. Table 7 shows the different combinations obtained using this multi-scale scheme

Table 7

n_patches derivations for dynamic_ceil_floor tiling scheme

n_patches	Dynamic_ceil_floor
64, 49	24
49, 36	454
36, 36	11
36, 25	1220
25, 16	284
16, 9	97
9, 4	15

The majority of the combinations lie in the [36,25] region, accounting for more than 50% of all plots. This, again, is inline with our findings for the fixed n strategy.

We additionally evaluated combinations that omit intermediate scale values, motivated by the observation that adjacent scales represent similar physical areas and may provide redundant rather than complementary signal. Results are reported in Tables 8 and 9.

Table 8

Effect of coarsest scale on the core _25_16_9 combination. Bold indicates best overall result across all tiling experiments.

Combination	Total patches	Private score
25_16_9 (baseline)	50	35.842
36_25_16_9	86	36.000
49_25_16_9	109	36.051
36_25_16 (baseline)	77	35.403
49_25_16	90	35.952
64_25_16	105	34.768

The upper half of Table 8 shows that replacing $n = 36$ with $n = 49$ as the coarsest scale above the _25_16_9 core improves performance (36.051 vs 36.000). The larger physical gap between $n = 49$ (≈ 7.14 cm) and $n = 25$ (≈ 10.0 cm) provides more complementary signal than the smaller gap between $n = 36$ (≈ 8.33 cm) and $n = 25$. The lower half confirms this without $n = 9$: 49_25_16 outperforms 36_25_16 by a larger margin (35.952 vs 35.403), while 64_25_16 degrades performance regardless.

Table 9

Effect of extending the scale range at the coarse ($n = 9$) and fine ($n = 64$) ends. Δ is the score difference from the baseline combination.

Baseline	Score	+ $n = 9$	Score	Δ	+ $n = 64$	Score	Δ
25_16	35.714	25_16_9	35.842	+0.128	64_25_16	34.768	-0.946
49_25_16	35.952	49_25_16_9	36.051	+0.099	64_49_25_16	35.044	-0.908
36_25_16	35.403	36_25_16_9	36.000	+0.597	—	—	—

Table 9 shows that adding $n = 9$ at the coarse end consistently improves performance across all anchor combinations, while adding $n = 64$ at the fine end consistently degrades it; the two boundaries are asymmetric in both direction and magnitude.

The contribution of $n = 9$ varies with how much coarse context the baseline already provides. The gain is largest for 36_25_16 (+0.597), whose physical range of 8.33–12.5 cm per patch side leaves the coarse end under-represented; adding ≈ 16.67 cm patches covers qualitatively new spatial territory. For 49_25_16 (+0.099), where $n = 49$ already extends to ≈ 7.14 cm, the marginal contribution of $n = 9$ is smaller. The coarse context is useful, but diminishing returns apply once the scale range is already wide.

The penalty from $n = 64$ tells a different story. The degradation is largest precisely where the baseline is strongest: -0.908 on 49_25_16 versus -0.946 on 25_16. This suggests that upsampling distortion at $n = 64$ does not merely add uninformative patches that the aggregation can ignore; it actively corrupts the confidence signal that the weighting scheme relies on, interfering with otherwise reliable predictions. $n = 64$ is excluded from all further consideration.

4.1.5. Sliding window

Sliding window extraction consistently underperforms fixed non-overlapping tiling at comparable granularities (Table 10). The best sliding window result ($n = 16$, 35.354%) falls below the corresponding fixed result (35.441%). We attribute this to redundancy in the patch set: at stride 259 each boundary region between adjacent fixed patches is captured as an additional patch, substantially increasing the total number of patches at the same spatial scale. Unlike multi-scale tiling, which introduces patches at

different physical scales, sliding window patches at a fixed n are all drawn from the same spatial scale. This redundancy amplifies noise in the aggregation rather than introducing complementary signal, consistent with the observation that the performance gap between sliding and fixed schemes widens as n increases and the patch count grows.

Table 10

Private scores for sliding window tiling (stride 259).

Scheme	Private score
Fixed $n = 36$	33.896
Fixed $n = 25$	34.966
Fixed $n = 16$	35.354
Dynamic ceiling	34.547
Dynamic floor	35.265
Dynamic round	34.458
Ceil & floor	35.095

4.1.6. JPEG recompression

JPEG recompression is harmful across all tested configurations in Table 11, dropping scores approximately 4–5 percentage points relative to uncompressed baselines despite substantially improving Espitalier’s submission [12]. Mode 2 performs worse than Mode 1 across all configurations.

As discussed in Section 3.2, recompression flattens the logit distribution, suppressing the variance signal that drives patch weighting and degrading aggregation quality. Preprocessing choices that benefit naive pipelines can actively harm uncertainty-aware ones.

Table 11

Private scores with both modes of recompression applied prior to tiling.

Configuration	Mode 1	Mode 2	No recompressions
$n = 16$	31.414	29.133	35.441
$n = 25$	30.767	27.041	35.672
$n = 36$	30.157	27.925	34.557
36_25_16_9	31.444	28.996	36.000
49_25_16_9	31.728	29.423	36.051
ceil_floor	31.849	27.109	35.993

5. Failure Mode Analysis

The preceding results establish a preprocessing ceiling: two independent approaches, namely Espitalier [12] and our current submission converge near 36%, and the aggregation pipeline from our prior work [11] adds no further headroom. We set aside the competition framing at this point.

The challenge is typically characterised by domain shift between single-species training imagery and multi-species test plots, and the absence of multi-label supervision. These characterisations are accurate but insufficiently specific. Domain shift is a symptom; it does not explain where the model fails, which predictions are recoverable by post-hoc methods, or what must change to recover them.

We reframe the problem as a three-tier inference hierarchy, patch, plot, and site, each introducing structurally distinct failure modes that persist regardless of backbone or aggregation strategy. The analysis is qualitative rather than quantitative. We note that without ground truth labels for the test set, species identifications are treated as presumptively correct where visually confirmable. The analysis is qualitative rather than quantitative, and its purpose is to motivate the specific model-centric interventions described in Section 7.

5.1. Patch Level

The patch is the fundamental unit of inference in our pipeline. At this tier, the primary challenge is signal extraction: determining which patches carry reliable species information and how that information should be weighted prior to aggregation.

5.1.1. Uncertainties at the patch level

The fundamental limitation of patch-level inference is that the model will always produce a prediction. Regardless of patch content, the backbone maps any input to a distribution over k classes; whether that distribution reflects genuine signal or visual confusion is not encoded in the output itself. This creates an immediate problem: an overgrown patch where multiple species compete and a barren patch where the model finds nothing to classify both produce diffuse distributions. Signal coherence metrics cannot distinguish between them.

We draw on the decomposition of Kendall and Gal [20], distinguishing *epistemic* uncertainty, reducible uncertainty in the model’s predictions addressable with more data or better training, from *aleatoric* uncertainty, irreducible uncertainty inherent in the observation itself. The distinction has a longer history in statistics and philosophy of probability [21], but we adopt the deep-learning framing here.

Prediction confidence serves as a proxy for epistemic uncertainty: a concentrated distribution indicates the model has committed to a prediction; a diffused one indicates it cannot resolve the patch content. Plantness, the estimated proportion of organic plant matter, serves as a proxy for reducing aleatoric uncertainty. It is derived independently of the model using an auxiliary regression model [11], producing a continuous score compatible with probability-based aggregation rather than the binary pixel masks of SAM-based approaches [18, 17], while avoiding SAM failure cases such as masking valid organic material like twigs.

The interaction between these two sources defines four qualitatively distinct patch types, summarised in Figure 3. The ideal case, high plantness, high confidence, requires no intervention. Low plantness with low confidence can be discarded outright. The two ambiguous cases are where the joint decomposition earns its value: a vegetated patch with low confidence retains genuine signal that a confidence-only filter would discard, while a barren patch with high confidence, the most dangerous case, exemplified by the *Posidonia oceanica* prediction in Section 5.3, would be treated as reliable without the plantness check.

		Confidence (higher is better)	
		HIGH	LOW
Plant score (higher is better)	HIGH	Ideal zone High plant matter and high confidence from model	Consider zone High plant matter, low model confidence. Signal is potentially from multiple sources but confused as noise
	LOW	Consider zone Low plant matter, high model confidence. Signal is potentially from lone signal in barren surrounding	Discard zone Low plant matter and low confidence from model

Figure 3: Decision matrix for patch weighting under joint epistemic and aleatoric uncertainty. Neither axis alone is sufficient.

Figure 4 illustrate the two problematic cases. Plot CBN-Pd1C-C1-20160726 shows patches from a near-full coverage of plant matter showing low distributions across patches; the model is overwhelmed by competing signal. Plot CBN-P1a-E5-20200630 is largely snow-covered, yet several patches produce distributions with comparable concentration to those in the vegetated plot. A confidence-only filter treats them as equally reliable; plantness scores near zero for the snow patches and down-weights them regardless of model output.

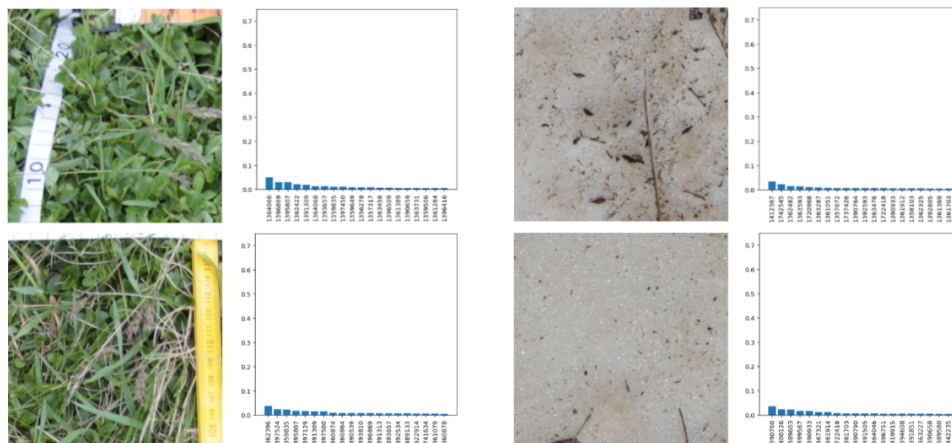


Figure 4: Patches from CBN-Pd1C-C1-20160726 (left) and CBN-P1a-E5-20200630 (right) with their probability distributions

Neither source alone is sufficient. Confidence alone conflates the overgrown and barren cases. Plantness alone conflates a sparse patch with a single identifiable species against an overgrown patch the model cannot resolve. Only their joint consideration treats qualitatively different patch types differently.

5.1.2. Attention-Guided Iterative Masking and Signal Suppression

Conventional image classifiers trained with cross-entropy objectives are optimised to rank the single most probable class, tending to lock onto the most visually discriminative features while suppressing weaker signals from co-occurring species. In single-label classification this is desirable; in quadrat analysis it constitutes a systematic failure mode.

The key question is whether this failure is *representational* or *attentional*. If suppressed species are absent from the learned representation, no inference-stage intervention can recover them. If the backbone encodes them but fails to surface them because attention is captured elsewhere, the information is present but inaccessible.

We investigate this through attention-guided iterative masking. If occluding the attended region responsible for the dominant prediction forces the model to redistribute attention, previously suppressed species may surface. Figure 5 shows the baseline: the model attends strongly to the petals of class 1362443, despite the flower occupying a small fraction of the patch, dominating the prediction distribution.

Following rectangular occlusion of the attended region, Figure 6 shows class 1400181 emerging as a prominent prediction, absent from the original ranking despite being visually present elsewhere in the patch.

This experiment provides evidence that attentional failure is a meaningful contributor to patch-level suppression: for this patch, a species absent from the original ranking was recoverable by occluding the dominant attended region, demonstrating that the backbone encoded sufficient information to identify it but did not surface that information under normal inference. The distribution patterns across leaf-dominated plots (Figures 7, 8) provide broader circumstantial support. The species was not invisible to the model; it was outcompeted for attention.

We do not claim this generalises quantitatively. The experiment is diagnostic rather than statistical, one patch, one suppressed species, and the proportion of suppressions that are attentional rather than

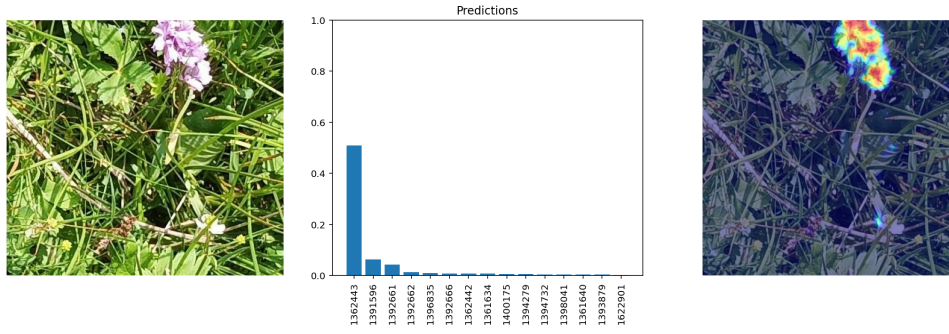


Figure 5: Attention heatmap and prediction distribution for a patch from plot 2024-CEV3-20240602. The model attends predominantly to the flower petals of class 1362443 despite the flower occupying a small fraction of the patch.

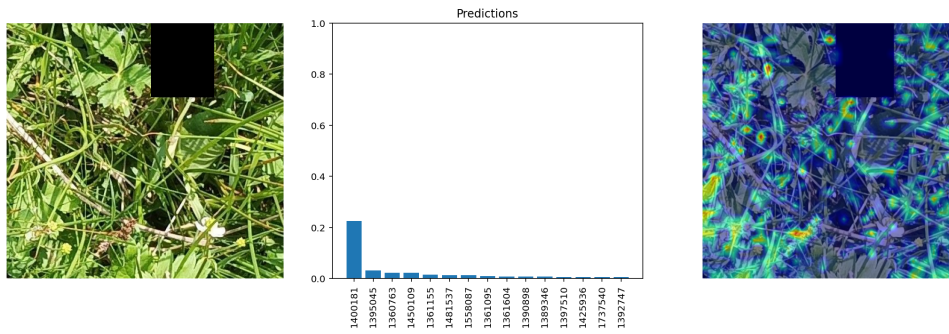


Figure 6: Same patch following occlusion of the dominant attended region. Class 1400181, absent from the original ranking, emerges as a prominent prediction.

representational remains an open question. The practical implication is narrower: if the backbone encodes more than it surfaces in at least a non-trivial fraction of cases, attention-targeted supervision has headroom that purely representational approaches would not. We present the masking experiment as evidence of failure type, not as a proposed solution.

5.2. Plot Level

Having extracted patch-level signals, the aggregation step must converge those signals into a coherent multi-species prediction for the plot. This tier introduces a different class of failure: not whether individual patches carry signal, but whether the aggregation recovers it faithfully.

We examine plot 2024-CEV3-20240602, which contains a combination of flowering plants and vegetative cover with near-complete plant coverage, making it representative of the dense, multi-species scenes the pipeline is designed for. Figure 7 visualises the prediction distributions for each patch under the 16-tile scheme, annotated with the species retained in the final aggregated prediction.

Three dominant prediction behaviours are observed, each corresponding to a distinct failure mode at the aggregation level:

1. **Dominant species capture:** high-confidence species control the aggregation outcome, biasing the final prediction toward visually salient structures.
2. **Visual similarity contamination:** visually similar species cluster in the prediction distributions due to embedding-space proximity, producing feature collision between taxa.
3. **Low-confidence spatial suppression:** species that appear consistently across multiple patches at low confidence are suppressed during aggregation despite their spatial prevalence, reflecting a fundamental limitation of signal-strength-based aggregation.

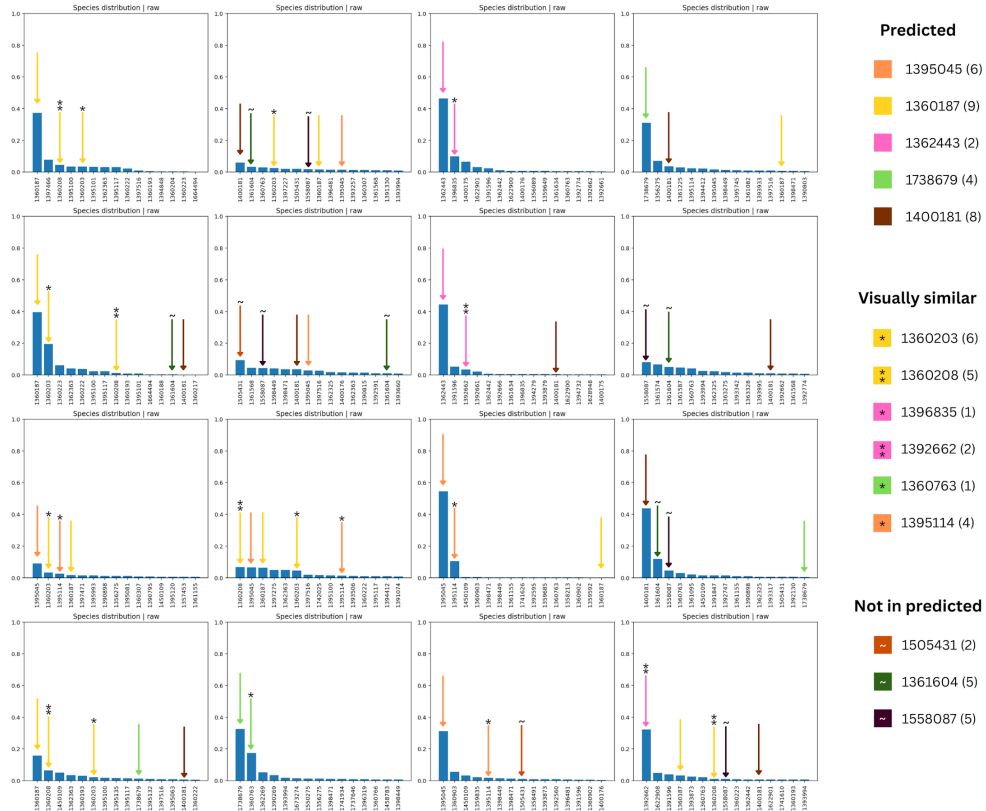
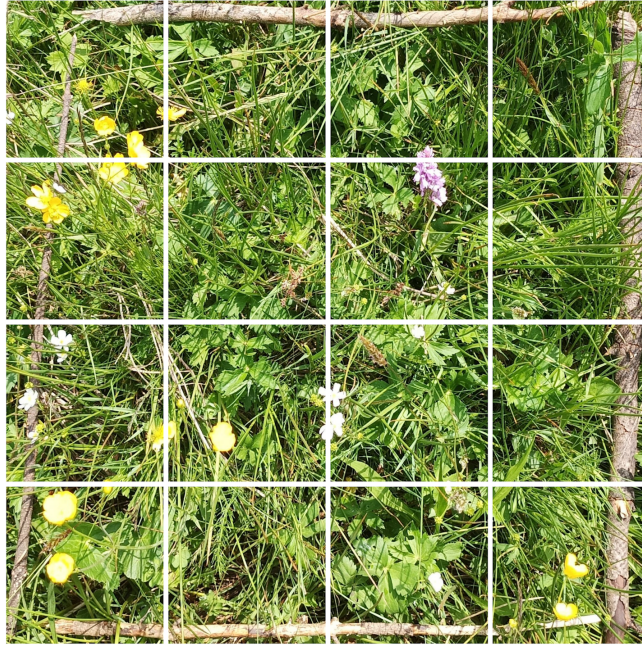


Figure 7: Prediction distributions for each patch in the 16-tile scheme for plot 2024-CEV3-20240602. Species with no marks indicate species retained in the final prediction, visually similar co-occurring species marked with (*), and species appearing consistently across patches but suppressed during aggregation (**).

5.2.1. Dominant Species Capture

The backbone exhibits a strong preference toward floral structures over vegetative organs, as introduced in Section 5.1.2. Flowers are among the most visually discriminative plant organs and are disproportionately represented as high-confidence predictions relative to their spatial coverage within a patch. In several cases, a flower occupying a small fraction of a patch is sufficient to dominate the entire prediction

distribution, while leaf structures occupying the majority of the patch remain weakly represented.

The leaf-dominated plot CBN-can-F5-20230706, shown in Figure 8, provides the complementary evidence. Across fifteen of the sixteen patches, prediction distributions are diffuse and low-confidence, consistent with the model failing to resolve vegetative features. The single exception is the patch containing a visible flower (bottom right), which produces a concentrated high-confidence distribution. The model is not failing because the plot lacks signal; it is failing because the signal it favours (floral structures) is largely absent.

5.2.2. Visual Similarity Contamination

Where dominant species capture reflects a failure of spatial attention, visual similarity contamination reflects a failure of discriminability in the embedding space. Species that are visually similar, particularly congeners sharing floral morphology, occupy proximate regions of the embedding space, producing clusters of co-occurring predictions that cannot be resolved without finer-grained feature supervision.

This is illustrated by the co-occurrence of *Ranunculus acris* L. (1360187), *Ranunculus flammula* L. (1360203), and *Ranunculus macrophyllus* Desf. (1360208) across prediction distributions for plot 2024-CEV3-20240602, shown in Figure 9. The three species produce nearly identical yellow flowers at the scale of a quadrat patch; the backbone cannot reliably separate them and surfaces all three as high-probability candidates whenever any one is present. Whether this represents genuine co-occurrence or feature collision is indistinguishable from prediction distributions alone.

The implication is that post-hoc aggregation cannot resolve visual similarity contamination: no combination of confidence weighting or rank fusion can distinguish between a species that is genuinely present and a species that appears because its congener is present. This failure mode requires an intervention at the representation level; specifically, supervision over morphological features fine-grained enough to separate visually similar taxa, or much simpler, an embedding space similarity matrix as a preliminary filter.

5.2.3. Low-Confidence Spatial Suppression

The third failure mode is in some respects the most recoverable, and the most informative about the limits of signal-strength-based aggregation. Several species appear consistently across multiple patches at low but non-zero confidence, yet are excluded from the final prediction because no individual patch provides a signal strong enough to survive the aggregation threshold.

In plot 2024-CEV3-20240602 (Figure 7), *Crepis aurea* (L.) Cass. (1505431) appears in 2 patches, *Triglochin palustris* L. (1361604) in 5 patches, and *Juncus jacquinii* L. (1558087) in 5 patches, none survive aggregation. The spatial consistency of these predictions is itself evidence of genuine presence: random misclassification noise does not produce the same species repeatedly across spatially distributed patches. The signal is real; the aggregation method cannot hear it over the dominant species.

This is the failure mode rank based aggregation methods, such as Reciprocal Rank Function [22], may solve. By rewarding species that appear consistently across patches at high rank rather than weighting by absolute confidence, RRF can surface spatially consistent low-amplitude signals that signal based methods suppress.

We expect that even RRF would struggle to fully recover these species, given that the signals are too weak for rank-based recovery to overcome the signal-to-noise ratio imposed by dominant species. The failure is partially aggregation-level and partially representational, which is why it sits at the boundary of what post-hoc methods can address.

5.3. Site Level

The patch and plot level analyses treat each quadrat in isolation. However, the test set encodes a spatial structure that extends beyond individual plots: plots sharing a site and transect were photographed in physical proximity on the same date and sample from the same underlying ecological community.

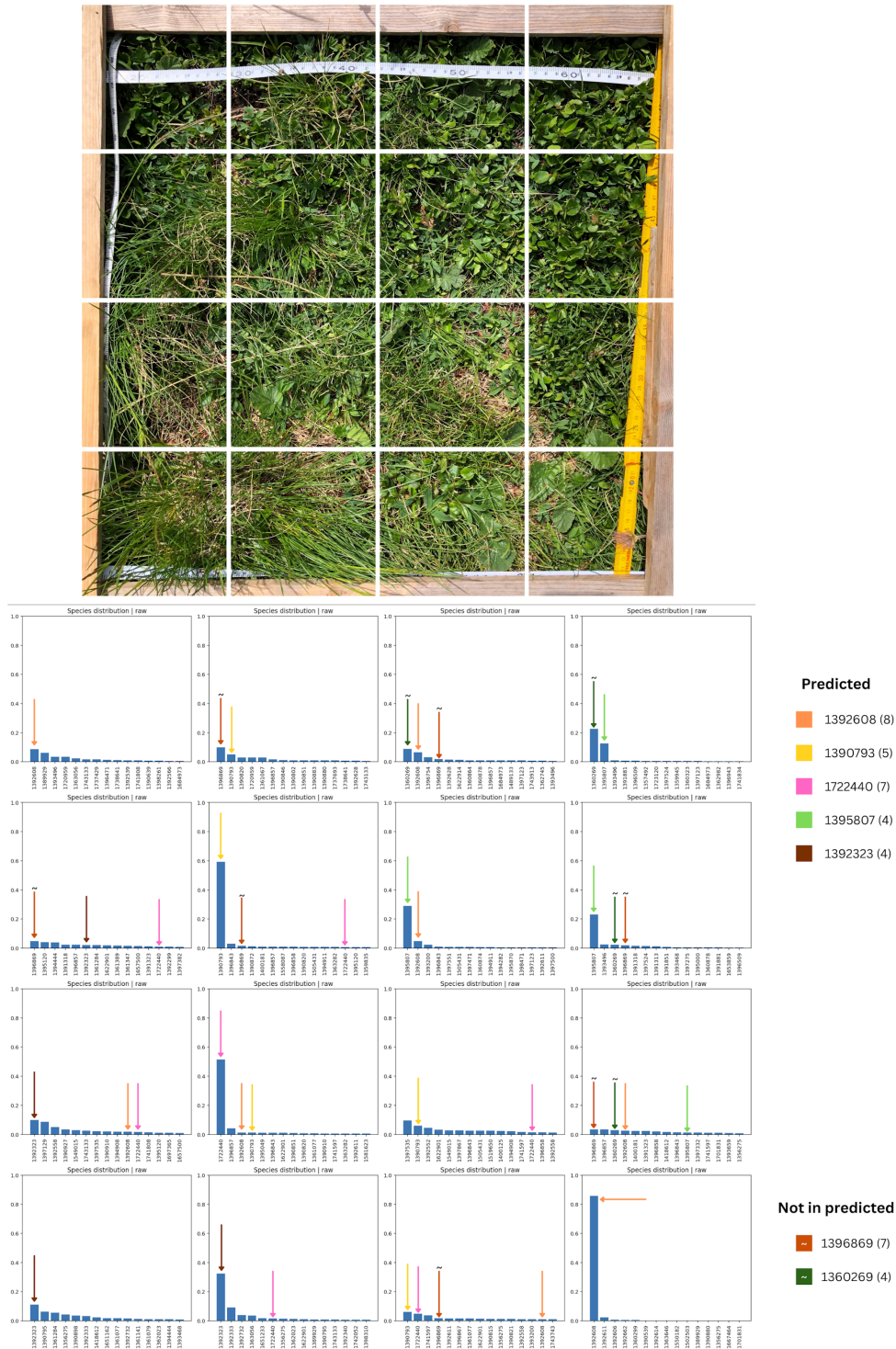


Figure 8: Prediction distributions for each patch in the 16-tile scheme for plot CBN-can-F5-20230706. The plot is predominantly vegetative with minimal floral content. Fifteen of sixteen patches produce diffuse low-confidence distributions; the single high-confidence patch corresponds to the only visible flower.

Exploiting this structure introduces a third tier of inference, operating not on visual features but on ecological plausibility.

In parsing the filename structure of the test set, we reconstructed the following four-level spatial hierarchy:

- **Site:** the top-level ecological location (e.g., CBN-Pd1C, LISAH, RNNB)
- **Transect:** a lettered or numbered section within a site (e.g., AB, CD, EF for CBN sites)



Figure 9: Classified species (denoted with an asterisk) and their visually similar co-occurring species for plot 2024-CEV3-20240602. Congeneric species with near-identical floral morphology cluster in the prediction distributions regardless of which species is actually present.

- **Plot:** an individual quadrat within a transect (e.g., A1, B3)
- **Date:** the capture date, with some sites containing repeated observations across multiple years

Table 12 summarises the resulting organisation across the full test set.

Table 12

Spatial structure of the test set reconstructed from filename conventions. CBN sites follow a consistent 2×8 transect grid; GUARDEN, LISAH, and OPTMix sites use numbered plots without a fixed grid structure, though spatial clusters are present. RNNB follows a 1×10 grid per transect.

Source	Locations	Transect structure	Images
CBN	4 (PdIC, can, Pla, Pyr)	3 per location (2×8 grid)	1476
GUARDEN	2 (CBNmed, AMB)	Numbered clusters	201
LISAH	4 (BOU, BVD, JAS, PEC)	Numbered clusters	208
OPTMix	12	Numbered clusters	78
RNNB	1	8 transects (1×10 grid)	141
2024-CEV3	Standalone	—	1
Total			2105

Figure 10 shows co-located plots from transect CBN-PdIC-[AB]-20200722 reconstructed into a spatial mosaic. Adjacent plots share continuous vegetation across their boundaries, confirming that they sample from a common ecological community. This is the empirical motivation for site-level inference: species confirmed in neighbouring plots constitute informative prior knowledge about what should, and should not, appear in any individual plot within the same transect.

Not all sources provide equally dense spatial coverage. CBN sites offer the most complete temporal and spatial records, with the same plots revisited across multiple years, providing both geographical and temporal laterality for community-level reasoning. GUARDEN, LISAH, and OPTMix sites provide spatial clusters without consistent grid structure, and not all dates are available across all sites.

5.3.1. Ecological Implausibility and Community-Level Filtering

The site-level failure mode is distinct from those at the patch and plot tiers: it is not a failure of signal extraction or aggregation, but a failure of ecological plausibility. The backbone will confidently predict species that are visually consistent with training data but ecologically impossible given the site context.

The clearest illustration is plot RNNB-1-2-20240117, shown in Figure 11. The plot is predominantly sandy and barren, with minimal vegetation. The model produces high-confidence predictions for

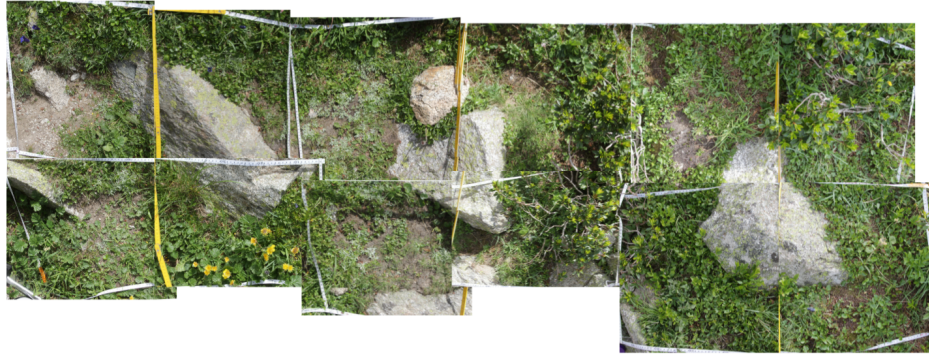


Figure 10: Spatial reconstruction of co-located plots from transect CBN-Pd1C-[AB]-20200722, produced by overlaying plots A1–A8 (top row) and B1–B8 (bottom row). Continuous vegetation is visible across plot boundaries, confirming shared ecological community membership.

Posidonia oceanica (L.) Delile across nearly every patch: a seagrass species whose training images are captured against sandy underwater substrates, texturally similar to the plot surface (Figure 12). No plot-level aggregation strategy can identify this as an error; the signal is strong, spatially consistent, and survives every confidence threshold.

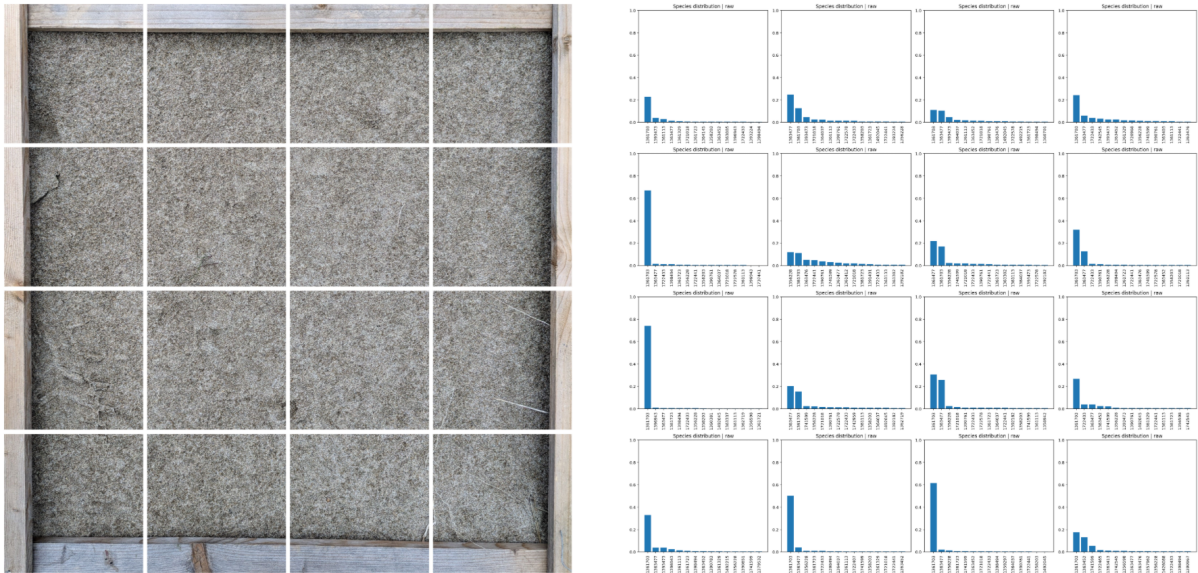


Figure 11: Prediction distributions for plot RNNB-1-2-20240117. The model produces high-confidence predictions for *Posidonia oceanica* across nearly all patches due to textural similarity between the sandy plot surface and the species' training imagery as shown in Figure 12

Community-level ecological filtering addresses this directly. The Ellenberg Indicator Values for Europe (EIVE) [23] encode five ecological gradients: light, temperature, soil moisture, soil reaction, and nitrogen supply, for over 14,000 taxa, expressed as per-species niche optima and niche breadths. A species confirmed at the site level occupies a region of this six-dimensional ecological space; a candidate species whose niche is incompatible with the confirmed community can be suppressed regardless of its visual confidence.

Rather than the species-to-species pairwise comparison of [18], we propose modelling the confirmed dominant species from co-located plots as a Gaussian Mixture Model (GMM) in EIVE space, inspired by the n-dimensional hypervolume framework [24] and community weighted mean approaches [25].

We note that Semenova [18] identified community-level ecological modelling as a direction for future work; our proposal was developed independently before this observation was encountered, and we



Figure 12: Species 1361703, *Posidonia oceanica* (L.) Delile

include the acknowledgement here in the interest of transparency. Independent convergence on the same direction provides additional motivation that moving from pairwise to community-level ecological reasoning is the natural next step for this problem.

Candidate species are evaluated against this mixture via the Bhattacharyya overlap [26], generalised from pairwise comparison to mixture-level overlap. The compatibility score admits an analytic closed form under the Gaussian product identity:

$$S(j) = \frac{1}{N} \sum_{i=1}^N \mathcal{N}(\mu_j \mid \mu_i, \Sigma_i + \Sigma_j) \quad (2)$$

Where $\mathcal{N}$ denotes the Gaussian density evaluated over the five EIVE axes, μ_i is the niche centroid of confirmed species i , μ_j is the niche centroid of candidate species j , and Σ_i, Σ_j are diagonal covariance matrices encoding per-axis niche breadth.

A candidate species, for example, *Posidonia oceanica* whose EIVE profile reflects marine salinity, minimal light competition, and absent terrestrial moisture gradients, would score near zero against a community mixture drawn from Pyrenean grassland plots. The visual confidence that defeats plot-level aggregation is irrelevant at this tier; the ecological signal does the work that the visual signal cannot.

This formulation also enables recovery of species missed at the plot level. Candidates suppressed by the z-score threshold or overwhelmed by dominant species can be re-evaluated for ecological compatibility with the confirmed community: if their EIVE profile is consistent with the broader site niche, their suppression is more likely to reflect aggregation failure than genuine absence. The site level therefore functions both as a filter, removing ecologically implausible predictions and as a rescue mechanism for ecologically plausible species that post-hoc aggregation could not surface.

6. Future Work

Two research directions follow directly from the failure mode analysis.

6.1. Concept Bottleneck Models for Morphological Supervision

The attentional failure and visual similarity contamination at the patch and plot levels share a common root: the backbone discriminates primarily through floral structures, and its embedding space is too coarse to separate congeneric taxa. Standard domain adaptation addresses distributional shift but not attentional bias; a model that encodes suppressed species but fails to surface them will not benefit from distribution alignment alone.

Concept Bottleneck Models (CBMs) [27] offer a targeted intervention by routing predictions through an intermediate layer of human-interpretable morphological concepts: leaf shape, growth form, venation pattern, and flower arrangement. This serves two functions.

First, concept-level supervision distributes attention across multiple discriminative features rather than locking onto floral structures. Second, morphological concepts are observable at both training

and quadrat scale even when raw visual appearance differs substantially, providing a domain-invariant intermediate representation. MONET [28] demonstrated in dermatology that concept-grounded predictions are both more accurate and more amenable to expert validation than saliency-based explanations; a standard ecological AI is increasingly expected to meet.

BioCLIP2 [29] and BioCAP [30] make concept grounding tractable without manual annotation: BioCAP derives morphological concept descriptions from Wikipedia entries refined by language models, and both models demonstrate substantially stronger zero-shot performance on plant identification tasks than general-purpose vision-language models [31], confirming their representational alignment with the target domain.

6.2. Unsupervised Domain Adaptation

The distributional mismatch between single-species close-up training imagery and top-down quadrat plots underlies all four failure modes. The fourth-place 2026 submission demonstrated a 6% gain from FixMatch [32] with only 5 epochs on unlabelled quadrat data. Direct evidence that pseudo-label adaptation is viable here even under weak supervision.

Pseudo-label reliability is the critical constraint under fine-grained visual confusion, where incorrect labels concentrate precisely at the congeneric decision boundaries identified as failure mode 2. Uncertainty-aware selection [33] and neighbourhood consistency filtering [34] improve label quality at these boundaries; the latter directly targets visual similarity contamination by confirming that a proposed pseudo-label is supported by its embedding-space neighbours before acceptance.

CBM and UDA are complementary: CBM addresses what the model attends to and how finely it discriminates; UDA addresses the distributional gap. A combined system with site-level GMM filtering as an ecological plausibility gate for pseudo-label selection represents the full proposed solution to the identified failure modes.

7. Conclusion

We presented a systematic ablation of tiling granularity for multi-species plant identification in vegetation quadrats, establishing that fixed multi-scale tiling centred around ecologically motivated spatial scales achieves 36.051% macro-averaged F1, 12th of 105 teams, without backbone modification. The key finding is that ecological spatial scale consistency is a more important determinant of tiling performance than pixel-level information fidelity, and that two independent preprocessing approaches converge on the same performance ceiling.

The central finding of this work is not the score. The standard framing attributes difficulty to domain shift and the single-to-multi-label mismatch. These are descriptions of the problem’s surface, not its structure. The failure mode analysis presented here goes one level deeper, identifying four mechanistically distinct failure modes across three tiers of inference:

- **Floral salience bias** at the patch level: the masking experiment demonstrates that suppressed species are encoded but not surfaced. The failure is attentional, not representational; representation learning alone is therefore insufficient.
- **Visual similarity contamination** at the plot level: the embedding space is too coarse to separate congeneric taxa. No aggregation strategy can resolve feature collision; finer-grained discriminative supervision is required.
- **Low-confidence spatial suppression** at the plot level: partially addressable post-hoc through rank-based aggregation, but ultimately bounded by representation quality. The fix is a stronger backbone, not a better aggregation rule.
- **Ecological implausibility** at the site level: invisible to every visual method. The GMM over EIVE space provides external ecological knowledge in a principled, probabilistic form that scales to full community structure.

Failure modes 1 and 2 are representational in origin and require model-centric interventions. Failure mode 3 sits at the boundary between aggregation and representation. Failure mode 4 is ecological in origin and requires external domain knowledge regardless of backbone quality. This decomposition moves beyond domain shift as a monolithic explanation and constrains the design space for future work.

Declaration on Generative AI

During the preparation of this work, the author(s) used ClaudeAI in order to check grammar, spelling and tone of voice. 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] M. Chytrý, Z. Otýpková, Plot sizes used for phytosociological sampling of european vegetation, *Journal of vegetation science* 14 (2003) 563–570.
- [2] S. K. Wiser, N. Spencer, M. De Caceres, M. Kleikamp, B. Boyle, R. K. Peet, Veg-x—an exchange standard for plot-based vegetation data, *Journal of Vegetation Science* 22 (2011) 598–609.
- [3] D. Ellenberg, D. Mueller-Dombois, *Aims and methods of vegetation ecology*, volume 547, Wiley New York, 1974.
- [4] J. S. Rodwell, J. N. C. C. (GB), *National vegetation classification: Users' handbook*, Joint nature conservation committee Peterborough, 2006.
- [5] J. Wäldchen, M. Rzánný, M. Seeland, P. Mäder, Automated plant species identification—trends and future directions, *PLoS computational biology* 14 (2018) e1005993.
- [6] 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.
- [7] 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.
- [8] G. Martellucci, I. Moummad, H. Goëau, P. Bonnet, F. Vinatier, A. Joly, Overview of PlantCLEF 2026: Identify multi-species plants in images of vegetation plots, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [9] L. Pícek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, Springer, 2026.
- [10] S. Chulif, H. A. Ishrat, Y. L. Chang, S. H. Lee, Patch-wise inference using pre-trained vision transformers: Neuron submission to plantclef 2024, in: *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, 2024.
- [11] H. A. Ishrat, S. H. Lee, Y. L. Chang, C. K. Chin, Post-hoc aggregation as a competitive alternative to model-centric pipelines: neuron submission to plantclef 2025, *Working Notes of CLEF (2025)*.
- [12] V. Espitalier, Preprocessing is all you need: Theheartofnoise submission to plantclef 2025, *Working Notes of CLEF (2025)*.
- [13] A. Affouard, H. Goëau, P. Bonnet, J.-C. Lombardo, A. Joly, Pl@ ntNet app in the era of deep learning, in: *ICLR: International Conference on Learning Representations*, 2017.
- [14] 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.

- [15] M. Gustineli, A. Miyaguchi, I. Stalter, Multi-label plant species classification with self-supervised vision transformers, arXiv preprint arXiv:2407.06298 (2024).
- [16] H. Herasimchyk, R. Labryga, T. Prusina, Multi-label plant species prediction with metadata-enhanced multi-head vision transformers, arXiv preprint arXiv:2508.10457 (2025).
- [17] S. Foy, S. McLoughlin, Utilizing dino v2 for domain adaptation in vegetation plot analysis, Working Notes of CLEF (2024).
- [18] N. Semenova, Optimizing a vision transformer with ecological context for multi-label plant species identification, Working Notes of CLEF (2025).
- [19] M. Gustineli, A. Miyaguchi, A. Cheung, D. Khattak, Tile-based vit inference with visual-cluster priors for zero-shot multi-species plant identification, arXiv preprint arXiv:2507.06093 (2025).
- [20] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, *Advances in neural information processing systems* 30 (2017).
- [21] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, *Structural safety* 31 (2009) 105–112.
- [22] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.
- [23] J. Dengler, F. Jansen, O. Chusova, E. Hüllbusch, M. P. Nobis, K. Van Meerbeek, I. Axmanová, H. H. Bruun, M. Chytrý, R. Guarino, et al., ecological indicator values for europe (eive) 1.0, *Vegetation Classification and Survey* 4 (2023) 7–29.
- [24] B. Blonder, C. Lamanna, C. Violle, B. J. Enquist, The n-dimensional hypervolume, *Global Ecology and Biogeography* 23 (2014) 595–609.
- [25] D. Zelený, Which results of the standard test for community-weighted mean approach are too optimistic?, *Journal of Vegetation Science* 29 (2018) 953–966.
- [26] A. Bhattacharyya, On a measure of divergence between two multinomial populations, *Sankhyā: the indian journal of statistics* (1946) 401–406.
- [27] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: *International conference on machine learning*, PMLR, 2020, pp. 5338–5348.
- [28] C. Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, S.-I. Lee, Transparent medical image ai via an image–text foundation model grounded in medical literature, *Nature medicine* 30 (2024) 1154–1165.
- [29] J. Gu, S. Stevens, E. Campolongo, M. Thompson, N. Zhang, J. Wu, A. Kopanav, Z. Mai, A. White, J. Balhoff, et al., Bioclip 2: Emergent properties from scaling hierarchical contrastive learning, *Advances in Neural Information Processing Systems* 38 (2026) 102778–102811.
- [30] Z. Zhang, X. Ma, A. Chowdhury, E. G. Campolongo, M. J. Thompson, N. Zhang, S. Stevens, H. Lapp, T. Berger-Wolf, Y. Su, et al., Biocap: Exploiting synthetic captions beyond labels in biological foundation models, arXiv preprint arXiv:2510.20095 (2025).
- [31] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, et al., Bioclip: A vision foundation model for the tree of life, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19412–19424.
- [32] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C.-L. Li, Fixmatch: Simplifying semi-supervised learning with consistency and confidence, *Advances in neural information processing systems* 33 (2020) 596–608.
- [33] M. N. Rizve, K. Duarte, Y. S. Rawat, M. Shah, In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning, arXiv preprint arXiv:2101.06329 (2021).
- [34] B. Tian, S. Lai, L. Li, Z. Shuai, R. Guan, T. Wu, Y. Yue, Pepl: Precision-enhanced pseudo-labeling for fine-grained image classification in semi-supervised learning, in: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.