Bridging Single-Label Training and Multi-Label Inference with Vision Transformers for Vegetation Monitoring

Carolin A. Rickert[1,*], Tobias Haar[1]

[1]MicroNova AG, Unterfeldring 6, 85265 Vierkirchen, Germany

Abstract

Automated vegetation monitoring is important for biodiversity assessment and ecosystem analysis. However, machine learning-based vegetation classification is often affected by strong domain shifts between training and inference data: Since creating annotated multi-species quadrat datasets is highly time-consuming, models are commonly trained on easier-to-obtain single-species plant images and later applied to complex high-resolution quadrat images containing multiple species and highly variable environmental conditions. In this working note, we investigate different strategies for multi-species image classification under such domain shift in the context of the PlantCLEF 2026 challenge. Using a pretrained DINOv2 Vision Transformer as a baseline model, we evaluate several tiling, pooling, and preprocessing strategies to improve robustness during inference. Our experiments show that tiling-based inference with max pooling and optimized confidence thresholds significantly improves classification performance compared to the provided baseline. Additionally, adapting jpg compression statistics during preprocessing further increases the F1 scores, while hierarchical multi-scale tiling achieved the best overall performance. We additionally explore LoRA-based domain adaptation and retrieval-augmented classification. However, under our computational constraints these methods did not outperform the strong tiling baseline. Overall, the experiments suggest that robust vegetation classification under strong domain shift can be substantially improved through tailored preprocessing and inference strategies, while the explored domain adaptation and retrieval-based approaches provide promising directions for future research.

Keywords

vegetation plots, DINOv2, domain shift, computer vision, biodiversity

1. Introduction

The world as we know it today is composed of complex ecosystems that build the foundation for life on our planet. Terrestrial ecosystems for instance play a key role in global and regional climate regulation as their ability to absorb atmospheric carbon dioxide during photosynthesis contributes to the stability of climate and directly counteracts global warming. In other words, the stability of terrestrial ecosystems is inseparably tied to the overall stability of the planet.[1] The resilience of terrestrial ecosystems, in turn, is strongly associated with the level of biodiversity they exhibit. Monitoring such biodiversity – in which plant diversity represents a particularly important indicator [2, 3]– is therefore essential for assessing the health and sustainability of life-support systems of the Earth.[4]

An essential part of vegetation monitoring for the assessment of biodiversity involves direct field measurements in which researchers assess species composition, plant abundance, biomass, canopy cover, leaf area, and other ecological indicators [5, 6]. They typically include dividing the monitored area into transects and quadrats of fixed size. Images are acquired for each quadrat and subsequently annotated with all observed plant species. Traditionally, those annotations are conducted fully manually, which is not only time consuming but also error prone and heavily dependent on expert knowledge. Hence, approaches involving automated plant identification, such as machine learning driven methods, are promising candidates for modern vegetation monitoring systems.

However, the diverse and complex composition of the data to be analyzed comes with several challenges: Due to the natural spatial composition of terrestrial ecosystems, plants may occupy only a small

CLEF Working notes, Jena, Germany, on September 21-24, 2026.

*Corresponding author.

carolin.rickert@micronova.de (C. A. Rickert); tobias.haar@micronova.de (T. Haar)

0000-0001-5532-3980 (C. A. Rickert)

fraction of the total quadrat image area. This results in an imbalance between plant and non-plant information, which can increase ambiguity during species identification and negatively affect classification performance. Furthermore, seasonal morphological variations among plant species introduce substantial intra-class variability across quadrat images, thereby increasing classification uncertainty.[6] This is aggravated by the afore-mentioned fact that annotation requires manual identification and verification of all species within each quadrat by an expert, making field data collection highly labor-intensive and resulting in very limited data availability and possible class-imbalance. Consequently, machine learning models may develop biased predictions, while common mitigation strategies such as undersampling or augmentation are often constrained by the limited amount of available test data and the presence of multiple species in a single image. [6]

Luckily, in contrast to the scarcity of annotated multi-species quadrat images, there do exist large-scale high-quality datasets of single-plant images. Indeed, recent machine learning approaches have demonstrated substantial progress in vegetation classification accuracy using those datasets.[7] In particular, self-supervised Vision Transformer architectures such as DINOv2 have gained increased attention for multi-class plant classification tasks based on quadrat imagery.[8] Interest in these approaches has also increased through annual PlantCLEF competitions organized by the LifeCLEF research initiative since 2024. In this context, DINOv2 models are commonly fine-tuned on single-plant images before being applied to multi-species quadrats for inference.[9, 10]

However, the use of different types of data in training and inference introduces a substantial domain shift: Whereas the training data consists of single-species, close-up, curated plant observations, the test data comes from dense vegetation plot images containing multiple interacting species in natural ecological scenes. This mismatch between object-centric, clean images and complex, multi-label ecosystem imagery leads to a significant distribution shift in visual appearance, context, and label structure across datasets. Additionally, differences in image resolution and experimental setups between collected quadrat datasets and single-plant training images has shown to have a significant effect on model performance when strong domain shift is present.[10, 11, 12]

In this study, we describe our contribution to the PlantCLEF 2026[6] competition, held within the framework of the LifeCLEF 2026 lab[10], where we participated under the team name *MicroNovians*. We show that based on a DINOv2 model pre-trained on single-label data, a well-tuned tiling approach can indeed partly bridge the gap of the domain shift and significantly increase the classification performance of the model. We additionally highlight two more advanced methods: namely, Low-Rank Adaption of the Vision Transformer and Retrieval-augmented classification. Those advanced methods, however, were tested and implemented under severely restricted computational resources, which is why their full potential could not completely be assessed. Overall, the experiments show that careful preprocessing and tiling strategies can already substantially improve vegetation classification under strong domain shift and form a solid base for more advanced approaches.

2. Methodology

2.1. Datasets

2.1.1. Single-label dataset (training set)

A mono-label dataset is provided that comprises images of individual plants from southwestern Europe. In total, it covers 1.4 million jpg images of 7,806 species. Since this dataset was used to pre-train the provided model (see section 2.2), we analyzed basic image encoding properties in more detail (**Table 1**): First, chroma sub-sampling was obtained without decoding the image. Therefore, the first 64 KB of each file were scanned for SOF0 (baseline) or SOF2 (progressive) segments and the horizontal and vertical sampling factors for the luminance (Y) and chrominance (Cb, Cr) components were read. The obtained values were then mapped to common factor combinations of standard labels such as 4:4:4, 4:2:0, or 4:2:2 (non-standard cases were marked n/a). Second, an estimate of the approximate jpg quality factor (q) was derived from the first luminance DQT (Define Quantization Table, marker 0xFFDB, 8-bit precision):

the 64 coefficients were compared to the JPEG Annex K luminance table scaled for each candidate $q \in \{1,\ldots,100\}$ using the usual libjpeg-style scaling rules (scale = 5000/q for $q < 50$, else $200 - 2q$). The q that minimized the sum of squared differences between the file's table and the scaled reference table was taken as quality estimate.

Table 1
Chroma subsampling and jpg compression quality of the labeled training images.

	scheme/quality	# of images
chroma subsampling	4:2:0	1,408,008
	4:4:4	2
	n/a	23
jpg quality factor	76	7
	78	1,102,476
	95	305,550

2.1.2. Multi-label dataset (test set)

The dataset to be classified is a collection of 2,105 high-resolution (mostly) quadrat images obtained during vegetation monitoring of various floristic environments, mainly Pyrenean and Mediterranean flora. Importantly, it comprises a high variance in angles, weather conditions, and experimental setups (such as wooden frames or measuring tape around the box). These images can naturally include multiple different species, which is in contrast to the previously described training dataset, that contains single-species images only. However, specific labels were not provided.

With sizes ranging from 1500×1500 px to mostly 3000×3000 px, the images of the test set are considerably larger than the training images. To even better understand the domain shift that has to be dealt with, we again assessed the chroma sub-sampling schemes and the jpg compression quality as described before for the training dataset (**Table 2**).

Table 2
Chroma sub-sampling and jpg compression quality of the unlabeled test images (quadrat images).

	scheme/quality	# of images
chroma subsampling	4:2:0	662
	4:2:2	1,438
	other	5
jpg quality factor	80–82	673
	95–98	1,398

2.1.3. Unlabeled pseudo-quadrat dataset

In addition to the datasets mentioned above, the organizers of PlantCLEF 2026 provided an unlabeled pseudo-quadrat dataset that they assume could be helpful for domain adaption. Those images are framed similarly to the test quadrats but the area size can differ. In total, 212,782 pseudo-quadrat images (individually selected from the LUCAS Cover Photos 2006-2018 collection) were provided without any species labels. Importantly, the image sizes of the included images were subject to a significant variance, ranging from 400×400 to up to 5000×5000 pixels. In alignment with the previous datasets, we analyzed the chroma sub-sampling and jpg quality (**Table 3**).

Table 3
Chroma sub-sampling and jpg compression quality of the pseudo-quadrat images.

	scheme/quality	# of images
chroma subsampling	4:2:0	178,172
	4:2:2	30,013
	other or n/a	4,597
jpg quality factor	∼48	25,535
	73–76	120,884
	>90	44,880
	other	21,483

2.2. Pre-trained baseline model

A pre-trained model was provided by the organizers of the competition. The referred model is a DINOv2 Vision Transformer (ViT-Base/14) with four registers and a patch size of 14×14 (vit_base_patch14_reg4_dinov2.lvd142m). The ViT-base architecture consists of 12 transformer encoder blocks with 12 attention heads each and 768-dimensional token embeddings; images are provided as RGB inputs, typically at 518×518 pixels. It was trained with the self-supervised learning (SSL) DINOv2 method and the full model (backbone and classification head) was then fine-tuned using the training dataset described before. For fine-tuning, an Adam optimizer (learning rate: 8×10^{-5}, weight decay: 0.0, momentum: 0.9) and cosine learning rate scheduling were used. The batch size was set to 144, with label smoothing (0.1) and strong data augmentations (mixup: 0.8, cutmix: 1.0, horizontal flip: 0.5). Training was performed on 7,806 classes with a fixed random seed (42) for reproducibility. Exponential Moving Average (EMA; decay: 0.9998) was used for stabilization.

2.3. Tiling approaches

In previous studies, it was shown that analyzing the full multi-species image by the model does not yield desired results – which was expected based on the significant differences of the training and testing datasets as illustrated before. To compensate this, different tiling approaches were tested in which the image is subdivided into multiple tiles, that are then analyzed individually and the prediction results are afterwards accumulated to derive a global prediction per image (see section 2.4).

2.3.1. Basic 4×4 tiling

For basic 4×4 tiling, the images were resized isotropically to a width of 2072 pixels (4x the model input width of 518 px), while preserving the original aspect ratio. If needed, padding was then applied using "reflect" mode to extend the image to a final size of 2072×2072 pixels (so that both dimensions are multiples of 518 px). The resized image was then split into a 4×4 grid, resulting in 16 tiles of 518×518 pixels without overlap. Tiling was performed after resizing and padding to ensure that each tile corresponded to a consistent spatial region across images.

2.3.2. Auto-grid tiling

For auto-grid tiling, the number of columns and rows was chosen independently from the original image width W and height H and the model tile size S of 518 px: along each axis, the tile count was set to $\mathrm{round}(W/S)$ and $\mathrm{round}(H/S)$ respectively (half-up rounding), lower-bounded by 1 and upper-bounded to 6 tiles per axis to avoid excessively large grids on very large scans. The image was then mapped to a rectangular canvas of size $(C \cdot S) \times (R \cdot S)$, where C and R are the chosen column and row counts: it was resized isotropically so that the full image fit inside that rectangle while preserving aspect ratio, and center padding was applied (default mode: reflect-style padding where valid, otherwise edge padding for degenerate sizes) until the canvas reached exactly $(C \cdot S) \times (R \cdot S)$. The padded

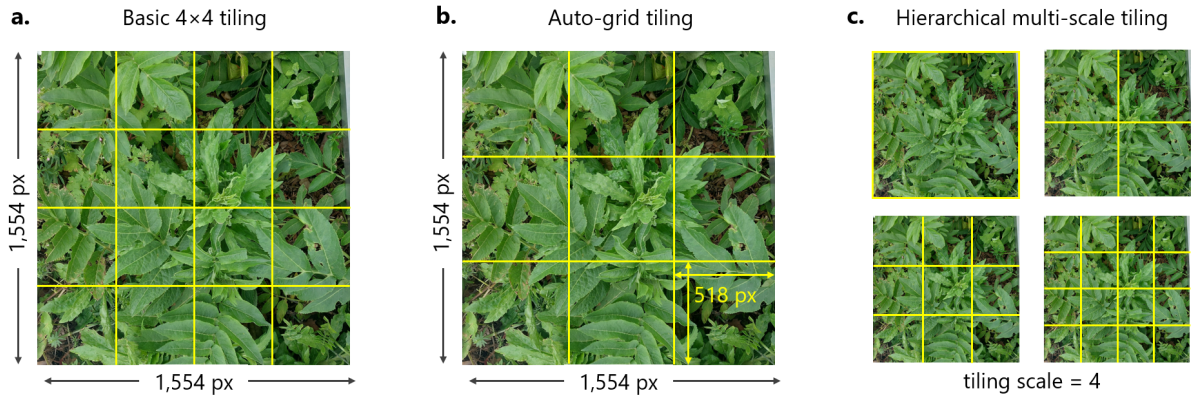

Figure 1: Visualization of different tiling approaches. For inference, quadrat images were either tiled with a fixed 4×4 grid (a), or with a grid pattern adapted to the image dimensions (b), or using hierarchical grids, starting with a fixed scale (i.e., the maximum number of tiles per dimension) and partitioning the image in progressively coarser grids until a 1×1 grid (full image) is reached (c).

canvas was partitioned into an $R \times C$ grid in row-major order; along each axis the intervals were as equal in length as possible (remainder pixels distributed over the leading strips).

2.3.3. Hierarchical multi-scale tiling

For hierarchical tiling, multiple square canvases were used at scales $s = 1, \ldots, N$: at scale s, the image was resized isotropically to fit inside an $(sS) \times (sS)$ square, then padded (same padding modes as above) to exactly $(sS) \times (sS)$. That square was split into an $s \times s$ grid of non-overlapping $S \times S$ tiles in row-major order (unless overlap was enabled, in which case intervals along each axis were widened symmetrically toward neighbors, analogous to the grid strategy). This procedure was repeated for every scale, yielding $1^2 + 2^2 + \cdots + N^2$ tiles in total: first the single global tile at scale 1, then the four tiles at scale 2, and so on. Because each scale uses its own resized-and-padded representation of the full image, coarse scales emphasize whole-quadrat context while finer scales subdivide the field of view at higher spatial resolution.

2.4. Pooling strategy

As described in section 2.3, tiling approaches were conducted to overcome the domain shift between single-label training and multi-label inference. However, this leads to the fact, that multiple prediction vectors are obtained for every full images (e.g. 4×4 tiling leads to 16 tiles, hence 16 prediction vectors). To derive one final prediction from those multiple vectors, different pooling approaches were used.

Max-over-tiles aggregation was implemented with the goal to promote a class to the image level if any tile assigns it a sufficiently high score. Therefore, tile-level logits $L_{t,c}$ were aggregated into a single image-level logit vector $\ell \in \mathbb{R}^C$ ($C :=$ number of classes) by element-wise max pooling over the tile dimension t (for each class c, $\ell_c = \max_t L_{t,c}$). The class probabilities were then computed as $p = \mathrm{softmax}(\ell)$. Based on those probabilities, a species was included in the full-image prediction if $p_c \geq \tau$ (τ being an adjustable threshold) up to a maximum list length of k classes. If not stated otherwise, $k = 8$ was used. In case no class met the probability threshold, fallback to the argmax class was used so that each image still received at least one prediction.

2.5. Image preprocessing

After tiling, each tile was obtained as an RGB PIL.Image and passed through a fixed torchvision pipeline: it was first resized to a square of a pre-defined size (if not stated otherwise, the input size was set to 518×518 px), converted to a tensor in [0,1], and standardized with ImageNet RGB mean and std. No

extra color space conversion was applied. If stated explicitly in the respective trials, an in-memory jpg round-trip was conducted before resizing the images. Therefore, each tile was encoded with Pillow as JPEG and immediately decoded back to RGB. Importantly, the jpg quality factor (q ∈ {1,...,100}) and the chroma sub-sampling scheme could be specified (e.g. 4:4:4, 4:2:2, or 4:2:0). When enabled, this step approximates lossy JPEG statistics similar to typical images used in training.

2.6. Evaluation

The overall performance is assessed using a macro-averaged F1 score. Generally, the F1 score provides a suitable balance between precision and recall in multi-label classification [13]. The F1 score of each image j is first calculated as

$$\mathrm{F1}_j = \frac{2\,\mathrm{Precision}_j\,\mathrm{Recall}_j}{\mathrm{Precision}_j + \mathrm{Recall}_j}. \tag{1}$$

Here,

$$\mathrm{Precision}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}, \qquad \mathrm{Recall}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j}, \tag{2}$$

with TP_j (true positives), FP_j (false positives), and FN_j (false negatives) computed for image j.

To account for imbalances in the number of images acquired from different ecological environments, a two-step macro-averaging is performed. First, the F1 scores are averaged over all quadrat images belonging to the same transect (where T_i is the number of quadrats in transect i). Second, the final score is derived by averaging across all N transects:

$$\mathrm{F1}_{\mathrm{final}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{j=1}^{T_i}\mathrm{F1}_j\right). \tag{3}$$

Importantly, the F1 scores were computed automatically after submitting a solution to the challenge; preliminary assessment of the scores was not possible due to the lack of ground-truth labels for the test dataset.

3. Results & Discussion

3.1. Preprocessing and pooling optimization

In a first set of trials, we aim at establishing a baseline for further experiments and at gaining an understanding of how different pipeline features influence the inference performance. Therefore, we combine the provided pre-trained Vision Transformer with a basic 4x4 tiling approach that was previously reported to be efficient [14]. After processing each tile individually, the final prediction for each image is then derived from the resulting 16 tiles by applying max pooling with an optimized threshold τ as recommended by multiple previous studies [15][16] (for details, please refer to section 2.4). Indeed, with this approach, we obtain a private F1 score of 0.31925 (**Fig.1a** blue cross), which is well above the reference tiling score of 0.21708 achieved with the example tiling approach provided by the organizers as a benchmark (**Fig. 1a** grey cross).

We assume that the improvement in performance of our approach can be attributed to several key differences in the inference and aggregation pipeline: First, the reference method applies a softmax normalization and quite strict top-k (k = 2) and threshold-based filtering at the individual tile level prior to aggregation. In contrast to this, our approach performs pooling directly on the unnormalized logits across tiles, which might avoid premature information loss caused by local competition between classes within each tile and preserve globally relevant but locally ambiguous signals. Second, we choose to perform tiling without stride to ensure controlled coverage with minimal correlation between tiles to avoid the amplification of false-positives or noise. Lastly, we use a tuned decision threshold, while we assume that the reference method relies on a fixed heuristic threshold at the tile level. We believe that

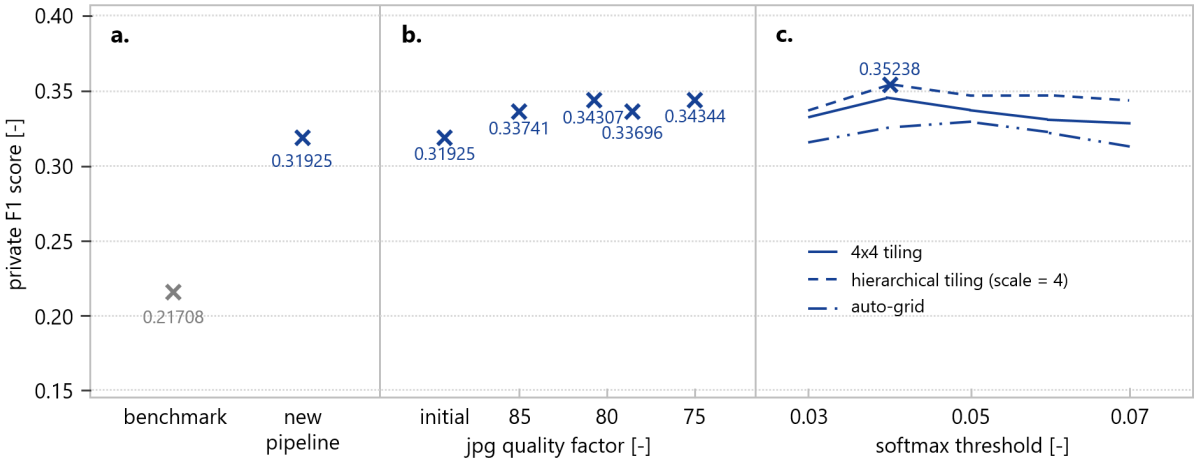

Figure 2: Private F1 scores for different tiling approaches. First, the new tiling pipeline based on 4x4 tiling and max pooling is compared to the benchmark provided by the challenge organizers (a). Then, the score sensitivity with respect to JPEG quality factor q is shown (b): starting from the initial setting, q is gradually decreased until q = 75. Last, different tiling approaches are compared over a fixed spectrum of softmax thresholds (c).

this combination of global calibration and threshold optimization further contributes to the observed performance gains.

In a next step, we want to tackle an aspect of the domain shift that is not immediately apparent. While the domain shift is commonly characterized by the discrepancy between single-image training data and multi-label images at inference time, a previous study [11] from the PlantCLEF 2025 competition pointed out that those datasets also exhibit a mismatch in jpg compression, more precisely, in how strongly images were lossily encoded. Following the procedure described in section 2.1, we previously estimated the jpg quality factors q (the standard encoder setting on the usual 1-100 scale) and the chroma sub-sampling scheme in both datasets. We found that the test dataset comprises mostly images with a quality factor between 95 and 98 with 4:2:2 chroma sub-sampling, whereas the training data cluster around q=78 with mainly 4:2:0 chroma sub-sampling (for details, see section 2.1).

To compensate for this difference, we include an optional preprocessing step into the inference pipeline, in which the image tiles are encoded from PIL. Images to JPG and directly decoded back to RGB. During this round-trip, we specify the jpg quality factor and the chroma sub-sampling scheme. For all trials, a sub-sampling scheme of 4:2:0 was chosen to match the predominant scheme of the training dataset. When including this preprocessing step and decreasing image quality of the tiles before inference ($q \in \{75, 78, 80, 85\}$), we observe increased F1 scores for all compression levels compared to the initial pipeline using the original image quality (**Figure 1b**). The overall best F1 scores of 0.34344 and 0.34307 are obtained at q = 75 and q = 80, whereas q = 78 (0.33696) and q = 85 (0.33741) perform similar to each other but remain below the previously mentioned options.

Intuitively, one might expect that q = 78 would lead to the best results, as it best matches the compression found in the training split (see section 2.1). However, performing re-encoding using q = 78 does not guarantee exact reproduction of the original compression, because some of the training images might have been produced with encoders and pipelines that might not fully coincide with the reference used for calculating the compression qualities. At the same time, the training distribution is not concentrated at a single q: Even though the majority of images lies near 78, a non-negligible fraction of images are estimated near . Hence, an inference-time encoding round-trip at a single q cannot fully match both regimes simultaneously. With this mixed distribution, it is plausible that a setting between the dominant q = 78 and the tail of higher q, such as q = 80, performs best on the evaluation metric without contradicting previous observations. Having said that, a compression of q = 80 will be applied for all following trials.

In a last step of tuning our base pipeline, we now investigate how different tiling approaches (see section 2.3) might impact the results. Across the tested softmax thresholds from 0.03 to 0.07, hierarchical tiling at scale 4 constantly achieves the highest scores out of the three methods (**Figure 1c**). At the optimal threshold, we achieve a private F1 score of 0.35238, which is a decent improvement compared to the performance achieved with basic 4×4 tiling. This suggests that enriching inference with multiple spatial resolutions stabilizes the interaction between tile-level predictions, pooling, and confidence filtering. However, multi-scale tiling comes with considerably increased computational costs compared to single-grid approaches. Interestingly, a fixed 4×4 grid outperforms auto-grid, an approach in which the grid is chosen individually for each in image in such a way that resizing is minimized. It was assumed that such an individualization might reduce errors that occur due to strong resizing and that it might tackle the variability of image sizes in the test set. However, these assumptions did not hold true, as the auto-grid approach is clearly lagging behind both fixed layout approaches (**Figure 1c**).

In conclusion, we find that both improved tiling strategies and compression-aware preprocessing significantly enhance performance over the baseline. A simple 4×4 logit-pooling approach already outperforms the reference pipeline, while JPEG round-trip simulation further mitigates domain shift effects. Hierarchical multi-scale tiling achieves the highest private F1 score of 0.35238 when using a scale of 4 (i.e. 10 tiles), demonstrating the effectiveness of combining information across multiple spatial resolutions. However, the PlantCLEF evaluation protocol ranks teams based on their five best submissions according to the public leaderboard. As a result, our best-scoring private submission was not among the five considered runs. The highest-ranked submission that contributed to the final team score employed hierarchical tiling with a scale factor of 6, achieving a private F1 score of 0.35009 (rank 16 of 105 in the challenge). While this configuration performed slightly worse on the private leaderboard and required substantially higher computational effort, it ranked more favorably on the public leaderboard. Overall, the results indicate that combining spatial aggregation with compression alignment is an effective strategy for robust inference under distribution shift.

3.2. Advanced approaches

Beyond exploring preprocessing and pooling strategies to use in a pipeline built around the provided pre-trained DINOv2 model, we implemented several more advanced approaches to improve the model performance. All of them reuse the same backbone checkpoint, preprocessing, and tiling configuration so that training, feature extraction, and pooling stay aligned. Unfortunately, we had to conduct our experiments under severely limited computational resources, which meant that we could only use a tiny fraction of the available training data. Under these conditions, none of the approaches was able to outperform the strong tiling baseline. However, we believe that these methods still represent promising directions and could have a much larger impact when trained and evaluated on the full dataset.

3.2.1. Low-Rank Adaption (LoRA)

First, we investigated whether the frozen backbone of the pretrained DINOv2 model could be adapted to the target domain using Low-Rank Adaption (LoRA) conducted with the unlabeled pseudo-quadrat images. The intention behind this was to use the pseudo-quadrats as a proxy domain to push the model toward field-quadrat statistics similar to the test dataset. In brief, trainable low-rank matrices are attached to both, attention and MLP layers and the classification head remains fully trainable. However, even though this approach is way more efficient (in terms of GPU memory and training time) compared to updating all backbone weights, we could only use 5000 of the 212000 images of the unlabeled dataset (randomly sampled), which corresponds to roughly 2.4% of the available data. To align with our previously described inference pipeline, those images were tiled using a fixed 4×4 grid. LoRA was then conducted using those 16 tiles per image (518 px input, ImageNet normalization) yielding 80,000 tiles in total.

In detail, we conducted one LoRA experiment in self-distillation mode (**Figure 3**). Here, a frozen copy of the pretrained DINOv2 acts as a teacher: For each training tile, the softmax distribution delivered by

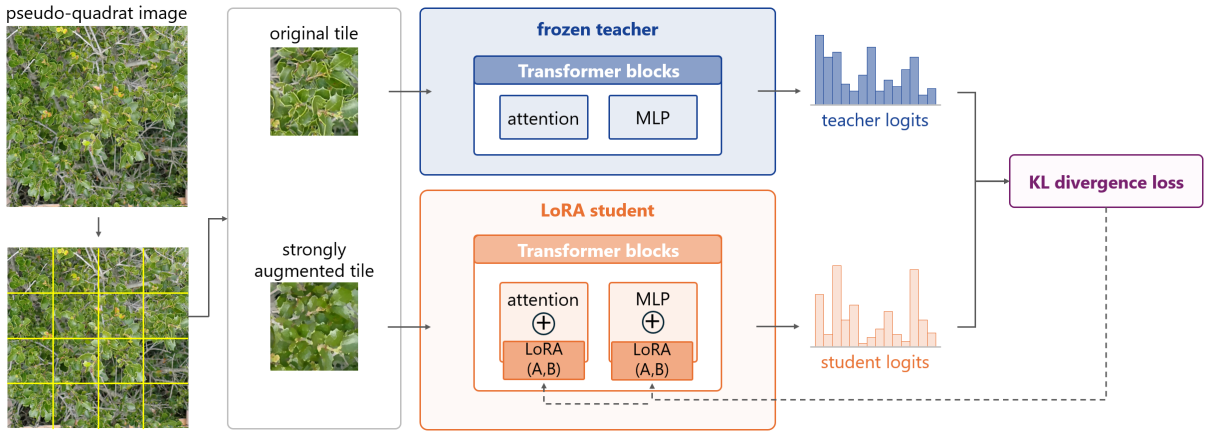

Figure 3: LoRA-based adaption of the DINOv2 using teacher-student distillation. Unlabelled pseudo-quadrat images are first tiled using a fixed 4x4 grid. A frozen teacher then generates target logits from image tiles, while a LoRA-augmented student is trained on strongly augmented tiles by minimizing the KL-divergence between teacher and student predictions.

the teacher was computed once and stored. A LoRA-wrapped student was then trained for 3 epochs to match these soft targets on strongly augmented tile views, using KL divergence with temperature scaling. Optimization used AdamW (learning rate 10^{-4}, weight decay 0.01), cosine schedule with linear warm-up, mixed precision, batch size 16 (tiles), and gradient clipping (norm 1.0).

As mentioned before, the idea behind our LoRA approach was to adapt to pseudo-quadrat / LUCAS-like image statistics while preserving the original PlantCLEF label space. However, we did not observe any improvement over the previously described tiling pipeline based on the pre-trained DINOv2 model provided by the organizers. We attribute this partly to the small training set used and the short training schedule, but also partly to the risk that proxy-domain self-distillation might reinforce teacher errors rather than correcting them. However, we still assume that scaling the self-distillation to the full image dataset of 212,000 images for a higher number of training epochs might yield a better domain adaption than what we were able to achieve. Additionally, we would propose to conduct subsequent supervised LoRA on labeled data at least for the classification head. We leave these extensions to future work when sufficient resources are available.

3.2.2. Retrieval-augmented classifier

Next, we tried a retrieval-augmented classification at inference time – no model training was involved in this trial. This approach was driven by the fact that our previously explained pipeline assigns species from pooled tile logits applying a fixed list cap and a global confidence threshold, with max pooling over tiles so that every tile contributes equally to the quadrat score. While this method seems to be efficient to a certain extent, it lacks adaptability to confidence variations on tile level. To compensate for this, we implemented a retrieval-augmented classification stage that adds nearest-neighbor species evidence to the classifier prediction and combines both outputs at tile level (**Figure 4**).

As in our 4x4 tiling pipeline, each test quadrat was split into 16 tiles and passed once through the frozen DINOv2 model; the resulting per-tile class logits and backbone embeddings for every quadrat were stored. The retrieval stage reads only these saved tensors together with a pre-built retrieval index and does not perform an additional ViT forward pass. The Faiss [17] index for retrieval was built based on the official PlantCLEF training images. To limit class imbalance, we randomly sampled at most 32 images per species (some species had less than 32 images in the training set). Each image was resized to 518 px, normalized with ImageNet statistics, embedded into a 768-dimensional pre-logit feature vector with the same pretrained DINOv2, L2-normalized, and stored in a Faiss inner-product index with a parallel vector of class indices (on unit vectors, inner product equals cosine similarity).

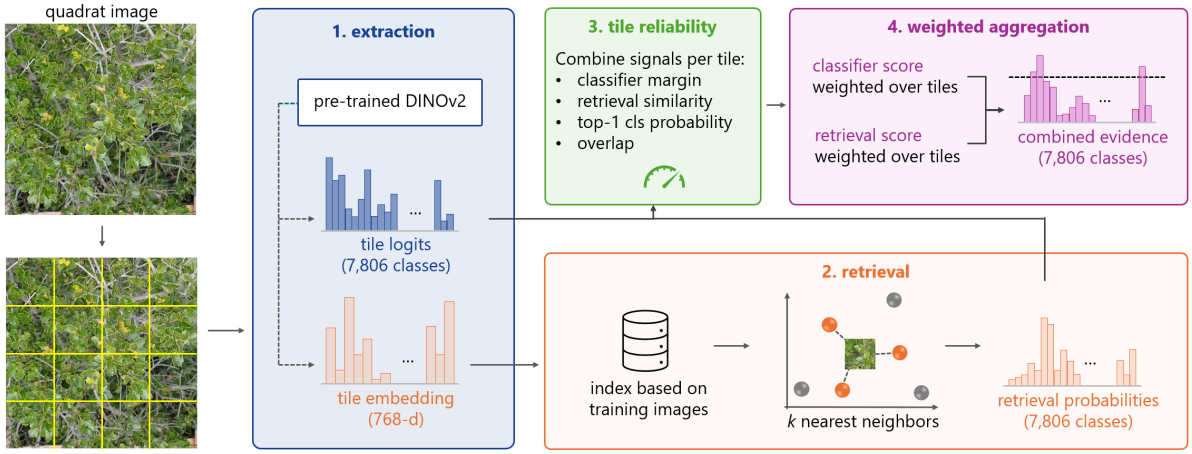

Figure 4: Retrieval-augmented classification at inference time. The test image is partitioned into tiles and processed by the pre-trained DINOv2 model to extract tile-level logits and 768-dimensional embeddings. The embeddings are queried against a FAISS index built from training images to retrieve the k nearest neighbors, yielding retrieval-based class probabilities. For each tile, classification and retrieval signals are combined with a reliability estimate derived from classifier confidence, retrieval similarity, and tile overlap. The resulting reliability scores are used to weight tile contributions, and classifier and retrieval evidence are aggregated across all tiles to obtain the final image-level prediction.

For each test tile, this index was queried for the 50 nearest vectors, each carrying the species label of the image it was created with. Following a similarity-weighted kNN vote, neighbor i with similarity sim_i received weight $w_i = exp(sim_i / \tau)$ with $\tau = 0.25$; we summed these weights per species and normalized the masses to a tile-level retrieval distribution over all 7,806 classes (only species that appear among the neighbors receive non-zero mass). This step creates an independent source of species information that will be combined with the classifier logits on the same tile.

Before using the two vectors for species prediction on quadrat level, we first used the newly gained information to assign each tile a reliability weight between 0.05 and 1.0; this step was included to account for the fact that not every tile is equally important in a quadrat image. The assigned weight reflects how decisive the classifier (pretrained DINOv2) was on that tile (margin between its top two logits), how strong the best retrieval match was, the top-1 softmax probability on that tile, and whether the classifier's leading species appeared among the top-12 retrieval species. These four cues were combined with coefficients 0.40, 0.35, 0.15, and 0.10, respectively, and scaled relative to the other tiles of the same quadrat. Classifier logits were aggregated with a weighted log-sum-exp over tiles (temperature 1.0), and tile-level retrieval distributions were aggregated with a weighted average using the same tile weights, yielding one classifier score vector and one retrieval score vector per quadrat. The two vectors were then merged into a base score: classifier evidence plus 0.08 times the logarithm of the retrieval probabilities (with $\varepsilon = 10^{-8}$ for numerical stability). Thus kNN contributes both directly to the species ranking and indirectly through tile reliability.

For the final submission, species per quadrat were selected from this fused base score using the same rule as in our baseline pipeline: softmax over all 7,806 classes, retention of up to eight species, and a minimum probability threshold, so that only the construction of the quadrat-level score differed from the baseline, not the list-length policy. In our experiments, this retrieval-augmented variant did not outperform max pooling with the organizer-provided checkpoint and the fixed submission rule.

Overall, despite not improving the final F1 score, we consider this retrieval-augmented formulation a promising direction as it introduces complementary, instance-based evidence beyond the classifier's learned decision boundaries; future work should focus on better calibration of the fusion scheme and reliability weighting, as well as on scaling the retrieval database to comprise a more extensive and ideally less class-imbalanced set of reference images.

4. Conclusion & Outlook

In this working note, we investigated different strategies for vegetation image analysis under strong domain shift in the context of the PlantCLEF 2026 challenge. We found that relatively simple changes in the inference pipeline already improved the results significantly compared to the provided baseline. In particular, tiling-based inference with max pooling and optimized confidence thresholds achieved strong performance improvements. Additionally, adapting the jpg compression statistics of the inference images to the training distribution further increased the F1 scores. Among the tested approaches, hierarchical multi-scale tiling achieved the best overall performance.

More advanced methods including LoRA-based adaption of the pretrained Vision Transformer or retrieval-augmented classification were also tested but did not outperform the strong tiling baseline. Importantly, the limited effectiveness of those methods most likely occurred due to the limited computational resources which made it necessary to drastically reduce training data. Hence, future work should investigate larger-scale domain adaptation using more data and longer training schedules. Additionally, adaptive tiling and improved aggregation strategies remain promising directions for future Vision Transformer-based vegetation monitoring systems.

Acknowledgments

The authors would like to thank the organizers of PlantCLEF 2026 and LifeCLEF 2026 for hosting the competition. Additionally, the authors would like to thank MicroNova AG for supporting this work by providing the opportunity to participate in the challenge.

Declaration of Conflicts of Interest

The authors declare that they have no conflict of interests.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Cursor for text refinement and coding assistance. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. A. Ali, M. Kamraju, Ecosystem Services, Springer Nature Switzerland, Cham, 2023, pp. 51–63. URL: https://doi.org/10.1007/978-3-031-46720-2_4. doi:10.1007/978-3-031-46720-2_4.

[2] C. Wagg, C. Roscher, A. Weigelt, A. Vogel, A. Ebeling, E. De Luca, A. Roeder, C. Kleinspehn, V. M. Temperton, S. T. Meyer, et al., Biodiversity–stability relationships strengthen over time in a long-term grassland experiment, Nature communications 13 (2022) 7752.

[3] Y. Li, A. Schuldt, A. Ebeling, N. Eisenhauer, Y. Huang, G. Albert, C. Albracht, A. Amyntas, M. Bonkowski, H. Bruelheide, et al., Plant diversity enhances ecosystem multifunctionality via multitrophic diversity, Nature Ecology & Evolution 8 (2024) 2037–2047.

[4] M. Loreau, S. Naeem, P. Inchausti, J. Bengtsson, J. P. Grime, A. Hector, D. U. Hooper, M. A. Huston, D. Raffaelli, B. Schmid, D. Tilman, D. A. Wardle, Biodiversity and ecosystem functioning: Current knowledge and future challenges, Science 294 (2001) 804–808. URL: https://www.science.org/doi/abs/10.1126/science.1064088. doi:10.1126/science.1064088. arXiv:https://www.science.org/doi/pdf/10.1126/science.1064088.

[5] J. Cavender-Bares, F. D. Schneider, M. J. Santos, A. Armstrong, A. Carnaval, K. M. Dahlin, L. Fatoyinbo, G. C. Hurtt, D. Schimel, P. A. Townsend, et al., Integrating remote sensing with ecology and evolution to advance biodiversity conservation, Nature Ecology & Evolution 6 (2022) 506–519.

[6] 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.
[7] M. Xu, S. Yoon, Y. Jeong, J. Lee, D. S. Park, Transfer learning with self-supervised vision transformer for large-scale plant identification., in: CLEF (Working Notes), 2022, pp. 2238–2252.
[8] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024). URL: https://openreview.net/forum?id=a68SUt6zFt, featured Certification.
[9] M. Gustineli, A. Miyaguchi, I. Stalter, Multi-label plant species classification with self-supervised vision transformers, arXiv preprint arXiv:2407.06298 (2024).
[10] L. Picek, 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, Springer, 2026.
[11] V. Espitalier, Preprocessing is all you need: Theheartofnoise submission to plantclef 2025, Working Notes of CLEF (2025).
[12] H. Venkateswara, S. Chakraborty, S. Panchanathan, Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations, IEEE Signal Processing Magazine 34 (2017) 117–129. doi:10.1109/MSP.2017.2740460.
[13] Y. Li, Y. Song, J. Luo, Improving pairwise ranking for multi-label image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3617–3625.
[14] M. Gustineli, A. Miyaguchi, A. Cheung, D. Khattak, Tile-based vit inference with visual-cluster priors for zero-shot multi-species plant identification, Working Notes of CLEF (2025).
[15] H. Herasimchyk, R. Labryga, T. Prusina, Multi-label plant species prediction with metadata-enhanced multi-head vision transformers, arXiv preprint arXiv:2508.10457 (2025).
[16] N. Semenova, Optimizing a vision transformer with ecological context for multi-label plant species identification, Working Notes of CLEF (2025).
[17] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, IEEE Transactions on Big Data 12 (2026) 346–361. doi:10.1109/TBDATA.2025.3618474.