使用元数据增强的多头视觉变换器进行多标签植物物种预测

2025 年 CLEF 会议 LifeCLEF 实验室笔记

Hanna Herasimchyk[1,*], Robin Labryga[1,*] and Tomislav Prusina[1,*]

[1]University of Hamburg, 177 Mittelweg, Hamburg, 20148, Germany

Abstract

我们提出了一种多头视觉 Transformer 方法，用于植被样方图像中的多标签植物种类预测，以应对 PlantCLEF 2025 挑战。该任务涉及在单一物种植物图像上训练模型，同时在多物种样方图像上进行测试，造成了显著的领域转变。我们的方法利用了预训练的 DINOv2 视觉 Transformer Base (ViT-B/14) 骨干网络，具有用于种、属和科预测的多个分类头，利用分类学层级。主要贡献包括多尺度平铺以捕捉不同尺度的植物，基于均值预测长度的动态阈值优化，通过袋装和 Hydra 模型架构的集成策略。该方法结合了多种推理技术，包括图像裁剪以去除非植物伪像，预测约束的 top-n 筛选，以及 logit 阈值策略。实验在大约 140 万幅涵盖 7806 种植物的训练图像上进行。结果显示出强劲的性能，使我们的提交在私人排行榜上名列第三。我们的代码可在 https://github.com/geranium12/plant-clef-2025/tree/v1.0.0 获取。

Keywords

Multi-Label Classification, DINOv2, Vision Transformer, Species Identification, Vegetation Plot Images, Biodiversity, PlantCLEF

1. 简介

植被样方调查在生态研究中是必不可少的，它使得生物多样性的采样和评估以及环境变化的监测成为可能。它们生成的宝贵数据支持生态系统分析、生物多样性保护和基于证据的环境决策。标准的植被调查检查小样方，这些样方是放置在地面上的大约半平方米的矩形框架，用来定义特定的采样区域。经过训练的植物学家记录发现的所有植物种类，并使用诸如生物量、生态评分或图像中观察到的覆盖率等指标量化其存在。

将机器学习方法整合到这一过程中可能会大大提高效率，使广泛的生态研究能够在减少专家参与的情况下进行。然而，开发能够在单张图像中识别数千种植物物种的模型仍然是一个重大的技术挑战。

拥有一个用所有存在的植物种类标注的样方图像数据集至关重要，但由于特定区域内众多物种的存在，这样的数据集既昂贵又具有挑战性。在这方面，相比之下，仅包含单一植物种类的图像的大量集合已经存在，这使得训练单种类分类模型更加容易。

PlantCLEF 2025 挑战 [1, 2, 3] 旨在通过评估设计用于预测高分辨率样方图像中多种植物的模型来解决这一差距。在这次竞赛中，模型使用单标签的单个植物图像进行训练，但在多标签的样方图像上进行测试，突出了训练和测试数据之间领域转移的挑战。

我们的主要方法使用了一种视觉 Transformer 架构 [4, 5]，并配备了多个分类头，使模型能够同时从共享的特征提取骨干中预测物种、属和科。这个多头设计有效地整合了分类学知识并利用了层次关系，显著增强了复杂植被图像中物种预测的鲁棒性。

我们工作对改进样方图像中植物种多标签分类的主要贡献包括：我们的代码可在 GitHub 上获取 [1]。

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*These authors contributed equally.

[1]https://github.com/geranium12/plant-clef-2025/tree/v1.0.0

2. 背景

2.1. 数据

训练数据集包含大约 140 万张植物个体的图像（约 281 GB），每张图像都有相应的元数据。这种大规模的数据给模型训练带来了显著的计算挑战。该数据集也用于 PlantCLEF 2024 竞赛，涵盖了 7806 个植物物种、1446 个属和 181 个科。

Figure 1: 训练图像示例。

物种间图像的分布如 **??** 所示，而属和科间的物种分布如 Fig. 2 所示。每张图像都标有单一的植物物种、单一的属和单一的科，并包括诸如器官类型和地理位置等元数据。属描述了一组植物物种，而科描述了一组植物属。示例训练图像显示在 Fig. 1 中。

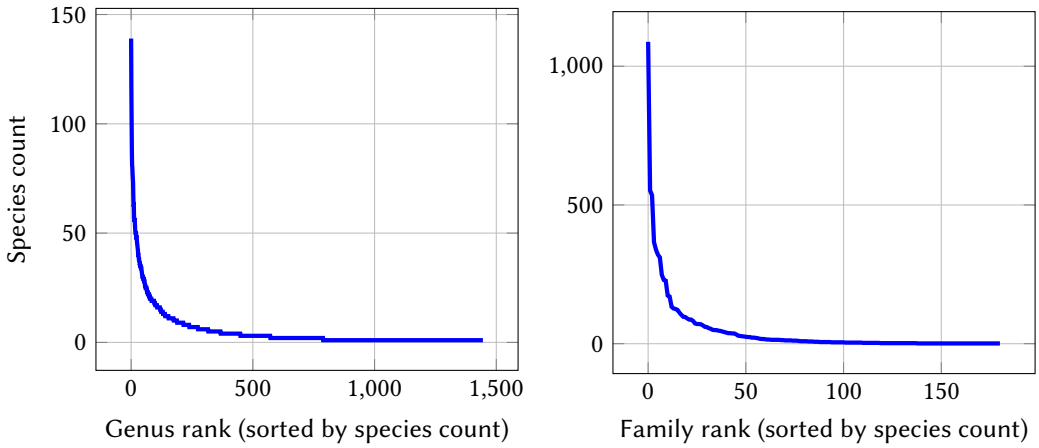

Figure 2: 分别在属和科中物种的分布。50 % 的物种分布在 113 个最大属中，而 90 % 的物种分布在 728 个最大属中。同样地，50 % 的物种仅属于最大科中的 9 个，而 90 % 的物种则包含在 49 个最大科中。与 **??** 类似，这些百分比表明物种在属和科中的分布存在偏倚。

Figure 3: 测试植被样地示例。

相比之下，测试数据集由包含多种植物物种的植被样方图像组成，而训练数据则专注于单一物种。测试图像中存在的物种数量没有限制。Fig. 3 中显示了测试图像示例。

与 PlantCLEF 2024 不同，本次竞赛使用了一个经过修改的评估指标。最终得分是测试集中每个样带的宏平均 F1 得分的平均值。样带是沿着在田野中定义的路径放置的一系列植被样方（样方），用于系统地记录物种出现情况。

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{j=1}^{T_i}\mathrm{F1}_{ij}\right)$$

其中：

2.2. DINOv2 模型

我们使用了由 PlantCLEF 组织者提供的 DINOv2 模型 [5, 8]，该模型在单一物种的训练图像上进行了预训练。其架构基于蒸馏的 Vision Transformer Base (ViT-B/14) 与寄存器 [9]，用作特征提取的主干网。对于每个输入图像，模型生成一个嵌入，然后通过由一个线性层构成的分类头以预测物种。更多细节可以参考 [6]。

我们选择使用 DINOv2 是基于来自 PlantCLEF 2024 挑战的实证证据，其中 ViT-B 架构比其他模型架构表现出更优越的性能 [6, 10, 11, 12]。此外，鉴于数据集的计算限制（140 万张图像，281 GB），从头开始训练大规模深度神经网络在计算上是不可能的。因此，我们使用了由组织者提供的已经预训练的 DINOv2 骨干，而无需额外的微调。

3. 方法论

3.1. 训练数据准备

对于我们的一些方法，必须训练或重新训练模型，包括新添加的属和科分类器，以及用于区分植物和非植物样本的模型。我们使用的训练程序如下所述。

在训练过程中，我们使用了多种数据增强技术来提高模型的鲁棒性和泛化能力。这些增强包括随机裁剪、随机水平翻转和垂直翻转、透视变换以及随机旋转。此外，我们应用了色彩抖动来引入亮度、对比度和饱和度的变化。

我们还应用了标准化和调整大小的程序，以确保输入图像符合 DINOv2 架构预期的分布和大小。这包括减去均值并除以标准差，以及将输入图像调整为 518×518 。

数据分割 提供的训练数据集已经预先分割。我们决定使用所有可用数据进行训练，包括那些没有用于预训练的图像。为了进行内部评估，我们对训练数据进行了分层分割，以确保物种的均衡表示。

LUCAS 数据集 组织者提供了一个额外的训练数据集，称为 LUCAS（Land Use/Cover Area frame Survey） [13]，包含 212,782 张未标注的地面植被图像，这些图像以垂直样方格式存储，总数据量为 170GB。我们尝试继续对 DINOv2 模型进行预训练以整合这些数据，其动机是认为在预训练期间接触特定领域的植被图像可以提升模型在后续分类任务中的表示能力。然而，由于硬件限制，这种方法被证明是不可行的。因此，我们在没有对 LUCAS 数据集进行额外预训练的情况下，继续使用原始的 DINOv2 权重。

3.2. 测试数据预处理

图像裁剪 对植被图像的初步目视检查显示，非植物物体如木质地块框架边缘、测量带和鞋类经常出现在图像边缘（见 Fig. 3 ）。为了减小这些非植物物体对模型的影响，我们尝试对所有四个图像边进行从 5 % 到 15 % 的居中裁剪。10 % 的裁剪策略在公共排行榜上获得了最佳结果，而 5 % 的策略在私人排行榜上更为有效，这表明 10 % 的方法可能过多。

为了应对植被地块中植物大小和密度变化的挑战，我们实施了一种多尺度平铺方法。这包括将图像划分为一个由多个小块组成的网格（2×2 ，3×3 ，...），使模型能够有效地捕捉到小型和大型植物物种。每个小块用作模型的输入图像。所有预处理步骤都相应地应用于每个小块。我们还尝试了重叠小块，以确保不遗漏位于小块边缘的植物。然而，我们发现不使用重叠的小块已足够，因为重叠并没有带来结果的改进。

3.3. 模型架构与训练

为了利用分类信息，在原始物种 MLP 分类头之外，我们在 DINOv2 ViT-B 骨干网络之上加入了用于属和科预测的额外 MLP 分类头。这些分类头利用了与每张图像相关的元数据。我们还对每个分类头中的层数进行了实验。

鉴于严格的层级关系——每个物种唯一地属于一个属，而每个属又属于一个科，我们将对物种、属和科的预测概率相乘，舍弃在提供的元数据中不存在的组合。这确保了只考虑有效的分类学分配。

除了分类学分类头，我们还训练了一个专门用于器官预测的分类头，旨在识别每张图像中所描绘的植物器官类型（例如，叶、花、茎）。然而，由于不同物种之间器官表现的固有差异，将基于器官的信息整合到整体预测流程中被证明是具有挑战性的。

此外，数据集包含一个“扫描”器官标签，指示通过扫描获得植物图像，而不是在自然环境中拍摄的图像。由于我们的主要关注点是植被样地分析，它依赖于现实环境中植物的照片，因此我们假设从训练数据集中移除这些图像可能会提高最终的准确性。

Hydra 模型架构 我们使用共享来自冻结主干的相同嵌入的独立分类头。每个头的几个版本，其中包含不同层数的版本同时被训练。在测试时，我们可以交换这些预训练的头，以从一个主要架构创建各种模型版本。我们将这种集成方法称为 Hydra 模型。我们训练的最佳 Hydra 模型包括用于种类分类的一层头和用于属和科分类的两层头，中间有 ReLU 激活函数。

DINOv2 ViT-L 我们通过基于 Vision Transformer Large (ViT-L/14 [4]) 骨干网络的 DINOv2 实现来探索模型架构的扩展。尽管与较小的变体相比，这种架构提供了更大的表示能力，但初步实验显示出显著的计算限制。在我们的 GPU 集群上，对完整 PlantCLEF 数据集进行单次训练迭代需要大约 30 小时（参见 **??** 中的模型训练）。鉴于需要至少约 50 次迭代才能实现收敛，总的训练时间将超过 1,500 小时（大约 62.5 天），这使得在项目的资源限制内，这种方法是不可行的。

为了减少来自无关前景杂物（例如岩石或土壤斑块）的误报，我们训练了一个二元分类器以区分植物和非植物区域。我们从公开渠道创建了一个独立的非植物图像数据集（主要是岩石），并训练了逻辑回归、随机森林和一个基于 ViT 的分类器。在这三种方法中，随机森林分类器达到了最高的总体准确率，在我们的验证数据中正确识别植物和非植物区域的准确率为 95 %。因此，我们在主要流程中采用了随机森林模型来过滤非植物对象。然而，该模型未能在植被地块图像上泛化，也没有提高最终预测质量。

我们在我们的 GPU 集群上训练了所描述的模型架构，每个实验使用 2× 个 NVIDIA A6000 GPUs。基于 ViT 的每个模型训练大约用了三天，具体时间视具体架构而变化。有关详细的技术规格和代码，请参阅我们公开的 GitHub 代码库（见 Section 1 ）。

3.4. 推理

我们实施了一个多步骤预测流程，将单一物种分类器适配到多物种方格预测任务中。多种策略被实际测试和整合，在公共和私有数据集上取得了不同行程度的成功。

由于每张植被样本图像通常不包含超过十二种不同的植物物种，我们通过限制允许的最大（前 n 种）预测，来约束每张图像的物种预测数量。通过实验，我们发现调节这个上限提高了公共排行榜上的评分。挑战结束后的相同实验表明，这通常会导致私人排行榜上的表现变差。此外，每张图像至少强制预测一种物种（底部 n）被证明是有益的。

对于每个图块，我们最多只允许一种物种对最终预测有贡献。为了确保仅包含最有信心的预测，我们应用了一个逻辑阈值策略。一种方法是为物种预测设置一个最小的逻辑值，从而过滤掉低信心的预测。另一种方法是根据所有测试图像的平均预测长度动态调整逻辑阈值。为了高效地执行这种动态调整，我们利用每个测试图像和图块的预计算逻辑值，并使用二分搜索算法找到合适的阈值。我们最后选择使用动态调整的阈值策略，每张图像平均有四种物种，因为它易于使用且性能明显。

元数据合并 测试集中一部分植被样方图像在其文件名中包含标识符和日期。我们研究了是否可以通过使用图像元数据，具体来说，是整合同一地块和年份拍摄的图像中的预测结果，来提高得分。例如，如果在同一地块的所有图像中某个物种被识别超过三次，它就会在该地块的每一张图像中被预测出来。这个想法是，通过整合相关样方的信息，这种方法可能会提高召回率。然而，我们没有使用这种方法，因为：首先，这种方法没有提高我们的得分；其次，元数据并未为整个测试集提供；第三，这种方法与挑战的目标相矛盾，挑战的目标是通过植被样方来发现生物多样性的变化。

为了进一步提高我们预测的鲁棒性，我们实施了一种 bagging 策略（参见 [14] ）。在生成最终预测之前，我们通过平均每个图像块来自多个模型的 logits 来结合多个模型。此方法通过使用来自不同模型的信息帮助减少变异性并提高我们结果的可靠性。

我们实现了一种基于核的平滑方法，应用于每个图像块的对数输出。具体来说，相邻块的对数被加权系数（例如，0.5）加入到每个块的预测对数中，使相邻块的预测能够互相影响。这个想法是植物可能跨越块的边界。然而，基于核的平滑的初步实验并没有在最终评估得分上取得改进。因此，我们没有尝试任何其他替代核。这种缺乏改进可能是由于我们使用了多尺度平铺法，它实际上起到了类似的作用。

我们探索了几种额外的策略，例如对 logits 进行 z-score 归一化，而不是阈值化或过滤稀有物种，但在不同的数据集上没有观察到一致的改进。由于收益甚微，这些方法最终未被纳入最终的流程中。

Table 1 显示了 PlantCLEF 挑战赛在公共和私人排行榜上的前三名位置。我们名为 “Chlorophyll Crew” 的团队分别在公共和私人排行榜上获得了第二和第三的最佳分数。

Table 1
公共和私有排行榜中的前三名最佳分数。我们的解决方案在公共和私有排行榜中分别取得了第二和第三的最佳分数，使用的团队名称是 “Chlorophyll Crew”。

Leaderboard	Team name	Rank	Score
Public	webmaking	1	0.38132
	Chlorophyll Crew	2	0.37555
	TheHeartOfNoise # Rust # Candle	3	0.35900
Private	TheHeartOfNoise # Rust # Candle	1	0.36479
	DS@GT PlantCLEF	2	0.34489
	Chlorophyll Crew	3	0.33655

Table 2 展示了我们在公共和私有 PlantCLEF 排行榜上的前五名提交，以及我们选择的五个预测。虽然所有模型在公共排行榜上都取得了较高的分数，但在私有排行榜上的所有提交中性能都有一致的下降。这种模式表明公共和私有测试集之间不平衡，且针对公共集合进行优化的模型可能无法很好地泛化到私有集合。私有排行榜上提交之间的分数差异相对较小，而公共排行榜上却存在较大差异，这进一步突显了这种不平衡。这些结果表明，排行榜驱动的优化可能导致对公共测试集的过拟合。特别是，与公共排行榜上的顶级解决方案相比，我们在私有排行榜上经历了最小的下降。

Table 2
在 PlantCLEF 排行榜上的前五名提交，显示了公共和私有分数、模型配置、关键超参数和 5 个选择的预测。所有提交都使用属-科信息，限制每个图块仅贡献一个物种，并要求每张图像至少有一个物种预测。加粗值表示在相应列中排名前五。Hydra 模型具有单层物种头以及用于属和科的双层头。另一方面，5h1l 模型使用单层头结构。Hydra $_{new}$ 模型是 Hydra 模型的扩展，经过额外的完整周期训练。Vitlarge 模型代表了一种配备双层头的 ViT-L 架构。最后，预训练的 Vit-B 模型[2]由组织者提供。

models	logit min	length mean	length max	tiling scales	crop %	score public	score private
Our Top 5 on the Public Leaderboard							
Hydra 5h1l		4.2	9	4,5	10	0.37555	0.33409
Hydra Hydra$_{new}$ 5h1l 5h1l$_{new}$		4.2	9	4,5	10	0.37543	0.33375
Hydra$_{new}$ 5h1l$_{new}$		4.2	9	4,5	10	0.37542	0.33104
Hydra 5h1l		4.175	9	4,5	10	0.37540	0.33142
Hydra 5h1l		4.15	9	4,5	10	0.37522	0.33253
Our Top 5 on the Private Leaderboard							
5h1l		4.0	∞	4,5	10	0.35134	0.34575
Hydra 5h1l		4.0	9	4,5	8,10,12	0.36263	0.34358
5h1l	0.01		10	4	5	0.33338	0.34352
Hydra 5h1l		4.0	∞	4,5	10	0.37100	0.34091
Hydra 5h1l		3.9	9	4,5	10	0.36972	0.34047
Our 5 selected submissions							
5h1l	0.02		10	4,5	10	0.36590	0.33947
Hydra 5h1l Vitlarge		4.2	9	4,5	10	0.37305	0.33655
Hydra 5h1l Vit-B		4.0	9	4,5	10	0.36730	0.33484
Hydra 5h1l		4.2	9	4,5	10	0.37555	0.33409
5h1l	0.02		10	1,2,4,5	10	0.36555	0.32074

由于训练数据和测试数据之间存在显著的领域转移，我们无法在本地验证我们的方法，这迫使我们依赖公共排行榜进行模型选择。尽管我们努力选择多样化的模型集，但我们选择的五个提交之一并未出现在私人排行榜的前五名中，这突出显示了测试数据分割所带来的挑战以及基

于排行榜评估的局限性。

我们的主要多头分类方法相较于依赖简单单头植物物种分类的基线取得了显著的提升。如 Table 2 所示，所有报告结果均使用多头分类，突显了这一提升。

我们评估了几种超参数配置，观察到 10 % 裁剪策略在公共测试集上产生了最有希望的结果，而 5 % 策略在私有测试集上表现更好，这表明前者可能导致信息丰富的视觉区域被过度裁剪。Top-9 和 top-10 过滤没有提高得分，并且 top-n 过滤通常会降低私有排行榜上的表现。始终预测至少一个物种能够提高得分。动态调整阈值并使每张图像平均包含四个物种提升了最终得分。我们最好的 Hydra 模型在物种分类中采用了一个单层头，对于属和科的分类，应用了一个在层之间具有激活函数的两层头。合并元数据没有改善结果，可能是因为并非整套测试数据都提供了元数据，并且这种方法与该挑战的目标相悖，即从植被图中发现生物多样性的变化。对于多尺度平铺，我们发现使用多个非重叠的 4 和 5 大小的平铺就足够了，因为重叠并未提供任何性能提升。虽然通过随机森林进行的植物/非植物过滤在我们的独立数据集上实现了 95 % 的验证精度，但未能推广至植被图像并未提高最终预测。虽然在公共排行榜上袋装显著提高了结果，但对私有排行榜得分有负面影响。然而，当袋装应用于使用不同裁剪参数的模型时，其确提高了私有成绩，这在我们在私有排行榜上的第二最佳提交中可见。最后，基于内核的平滑初步实验未能改善最终评价得分，可能是因为多尺度平铺已经提供了类似的效果。

4. 相关工作

深度学习和计算机视觉方法已被广泛应用于植物物种识别和植被分析。早期的工作主要集中在使用卷积神经网络（CNNs）进行遥感和植被制图，如 Kattenborn 等人所述 [15] 。最近，以变压器为基础的架构在植物相关任务中表现出色，例如无人机图像中的杂草检测 [16] ，而我们的工作则利用视觉变压器骨干网进行多标签植物物种预测，以延续这一趋势。

针对图像中对象大小变化的挑战，已经探索了基于补丁和多尺度的方法。Adelson 等人 [17] 引入了图像金字塔方法，这与我们使用多尺度瓦片相似，能够捕捉不同空间分辨率的信息。

层次分类利用了例如分类关系，在各个领域都有研究。Silla 和 Freitas [18] 提供了关于层次机器学习的全面调查。在分类学背景下的层次分类的一个例子是 Colonna 等人 [19] 的工作，他们使用自上而下的方式来预测青蛙的科、属和种。几项工作 [20, 21] 提出沿分类层次结构乘以概率，有些在每个层次层使用一个分类器，有些在层次结构中的每个内部节点使用一个分类器。这类似于我们的多头架构，它独立地预测种、属和科并融合它们的输出。

数据增强仍然是提高模型鲁棒性的重要技术。Shorten 和 Khoshgoftaar [22] 提供了一份关于图像增强方法的全面调查，其中许多方法被我们用于训练管道中。

以前在 PlantCLEF2024 挑战赛中，[6, 7] 展示了多种用于植物物种识别的深度学习方法。Foy 和 McLoughlin [11] 利用视觉 Transformer（ViT）架构与 Segment-Anything Model（SAM）结合，有效地抑制了非植物图像区域中的误报。Gustineli 等人 [10] 探索了基于 ViT 的多种嵌入方法和分类器架构，而 Chulif 等人 [12] 将卷积神经网络（CNNs）与 ViT 结合，使用贝叶斯模型平均进行更好的预测。这些方法凸显了一种趋势，即向视觉 Transformer 和先进的后处理技术发展，以实现强健的植物物种识别。

5. 结论

我们提出了一种元数据增强的多头视觉变换器，用于多标签植物物种预测，通过分类融合结合物种、属和科的输出。使用多尺度切片、动态阈值以及集合策略（Hydra），我们的模型在公开排行榜上取得了优异的成绩。

然而，在私有测试集上的性能下降，揭示了对领域转移的敏感性以及基于排行榜调整的局限性，但结果仍具有竞争力。

未来的工作应该解决领域适应问题，结合器官特定的线索，并探索微调策略以提高真实世界的鲁棒性。

[2]https://www.kaggle.com/models/juliostat/dinov2_patch14_reg4_onlyclassifier_then_all/PyTorch/default

Acknowledgments

我们要感谢 PlantCLEF 2025 和 LifeCLEF 2025 的组织者举办这次比赛。

6.

生成式人工智能宣言 在准备这项工作期间，作者使用了 ChatGPT、GitHub Copilot、Grammarly 进行：语法和拼写检查，改写和重述。在使用这些工具或服务后，作者根据需要审核并编辑了内容，并对出版物的内容承担全部责任。

References

[1] H. Goëau, G. Martellucci, P. Bonnet, F. Vinatier, A. Joly, PlantCLEF2025 @ LifeCLEF & CVPR-FGVC, https://kaggle.com/competitions/plantclef-2025, 2025. .

[2] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2025.

[3] K. Janouskova, J. Matas, L. Picek, Overview of FungiCLEF 2025: Few-shot classification with rare fungi species, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[4] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, X. Zhai, An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations (2021).

[5] M. Oquab, T. Darcet, T. Moutakanni, H. 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).

[6] H. Goëau, V. Espitalier, P. Bonnet, A. Joly, Overview of PlantCLEF 2024: multi-species plant identification in vegetation plot images, Conference and Labs of the Evaluation Forum (2024).

[7] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of lifeclef 2024: Challenges on species distribution prediction and identification, Conference and Labs of the Evaluation Forum (2024).

[8] H. Goëau, J. 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.

[9] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision Transformers Need Registers, International Conference on Learning Representations (2024).

[10] M. Gustineli, A. Miyaguchi, I. Stalter, Multi-Label Plant Species Classification with Self-Supervised Vision Transformers, Conference and Labs of the Evaluation Forum (2024).

[11] S. Foy, S. McLoughlin, Utilizing Dino V2 for Domain Adaptation in Vegetation Plot Analysis, Conference and Labs of the Evaluation Forum (2024).

[12] S. Chulif, H. Ishrat, Y. Chang, S. Lee, Patch-wise inference using pretrained vision transformers: Neuon submission to plantclef2024, Conference and Labs of the Evaluation Forum (2024).

[13] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, P. Haub, O. Buck, C. Haub, B. Eiselt, M. van der Velde, Lucas cover photos 2006–2018 over the eu: 874 646 spatially distributed geo-tagged close-up photos with land cover and plant species label, Earth System Science Data (2022).

[14] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, 2009.
[15] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (cnn) in vegetation remote sensing, ISPRS Journal of Photogrammetry and Remote Sensing (2021).
[16] R. Reedha, E. Dericquebourg, R. Canals, A. Hafiane, Transformer neural network for weed and crop classification of high resolution uav images, Remote sensing (2022).
[17] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, J. M. Ogden, Pyramid methods in image processing, RCA Engineer (1984).
[18] C. N. Silla, A. A. Freitas, A survey of hierarchical classification across different application domains, Data mining and knowledge discovery (2011).
[19] J. G. Colonna, J. Gama, E. F. Nakamura, A comparison of hierarchical multi-output recognition approaches for anuran classification, Machine Learning (2018).
[20] J. N. Hernandez, L. E. Sucar, E. F. Morales, A hybrid global-local approach for hierarchical classification., Florida Artificial Intelligence Research Society (2013).
[21] L. Fiaschi, M. Cococcioni, Informed deep hierarchical classification: a non-standard analysis inspired approach, IEEE Transactions on Neural Networks and Learning Systems (2024).
[22] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data (2019).