Self-Supervised Learning of Plant Image Representations

Ilyass Moummad[1][0009−0003−2925−2500], Kawtar Zaher[1,2][0009−0008−9920−9483], Hervé Goëau[3][0000−0003−3296−3795], Jean-Christophe Lombardo[1][0000−0002−9656−4219], Pierre Bonnet[3][0000−0002−2828−4389], and Alexis Joly[1][0000−0002−2161−9940]

[1] INRIA, LIRMM, Université de Montpellier, Montpellier, France
[2] Institut National de l'Audiovisuel, Paris, France
[3] AMAP, Université de Montpellier, CIRAD, INRAE, IRD, Montpellier, France
{ilyass.moummad,alexis.joly}@inria.fr

Abstract. Automated plant recognition plays a crucial role in biodiversity monitoring and conservation, yet current approaches rely heavily on supervised learning, which is limited by the availability of expert-labeled data. Self-supervised learning (SSL) offers a scalable alternative, but existing methods and training protocols are largely designed for coarse-grained visual tasks and may not transfer well to fine-grained domains such as plant species recognition. In this work, we investigate SSL for plant image representation learning. We show that commonly used augmentations in SSL pipelines—such as Gaussian blur, grayscale conversion, and solarization—are detrimental in the context of plant images, as they remove subtle discriminative cues essential for fine-grained recognition. We instead identify alternative transformations, including affine and posterization, that are better suited to this domain. We further demonstrate that training SimDINOv2 on the iNaturalist 2021 Plantae subset yields significantly stronger representations than training on ImageNet-1K, highlighting the importance of domain-specific data for SSL. Our findings are consistent across both ViT-Base and ViT-Large architectures. Moreover, our models achieve competitive performance and sometimes outperform strong supervised baselines Pl@ntCLEF and BioCLIP, on downstream plant recognition tasks in few-shot settings. Overall, our results highlight the critical importance of domain-adapted augmentation strategies and dataset selection in self-supervised learning, and provide practical guidelines for building scalable models for biodiversity monitoring.[4]

Keywords: Self-supervised learning, plant identification, fine-grained visual categorization, data augmentation, biodiversity monitoring.

1 Introduction

Automatic plant recognition plays a crucial role in biodiversity monitoring, ecological research, and precision agriculture [3]. Large scale platforms such as

[4] Code and models: https://github.com/ilyassmoummad/sslplant

Pl@ntNet [12] have demonstrated the potential of computer vision systems to assist both experts and citizens in identifying plant species in the wild. However, these systems rely heavily on supervised learning, which requires large amounts of labeled data. In the context of plant species recognition, acquiring such annotations is particularly challenging, as it demands domain expertise and significant manual effort [11]. This limitation raises important concerns regarding scalability and coverage, especially for under-documented ecosystems.

Self-supervised learning (SSL) [1] offers a promising alternative by enabling representation learning without manual annotations. While SSL has achieved remarkable success in general-purpose visual representation learning [15, 7, 4, 14, 8, 28, 5, 22, 26, 27], its application to plant image analysis remains relatively underexplored. Most state-of-the-art SSL methods are designed and evaluated on coarse-grained datasets such as ImageNet [9], where categories differ significantly (e.g., distinguishing between animals and objects). In this setting, invariance to strong data augmentations such as aggressive color transformations, blurring, or solarization—has proven beneficial [28, 5]. However, plant species recognition is inherently a fine-grained task, where subtle visual cues such as texture, venation patterns, and color distributions are critical for discrimination. As a result, augmentation strategies that are effective for coarse-grained tasks may be detrimental in this domain. For example, color-based transformations like solarization or grayscale conversion can distort or remove key discriminative features, as illustrated in Figure 1. This highlights the need for domain-adapted augmentation policies in SSL for biodiversity applications.

In this work, we investigate the recently proposed SimDINOv2 [26] framework for self-supervised representation learning on plant images. SimDINOv2 builds upon DINOv2 [22] while simplifying its training pipeline by removing several heuristic components and introducing an explicit coding rate regularization to prevent representation collapse. We conduct a systematic ablation study of commonly used augmentations and demonstrate that widely adopted transformations such as Gaussian blur, grayscale conversion, and solarization negatively impact performance in the context of plant recognition.

Furthermore, we introduce and evaluate two alternative transformations, posterization and affine transformations, which have been largely overlooked in SSL pipelines [20]. Our results show that these augmentations are better suited to preserving fine-grained visual characteristics of plant species, leading to improved representation quality. This work emphasizes the importance of domain-specific design choices in self-supervised learning and provides practical insights for developing more effective models for biodiversity monitoring.

2 Related Work

Automatic plant recognition plays a central role in biodiversity monitoring [3] and has been significantly advanced by citizen science platforms that collect large-scale visual observations [12]. Applications such as ObsIdentify [21], iNaturalist [17], Flora Incognita [19], and Pl@ntNet [12] allow users to iden-

tify plant species from images while continuously enriching training datasets through user contributions. Among these, Pl@ntNet [12] stands out as a large-scale system dedicated to plant identification, covering over 80,000 species and supporting fine-grained recognition, including intra-species variability as well as plant diseases and pests. Despite these advances, most existing systems rely on supervised learning, which limits scalability due to the high cost and expertise required for annotation.

Large-scale plant datasets have been introduced to support plant recognition, covering both species identification and plant pathology. Pl@ntNet-300K [10] contains approximately 300,000 images across 1,000 species and is widely used as a benchmark for supervised classification. Larger collections are available through citizen science initiatives, notably the iNaturalist 2021 dataset [16], whose Plantae subset comprises around 1.1 million images spanning more than 4,000 species. Its scale and diversity, together with its similarity to ImageNet-1K, make it particularly well suited for training data-intensive self-supervised models. Other datasets, such as Deep-Plant-Disease [6], focus on plant pathology and provide labeled examples of diseases across multiple crops. In this work, we leverage the iNaturalist 2021 Plantae subset due to its scale and relevance for self-supervised pretraining.

Self-supervised learning has emerged as a powerful paradigm for learning visual representations without manual annotations. A dominant approach is instance discrimination, where models are trained to produce invariant representations across augmented views of the same image. Early methods rely on contrastive objectives with explicit negative samples (e.g., SimCLR [7], MoCo [15]), while clustering-based approaches such as SwAV [4] improve scalability. More recent methods eliminate the need for negatives and instead prevent representation collapse through implicit regularization, including asymmetric self-distillation (BYOL [14], SimSiam [8], DINO [5]) and redundancy reduction objectives (Barlow Twins [28], VICReg [2]).

These approaches primarily focus on learning global representations. DINOv2 [22] extends this paradigm by incorporating a patch-level objective, enabling the model to capture both global semantics and local structures. Building on this, SimDINOv2 [26] simplifies DINOv2 [22] by removing heuristic mechanisms such as centering and sharpening, and instead enforces representation diversity through coding-rate regularization [18], while relying on a simple cosine similarity objective for alignment.

Despite their strong performance on generic benchmarks, SSL methods rely heavily on augmentation invariance designed for coarse-grained recognition. Their application to fine-grained domains such as plant species recognition remains underexplored, particularly with respect to augmentation design, which can critically impact the preservation of subtle discriminative features.

3 PlantSSL: Self-Supervised Representation Learning of Plant Images

3.1 Preliminaries: Contrastive Learning

Contrastive learning aims to learn representations that are invariant across different augmented views of the same image while remaining discriminative across different images.

SimCLR. Given an input image x, a set of augmented views $\mathcal{V}(x)$ is generated using a stochastic augmentation pipeline. Let $z(v) \in \mathbb{R}^d$ denote the ℓ_2-normalized representation of view v, obtained from an encoder f_θ.

Let $\mathcal{B}$ be a batch of images, and define the set of all views in the batch as

$$\mathcal{V}_\mathcal{B} = \bigcup_{x \in \mathcal{B}} \mathcal{V}(x).$$

For a given anchor view $v \in \mathcal{V}(x)$, we denote by v^+ another view of the same image x, and by $\mathcal{A}(v) = \mathcal{V}_\mathcal{B} \setminus \{v\}$ the set of all other views in the batch.

The SimCLR loss (also known as InfoNCE or NT-Xent) is defined as:

$$\mathcal{L}_{\text{SimCLR}} = \frac{1}{|\mathcal{V}_\mathcal{B}|} \sum_{v \in \mathcal{V}_\mathcal{B}} -\log \frac{\exp\left(\langle z(v), z(v^+)\rangle/\tau\right)}{\sum_{v' \in \mathcal{A}(v)} \exp\left(\langle z(v), z(v')\rangle/\tau\right)}, \tag{1}$$

where τ is a temperature parameter. This objective pulls together representations of different views of the same image while pushing apart representations from other images in the batch.

SupCon. Supervised Contrastive Learning extends SimCLR by leveraging label information to define multiple positive pairs.

Let $y(x)$ denote the label of image x. For an anchor view $v \in \mathcal{V}(x)$, we define the set of positives as $\mathcal{P}(v) = \{v' \in \mathcal{V}_\mathcal{B} \setminus \{v\} \ : \ y(v') = y(x)\}$, i.e., all views in the batch whose underlying images share the same label.

The SupCon loss is then given by:

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{|\mathcal{V}_\mathcal{B}|} \sum_{v \in \mathcal{V}_\mathcal{B}} \frac{-1}{|\mathcal{P}(v)|} \sum_{v^+ \in \mathcal{P}(v)} \log \frac{\exp\left(\langle z(v), z(v^+)\rangle/\tau\right)}{\sum_{v' \in \mathcal{A}(v)} \exp\left(\langle z(v), z(v')\rangle/\tau\right)}. \tag{2}$$

This formulation encourages representations of samples from the same class to cluster together, providing an upper bound on the performance achievable by purely self-supervised methods.

Role of data augmentation. The definition of positive pairs, and therefore the invariances learned by the model, is entirely determined by the augmentation pipeline $\mathcal{A}$. As a result, the choice of augmentations plays a critical role in shaping the learned representation, particularly in fine-grained settings where subtle visual cues must be preserved.

3.2 Data Augmentation for Plant Images

Data augmentation plays a central role in SSL [1], as the learning objective enforces invariance between different views of the same image. Standard augmentation pipelines, popularized by methods such as SimCLR [7], BYOL [14], and DINO [5], include strong appearance transformations such as color jittering, grayscale conversion, Gaussian blur, and solarization. While effective for coarse-grained recognition, these transformations can significantly alter or remove subtle visual cues—such as color distributions, textures, and venation patterns—that are essential for plant species discrimination (Figure 1, top row).

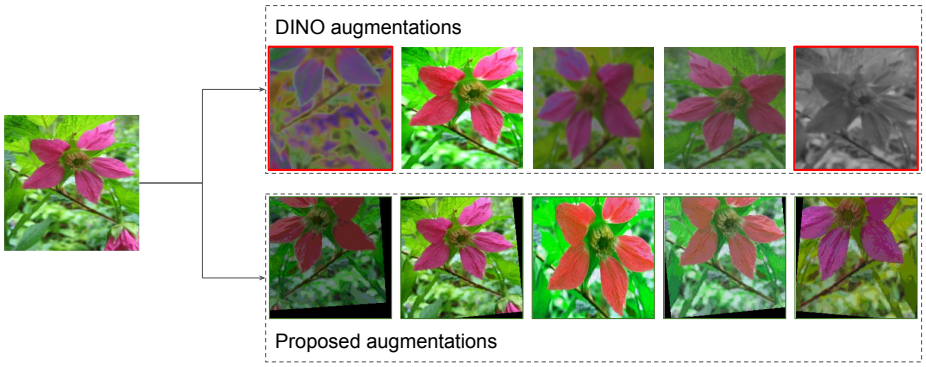

Fig. 1. Comparison of general-purpose data augmentations from self-supervised visual learning frameworks (e.g., SimCLR, BYOL, DINO) with our proposed augmentations for fine-grained plant species recognition. Standard techniques such as Gaussian blur, grayscale, and polarize destroy crucial plant information, whereas posterize and affine transformations preserve distinguishing features while still allowing visual diversity.

To address this limitation, we propose a plant-adapted augmentation strategy that preserves fine-grained characteristics while still generating sufficiently diverse views. Specifically, we replace grayscale conversion, Gaussian blur, and solarization with two alternative transformations: *posterization* and *affine transformations*. Posterization reduces the number of color levels, producing appearance variations while retaining global color structure. Affine transformations (including rotation, translation, scaling, and shearing) introduce geometric variability without degrading local textures. As illustrated in Figure 1 (bottom row), these transformations maintain discriminative plant features while increasing view diversity.

Experimental validation. To isolate the impact of data augmentation independently of architectural and optimization choices, we conduct a controlled study using contrastive learning (SimCLR) and its supervised counterpart (SupCon). This setup allows us to evaluate how augmentation choices affect representation quality relative to label supervision.

All models are trained from scratch using a ResNet-18 backbone on a subset of the iNaturalist 2021 Plantae dataset, comprising 213,550 training images across 4,271 species. We use random resized crop and horizontal flip as base augmentations, denoted $\mathcal{A}_{\text{base}}$, to ensure sufficient variability and avoid collapsed representations. Models are trained for 100 epochs, and the learned representations are evaluated using k-nearest neighbors classification on the validation subset containing 42,710 images (10 images per species).

Table 1. Impact of data augmentations for self-supervised learning of plant image representations on the iNaturalist 2021 Plantae subset (4,271 species). We train ResNet18 from scratch on training mini subset (213,550 images) for 100 epochs using contrastive learning and evaluate on the validation subset (42,710 images, 10 images per species). $\mathcal{A}_{\text{base}}$ denotes Random Resized Crop and Horizontal Flip.

Method	**Augmentations**	**Acc. (%)**
Reference		
SupCon (w/ labels)	$\mathcal{A}_{\text{base}}$	10.61
SimCLR	$\mathcal{A}_{\text{base}}$	3.70
DINO augmentations		
SimCLR	$\mathcal{A}_{\text{base}}$ + Jitter + Grayscale + Blur + Solarize	9.70
SimCLR	$\mathcal{A}_{\text{base}}$ + Jitter + Grayscale + Blur	9.96
Plant-adapted augmentations (ours)		
SimCLR	$\mathcal{A}_{\text{base}}$ + Jitter + Posterize + Affine	**10.35**
SimCLR	$\mathcal{A}_{\text{base}}$ + Jitter + Posterize + Affine + Grayscale	9.83
SimCLR	$\mathcal{A}_{\text{base}}$ + Jitter + Posterize + Affine + Blur	8.76

Results are reported in Table 1. Using only $\mathcal{A}_{\text{base}}$, SimCLR performs poorly compared to its supervised counterpart, highlighting the difficulty of learning fine-grained discriminative features for plant species recognition. Incorporating standard augmentations from prior SSL methods (e.g. DINO) significantly improves performance; however, removing solarization yields further gains, suggesting that this transformation is detrimental in the plant domain. Our proposed augmentation strategy achieves the best performance, nearly matching the supervised SupCon upper bound. Importantly, ablation results show that reintroducing grayscale or Gaussian blur consistently degrades performance. These findings confirm that commonly used SSL augmentations are not well suited for fine-grained plant recognition, and that carefully designed, domain-adapted transformations are critical for learning discriminative representations.

These observations motivate the use of plant-adapted augmentations in modern SSL frameworks, which we explore next.

3.3 SimDINOv2 for Plant Representation Learning

Building on the importance of augmentation design highlighted in the previous section, we integrate plant-adapted augmentations into a modern multi-

view self-supervised framework. We adopt SimDINOv2, which combines self-distillation with explicit regularization to prevent representation collapse, while jointly learning global and local features.

Teacher–student framework. Given an input image x, we generate a set of augmented views. We denote by $\mathcal{V}_g$ the set of global views and by $\mathcal{V}_l$ the set of local views. The teacher network f_{θ_t} processes only global views $v \in \mathcal{V}_g$, while the student network f_{θ_s} processes all views $v \in \mathcal{V}_g \cup \mathcal{V}_l$.

For a given view v, a Vision Transformer produces a global CLS token representation denoted by $z^{\text{cls}}(v)$, and a set of patch-level token representations denoted by $\{z^i(v)\}_{i=1}^N$. All representations are ℓ_2-normalized, so that dot products correspond to cosine similarity.

Global (CLS) alignment. The student is trained to match the teacher's CLS representations from global views across all student views:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \frac{1}{|\mathcal{V}_g|\,|\mathcal{V}_g \cup \mathcal{V}_l|} \sum_{v_t \in \mathcal{V}_g} \sum_{v_s \in \mathcal{V}_g \cup \mathcal{V}_l} \left(1 - \langle z_s^{\text{cls}}(v_s), z_t^{\text{cls}}(v_t) \rangle\right). \tag{3}$$

Patch-level alignment. In addition, a masked patch prediction objective inspired by iBOT is applied to patch tokens. For each image, a global view $v \in \mathcal{V}_g$ is processed by both networks. The teacher receives the full (unmasked) sequence of patch tokens, while the student operates on a masked version of the same view, where a subset of patch tokens is replaced by mask tokens after patch embedding.

Let $\mathcal{M}(v) \subset \{1, \ldots, N\}$ denote the set of masked patch indices. The student is trained to match the teacher representations at these locations:

$$\mathcal{L}_{\text{patch}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \frac{1}{|\mathcal{M}(v)|} \sum_{i \in \mathcal{M}(v)} \left(1 - \langle z_s^i(v), z_t^i(v) \rangle\right). \tag{4}$$

This objective encourages the model to infer missing local information from context while producing consistent and semantically meaningful patch-level representations.

Coding rate regularization. To further prevent representation collapse, SimDINOv2 incorporates a coding rate regularization term applied to the student CLS representations. Let $Z_s \in \mathbb{R}^{B \times d}$ denote the matrix of ℓ_2-normalized CLS features for a batch of size B. The regularization term is defined as:

$$\mathcal{L}_{\text{reg}} = -\frac{1}{2} \log \det \left(I + \frac{d}{B\varepsilon} Z_s^\top Z_s \right), \tag{5}$$

where $\varepsilon > 0$ controls the precision of the representation in the rate–distortion sense. This term encourages the representations to span a large volume in feature space, promoting diversity and preventing collapse.

Final objective. The overall training objective is a weighted combination of the different terms:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{patch}}\mathcal{L}_{\text{patch}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{6}$$

The teacher parameters are updated as an exponential moving average (EMA) of the student:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s. \tag{7}$$

Figure 2 illustrates SimDINOv2 framework, showing both global and patch-level self-distillation using cosine similarity.

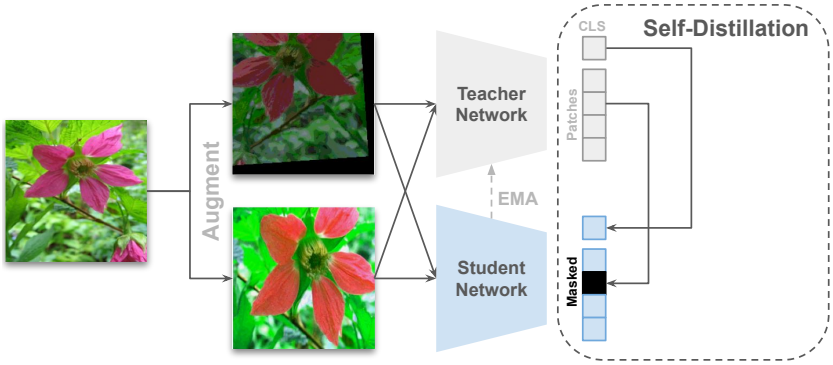

Fig. 2. Self-supervised learning of plant image representations using SimDINOv2 framework. An input image is augmented to produce two views: one passes through a teacher network to generate target representations, while the other passes through a student network that learns to match the teacher. Both global CLS token and patch-level representations are trained via self-distillation with cosine similarity. The student is optimized with backpropagation, and the teacher is updated using an exponential moving average of the student. The coding rate regularization is applied to the student model to prevent collapsed representations.

4 Experiments

We evaluate the effectiveness of our approach for self-supervised learning of plant image representations through few-shot classification on diverse plant benchmarks. Our experiments aim to answer two main questions: (i) whether domain-specific pretraining on plant data improves representation quality compared to generic datasets such as ImageNet, and (ii) whether our approach can compete with strong supervised baselines.

4.1 Experimental Setup

Pretraining. We consider two sources of pretrained models. First, we use publicly available SimDINOv2 models trained on ImageNet-1K. Second, we train

Table 2. Summary of evaluation datasets from MetaAlbum.

Dataset	Description	#Samples	#Classes
PlantNet	Citizen science species-labeled plant images	1,000	25
PlantVillage	Museum-style leaf specimens	1,520	38
Med. Leaf	Leaves from mature healthy medicinal plants	1,040	26
PlantDoc	17 diseases across 13 plant species	1,080	27

SimDINOv2 from scratch on the iNaturalist 2021 Plantae subset, which contains approximately 1.1M images across 4,271 plant species and is comparable in scale to ImageNet. This setting allows us to isolate the impact of domain-specific data. We evaluate both ViT-Base (B) and ViT-Large (L) architectures.

Evaluation protocol. We adopt the few-shot evaluation protocol introduced in BioCLIP using the SimpleShot [25] classifier. Given a small support set (1-shot or 5-shot), class prototypes are computed as the mean of L2-normalized feature embeddings, and query samples are classified via nearest-neighbor search in the embedding space.

Datasets. While BioCLIP evaluates on a diverse set of biological domains, we restrict our study to plant-related datasets from the MetaAlbum benchmark. Specifically, we consider Pl@ntNet, PlantVillage, Medicinal Leaf, and PlantDoc, covering both plant species recognition and plant disease classification tasks. A summary of the datasets is provided in Tab. 2.

Baselines. We compare our approach against strong baselines commonly used for plant recognition. Pl@ntCLEF [13] is a DINOv2 [22] model fine-tuned in a supervised manner on the Pl@ntCLEF-2024 dataset, which represents the flora of southwestern Europe with 1.4M images covering 7,800 plant species. BioCLIP [24] is a vision–language model based on CLIP [23], adapted to biodiversity data and fine-tuned on large-scale image–text pairs from TreeOfLife-10M. On the self-supervised side, we consider SimDINOv2 [26] models pretrained on ImageNet-1K as a reference, allowing us to isolate the impact of domain-specific pretraining. Altogether, these baselines provide a comprehensive comparison across supervised, multimodal, and self-supervised paradigms. Table 3 reports few-shot classification accuracy.

Implementation details. For self-supervised training, we closely follow the SimDINOv2 training protocol [26]. We train both ViT-Base and ViT-Large models using the AdamW optimizer, adopting the same optimization hyperparameters, including the learning rate schedule and warm-up strategy. We employ the same multi-crop augmentation scheme consisting of 2 global views with resolution 224×224 and 10 local views with resolution 96×96.

4.2 Few-Shot Classification Results

Impact of domain-specific pretraining. Training SimDINOv2 on iNaturalist Plantae significantly improves performance over ImageNet-pretrained models across all datasets and shot settings. For example, on Pl@ntNet, ViT-L improves from 35.6% to 77.4% in the 1-shot setting. Similar gains are observed across all benchmarks, highlighting the importance of domain-specific data for fine-grained plant recognition.

Comparison with supervised baselines. Self-supervised models trained on iNaturalist achieve competitive performance and often surpass supervised baselines such as Pl@ntCLEF and BioCLIP. In particular, the ViT-L model achieves state-of-the-art results on Pl@ntNet (1-shot) and PlantVillage (both 1-shot and 5-shot), and consistently outperforms BioCLIP on the more challenging PlantDoc dataset, which contains plant disease images unseen during training.

Effect of model scale. Scaling from ViT-Base to ViT-Large yields consistent performance improvements. However, these gains remain smaller than those obtained by switching from generic to domain-specific pretraining, indicating that dataset relevance is a more critical factor than model capacity in this setting. At this scale, the ViT-L model would likely benefit further from additional training data, suggesting that its capacity is not yet fully utilized.

Table 3. Few-shot classification accuracy (mean ± std) on plant image datasets from MetaAlbum. PC and TOL denote the Pl@ntCLEF and TreeOfLife datasets.

			Pl@ntNet		PlantVillage		Med. Leaf		PlantDoc	
Model	**Arch**	**Pretraining**	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Supervised Models										
Pl@ntCLEF	B	LVD-142M → PC'24	72.9 ± 1.7	91.2 ± 1.3	67.1 ± 2.0	85.6 ± 1.4	93.6 ± 0.8	99.4 ± 0.3	38.8 ± 1.2	52.6 ± 1.0
BioCLIP	B	OpenAI → TOL-10M	61.2 ± 2.0	83.3 ± 1.5	57.6 ± 2.2	79.6 ± 0.8	81.8 ± 1.2	95.5 ± 1.3	26.6 ± 2.2	42.3 ± 1.2
Self-Supervised Models										
SimDINOv2	B	ImageNet-1K	32.8 ± 2.9	54.1 ± 0.7	55.1 ± 1.4	79.2 ± 1.0	77.6 ± 2.4	93.1 ± 1.0	23.9 ± 2.0	37.7 ± 1.6
SimDINOv2	L	ImageNet-1K	35.6 ± 3.2	56.5 ± 0.8	57.3 ± 1.4	80.1 ± 1.5	76.1 ± 2.7	93.9 ± 0.7	23.5 ± 1.5	38.0 ± 1.6
Ours										
SimDINOv2	B	iNat21 Plantae	73.4 ± 4.4	87.9 ± 0.7	64.7 ± 1.9	85.4 ± 0.5	85.9 ± 1.6	97.8 ± 0.7	**31.5** ± 1.3	47.8 ± 1.3
SimDINOv2	L	iNat21 Plantae	**77.4** ± 3.3	**88.2** ± 1.6	**68.0** ± 1.9	**86.6** ± 0.1	**87.8** ± 1.7	**98.5** ± 0.5	31.2 ± 2.0	**49.0** ± 1.1

Feature space visualization. We examine the geometry of the learned representations using t-SNE projections on the PlantNet dataset (Figure 3). Representations obtained from ImageNet pretraining exhibit weak class structure, with samples from different species intermingled in the embedding space. In contrast,

domain-specific pretraining on iNaturalist Plantae with adapted augmentations results in clearly structured clusters, where samples from the same species are grouped more coherently and separated from others. Interestingly, this organization is comparable to that observed for supervised models such as BioCLIP and Pl@ntCLEF. These results suggest that the performance gains observed in few-shot classification stem from a better alignment of the feature space with underlying semantic classes.

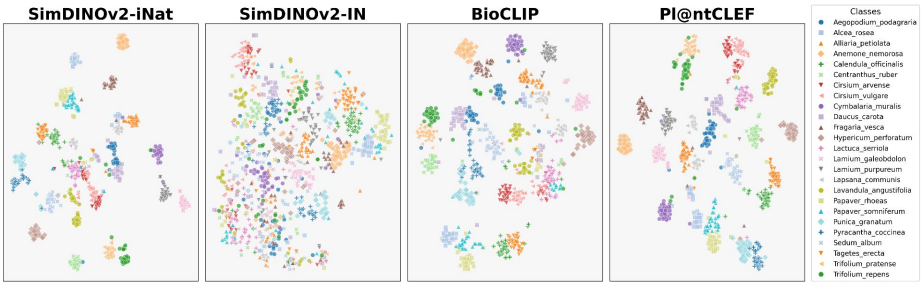

Fig. 3. t-SNE visualization of feature embeddings on the Pl@ntNet dataset for different pretrained models. SimDINOv2 pretrained on ImageNet produces dispersed and overlapping class clusters, indicating poor separability of plant species. In contrast, SimDINOv2 pretrained on iNaturalist Plantae forms compact and well-separated clusters, comparable to supervised approaches BioCLIP and Pl@ntCLEF. This highlights the importance of domain-specific pretraining for fine-grained plant representation learning.

Discussion. These results demonstrate that domain-specific self-supervised pretraining is key to learning discriminative plant representations. Despite using no labels, our approach matches or exceeds supervised methods, highlighting the potential of SSL for scalable plant recognition.

5 Conclusion

In this work, we investigated self-supervised learning for plant image representation learning, a fine-grained visual recognition problem where subtle visual cues are critical for discrimination. We showed that standard augmentation strategies commonly used in SSL pipelines are not well suited to this domain, as they tend to remove or distort key discriminative features. Instead, we demonstrated that carefully designed, plant-adapted augmentations, such as posterization and affine transformations, are more suitable for learning meaningful and discriminative representations.

Building on these insights, we applied SimDINOv2 to plant images and showed that domain-specific pretraining on the iNaturalist 2021 Plantae subset

leads to substantial improvements over generic ImageNet pretraining. Our approach achieves competitive and often superior performance compared to strong supervised baselines in few-shot settings, highlighting the effectiveness of SSL when combined with appropriate data and augmentation design.

Overall, our results show that self-supervised learning in fine-grained domains depends on both augmentation strategies and dataset relevance, as confirmed both quantitatively through few-shot performance and qualitatively through the structure of the learned feature space. Future work should focus on scaling with larger and more diverse plant datasets to enable annotation-free plant recognition for biodiversity monitoring.

Acknowledgements This project was funded by the French National Research Agency (ANR) through the grant Pl@ntAgroEco 22-PEAE-0009. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0191011389 made by GENCI.

References

1. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A Cookbook of Self-Supervised Learning. arXiv preprint arXiv:2304.12210 (2023)
2. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In: International Conference on Learning Representations (2022)
3. Bonnet, P., Joly, A., Faton, J.M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J.C., Mary, L., et al.: How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. Ecological Solutions and Evidence **1**(2), e12023 (2020)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
6. Chai, A.Y.H., Jee, K.L.Z., Lee, S.H., Tay, F.S., Vandeputte, J., Goeau, H., Bonnet, P., Joly, A.: Deep-plant-disease dataset is all you need for plant disease identification. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 12578–12584. MM '25, Association for Computing Machinery, New York, NY, USA (2025)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning (ICML) (2020)
8. Chen, X., He, K.: Exploring Simple Siamese Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

10. Garcin, C., Joly, A., Bonnet, P., Affouard, A., Lombardo, J.C., Chouet, M., Servajean, M., Lorieul, T., Salmon, J.: Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (2021)
11. Goëau, H., Bonnet, P., Joly, A.: Overview of PlantCLEF 2023: Image-based Plant Identification at Global Scale. In: Working notes of the Conference and Labs of the Evaluation Forum (CLEF 2023). CEUR Workshop Proceedings, vol. 3497, pp. 1972–1981. CEUR-WS (Sep 2023)
12. Goëau, H., Bonnet, P., Joly, A., Bakić, V., Barbe, J., Yahiaoui, I., Selmi, S., Carré, J., Barthélémy, D., Boujemaa, N., et al.: Pl@ ntnet mobile app. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 423–424 (2013)
13. Goeau, H., Espitalier, V., Bonnet, P., Joly, A.: 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. CEUR Workshop Proceedings (2024)
14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
16. iNaturalist 2021: inaturalist 2021 competition dataset. https://github.com/voxel51/inaturalist-2021 (2021)
17. iNaturalist community: inaturalist. https://www.inaturalist.org (2025)
18. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9), 1546–1562 (2007)
19. Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H.C., Deggelmann, A., Wäldchen, J.: The Flora Incognita app – Interactive plant species identification. Methods in Ecology and Evolution **12**(7), 1335–1342 (2021)
20. Moutakanni, T., Oquab, M., Szafraniec, M., Vakalopoulou, M., Bojanowski, P.: You Doné Need Domain-Specific Data Augmentations When Scaling Self-Supervised Learning. In: Advances in Neural Information Processing Systems. vol. 37, pp. 116106–116125 (2024)
21. Observation International: Obsidentify: Wildlife and plant identification app. https://observation.org/apps/obsidentify/ (2026)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Bojanowski, P., LeCun, Y., Caron, M.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (TMLR) (2023)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021)
24. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., 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. pp. 19412–19424 (2024)

25. Wang, Y., Chao, W.L., Weinberger, K.Q., Van Der Maaten, L.: SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning. arXiv preprint arXiv:1911.04623 (2019)
26. Wu, Z., Zhang, J., Pai, D., Wang, X., Singh, C., Yang, J., Gao, J., Ma, Y.: Simplifying DINO via Coding Rate Regularization. In: Proceedings of the International Conference on Machine Learning (ICML) (2025)
27. Zaher*, K., Moummad*, I., Buisson, O., Joly, A.: Self-Supervised Learning as Discrete Communication. In: International Conference on Machine Learning (ICML) (2026)
28. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: International Conference on Machine Learning (ICML) (2021)