Dual-Margin Embedding for Fine-Grained Long-Tailed Plant Taxonomy

Cheng-Yaw Low[1,2*], Heejoon Koo[2*], Jaewoo Park[3], and Meeyoung Cha[2]

[1] Changwon National University, Republic of Korea
[2] Max Planck Institute for Security and Privacy (MPI-SP), Germany
[3] AIVEX Inc., Republic of Korea

Abstract. Taxonomic classification of ecological families, genera, and species underpins biodiversity monitoring and conservation. Existing computer vision methods typically address fine-grained recognition and long-tailed learning in isolation. However, additional challenges such as spatiotemporal domain shift, hierarchical taxonomic structure, and previously unseen taxa often co-occur in real-world deployment, leading to brittle performance under open-world conditions. We propose TaxoNet, an embedding learning framework with a theoretically grounded dual-margin objective that reshapes class decision boundaries under class imbalance to improve fine-grained discrimination while strengthening rare-class representation geometry. We evaluate TaxoNet in open-world settings that capture co-occurring recognition challenges. Leveraging diverse plant datasets, including Google Auto-Arborist (urban tree imagery), iNaturalist (Plantae observations across heterogeneous ecosystems), and NAFlora-Mini (herbarium collections), we demonstrate that TaxoNet consistently outperforms strong baselines, including multimodal foundation models.

Keywords: Fine-grained and Long-tailed Recognition · Plant Taxonomy · Dual-Margin Embedding Learning

1 Introduction

Biodiversity forms the foundation for the stability and resilience of ecosystems, ensuring essential ecological functions, regulating the climate, and supporting human well-being [25]. Its conservation is recognized by the United Nations as fundamental, offering both a scientific rationale and a policy framework for global action. This commitment is reflected in several United Nations Sustainable Development Goals (SDGs); SDG 15 'Life on Land' which directly targets biodiversity loss, alongside SDG 13 'Climate Action', and SDG 11 'Sustainable Cities and Communities' [26]. In parallel, machine learning and computer vision have become transformative tools for biodiversity monitoring, particularly for ecological taxonomic classification [8, 34]. However, practical deployment remains challenging due to environmental heterogeneity and data scarcity. These

* This work was done when the first and second authors were with MPI-SP.

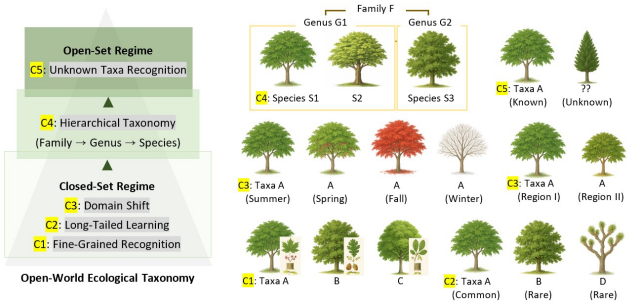

Fig. 1: Prior work in ecological taxonomy studies fine-grained and long-tailed recognition under closed-set assumptions, whereas real-world settings involve spatiotemporal shifts, hierarchical taxonomic structure, and unseen taxa.

complexities necessitate robust systems capable of generalizing to real-world ecological conditions.

Real-world ecological data often exhibit long-tailed distributions, where only a few common (*i.e.* head) taxa dominate observations, while rare (*i.e.* tail) taxa—many of which are endemic, endangered, or invasive—remain underrepresented [2, 24, 33]. This bias causes recognition models to favor dominant taxa, limiting performance on those most critical for conservation and management. Moreover, taxonomic classification requires fine-grained recognition, as many taxa differ only by subtle morphological traits. For example, species within the same genus (*e.g.* *Acer rubrum* vs. *Acer saccharum*) may exhibit high inter-taxa similarity and intra-taxa variability, leading to misclassification [33].

Beyond long-tailed class distribution, ecological data exhibit test-time shifts driven by seasonal, habitat, and geographic variability. These factors introduce mismatches between training and deployment conditions that degrade generalization [2]. Biodiversity monitoring is fundamentally an open-set problem due to undocumented taxa in the wild, yet most models rely on closed-set assumptions [19]. This results in the misclassification and poses risks to conservation efforts [23]. While existing research tackles long-tail imbalance via resampling, loss reweighting [5, 13, 17], and more recently, logit adjustments [22] or margin-based classifiers [4], these targeted solutions limit generalizability in realistic conditions, where class imbalance, fine-grained variability, open-set dynamics, and distribution shifts intersect.

We study ecological taxonomy under realistic deployment conditions, where fine-grained recognition, long-tailed class distributions, spatiotemporal variation, and open-set scenarios co-occur. Departing from the standard practice of examining these challenges independently, we analyze model robustness in-the-wild where these factors jointly manifest (see Figure 1). Although ecological taxonomy spans diverse groups, including insects, birds, mammals, and fungi commonly represented on large-scale biodiversity platforms such as iNaturalist [33], our work focuses on plant-level taxa across urban, natural, and specimen-based environments. This focus is motivated by the availability of large-scale annotated datasets and the foundational role of plants in biodiversity monitoring and conservation efforts [3].

Our primary contributions are as follows:

- We propose a theoretically grounded dual-margin loss that models class imbalance for fine-grained and long-tailed recognition, forming the foundation of our embedding learning framework, TaxoNet. Specifically, the dual-margin formulation suppresses repulsive gradients on tail-class prototypes induced by dominant head classes, leading to more balanced decision boundaries.
- We introduce a novel sample selection strategy that prioritizes tail taxa and morphologically diverse instances reflecting broader ecological variation.
- Evaluation shows advances in fine-grained recognition under long-tailed distributions, domain shift, and hierarchical structure against strong baselines and multimodal foundation models.

Our contributions support ecological AI applications for robust plant biodiversity observation. Henceforth, the term *taxa* refers to taxonomic ranks including family, genus, and species, unless otherwise specified.

2 Related Work

We review prior work along five axes: fine-grained recognition, long-tailed learning, domain generalization, hierarchical taxonomy, and open-set recognition. While fine-grained and long-tailed recognition have been extensively studied in isolation under closed-set assumptions, their joint manifestation in plant taxonomic recognition remains underexplored.

Fine-Grained Classification. Fine-grained ecological taxonomy requires distinguishing visually similar species based on subtle morphological cues (*e.g.*, petal venation, leaf shape, or bark texture in plants), often relying on high-resolution imagery to preserve discriminative details [6, 33]. Recent advances have extended this task to the vision–language paradigm, where visual and textual prompts jointly guide classification [16, 30, 31]. Although margin-based softmax losses—originally developed for face recognition—offer strong class separation properties [7, 37], they remain underexplored in ecological classification, largely due to instability under long-tailed distributions and limited inductive capacity for rare taxa. These challenges are exacerbated in ecological datasets, where inter-species similarity and intra-species variation are both high.

Long-Tailed Learning. Ecological datasets, including plant taxa, suffer from severe class imbalance where a few common species dominate and many rare and underrepresented taxa. Data-level strategies, such as oversampling rare taxa or undersampling common taxa, are conventional but risk overfitting or discarding informative data [13]. Loss reweighting offers an alternative by scaling gradients based on class priors, though naïve implementations can cause training instability [40]. More principled approaches include Focal Loss [17] for hard examples; Class-Balanced Loss [5] based on effective samples size; Label-Distribution-Aware Margin (LDAM) loss [4] for larger margins on rare classes; and Logit Adjustment [22] to recalibrate decision boundaries. Despite their efficacy against imbalance, these methods do not address specific ecological challenges like fine-grained discrimination, distributional shift, or taxonomic novelty.

Domain Generalization. Ecological deployment involves distributional shifts across geographic regions and temporal conditions. Despite its practical importance, domain generalization has received limited attention in biodiversity recognition. The Google Auto-Arborist dataset [2] is one of the few large-scale, street-level tree repositories curated to evaluate inter-city generalization in North America. In contrast, iNaturalist-2019 [33], while taxonomically diverse through citizen science contributions, lacks structured test splits for assessing geographic or temporal robustness. Hence, evaluating model robustness under distributional shift remains an open challenge in ecological AI.

Hierarchical Taxonomy. Ecological labels are inherently structured by biological taxonomy (family–genus–species), forming nested semantic relationships. Leveraging hierarchical metadata, biodiversity datasets such as iNaturalist-2019 [33] and BIOSCAN-1M [9], recent foundation models such as BioCLIP [30] demonstrate improved taxonomic separability across multiple ranks. However, most benchmarks focus solely on genus or species levels. Consequently, explicit evaluation of hierarchical robustness in ecological recognition remains limited.

Open-Set Recognition. Most ecological models operate under a closed-set assumption, recognizing only taxa seen during training. This is reflected in mobile applications such as *Seek Camera* by iNaturalist, *Pl@ntNet*, and *Flora Incognita*, which support a bounded set of recognizable plant taxa [34]. However, real-world biodiversity monitoring often violates this assumption, as hundreds of thousands of vascular plant species plant species have been documented globally, including many rare, endemic, or geographically localized taxa [1,10]. This gap hinders initiatives like the Kunming-Montreal Global Biodiversity Framework, which calls for comprehensive monitoring by 2030. Addressing this challenge requires open-set recognition frameworks that can detect taxonomic novelty [23,27]. Although recent studies have explored open-set challenges in ecological domains such as marine and bird species [29,38], their application to large-scale plant taxonomy remains limited.

3 Methodology

We propose TaxoNet, a convolutional embedding framework with a dual-margin loss that reshapes class decision boundaries to improve generalization for fine-grained, long-tailed plant taxonomic recognition.

3.1 Preliminary

Problem Definition. Given a C-class plant taxonomic classification problem, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training dataset of N examples, where x_i is the i-th input image and $y_i \in \mathcal{C} = \{1, \ldots, C\}$ is the corresponding class label (family, genus, or species). Let $\phi : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^d$ be a feature encoder that maps each image x_i to a d-dimensional embedding vector $\mathbf{x}_i = \phi(x_i)$. For notational simplicity, we omit the sample index i in subsequent derivations unless otherwise specified.

The additive margin-based softmax cross-entropy loss [35] is formulated as:

$$\mathcal{L} = -\log \frac{e^{s(z_y - m)}}{e^{s(z_y - m)} + \sum_{j \neq y} e^{s z_j}}, \tag{1}$$

where s is a scaling factor and m is an additive margin term. Each embedding vector $\mathbf{x}$ is compared with a set of class prototypes $\{\mathbf{w}_j \in \mathbb{R}^d\}_{j=1}^{c}$, with logits defined as $z_j = \mathbf{w}_j^\top \mathbf{x}$, where $\mathbf{w}_y$ denotes the prototype corresponding to the ground-truth class y.

Setting $s = 1$ and $m = 0$ reduces Eq. (1) to the standard softmax cross-entropy loss. While effective for general classification tasks, it lacks explicit mechanisms to enforce intra-class compactness and inter-class separation—both essential for distinguishing fine-grained plant taxa with subtle morphological traits such as leaf shape, texture, or color. Moreover, it does not account for the long-tailed class distributions prevalent in plant and other ecological datasets [2, 33].

An alternative formulation introduces additive margin-based softmax classifiers (AM-Softmax) [36, 37], which incorporate $s > 1$ and $m > 0$. Specifically, logits are defined as cosine similarities between normalized embeddings and class prototypes: $z_j = \cos(\theta_j) = \hat{\mathbf{w}}_j^\top \hat{\mathbf{x}}$, given $\|\hat{\mathbf{x}}\| = \|\hat{\mathbf{w}}_j\| = 1$. Although AM-Softmax was not originally designed for long-tailed classification, it has been shown to outperform standard softmax variants in fine-grained recognition tasks such as biometrics [7, 37], making it a compelling foundation for ecological taxonomy.

Motivation. Softmax-based classifiers like AM-Softmax variants treat all classes with equal importance, implicitly assuming balanced label distributions. Under long-tailed distributions, tail-class examples are rarely observed, causing head-class gradients to dominate the training dynamics. This imbalance induces strong attractive (pull) forces that draw head-class embeddings toward their prototypes, while tail-class prototypes experience cumulative repulsive forces from dominant classes. As a result, tail-class prototypes may become persistently misaligned with their corresponding embeddings, a phenomenon we refer to as *prototype misalignment*. This misalignment disrupts intra-class compactness for tail classes, thereby degrading the overall model's performance (see Figure 2).

Building upon AM-Softmax, we introduce a specialized objective termed *dual-margin penalization loss*. Our formulation explicitly rebalances the influence of head and tail classes through a dual-margin penalization mechanism. In doing so, we extend AM-Softmax to better accommodate fine-grained, long-tailed plant taxonomy. Since AM-Softmax promotes compact intra-class clustering and enhanced inter-class separability, these geometric properties further contribute to improved robustness under domain shifts and open-set conditions, as demonstrated in our experiments.

3.2 Dual-Margin Penalization Loss

We propose a dual-margin penalization loss that operates at the logit level, applying margin constraints to both target and non-target classes as follows:

$$\mathcal{L}_{\text{ours}} = -\log \frac{e^{s(z_y - m_y)}}{e^{s(z_y - m_y)} + \sum_{j \neq y} e^{s(z_j - m_j)}}. \tag{2}$$

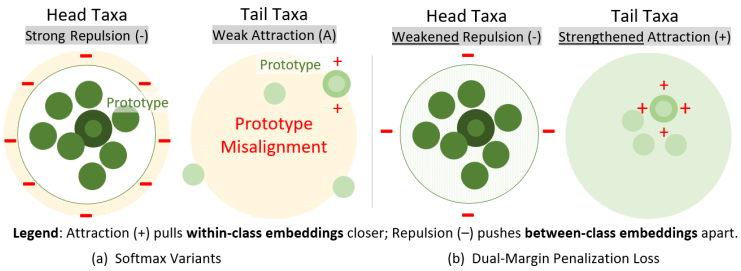

Fig. 2: Schematic comparison of softmax-based losses and the proposed dual-margin penalization loss: (a) Softmax losses cause *prototype misalignment* as head classes strongly repel tail embeddings; (b) the proposed dual-margin loss regulates intra-class attraction and inter-class repulsion, inducing compact tail-class embeddings.

where $m_y = m_0 + \Delta_y$ is the *target margin* for the ground-truth class y, and $m_j = \Delta_j$ is the *non-target margin* for each competing class $j \neq y$. Here, m_0 is a pre-configured base margin, while Δ_y and Δ_j are class-relative margins determined according to Eq. (3). Specifically, for the c-th class, the class-relative margin Δ_c is defined as

$$\Delta_c = m_0 \cdot \left(\frac{\rho_c}{m_0} \right)^{\zeta(\gamma)}, \tag{3}$$

where ρ_c is the pre-computed class prior and γ is a learnable parameter regularized via

$$\mathcal{L}_{\text{reg}} = \sum_{c \in \mathcal{C}} (\Delta_c - \rho_c)^2 . \tag{4}$$

Here, ρ_c is the log-normalized class statistic defined by $\rho_c \propto -\log \frac{N_c}{N}$, normalized such that $\sum_c \rho_c = m_0$, where N_c is the number of samples in class c, N is the total number of samples across C classes, and $c \in \mathcal{C}$. Accordingly, $\Delta_c \in [0, m_0]$ increases as N_c decreases, assigning larger margin adjustments to rare classes. The function $\zeta(\cdot)$ is a smooth, monotonic transformation; in our experiments, we adopt the soft-plus parameterization $\zeta(\gamma) = \log(1 + e^{\gamma})$.

In principle, when a tail-class example is sampled as the target, a larger m_y strengthens the attractive gradient, pulling its embeddings closer to the corresponding prototype and thereby improving intra-class compactness. On the other hand, since head-class samples dominate the training batches, tail classes frequently serve as non-targets. In this case, the non-target margin $m_j = \Delta_j$, which is larger for tail classes, reduces the effective logit $z_j - m_j$ in the softmax denominator, suppressing the repulsive gradient that would otherwise push the tail-class prototype away from its class mean. Without such regulation, the cumulative repulsive forces from abundant head-class samples may persistently shift tail-class prototypes, destabilizing optimization. We formally analyze this effect in Section 4.

Training Objective. The overall training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{ours}} + \lambda \mathcal{L}_{\text{reg}}, \tag{5}$$

where λ controls the strength of regularization. In practice, training TaxoNet requires tuning only the base margin m_0, which governs embedding compactness and separability, while each class-relative margin Δ_k is determined by the empirical class prior ρ_k and the single learnable scalar γ. The scaling factor s is fixed to 32 as large values of s may exacerbate gradient imbalance under long-tailed distributions, potentially hindering convergence for rare classes. The training algorithm and parameter configurations are provided in the Appendix.

3.3 Norm-Guided Sample Selection

Naïve oversampling of extremely scarce tail-class examples often induces memorization, leading to overfitting. Instead, we prioritize informative samples (i.e., those with low embedding norm magnitude), drawing inspiration from margin-based face representation learning [20, 21]. Low-norm samples primarily arise from two scenarios: (1) underrepresented tail classes, and (2) samples exhibiting high intra-class variation due to seasonal or phenotypic changes (e.g., canopy structure, leaf morphology, or fruiting status), which may also occur within head classes. By emphasizing such samples, the model focuses on harder and more diverse instances, thereby improving generalization across ecological conditions while mitigating memorization from repeated tail-class oversampling. Importantly, this sampling strategy is adaptive: as embeddings evolve during training, previously emphasized samples tend to increase in norm magnitude and thus become less likely to be reselected under the norm-guided criterion.

Oversampling with Augmented Diversity. To enhance tail-class representation, we augment each training batch of B images by sampling an additional b tail-class instances using **AugMix** [11], applied stochastically with Bernoulli probability p. From the $B + b$ candidates, we retain the B samples with the lowest embedding norms. This norm-guided criterion is particularly compatible with margin-based softmax formulations, including our proposed dual-margin loss, where embedding geometry correlates with classification confidence: lower norms indicate less confident predictions due to weaker alignment with class prototypes. However, standard softmax loss does not impose such geometric calibration during optimization [21, 37]. We show in Section 5 that low-norm samples are strongly associated with tail classes or instances exhibiting substantial intra-class variability.

4 Theoretical Analysis

We establish theoretical grounds showing that the proposed dual-margin loss mitigates prototype deviation from class-conditional means by characterizing the mean-seeking behavior of class prototypes and analyzing how head-class dominance causes misalignment of tail classes.

Definitions. Our task is to partition a class index set $\mathcal{C}$ into $\mathcal{C} = \mathcal{C}_H \cup \mathcal{C}_T$, where $\mathcal{C}_H$ is the index set for head classes and $\mathcal{C}_T$ is for tail classes. We only require that both index sets are non-empty and that every head class has more samples than any tail class. Accordingly, we partition the sample index set into $\{1, \ldots, N\} = I_H \cup I_T$, where I_H corresponds to head-class samples and I_T to tail-class samples, and we write $N_c = |I_c|$ for the number of samples in class c. For notational simplicity, we let $\mathcal{L} = \mathcal{L}_{\text{ours}}$.

4.1 Prototype Alignment

Under the update rules based on stochastic gradient descent (SGD), each class prototype $\hat{\mathbf{w}}_c$ is updated in the direction of the negative gradient of the loss $\mathcal{L}$,

$$-\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{w}}_c} = -\sum_{i \notin I_c} \frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c} - \sum_{i \in I_c} \frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c}, \tag{6}$$

where $\mathcal{L}_i$ is the loss for i-th sample, and I_c is the set of indices of c-th class samples. The first term on the right-hand side (RHS) of Eq. (6) contributes to the undesired deviation of the prototype $\hat{\mathbf{w}}_c$. When $\hat{\mathbf{w}}_c$ is a prototype of the tail class, its gradient is dominated by the first term as there are a much greater number of samples that are not of the c-th class. Without being properly handled, this causes unstable training and, therefore, poor representation. This will be discussed in the next part. The second term of the RHS of Eq. (6), on the other hand, aligns the prototype $\hat{\mathbf{w}}_c$ to be in the direction of the class-wise embedding mean $\boldsymbol{\mu}_c$ based on the following:

Proposition 1 *Let I_c be the set of indices of samples in the c-th class. Then,*

$$-\sum_{i \in I_c} \frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c} \to \alpha \boldsymbol{\mu}_c \tag{7}$$

as $\sigma_c \to 0$ where $\boldsymbol{\mu}_c$ is class-wise mean

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i \in I_c} \hat{\mathbf{x}}_i, \tag{8}$$

σ_c is the class-wise standard deviation of the predicted probabilities

$$\sigma_c^2 = \frac{1}{N_c - 1} \sum_{i \in I_c} |p_{i,c} - \overline{p}_c|^2 \tag{9}$$

with $\overline{p}_c = \frac{1}{N_c} \sum_{i \in I_c} p_{i,c}$, and $\alpha = N_c(1 - \overline{p}_c)$

Proof. As

$$\frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c} = -(1 - p_{i,c})\hat{\mathbf{x}}_i,$$

we examine the norm

$$\left\| -\sum_{i \in I_c} \frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c} - \alpha \boldsymbol{\mu}_c \right\| = \left\| \sum_{i \in I_c} (\overline{p}_c - p_{i,c}) \hat{\mathbf{x}}_i \right\|.$$

By the triangle inequality:

$$\begin{aligned} \left\| \sum_{i \in I_c} (\overline{p}_c - p_{i,c}) \hat{\mathbf{x}}_i \right\| &\leq \sum_{i \in I_c} |\overline{p}_c - p_{i,c}| \, \|\hat{\mathbf{x}}_i\| \\ &\leq N_c \cdot \max_i |\overline{p}_c - p_{i,c}| \,, \end{aligned} \tag{10}$$

which converges to 0 as the standard deviation σ_c of the predicted probabilities tends to 0, since $\max_i |\overline{p}_c - p_{i,c}| \leq \sqrt{N_c - 1}\, \sigma_c$.

4.2 Tail-Class Prototype vs Class-Wise Mean

As indicated in the gradient decomposition of Eq. (6), when $c \in \mathcal{C}_T$ is of a tail class, its prototype $\hat{\mathbf{w}}_c$ deviates from the class-wise mean $\boldsymbol{\mu}_c$ as a result of frequent updates from head class samples, through minimizing $\mathcal{L}_i$ with $i \in I_H$. In particular, the tail class prototype $\hat{\mathbf{w}}_c$ deviates from the class-wise mean $\boldsymbol{\mu}_c$ due to the head-class gradients $-\frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c}$ for $i \in I_H$. The impact of this undesired deviation on the tail-class prototype is determined by the magnitude of the gradient, that is, $\|\frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c}\|$. To learn properly on the tail classes, this impact must be constrained.

Here, we demonstrate that under the proposed dual-margin penalization loss, its influence is bounded by an exponential term that decays with the class margin m_c, leading to the following proposition:

Proposition 2 *For head-class loss $\mathcal{L}_i$ with $i \in I_H$ and tail-class prototype $\hat{\mathbf{w}}_c$ with $c \in \mathcal{C}_T$, we have*

$$\|\frac{\partial \mathcal{L}_i}{\partial \hat{\mathbf{w}}_c}\| \leq \exp(m_{y_i} - m_c) \tag{11}$$

if $y_i = \arg\max_k p_{i,k}$.

This condition in the proposition holds for most samples, even during early training, as minimizing loss $\mathcal{L}_i$ inherently maximizes p_{i,y_i}. The assumption $y_i = \arg\max_k p_{i,k}$ is made only for simplicity; the proposition holds under the weaker condition $\arg\max_k p_{i,k} \in \mathcal{C}_H \setminus \{c\}$.

Proof. Observe

$$p_{i,c} \leq \frac{\exp(z_{i,c} - m_c)}{\exp(z_{i,y_i} - m_{y_i})} \leq \exp(m_{y_i} - m_c) \tag{12}$$

this holds because the sample loss $\mathcal{L}_i$ corresponds to the head classes (i.e., $i \in I_H$). Consequently, the gradient is $\partial \mathcal{L}_i / \partial \hat{\mathbf{w}}_c = p_{i,c} \hat{\mathbf{x}}_i$ with $\|\hat{\mathbf{x}}_i\| = 1$, so $\|\partial \mathcal{L}_i / \partial \hat{\mathbf{w}}_c\| = p_{i,c}$.

Therefore, for tail classes, the prototype is less influenced by head-class learning.

5 Evaluation

5.1 Experimental Setup

Datasets. We evaluate on three plant datasets: Google Auto-Arborist (AA) [2], iNat-Plantae (iNaturalist-2019 restricted to the only Plantae kingdom) [33], and NAFlora-Mini [24]. AA contains urban street-tree images from three North American regions: AA-West, AA-Central, and AA-East; iNat-Plantae includes citizen-science plant observations across global ecosystems; and NAFlora-Mini comprises herbarium specimens of North American vascular plants. Dataset statistics are provided in the Appendix.

Implementation Details. We employ the ResNet-101 backbone pretrained on ImageNet as the embedding encoder. For dual-margin penalization, we set the scaling factor to $s = 32.0$ and the base margin to $m_0 = 0.15$. All input images are resized to $614 \times 512 \times 3$. The batch size is set to $B = 32$, oversampled with $b = 8$ additional tail-class examples using **AugMix** [11]. We set the Bernoulli probability to $p = 0.1$ for stochastic oversampling.

Comparison and Evaluation Metrics. We compare our model with the standard cross-entropy (CE) classifier, the margin-based AM-Softmax [36], and established state-of the-art (SOTA) models: class-balanced loss (CBL) [5], logit adjustment (LA) [22], and label-distribution-aware margin loss (LDAM) [4]. We also benchmark against multimodal large language models (MLLMs) such as GPT-4o [12] and Gemini-2.5-Flash [32], as well as vision–language foundation model (VLFM), BioCLIP [30], for zero-shot plant classification.

We report rank-1 accuracy (overall performance) and macro-recall (average per-class accuracy) to capture performance across both head and tail taxa. For open-set recognition, we use true negative rate (TNR) and overall accuracy (ACC), both evaluated at a fixed true positive rate (TPR) of 95%.

5.2 Results

Performance Comparison. Table 1 summarizes that TaxoNet achieves significant macro-averaged recall gains, outperforming LDAM (the strongest SOTA baseline) by 6% on Google AA, 3% on iNat-Plantae, and 1% on NAFlora-Mini. Although rank-1 accuracy shows marginal decline, this trade-off reflects improved balance between head and tail classes. Built on AM-Softmax [37], our model explicitly captures fine-grained class geometry in the embedding space. Meanwhile, LDAM, LA, and CBL improve tail-class decision boundaries through margin penalties, logit adjustments, and reweighting strategies, but they lack mechanisms for enforcing within-class compactness and between-class separation, often resulting in embedding drift. We examine other open-world challenges next.

Domain Generalization. We examine regional distribution shift by transferring the AA-Central models of Auto-Arborist to other regions for our model, CBL, and LDAM. Due to taxonomic divergence, performance is evaluated using instance-level matching. As shown in Table 2, our model demonstrates stronger

Table 1: Performance comparison across benchmarking datasets in terms of rank-1 accuracy (R@1) and macro-averaged recall (%), including ablations of TaxoNet under varying configurations. For iNat-Plantae, R@1 is effectively equivalent to recall as the number of test sample is similar across classes. The best and second-best results are highlighted in **bold** and underlined, respectively, across all tables in this paper.

Models	AA-Central		AA-West		AA-East		iNat-Plantae	NAFlora-Mini	
	R@1	Recall	R@1	Recall	R@1	Recall	R@1=Recall	R@1	Recall
Standard CE	91.59	63.55	83.44	59.24	82.87	56.21	79.28	89.48	86.29
CBL [5]	91.46	67.59	84.36	62.72	84.12	63.44	81.38	90.88	90.08
LDAM [4]	**92.31**	67.85	**85.96**	62.82	**85.32**	62.21	81.57	91.26	89.96
LA [22]	90.30	66.90	84.08	61.95	83.32	61.21	80.30	90.73	89.98
AM-Softmax [37]	92.07	64.77	84.99	60.88	83.33	59.22	80.25	89.84	88.85
TaxoNet w/ A, C	92.12	63.91	84.80	60.81	83.77	60.72	78.91	89.84	88.85
TaxoNet w/ B, C	91.30	67.59	84.37	63.15	83.17	62.25	80.01	90.55	89.56
TaxoNet w/ B, D	92.00	67.56	84.36	63.12	84.42	62.47	81.08	91.10	89.83
TaxoNet w/ B, E	91.84	69.48	85.72	65.49	84.53	63.46	81.87	91.48	90.16
TaxoNet w/ B, E, F	91.92	**72.90**	85.94	**67.67**	84.32	**64.96**	**83.21**	**91.52**	**90.40**

Note. A: Base Margin Only; B: Dual-Margin; C: No Oversampling (OS); D: OS with Random Selection; E: OS with Norm-Guided Selection; F: Regularization.

domain generalization, attributed to norm-guided selection, whereas CBL and LDAM rely on random selection.

Table 2: Rank-1 accuracy and macro-recall (%) comparison evaluating regional distribution shift for the model trained on **AA-Central**.

	AA-Central		AA-West		AA-East	
	R@1	Recall	R@1	Recall	R@1	Recall
CBL	91.18	65.93	62.73	44.15	47.19	37.71
LDAM	**91.41**	68.37	**65.95**	46.50	49.26	39.95
TaxoNet	91.14	**70.18**	64.43	**48.10**	**52.13**	**40.53**

Open-Set Recognition. In open-set evaluation, the goal is to flag previously unseen taxa for expert review. Taxonomists assess whether the sample belongs to a known class or represents a novel taxon. Our results in Table 3 shows that our model achieves over 90% TNR on Regions West and Central of Auto-Arborist, outperforming CBL and LDAM in unknown-class rejection. This comes from its base margin, which improves embedding separation. Increasing the margin can further improve unknown detection, though it may marginally reduce performance on known classes.

Additional Insights. We compare the performance of our model with CE, CBL [5], and LDAM [4] at both family and genus levels in Table 4. Our model marks the top performance, especially in tail groups, while LDAM outperforms in head

Table 3: Performance comparison in terms of TPR, TNR, and acc. for open-set evaluation on **AA-Central**, with the threshold calibrated to 95% TPR. An unknown set (88 novel taxa, 8,845 samples) is used to simulate open-set conditions.

	AA-Central			**AA-West**		
	TPR	TNR	Acc.	TPR	TNR	Acc.
CBL	91.83	86.34	89.08	86.93	84.76	85.84
LDAM	92.29	88.59	<u>90.44</u>	88.84	87.09	<u>87.96</u>
TaxoNet	93.25	91.28	**92.26**	90.91	89.84	**90.38**

Table 4: Performance comparison using macro-averaged recall (%) across taxonomic levels for TaxoNet, CE, CBL, and LDAM on **AA-Central**. **Taxa group** is defined by training cardinality: head (>2,000 samples), between (100–2,000 samples), and tail (<100 samples). Class-level recall results are reported in the Appendix.

	Taxa Group	**TaxoNet**	**CE**	**CBL**	**LDAM**
Family	Overall	**78.70**	72.67	<u>74.57</u>	72.09
Genus	Head	92.37	92.60	<u>93.34</u>	**94.13**
	Between	<u>86.13</u>	81.60	84.80	**86.80**
	Tail	**57.06**	43.18	<u>50.27</u>	44.18
	Overall	**72.90**	63.55	<u>67.59</u>	66.85

classes. This reflects the effect of norm-guided sampling, which prioritizes tail examples, unlike random sampling in LDAM. Interestingly, our embeddings reflect taxonomic relationships for hierarchical classification—a behavior that we leave for future exploration.

5.3 Analysis

Ablation Studies. We analyze the contribution of each model component: base margin, class-relative margin (with and without power-based regularization), and norm-guided sample selection. As shown in Table 1, their combination yields the highest gain. Notably, power-based regularization stabilizes the base margin by adapting class-relative margins to prior statistics and learned signals, improving generalization in class imbalanced and fine-grained settings.

Figure 3 shows that high-norm instances primarily originate from head classes or confidently learned examples with distinct taxonomic traits. Conversely, low-norm instances typically arise from tail classes or samples with high within-class variation due to seasonal and morphological changes. As reported in Table 1, these findings validate our norm-guided selection strategy, which prioritizes low-norm samples to enhance tail-class generalization (Table 4) and mitigate overfitting from naïve oversampling. TaxoNet with dual-margin and norm-guided selection consistently outperforms random selection. We reemphasize that *the norm proxy applies only to models built on margin-based softmax variants [21], including our dual-margin penalization loss, and is not compatible with baselines such as CBL, LDAM, or LA.*

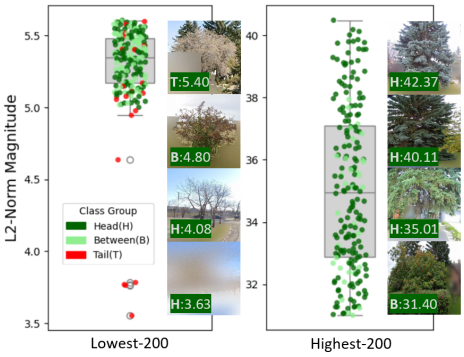

Fig. 3: Embedding norm distribution for the 200 highest-norm and 200 lowest-norm samples in **AA-Central**. High-norm examples are concentrated in head and between classes; low-norms are from tail or those with greater within-class variations (e.g., due to seasonal changes). **Class Group** is defined by training cardinality: head > 2,000 samples, tail < 100 samples, between otherwise.

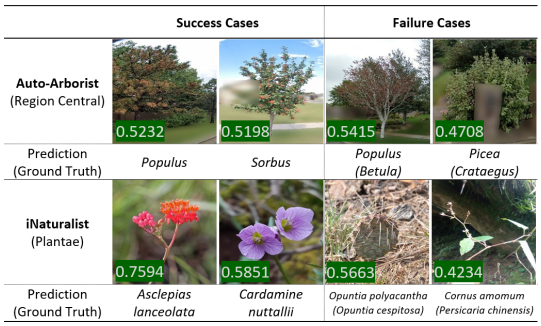

Fig. 4: Success and failure cases of model predictions on AA-Central and iNat-Plantae. Softmax prediction scores are shown for each test image.

Success and Failure Cases. We analyze success and failure cases in Figure 4. Our model performs well under challenging conditions such as seasonal variation in morphology and flower-only inputs that lack contextual cues like leaves or growth form. Failure cases often involve visually ambiguous samples: *Opuntia polyacantha* is misclassified as *Opuntia cespitosa*, a closely related species from the same genus, and *Cornus amomum* as *Persicaria chinensis*, likely due to leaf-only views. These limitations underscore the need for future multi-view, multimodal models, potentially leveraging advances in MLLMs or VLFMs, to support reasoning.

5.4 Further Benchmarking

Comparison with MLLMs. Given general-purpose Multimodal Large Language Models (MLLMs) excel at broad tasks [12, 32], we test whether they can handle biodiversity recognition. Table 5 compares our model against GPT-4o [12] and Gemini-2.5-Flash [32] on the iNat-Plantae dataset. To evaluate these models, we used zero-shot (ZS) persona-based prompts and asked the AI to act as a botanical expert [28] and a chain-of-thought (CoT) approach reasoning [39]. This guided the models to reason through the taxonomic hierarchy, first identifying the genus before narrowing down to the specific species.

TaxoNet significantly outperforms MLLMs across both genus and species levels, achieving 2× higher performance in species-level accuracy. Although CoT

Table 5: Comparative analysis of taxonomic classification performance. We report macro-averaged recall (%) on the iNat-Plantae dataset, where taxa group is defined by training cardinality: 10 head taxa (>500 samples), 10 between taxa (50–500 samples), and 20 tail taxa (<50 samples). TaxoNet consistently outperforms both general-purpose MLLMs and domain-specific vision-language models, particularly in the rare-taxa (tail) regime. See Appendix for the class-level results.

	Taxa Group	TaxoNet	GPT-4o		Gemini-2.5		BioCLIP
			ZS	CoT	ZS	CoT	ZS
Genus	Head(10)	**96.53**	92.16	91.20	91.67	92.46	60.34
	Between(10)	**94.91**	87.23	88.62	88.75	88.91	48.00
	Tail(20)	**96.71**	91.68	91.87	92.60	93.11	51.92
	Overall	**95.96**	89.08	89.91	90.86	91.50	53.42
Species	Head(10)	**96.67**	13.00	13.00	13.33	16.67	63.33
	Between(10)	**93.33**	43.33	50.00	46.67	53.34	76.67
	Tail(20)	**76.67**	16.67	11.67	26.67	31.67	68.33
	Overall	**83.21**	35.28	32.51	40.39	41.46	70.09

reasoning improves over ZS prompting in Gemini-2.5-Flash's results, it yields negligible gains for GPT-4o, highlighting the inherent limitations of general-purpose models in fine-grained botanical discrimination. We provide additional results for AA-Central and prompting templates in the Appendix.

Comparison with VLFM. Unlike open-domain MLLMs, BioCLIP leverages large-scale biodiversity resources spanning plants, animals, fungi, and microbes, and is therefore not strictly domain-specialized. As BioCLIP includes training examples overlapping with iNat-Plantae, it attains higher species-level performance than MLLMs. However, MLLMs, aided by text-guided CoT reasoning, enhances interpretability at coarser taxonomic levels. Overall, our proposed domain-specialized model establishes a strong foundation for open-world taxonomic monitoring when deployed in the plant domain.

6 Conclusion

We propose a dual-margin penalization objective that balances learning signals across dominant and rare taxa, forming the basis of *TaxoNet*, an embedding framework for plant taxonomy recognition. Validated across urban, citizen-science, and herbarium plant datasets, TaxoNet consistently outperforms baselines, including state-of-the-art multimodal foundation models.

TaxoNet represents a significant advancement in fine-grained and long-tailed representation learning tailored to the structural complexities of ecological data. By addressing additional open-world deployment challenges such as spatiotemporal shifts, hierarchical taxonomic structure and novel taxa, our framework enables scalable monitoring of plant diversity across urban and natural forest ecosystems. Furthermore, it provides a robust backbone for citizen science platforms such as *iNaturalist*, *Flora Incognita*, *Pl@ntNet*). We anticipate that our model can be extended to broader biodiversity monitoring tasks.

Acknowledgements

We thank Kaleb Mesfin Asfaw, currently at the University of Calgary, for his insightful comments, particularly on the preprocessing of the Google Auto Arborist dataset. This work was supported by the 2026 Academic Research Promotion Support Project of Changwon National University, South Korea.

References

1. Barajas-Barbosa, M.P., Craven, D., Weigelt, P., et al.: Global patterns of vascular plant alpha diversity. Nat. Commun. **13**(1), 1–9 (2022)
2. Beery, S., Wu, G., Edwards, T., Pavetic, F., Majewski, B., Mukherjee, S., Chan, S., Morgan, J., Rathod, V., Huang, J.: The auto arborist dataset: a large-scale benchmark for multiview urban forest monitoring under domain shift. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21294–21307 (2022)
3. Blackmore, S.: Botanic gardens are vital for delivering the kunming-montreal global biodiversity framework. Biological Diversity **1**(3-4), 120–123 (2024)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
5. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
6. Cui, Y., Song, Y., Sun, C., Howard, A.G., Belongie, S.J.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4109–4118 (2018), https://api.semanticscholar.org/CorpusID:43993788
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
8. Duraiappah, A.K., Rogers, D.: The intergovernmental platform on biodiversity and ecosystem services: opportunities for the social sciences. Innovation: The European Journal of Social Science Research **24**(3), 217–224 (2011)
9. Gharaee, Z., Gong, Z., Pellegrino, N., Zarubiieva, I., Haurum, J.B., Lowe, S.C., McKeown, J.T.A., Ho, C.Y., McLeod, J., Wei, Y.C., Agda, J., Ratnasingham, S., Steinke, D., Chang, A.X., Taylor, G.W., Fieguth, P.: A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. In: Advances in Neural Information Processing Systems. vol. 36, pp. 43593–43619 (2023)
10. Govaerts, R., Nic Lughadha, E., et al.: The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. Scientific Data **8**(1), 1–10 (2021)
11. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple method to improve robustness and uncertainty under data shift. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=S1gmrxHFvB
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

13. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of big data **6**(1), 1–54 (2019)
14. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18729–18738 (2022), https://api.semanticscholar.org/CorpusID:247940176
15. Koo, H.: Next visit diagnosis prediction via medical code-centric multimodal contrastive ehr modelling with hierarchical regularisation. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 41–55 (2024)
16. Lewis, K.M., Mu, E., Dalca, A.V., Guttag, J.: Gist: Generating image-specific text for fine-grained object classification. arXiv preprint arXiv:2307.11315 (2023)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Low, C.Y., Cha, M., Wäldchen, J., Gummadi, K.P.: Open-set classification for rare and unknown urban tree taxa. In: International Conference on Information Technology for Social Good (GoodIT '25). pp. 1–7. ACM, Antwerp, Belgium (2025)
20. Low, C.Y., Chai, J.C.L., Park, J., Ann, K., Cha, M.: Slackedface: Learning a slacked margin for low-resolution face recognition. In: Proc. of the BMVC (2023)
21. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14220–14229 (2021)
22. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
23. Park, J., Low, C.Y., Teoh, A.B.J.: Divergent angular representation for open set image recognition. IEEE Transactions on Image Processing **31**, 176–189 (2021)
24. Park, J., de Lutio, R., Rappazzo, B., Ambrose, B., Michelangeli, F., Watson, K., Belongie, S., Little, D.: Naflora-1m: Continental-scale high-resolution fine-grained plant classification dataset. Journal of Data-centric Machine Learning Research (2024)
25. Sala, O.E., Stuart Chapin, F., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., et al.: Global biodiversity scenarios for the year 2100. science **287**(5459), 1770–1774 (2000)
26. SCBD: Biodiversity and the 2030 agenda for sustainable development. Tech. rep., Secretariat of the Convention on Biological Diversity (2017)
27. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence **35**(7), 1757–1772 (2012)
28. Shanahan, M., McDonell, K., Reynolds, L.: Role play with large language models. Nature **623**(7987), 493–498 (2023)
29. Smith, J., Patel, S.: Open-set classification strategies for long-term acoustic biodiversity monitoring. Journal of the Acoustical Society of America **151**(6), 4028–4042 (2024)
30. 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., Chao, W.L., Su, Y.: Bioclip: A vision foundation model for the tree of life. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19412–19424 (2024)

31. Sun, H., He, X., Zhou, J., Peng, Y.: Fine-grained visual prompt learning of vision-language models for image recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5828–5836 (2023)
32. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
33. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
34. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLoS computational biology **14**(4), e1005993 (2018)
35. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
36. Wang, F., Cheng, J., Liu, W., Liu, H.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia (ACM MM). pp. 1041–1049 (2017)
37. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
38. Wang, Y., Zhao, Q.: Open-set fish species recognition with non-parametric methods. Sensors **25**(5), 1570 (2023)
39. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
40. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)