Proceedings of Engineering and Technology Innovation, vol. 34, 2026, pp. 47-62

Transformer-Based Hybrid Convolution Attention Framework for Plant Species Classification

Karnan Arunasalam*, Ragupathy Rengaswamy

Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu, India

Received 16 February 2026; revised 11 June 2026; accepted 18 June 2026

DOI: https://doi.org/10.46604/peti.2026.16186

Abstract

Fine-grained plant species identification remains challenging because of high intra-class variation and strong inter-class similarity among closely related taxa. Existing hybrid convolutional neural network (CNN)-Transformer architectures progressively attenuate discriminative texture cues during deep attention processing. To address this limitation, this study aims to develop the Transformer-based hybrid convolution attention framework (TB-HCAF), a targeted extension of CoAtNet that preserves complementary local morphological and global structural representations through a dual-path architecture, a parameter-efficient spatial alignment block, and a channel-wise hybrid fusion. The framework is evaluated using a leakage-safe, group-stratified k-fold protocol eliminates specimen-level data leakage and validated through comparative experiments, ablation studies, effective receptive field (ERF) analysis, and cross-dataset evaluation. TB-HCAF achieves high Top-1 accuracy on PlantCLEF 2015, surpassing representative baseline models, and further demonstrates strong zero-shot cross-dataset generalization on Oxford 102 Flowers and iNaturalist 2018. ERF analysis confirms improved discrimination of morphologically similar species and robustness across long-tailed distributions.

Keywords: fine-grained visual categorization, hybrid inductive bias, plant species identification, leakage-safe evaluation, CoAtNet architecture

1. Introduction

The accelerating biodiversity crisis has created an urgent demand for scalable and automated ecological monitoring systems. Plant diversity is a key indicator of ecosystem stability; however, accurate species identification remains challenging owing to the large number of species and the high visual similarity among closely related taxa [1-2]. Automated plant recognition is typically framed as a fine-grained visual categorization (FGVC) task, requiring models to distinguish closely related species using subtle morphological cues such as venation patterns, leaf margins, and surface textures [3-4]. Two fundamental challenges complicate this task: substantial intra-class variation caused by environmental conditions and developmental stages, and high inter-class similarity among taxonomically similar species. Together, these challenges necessitate models capable of achieving both robustness and fine-grained discrimination.

Early approaches relied on CNNs for their strong ability to capture localized spatial features [5-6], with successful applications in plant disease detection [7] and phenological analysis [8]. Yet, CNNs suffer from limited receptive field growth, which constrains their ability to model long-range structural dependencies essential for representing complex plant morphology. Transformer-based architectures [9-10] address this limitation by leveraging self-attention mechanisms to capture global contextual relationships. Despite these advances, transformer models lack strong spatial inductive priors, which reduces their effectiveness for fine-grained texture discrimination, particularly under the data-scarce and class-imbalanced conditions typical of botanical datasets.

* Corresponding author. E-mail address: mekarnan@gmail.com

Hybrid CNN–Transformer architectures combine local feature extraction with global contextual reasoning and have demonstrated promising results in visual recognition tasks [11-12]. However, a persistent limitation remains: fine-grained local features are progressively attenuated in deeper attention stages of these architectures [13]. Furthermore, existing studies largely overlook specimen-level data leakage, in which multiple images of the same biological individual appear across training and evaluation splits, leading to inflated performance estimates and compromised generalization. These limitations reveal two research gaps that remain largely unaddressed:

(1) Architectural limitation: Existing hybrid models progressively attenuate fine-grained local texture representations during deep transformer processing, thereby hindering the discrimination of morphologically similar species under conditions of data scarcity and class imbalance [11-13].
(2) Methodological limitation: Specimen-level data leakage remains largely unaddressed in botanical image datasets, resulting in systematically biased evaluation outcomes and overestimated classification performance.

In response to these challenges, this study proposes TB-HCAF, a domain-specific structural adaptation of CoAtNet for botanical fine-grained visual categorization. Rather than introducing a fundamentally new architecture, TB-HCAF refines the CoAtNet baseline through three targeted structural modifications. These modifications collectively mitigate the progressive attenuation of discriminative morphological features during deep attention processing without incurring excessive parameter overhead. The overall system operates within a leakage-safe experimental pipeline that enforces strict specimen-level separation across all data splits, as illustrated in Fig. 1.

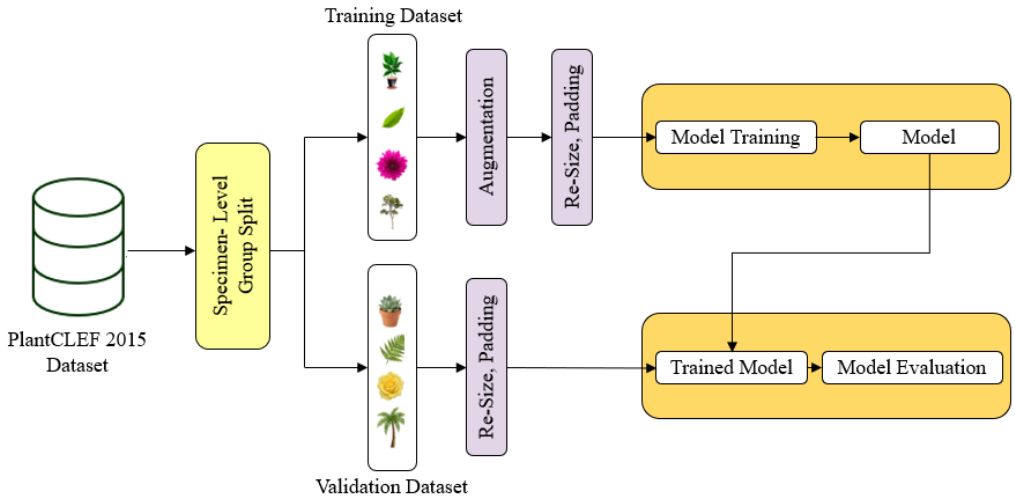


Fig. 1 Overview of leakage-safe experimental pipeline for hybrid plant species classification

The bifurcated architecture and its stage-wise processing mechanisms are illustrated in Fig. 2. The complete end-to-end pyramidal architecture is presented in Fig. 3. To ensure experimental validity, a leakage-safe group-stratified k-fold protocol is introduced to enforce specimen-level isolation between training and evaluation sets. This prevents models from exploiting observation-specific artifacts and ensures that reported performance reflects genuine species-level generalization. All comparative baselines are evaluated using architecture-specific hyperparameters under an identical leakage-safe data pipeline to ensure fair comparison. The main contributions of this study are as follows:

(1) TB-HCAF architecture: A domain-specific adaptation of the CoAtNet baseline, engineered to preserve fine-grained morphological features without constituting a standalone architecture. Specifically, three targeted modifications are introduced: (i) a shortcut convolutional path to retain high-resolution texture representations; (ii) a parameter-efficient spatial alignment block (SAB) for resolution alignment without learnable parameters; and (iii) a channel-wise hybrid fusion mechanism to produce a unified 1,920-dimensional representation that jointly encodes local textures and global structural context. ERF analysis is further employed to investigate the interpretability of the proposed modifications.

(2) Leakage-safe evaluation protocol: A group-stratified k-fold cross-validation strategy designed for specimen-correlated botanical datasets, with quantitative analysis demonstrating that conventional random splits overestimate classification performance by 0.81% – 6.31%, exposing the previously overlooked impact of specimen-level data leakage on reported accuracy.

(3) Cross-dataset generalization validation: Comprehensive experiments demonstrate statistically significant performance improvements and strong zero-shot cross-dataset generalization. The proposed framework consistently outperforms representative baselines under rigorously tuned evaluation conditions, demonstrating substantial robustness against long-tailed minority classes.

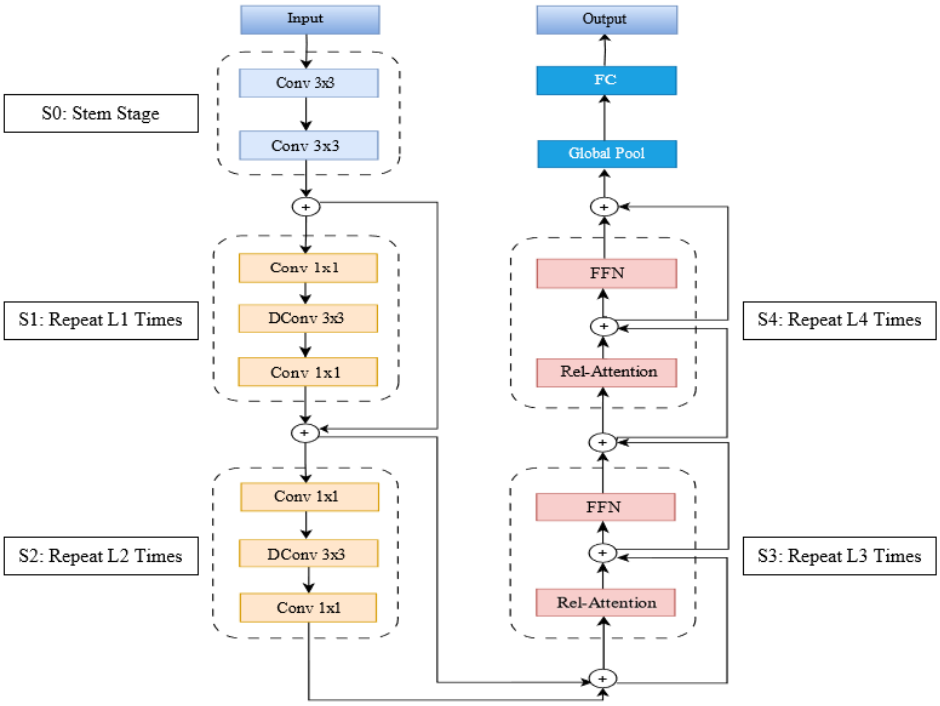


Fig. 2 Hierarchical transition from convolutional feature extraction (Stages 0–1) to attention-based global reasoning (Stages 3–4) in CoAtNet baseline architecture

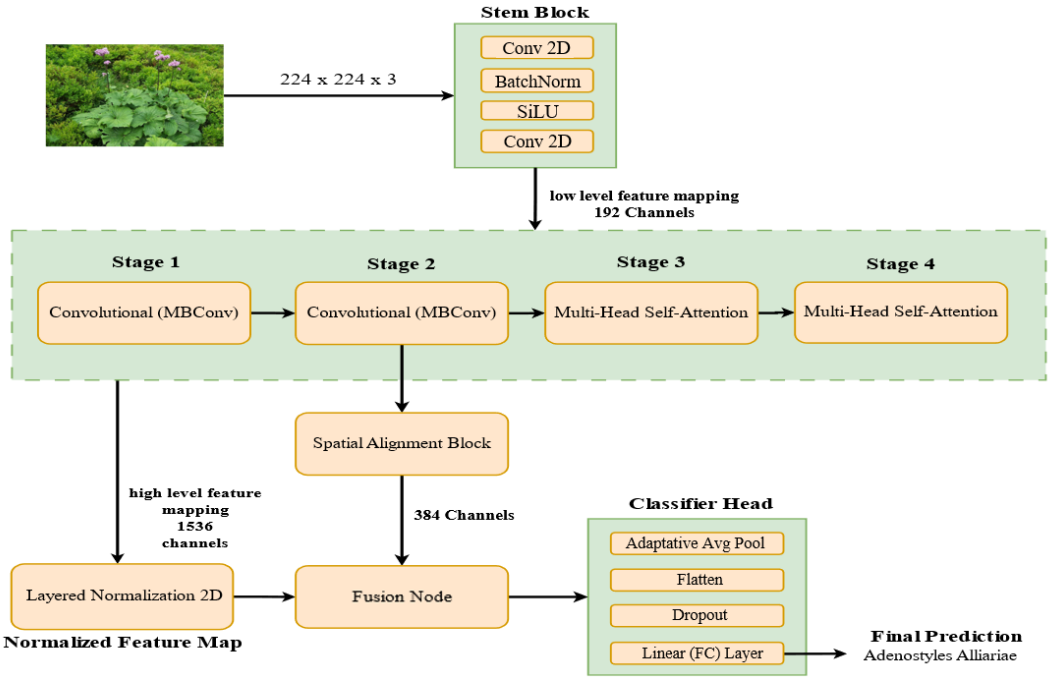


Fig. 3 End-to-end four-stage pyramidal hybrid convolution attention network architecture

The remainder of this paper is organized as follows. Section 2 details the methodology, covering the bifurcated hybrid architecture, the TB-HCAF design, and the leakage-safe evaluation protocol. Section 3 presents results and discussion, including comparative performance analysis, training dynamics, ablation studies, ERF interpretability, and cross-dataset generalization. Section 4 summarizes the conclusions, discusses limitations, and outlines directions for future work.

2. Methodology

Fine-grained recognition of plant species requires a learning framework that is both sensitive to subtle morphological variation and robust to dataset-specific biases. Discriminative features separating closely related species are often distributed across multiple plant organs and captured under diverse environmental conditions [1]. Moreover, botanical datasets frequently contain multiple correlated images of the same biological specimen, introducing hidden dependencies that can compromise the reliability of evaluations [2]. The proposed methodology addresses these challenges by integrating a hybrid inductive-bias architecture with a leakage-safe data-partitioning strategy and a carefully designed optimization framework.

2.1 Hybrid inductive bias: Motivation and theoretical basis

Deep neural networks learn a mapping function f: X→Y within a hypothesis space defined by architectural inductive biases. These biases influence how models generalize, particularly under limited and imbalanced data conditions commonly observed in botanical datasets. Plant imagery exhibits both local and global structural characteristics. At the local level, morphological patterns such as venation networks, surface textures, and leaf margin structures are spatially consistent and shared across individuals of the same species. At the global level, accurate species recognition requires interpreting these local features within a broader anatomical context, such as the spatial relationship between leaves, flowers, and branching structures. This hierarchical organization motivates a hybrid inductive bias in which local feature extraction and global contextual reasoning are treated as complementary processes rather than competing components.

This hybrid formulation aligns with botanical taxonomic practice, in which fine-grained visual details such as venation topology, epidermal surface features, and margin morphology must be interpreted alongside overall plant structure to achieve reliable species discrimination. Unlike generic hybrid CNN-Transformer models, the proposed TB-HCAF framework adopts a domain-aligned inductive bias in which convolutional layers are expected to capture local morphological patterns, whereas transformer layers are expected to model global structural relationships, together supporting a more biologically relevant feature representation.

2.2 Architectural realization: From convolution to attention

To implement the hybrid inductive bias, the present study builds upon the CoAtNet architecture [15], which implements the transition from convolution-dominated processing to attention-based reasoning as network depth increases. A structural summary of this hierarchical progression is illustrated in Fig. 2, while the complete end-to-end architecture is provided in Fig. 3. In the initial processing stages, convolutional operations encode translation equivariance and local spatial continuity. For an input feature tensor $\boldsymbol{X} \in \mathbb{R}^{(H\times W\times C)}$, the convolutional activation at spatial position (i, j) is computed as:

$$Y_{i,j} = \sum_{(a,b)\in\Delta} K_{a,b} \cdot X_{i+a,\,j+b} \tag{1}$$

where Δ denotes the kernel's spatial neighborhood, and K represents learnable weights. Convolution is effective in capturing local patterns such as venation topology and leaf-margin morphology, which are essential for species discrimination. However, convolution alone is limited in capturing long-range spatial dependencies. Consequently, deeper stages employ transformer blocks, incorporating relative self-attention [16] to model interactions between all spatial positions. The pre-softmax attention score between positions i and j is defined as:

$$A_{i,j} = x_i \cdot x_j + w_{i-j} \tag{2}$$

where x_i and x_j are feature vectors, and $w_{i,j}$ represents a learnable relative positional bias. This formulation enables the attention mechanism to emphasize morphologically informative regions such as venation junctions, leaf boundaries, and reproductive structures, therefore improving discrimination of visually similar species compared to conventional CNNs with fixed receptive fields [15-16]. A critical limitation of this single-stream design is the progressive attenuation of discriminative micro-features extracted in early convolutional stages. This occurs as feature maps pass through successive attention layers operating on progressively downsampled spatial representations. In fine-grained plant classification, where subtle morphological cues are decisive for species discrimination, this attenuation may degrade model performance, particularly for morphologically similar species and minority classes.

To mitigate this issue, TB-HCAF introduces a bifurcated dual-path design that retains high-resolution convolutional features via a dedicated shortcut convolutional path, preserving fine-grained local representations at the fusion stage alongside deep transformer features. This design also provides a direct gradient pathway from the classification objective to early convolutional layers, helping alleviate gradient attenuation and improve convergence stability in deep hybrid architectures.

2.3 Hierarchical backbone and classification head

As shown in Fig. 3, the architecture follows a four-stage pyramidal design, beginning with a convolutional stem that downsamples the input from 224×224 to 56×56, reducing low-level redundancy while preserving structural edges and texture boundaries. Intermediate stages employ MBConv blocks [5-17] with squeeze-and-excitation (SE) modules. These modules perform adaptive channel recalibration and suppress background-dominated features, improving robustness to intra-class variation arising from environmental factors such as clutter, illumination changes, and phenological differences. To prevent attenuation of fine-grained morphological cues in deeper attention stages, three targeted architectural modifications are introduced over the baseline CoAtNet. Such attenuation has been insufficiently addressed in prior hybrid CNN-Transformer models for botanical domains. First, a bifurcated dual-path design diverges at Stage 2, where a shortcut convolutional path preserves high-resolution local features in parallel with the main transformer path, which proceeds through Stage 3 and Stage 4 multi-head self-attention blocks to model long-range inter-organ dependencies.

This dual-path design is intended to preserve discriminative micro-features (such as venation patterns and margin serration) despite successive spatial downsampling in the attention layers. By maintaining species-discriminative local cues alongside global structural context, this design directly addresses high intra-class variation. Second, a parameter-efficient SAB aligns the Stage 2 feature map from 28×28 to 7×7 spatial resolution via non-parametric adaptive average pooling. For an input feature map $X \in \mathbb{R}^{(C\times H_{in}\times W_{in})}$, the spatially aligned output Y is computed as:

$$Y_{c,i,j} = \frac{1}{k_h . k_w} \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X_{c,\lfloor i.s_h \rfloor+m,\lfloor j.s_w \rfloor+n} \tag{3}$$

where c denotes the channel index, and k_h, k_w, s_h, s_w represent the adaptive kernel sizes and strides computed to map the 28×28 Stage 2 feature map to a 7×7 resolution. This operation introduces no learnable parameters and preserves computational efficiency. Third, a channel-wise hybrid fusion mechanism concatenates layer-normalized transformer features with SAB-aligned convolutional features to form a unified 1,920-dimensional representation:

$$F_{fused} = [LayerNorm(S_4) \oplus SAB(S_2)] \in \mathbb{R}^{1920} \tag{4}$$

where $\oplus$ denotes channel-wise concatenation. The resulting representation comprises 384 channels of high-resolution convolutional features and 1,536 channels of global transformer features. This joint encoding preserves both fine-grained local texture and global structural context, enabling the classifier to weight each pathway independently, which may enhance robustness to intra-class variation and inter-class similarity. The classification head applies global average pooling, layer normalization, and dropout to promote translation invariance and reduce overfitting.

2.4 Training configuration and hyperparameter selection

All baseline models and the proposed TB-HCAF architecture were trained under a unified evaluation protocol, encompassing consistent data splits, input resolution, augmentation strategy, optimizer configuration, learning rate scheduling, batch size, and early stopping criteria to ensure fair comparison. Training was conducted under the severe class-imbalance conditions characteristic of large-scale botanical datasets. To mitigate bias toward dominant species, label smoothing was applied to soften target distributions [18]:

$$y_k^{LS} = (1-\varepsilon)y_k + \frac{\varepsilon}{C} \tag{5}$$

where ε = 0.1 and C denotes the total number of classes. This reduces overconfidence and may improve generalization across species. Optimization was performed using AdamW [19], whose decoupled weight decay improves stability in hybrid convolution-transformer architectures. A one-cycle learning rate schedule was employed to encourage rapid exploration during early training stages, followed by controlled convergence. A batch size of 32 balances gradient estimation quality against the 16 GB GPU memory constraints. Data augmentation was performed using TrivialAugmentWide [20], while transfer learning was applied using ImageNet-pretrained initialization. TB-HCAF specifically utilized ImageNet-21K pretraining [21]. The complete hyperparameter configuration is summarized in Table 1.

Table 1 Unified training configuration for all models

Parameter	Value	Description
Optimizer	AdamW [19]	Adaptive moment estimation with decoupled weight decay
Learning rate	1 × 10−3	Peak learning rate under the one-cycle policy
Weight decay	0.05	Regularization strength
Batch size	32	Optimized for 16 GB GPU memory
Epochs	100	Early stopping with patience of 5 epochs
Loss function	Label Smoothing CE [18]	Smoothing factor ϵ = 0.1
Precision	AMP (FP16)	Automatic mixed precision training

Training stability was further supported through several complementary mechanisms. First, layer normalization within transformer blocks helped stabilize activation distributions across attention layers and mitigated gradient degradation. Second, AdamW [19] provided adaptive gradient updates with decoupled regularization. Third, automatic mixed precision (AMP, FP16) training improved computational efficiency and memory utilization while maintaining stable numerical behavior. In addition, the shortcut convolutional pathway facilitated direct gradient propagation to earlier convolutional layers, helping alleviate vanishing-gradient effects in deep hybrid architectures.

2.5 Leakage-safe data preparation and validation protocol

A critical challenge in botanical image classification is specimen-level data leakage, wherein multiple correlated images originating from the same biological specimen are distributed across both training and evaluation splits. Such leakage enables models to exploit observation-specific artifacts rather than learn genuine species-discriminative morphological features, thereby producing inflated performance estimates. To prevent such leakage, a leakage-safe group-stratified k-fold validation protocol is employed.

All images were first grouped at the specimen level. When available, official observation IDs provided in the PlantCLEF 2015 metadata were used directly. Otherwise, a deterministic composite specimen ID was constructed from the tuple (author, date, GPS location), serving as a conservative proxy for biological identity. A systematic metadata audit was conducted across all 113,205 images to assess the reliability of specimen grouping. Approximately 78.3% of images contained valid official observation IDs (extracted from the XML metadata field observation IDs) and were assigned unambiguously to specimen buckets. The remaining 21.7% required composite specimen IDs derived from the (author, date, latitude, and longitude) fields.

Among these composite-ID images, approximately 64.2% contained complete author, date, and GPS fields, while 35.8% exhibited at least one missing field, most commonly latitude and longitude. This observation is consistent with the dataset documentation, which indicates that GPS coordinates are frequently absent for remote or rural collection sites.

For incomplete entries, grouping was performed conservatively using only the available fields, intentionally erring toward over-grouping to prevent leakage. Specimen bucket sizes ranged from 1 to 47 images, with a median of 4 and a mean of 2.71 images per bucket, reflecting the multi-organ acquisition structure of PlantCLEF 2015 in which a single plant individual may contribute branch, entire, flower, fruit, leaf, leaf-scan, and stem images. Although the heuristic may occasionally group distinct individuals photographed concurrently at the same location, this represents a conservative bias: it prevents leakage from spatially and temporally correlated artifacts without artificially inflating performance estimates. Conversely, under-grouping is considered unlikely given the strict spatio-temporal matching constraints applied.

All images sharing the same specimen identifier are treated as an atomic unit regardless of organ type or acquisition condition, ensuring that reported performance reflects genuine taxonomic generalization rather than memorization of observation-specific cues. A group-stratified k-fold cross-validation scheme is subsequently applied under two constraints: (i) all images within a specimen bucket are assigned exclusively to either the training or validation partition, and (ii) the overall species distribution is approximately preserved across folds. The optimal number of folds is determined through sensitivity analysis, as summarized in Table 2.

Table 2 Sensitivity analysis of group k-fold parameters on PlantCLEF 2015

Metric	k = 3	k = 4	k = 5	k = 6	k = 8	k = 10
Mean Validation Accuracy (%)	97.80	98.20	98.65	98.68	98.70	98.71
Standard Deviation (σ, %)	1.20	0.80	0.30	0.28	0.25	0.24
Total Training Time (hours)	10.8	14.5	18.2	21.9	29.4	36.8
Incremental Gain (%)	-	+0.40	+0.45	+0.03	+0.02	+0.01

Validation accuracy increases markedly at lower fold counts but saturates beyond k = 5. The increment from k = 5 to k = 10 yields only a marginal 0.06% gain while nearly doubling total training time from 18.2 to 36.8 hours. The trade-off between validation accuracy and training cost exhibits diminishing returns as k increases. The fold-wise standard deviation stabilizes below 0.3% at k = 5, indicating reliable and statistically stable performance estimation. Consequently, k = 5 was selected as the optimal balance between statistical reliability and computational efficiency.

2.6 Experimental setup and dataset description

All experiments were conducted on a workstation equipped with an NVIDIA Tesla T4 GPU (16 GB) and an Intel Xeon processor, running Ubuntu Linux and CUDA 11.x. The TB-HCAF model is based on the CoAtNet-3 architecture pretrained on ImageNet-21K and implemented using the timm library in PyTorch. Data augmentation was performed using Albumentations, and automatic mixed precision training was employed throughout. Experiments were primarily conducted on the PlantCLEF 2015 dataset [1], which contains 113,205 images spanning 1,000 plant species, organized into 41,794 multi-view observation events covering multiple plant organs, including branch, flower, fruit, leaf, leaf scan, and stem. For cross-dataset evaluation, the Oxford 102 Flowers dataset [14] (8,189 images, 102 classes) and the iNaturalist 2018 dataset [2] (437,513 images, 8,142 species) were used to assess fine-grained recognition and robustness under long-tailed distributions, respectively.

All baseline models were trained under the unified protocol described in Table 1, including data splits, augmentation strategies, input resolutions, pretraining sources, and early stopping criteria, with no separate tuning performed per model, ensuring fair and reproducible comparison. The visual diversity of PlantCLEF 2015 images across different plant organs and varying quality levels is illustrated in Fig. 4.

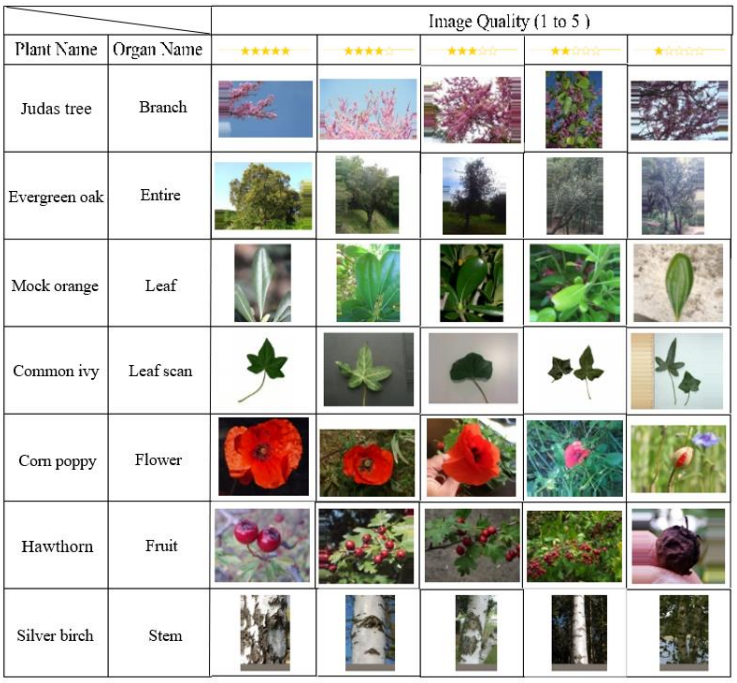


Fig. 4 Visual diversity of PlantCLEF 2015 images across organs and quality levels

3. Results and Discussion

This section presents a comprehensive empirical evaluation of the proposed TB-HCAF framework for fine-grained plant species classification. The analysis examines predictive performance, optimization behavior, robustness under different training regimes, interpretability through receptive field analysis, error characteristics, computational requirements, and cross-dataset generalization. All experiments were conducted using the leakage-safe group k-fold protocol described in Section 2, ensuring that reported results reflect generalization to unseen plant specimens rather than memorization of observations.

3.1 Comparative performance analysis

TB-HCAF was evaluated against six representative CNN and Transformer baselines, namely MobileNetV2 [5], BiT (ResNet-50) [22], ViT-B/16 [23], Swin-B [24-25], EfficientNetV2-M [17], and ConvNeXt V2-Base [26], on PlantCLEF 2015 under a two-level experimental control. At the first level, dataset splits, input resolution (224×224), augmentation pipeline (TrivialAugmentWide), optimizer (AdamW), weight decay, batch size (32), learning rate schedule, and early stopping criteria were held identical across all models, as detailed in Table 1. At the second level, architecture-internal hyperparameters known to cause divergence when applied uniformly, specifically the base learning rate for ViT-B/16 [23] and Swin-B (5×10−5) and the Swin-B drop-path rate (0.3) [24], were set according to the original publications. These are architectural necessities, not tuning choices; all other hyperparameters were kept identical.

TB-HCAF (168.5 M parameters) has substantially higher model capacity than ConvNeXt V2-Base (89.2 M parameters) and exhibits the lowest accuracy-per-parameter ratio among all evaluated models in Table 3. However, as shown in Fig. 5, the scaling analysis indicates that smaller TB-HCAF variants consistently outperform larger baselines at matched counts, suggesting that the observed performance gains are attributable to inductive-bias alignment rather than model capacity.

Table 3 Model capacity and computational cost at 224×224 resolution.

Model	Params (M)	GFLOPs	Top-1 Accuracy (%)	Accuracy per Parameter ($\times 10^{-3}$)
MobileNetV2 [5]	3.5	0.30	90.19	25.8
BiT (ResNet-50) [22]	25.5	4.10	92.49	3.63
ViT-B/16 [23]	54.1	12.2	93.59	1.73
Swin-B [24-25]	86.4	18.8	94.39	1.09
EfficientNetV2-M [17]	87.8	16.5	95.80	1.09
ConvNeXt V2-Base [26]	89.2	12.5	97.50	1.09
TB-HCAF (Proposed)	168.5	24.8	98.65	0.59

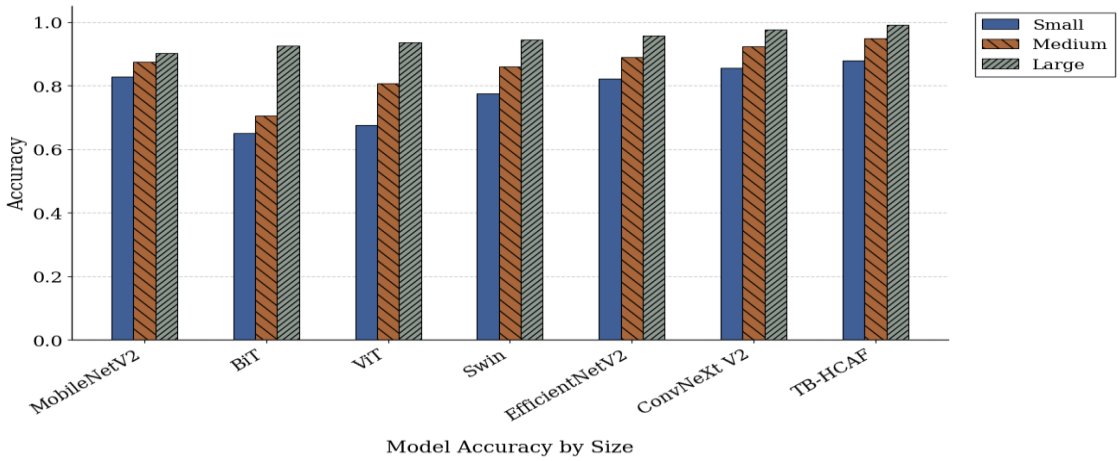


Fig. 5 Accuracy scaling of different architectures across model sizes

Table 4 Performance comparison and statistical significance analysis on PlantCLEF 2015 under the leakage-safe protocol

Model Architecture	Inductive Bias	Type	Pretraining	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Macro-F1	p-value
MobileNetV2 [5]	Local / Efficient	CNN	ImageNet-21K	90.19 ± 0.65	96.72	0.8617	$p < 0.001$
BiT (ResNet-50) [22]	Local / Standard	CNN	ImageNet-21K	92.49 ± 0.58	97.38	0.8949	$p < 0.001$
ViT-B/16 [23]	Global	Transformer	ImageNet-21K	93.59 ± 0.55	98.04	0.9298	$p < 0.001$
Swin-B [24-25]	Hierarchical	Transformer	ImageNet-22K	94.39 ± 0.48	98.26	0.9399	$p < 0.001$
EfficientNetV2-M [17]	Optimized	CNN	ImageNet-21K	95.80 ± 0.38	98.10	0.9260	0.005
ConvNeXt V2-Base [26]	Expanded Receptive Field	CNN	ImageNet-22K	97.50 ± 0.42	99.10	0.9740	0.041
TB-HCAF (Proposed)	Convolution + Attention	Hybrid	ImageNet-21K	98.65 ± 0.30	99.48	0.9842	-

As summarized in Table 4, TB-HCAF achieves a Top-1 accuracy of 98.65% and a Macro-F1 of 0.9842, outperforming ConvNeXt V2-Base by 1.15%. A paired two-tailed t-test on the five-fold-wise accuracy differences yields $p = 0.041$ (95% CI: [0.08%, 2.22%]; Cohen's $d = 3.15$). Because only five folds are available, Cohen's d is sensitive to the small within-pair standard deviation (0.26%) and should not be interpreted as a population-level effect size. A Wilcoxon signed-rank test yields $p = 0.043$, and a percentile bootstrap ($B = 10{,}000$) produces a 95% CI of [0.87%, 1.63%], entirely excluding zero and further supporting the statistical reliability of the observed improvement. Fold-wise results in Table 5 show that all five accuracy differences are positive (range: +0.73% to +1.32%), demonstrating consistent directional improvement. TB-HCAF achieves the lowest fold-wise standard deviation (± 0.30%), indicating consistent performance across cross-validation folds under the leakage-safe protocol.

Table 5 Fold-wise Top-1 accuracy for TB-HCAF and ConvNeXt V2-Base under leakage-safe group-stratified cross-validation

Fold	TB-HCAF (%)	ConvNeXt V2-Base (%)	Difference (%)
Fold 1	98.41	97.12	+1.29
Fold 2	98.72	97.48	+1.24
Fold 3	98.95	97.63	+1.32
Fold 4	98.61	97.44	+1.17
Fold 5	98.56	97.83	+0.73
Mean	98.65	97.50	+1.15
Std Dev	0.30	0.42	0.24

Table 6 Impact of data leakage under random and group-stratified splits

Model Architecture	Random Split (%)	Group Split (%)	Performance Gap (%)
MobileNetV2 [5]	96.50	90.19	−6.31
BiT (ResNet-50) [22]	97.78	92.49	−5.29
ViT-B/16 [23]	98.45	93.59	−4.86
Swin-B [24-25]	98.83	94.39	−4.44
EfficientNetV2-M [17]	98.99	95.80	−3.19
ConvNeXt V2-Base [26]	99.20	97.50	−1.70
TB-HCAF (Proposed)	99.46	98.65	−0.81

The impact of specimen-level data leakage is quantified in Table 6. Conventional models exhibit substantial accuracy drops under group-stratified splits (e.g., −6.31% for MobileNetV2 [5]), whereas TB-HCAF maintains a minimal gap of −0.81%, reflecting genuine cross-specimen generalization. This is consistent with the grouping strategy in Section 2.5.

3.2 Training dynamics and convergence stability

The optimization behavior of TB-HCAF is evaluated through training and validation curves in Figs. 6–8. The model exhibits stable and consistent convergence, reaching near-optimal performance within approximately 20 epochs, a behavior facilitated by pretrained initialization on ImageNet-21K. Compared with ConvNeXt V2, which exhibits validation fluctuations, and pure transformer models such as ViT and Swin, which exhibit greater early-stage instability owing to weaker spatial inductive priors, TB-HCAF maintains close alignment between training and validation curves, indicating improved generalization and reduced overfitting. The convergence stability of TB-HCAF arises from the integration of convolutional inductive biases and transformer-based global attention.

The convolutional stem may provide local feature priors that support stable early optimization, while residual connections and the hierarchical architecture facilitate gradient propagation, mitigating gradient vanishing in deeper layers. The fusion of local and global pathways further balances fine-grained feature extraction with high-level contextual understanding, resulting in smooth convergence and robust performance across diverse botanical imaging conditions.

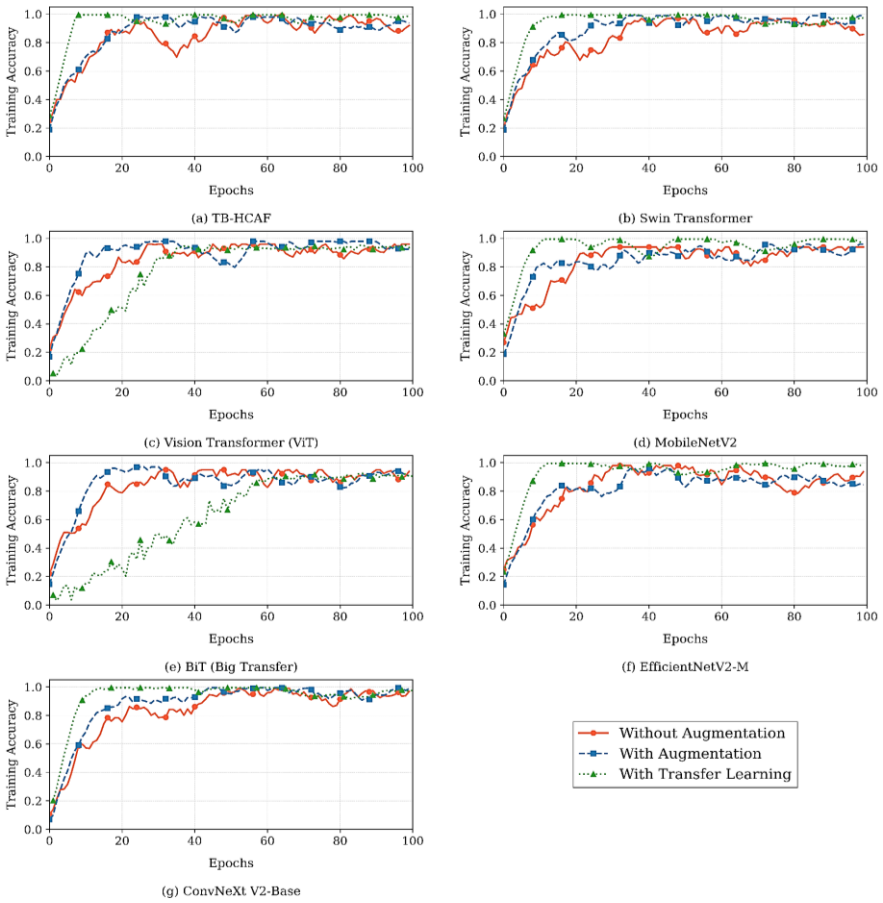
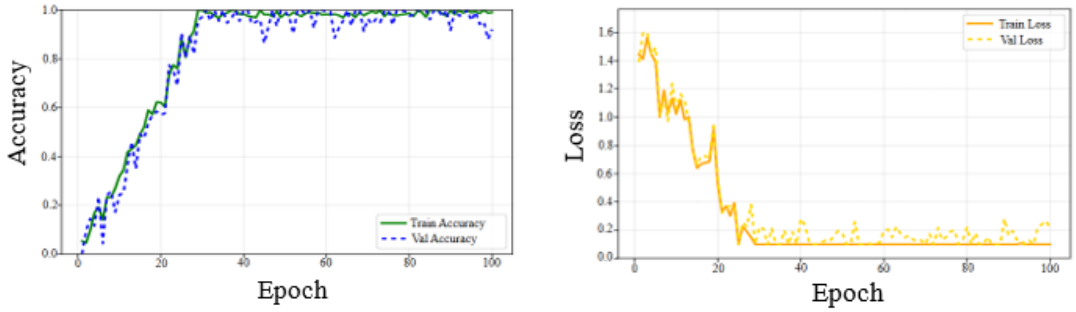
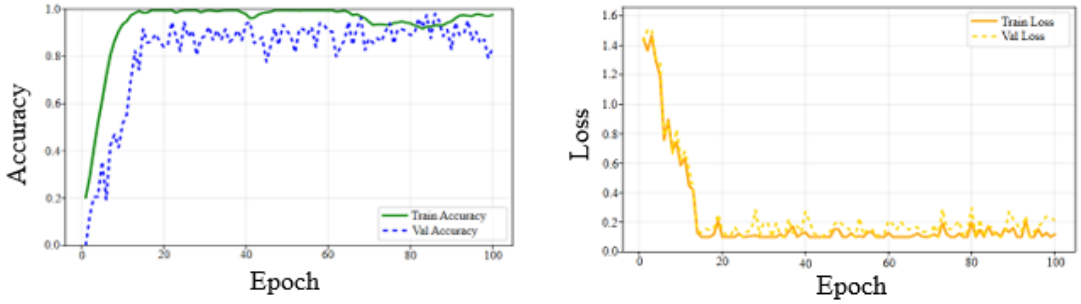


Fig. 6 Comparative training accuracy dynamics under different training regimes



(a) Accuracy curves (b) Loss curves

Fig. 7 Training and validation performance of TB-HCAF



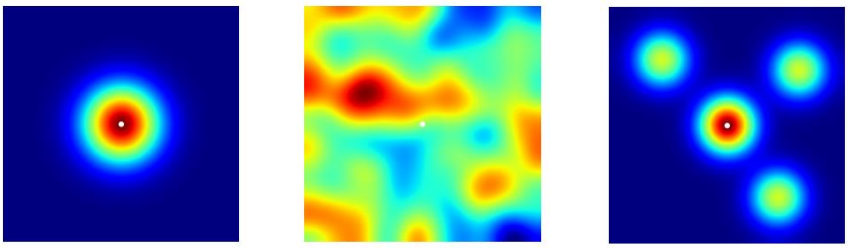
(a) Accuracy curves (b) Loss curves

Fig. 8 Training and validation performance of ConvNeXt V2

3.3 Interpretive analysis of inductive bias and receptive field dynamics

Although Table 4 reports improved accuracy metrics, it does not fully capture how spatial information is utilised. The effective receptive field (ERF) is therefore analyzed to characterize spatial response patterns [6], as shown in Fig. 9. BiT (ResNet-50) exhibits a localized, Gaussian-like ERF, indicating a strong locality bias, while ViT demonstrates a broadly distributed ERF with weaker spatial selectivity. In contrast, TB-HCAF demonstrates an adaptive, intermediate ERF that integrates local and non-local responses, enabling dynamic focus based on image content. This enhances robustness to intra-class variation and improves discrimination under high inter-class similarity. To further interpret model behavior, activation and attention heatmaps from intermediate layers are used instead of gradient-based methods.

These visualizations reveal that convolutional layers capture fine-grained morphological details (e.g., venation patterns and leaf margins), while transformer layers model global structural patterns (e.g., overall plant morphology and inter-organ spatial relationships). The fused representations consistently emphasize biologically meaningful regions, suggesting that TB-HCAF learns interpretable and morphologically significant features for accurate fine-grained classification.



(a) ResNet-50 (b) ViT-B/16 (c) TB-HCAF

Fig. 9 ERF visualization of local (BiT), global (ViT), and hybrid (TB-HCAF) architectures

3.4 Ablation study on training strategy

To quantify the contribution of each training component, an ablation study evaluates the effects of data augmentation and transfer learning on TB-HCAF, as reported in Table 7. Training from scratch yields substantially lower accuracy, indicating

that fine-grained botanical classification benefits from strong prior knowledge of natural image statistics and textures. Data augmentation yields moderate gains in accuracy by improving model invariance to variations in orientation and illumination [27]. The highest performance is achieved by combining ImageNet-21K pretraining with data augmentation, demonstrating that transfer learning and training-set diversity provide complementary benefits under the leakage-safe evaluation protocol. Progressive ablation confirms the contribution of each architectural component: the baseline CoAtNet achieves 97.84% on PlantCLEF 2015, improving to 98.21% with the addition of the bifurcation path, 98.43% with the SAB, and 98.65% with full channel-wise hybrid fusion. Consistent gains are observed on Oxford 102 Flowers, from 98.92% to 99.65%.

Table 7 Ablation study of training strategies on TB-HCAF (CoAtNet)

Training Configuration	Top-1 Accuracy (%)	Top-1 Error (%)	Macro-F1 Score
Training from Scratch	89.19	10.81	0.87
Scratch + Augmentation	93.39	6.61	0.91
Transfer Learning + Augmentation	98.65	1.35	0.98

3.5 Error analysis and confusion characteristics

Fig. 10 presents a condensed confusion matrix for the 1,000-class PlantCLEF 2015 test set, highlighting the first 12 and last 10 species categories. The strong diagonal alignment indicates high per-class precision across diverse taxa, from gymnosperms (Abies) to angiosperms (Viola). Residual errors are structured and confined to morphologically similar genera; for example, minor confusion within Viola (e.g., V. calcarata vs V. odorata) likely reflects visual discrimination of this genus. The model successfully resolves finer distinctions, such as Acer leaf-margin serration patterns, suggesting that SE recalibration enhances texture sensitivity while self-attention captures global structure context. Overall, joint modeling of local and global features contributes to reduced inter-class confusion and improved robustness to intra-class variation.

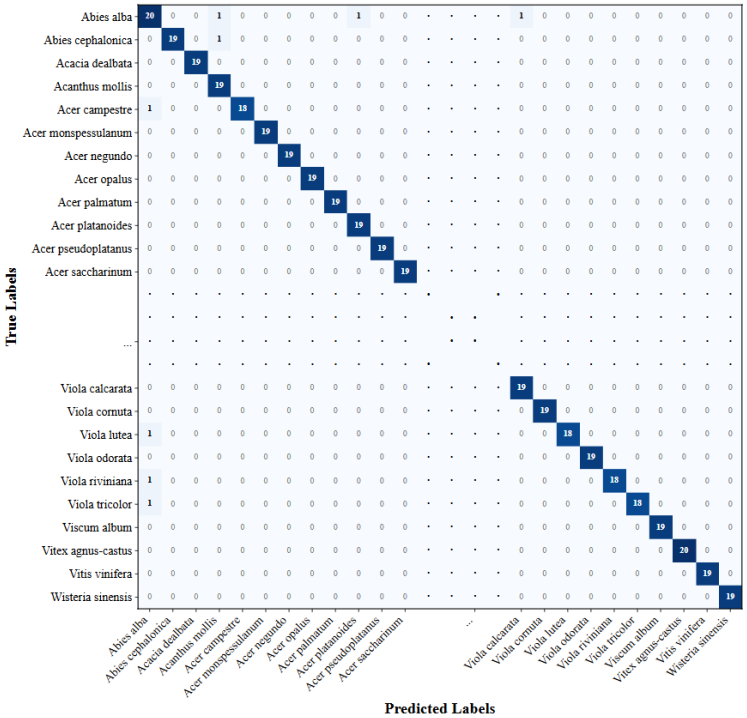


Fig. 10 Condensed confusion matrix for selected PlantCLEF 2015 classes

3.6 Computational efficiency and trade-off analysis

As detailed in Table 8 and illustrated in Fig. 11 and Fig. 12, the proposed TB-HCAF model incurs a substantial computational footprint (168.5M parameters, 24.8 ms latency, 2950 MB VRAM). While its computational complexity is lower than that of fully attention-based models such as ViT because self-attention is applied only in the deeper stages, it substantially exceeds that of lightweight CNNs such as MobileNetV2. In scenarios where strict latency, memory, or energy efficiency is

paramount, lightweight architectures such as MobileNetV2 (3.5M parameters, 4.2 ms latency, 450 MB VRAM) may remain preferable despite lower accuracy. Accordingly, TB-HCAF is better suited to accuracy-critical server-side or cloud-based biodiversity analysis, where a Top-1 accuracy improvement of +8.46% over MobileNetV2, as shown in Table 4, may justify the computational trade-off.

Beyond computational considerations, class-balanced performance was assessed using Macro-F1. Regarding minority-class performance, TB-HCAF achieves a Macro-F1 score of 0.9842, compared with 0.8617 for MobileNetV2, representing a 14.2% relative improvement (0.1225 absolute gain). Since Macro-F1 averages performance equally across all 1,000 plant taxa regardless of sample size, this gain reflects substantially reduced bias toward majority classes and improved recognition of rare and underrepresented species under long-tailed taxonomic distributions. Notably, morphologically challenging species such as Viola lutea, Viola riviniana, and Viola tricolor each achieve a per-class recall of 90% (18/20), demonstrating robust minority-class discrimination even among visually similar taxa.

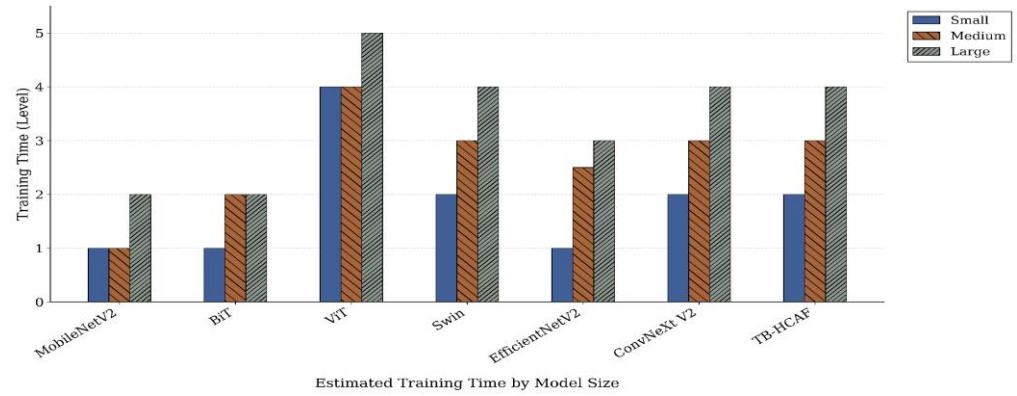


Fig. 11 Training time comparison across architectures and model sizes

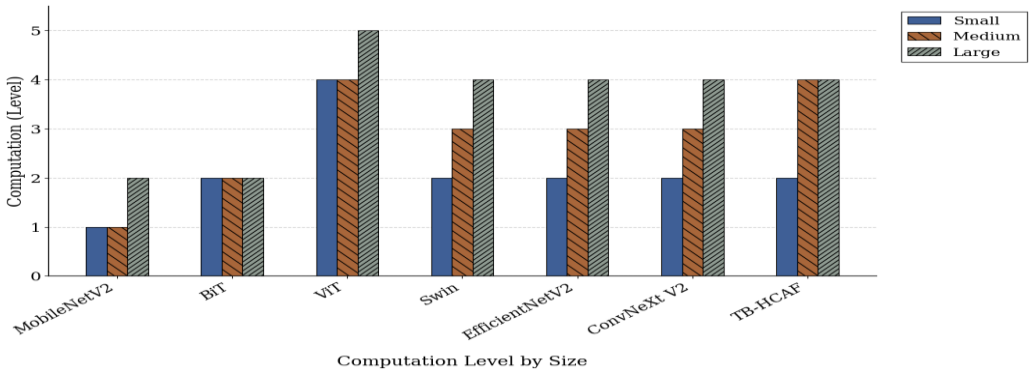


Fig. 12 Computational complexity (flops) comparison across model sizes

Table 8 Hardware and deployment metrics

Models	Params (M)	Latency (ms)	VRAM (MB)	Disk (MB)
MobileNetV2 [5]	3.5	4.2	450	14
BiT [22]	25.5	8.4	980	98
ViT-B/16 [23]	54.1	12.2	1450	210
Swin-B [24-25]	86.4	18.8	2400	330
EfficientNetV2-M [17]	87.8	16.5	2100	340
ConvNeXt V2-Base [26]	89.2	12.5	1300	350
TB-HCAF (Proposed)	168.5	24.8	2950	645

A quantitative failure analysis of the 290 misclassified samples from the 21,447-image test set (PlantCLEF 2015, 1,000 classes, error rate = 1.35%) reveals that the dominant error pattern occurs at the intra-genus level, as directly evidenced by the confusion matrix. These species share overlapping leaf morphology, flower structure, and coloration, making fine-grained

discrimination inherently challenging. The remaining misclassifications are distributed across visually difficult conditions, including background clutter and variable illumination, which are well-known challenges in field-collected plant image datasets. These results show that residual errors arise from extreme morphological similarity among closely related species. Future work should focus on fine-grained discriminative learning for visually ambiguous species pairs and targeted augmentation for rare taxa. Overall, TB-HCAF demonstrates a balanced trade-off between computational demands and predictive performance, supporting its suitability for high-precision biodiversity analysis in high-performance computing environments.

3.7 Cross-dataset generalization

To evaluate the transferability of botanical representations learned by TB-HCAF beyond PlantCLEF 2015, cross-dataset experiments were conducted under strict zero-shot conditions: the model was trained exclusively on PlantCLEF 2015 with no fine-tuning or domain adaptation applied to the target datasets. The Oxford 102 Flowers evaluation was conducted on a taxonomy-constrained 74-class subset (72.5% of 102 categories); 28 classes (27.5%) could not be mapped to any PlantCLEF 2015 species under the APG IV classification system [28] and were therefore excluded. This result is therefore not directly comparable to standard supervised Oxford 102 Flowers benchmarks and should be interpreted strictly within the zero-shot, taxonomy-constrained evaluation setting. All competing models were evaluated on the identical 74-class subset under the same predefined taxonomic mapping protocol to ensure unbiased comparison.

Since the label spaces across datasets are disjoint, direct species-level evaluation is infeasible. A taxonomy-aware mapping based on the APG IV system [28] was therefore employed, aligning each target class to the nearest PlantCLEF 2015 proxy at the species, genus, or family level. Of the 102 Oxford categories, 74 were successfully mapped (31 at the genus level and 43 at the family level), as detailed in Table 9. Within the taxonomy-constrained 74-class Oxford 102 subset, TB-HCAF achieved 99.65% Top-1 accuracy, outperforming BiT (ResNet-50) by 2.55% (97.10%) and ViT-B/16 by 1.45% 98.20%0, as reported in Table 10.

Table 9 Zero-shot taxonomic mapping between Oxford 102 and PlantCLEF 2015

Oxford 102 Species	APG IV Family	PlantCLEF Proxy	Mapping Level
Passiflora incarnata	Passifloraceae	Passiflora caerulea	Genus
Digitalis purpurea	Plantaginaceae	Digitalis lutea	Genus
Erysimum cheiri	Brassicaceae	Brassica napus	Family
Helianthus annuus	Asteraceae	Arnica montana	Family
Viola tricolor	Violaceae	Viola riviniana	Genus
Malva sylvestris	Malvaceae	Hibiscus syriacus	Family

Table 10 Cross-dataset generalization performance on Oxford 102 flowers and iNaturalist 2018

Model	Oxford 102 Flowers (%)	iNaturalist 2018 (%)
BiT (ResNet-50)	97.10	66.40
ViT-B/16	98.20	79.50
TB-HCAF (Proposed)	99.65	84.12

This result indicates effective transfer of fine-grained morphological representations across botanical domains; however, the near-perfect accuracy is partly attributable to the controlled visual characteristics and reduced background complexity of the Oxford 102 dataset within the constrained subset, and it should not be extrapolated to the full 102-class benchmark. On iNaturalist 2018, TB-HCAF attained 84.12% Top-1 accuracy without any target-domain training, exceeding BiT by 17.72% and ViT-B/16 by 4.62%. Given the dataset's large scale, long-tailed species distribution, and uncontrolled natural image conditions, this result provides further evidence of cross-domain generalization beyond the taxonomy-mapping constraints or dataset-curation effects.

3.8 Limitations and failure cases of the Proposed Model

The proposed model has several limitations. Classification performance on minority classes remains lower than that on well-represented classes because of the limited number of training samples. The model also struggles to distinguish visually similar species, particularly at the intra-genus level. In addition, performance degrades under challenging imaging conditions such as occlusion, motion, blur, background clutter, and low illumination. Finally, accuracy decreases on highly diverse datasets such as iNaturalist 2018 due to increased visual variability and uncontrolled image-capture conditions [2].

4. Conclusion

This study presented TB-HCAF, a targeted domain-specific extension of the CoAtNet baseline for fine-grained plant species classification, designed to jointly capture complementary local morphological and global structural representations. Overall, the results demonstrated that TB-HCAF achieved competitive performance compared to the baseline and existing methods. The key findings are summarized as follows:

(1) The proposed architecture introduces three targeted modifications to CoAtNet's baseline that integrate convolutional and transformer components to better preserve fine-grained morphological features for botanical species classification.
(2) A group-stratified k-fold cross-validation strategy ensured reliable evaluation by eliminating specimen-level data leakage.
(3) The model achieved 98.65% Top-1 accuracy on PlantCLEF 2015 and 84.12% on iNaturalist 2018 under strict zero-shot transfer conditions. On Oxford 102 Flowers, it achieved 99.65% on a taxonomy-constrained 74-class subset (not comparable to standard supervised benchmarks).
(4) ERF analysis demonstrated improved discrimination of morphologically similar species by directing attention toward biologically meaningful regions.

Future work will explore open-set recognition for detecting previously unseen species, few-shot learning for rare and unrepresented taxa, and uncertainty-aware prediction to improve robustness under challenging imaging conditions. These directions will further enhance the model’s applicability in real-world botanical identification tasks.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, et al., “LifeCLEF 2015: Multimedia Life Species Identification Challenges,” Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. 9283, pp. 462-483, 2015.
[2] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, et al., “The iNaturalist Species Classification and Detection Dataset,” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8769-8778, 2018.
[3] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN Models for Fine-Grained Visual Recognition,” Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 1449-1457, 2015.
[4] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition,” Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 5209-5217, 2017.
[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint arXiv:1704.04861, 2017.
[6] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the Effective Receptive Field in Deep Convolutional Neural Networks,” Proceedings of Advances in Neural Information Processing Systems, vol. 29, pp. 4898-4906, 2016.
[7] Y. S. Tong, T. H. Lee, and K. S. Yen, “Recognition of Ginger Seed Growth Stages Using a Two-Stage Deep Learning Approach,” Proceedings of Engineering and Technology Innovation, vol. 26, pp. 1-17, 2024.
[8] V. Choudhary and A. Thakur, “Multiclass Plant Leaf Disease Prediction Using Fuzzy Multimodal Feature Extraction,” Advances in Technology Innovation, vol. 10, no. 4, pp. 370-382, 2025.

[9] F. O. Isinkaye, M. O. Olusanya, and A. A. Akinyelu, "A Multi-Class Hybrid Variational Autoencoder and Vision Transformer Model for Enhanced Plant Disease Identification," Intelligent Systems with Applications, vol. 26, article no. 200490, 2025.

[10] M. A. K. Sergio, A. B. S. Siallagan, and E. Joelianto, "Potato Leaf Disease Detection by Means of Vision Transformer and CoAtNet Architecture," Proceedings of IEEE International Conference on Technology, Informatics, Management, Engineering and Environment (TIME-E), pp. 154-159, 2024.

[11] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and Construction Learning for Fine-Grained Image Recognition," Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5157-5166, 2019.

[12] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to Navigate for Fine-Grained Classification," Proceedings of European Conference on Computer Vision (ECCV), pp. 420-435, 2018.

[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9268-9277, 2019.

[14] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," Proceedings of the Sixth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), pp. 722-729, 2008.

[15] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," Advances in Neural Information Processing Systems, vol. 34, pp. 3965-3977, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.

[17] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 10096-10106, 2021.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.

[19] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Proceedings of International Conference on Learning Representations (ICLR), 2019.

[20] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation," Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 774-782, 2021.

[21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, et al., "A Comprehensive Survey on Transfer Learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, 2021.

[22] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, et al., "Big Transfer (BiT): General Visual Representation Learning," Proceedings of European Conference on Computer Vision (ECCV), pp. 491-507, 2020.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," Proceedings of International Conference on Learning Representations (ICLR), 2021.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022, 2021.

[25] H. Zhang and G. Ren, "Intelligent Leaf Disease Diagnosis: Image Algorithms Using Swin Transformer and Federated Learning," The Visual Computer, vol. 41, pp. 4815-4838, 2025.

[26] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, et al., "ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16133-16142, 2023.

[27] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical Automated Data Augmentation with a Reduced Search Space," Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3008-3017, 2020.

[28] The Angiosperm Phylogeny Group, "An Update of the Angiosperm Phylogeny Group Classification for the Orders and Families of Flowering Plants: APG IV," Botanical Journal of the Linnean Society, vol. 181, no. 1, pp. 1-20, 2016.



Copyright© by the authors. Licensee TAETI, Taiwan. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/).