Transformer Based Hybrid Convolution-Attention Framework for Plant Species Classification

Author Name(s)

***Abstract*—Fine-grained plant species identification involves the concurrent recognition of high-frequency local textures, such as leaf venation and margin serration, and long-range global structures, including branching patterns and phyllotaxis. Prior approaches have predominantly relied on either Convolutional Neural Networks (CNNs), which emphasize localized feature extraction, or Vision Transformers (ViTs), which model global context but often require large-scale data to learn fine-grained textures effectively. In this work, we investigate the effectiveness of a Hybrid Inductive Bias that integrates convolutional and attention-based mechanisms for plant species classification. We conduct a comprehensive evaluation of CoAtNet, a hybrid architecture that stacks convolutional stages with transformer blocks to combine translation-equivariant feature extraction with input-adaptive global attention. To ensure a leakage-safe assessment, we employ a strict Group K-Fold cross-validation protocol stratified by specimen identity, explicitly preventing shared biological individuals across training and evaluation splits. Under this evaluation regime, the hybrid model achieves a Top-1 accuracy of 98.65% and a Top-5 accuracy of 99.48% on the PlantCLEF 2015 benchmark. Generalization performance is further assessed on the Oxford 102 Flowers dataset, where the model attains 99.65% accuracy, and on the highly imbalanced iNaturalist 2018 dataset, achieving 84.12% accuracy, consistently outperforming representative CNN and Transformer baselines. Qualitative analysis using Effective Receptive Field visualization illustrates that the hybrid architecture captures both localized texture information and distributed global contextual dependencies, providing insights into its improved discrimination of morphologically similar species.**

***Index Terms*—Hybrid Inductive Bias, Fine-Grained Visual Categorization, Plant Species Identification, Leakage-Safe Evaluation, CoAtNet, Effective Receptive Field.**

I. Introduction

The accelerating crisis of global biodiversity loss, driven by climate change, habitat fragmentation, and the spread of invasive species, has created an urgent demand for scalable and automated ecological monitoring [15]. As primary producers in terrestrial ecosystems, plants form the structural and functional foundation of biological networks, and their accurate identification is a prerequisite for downstream conservation actions like invasive species management and carbon stock assessment. However, the immense scale of botanical diversity, with nearly 400,000 described plant species, combined with the taxonomic impediment caused by the global decline in expert taxonomists, has produced a severe bottleneck in large-scale flora cataloguing [25]. Although computer vision offers a promising pathway to broaden access to botanical expertise, transferring success from controlled herbarium imagery to "in the wild" plant identification remains deeply challenging.

Plant species recognition is a prototypical Fine-Grained Visual Categorization (FGVC) problem [26]. Unlike generic object recognition, which separates semantically distinct categories like vehicles or animals, botanical classification requires distinguishing among visually similar subclasses that share a common structural blueprint but differ in subtle and localized traits. This difficulty is amplified by high intra-class variability, where individuals of the same species exhibit pronounced visual differences due to phenotypic plasticity and developmental stage, alongside low inter-class variability, where closely related taxa may be separable only by minute distinctions in vein topology, trichome density, or leaf margin morphology.

The methodological landscape for addressing these challenges has evolved substantially over the past two decades. Early computational botany adopted a biomimetic philosophy, attempting to operationalize the manual identification keys used by taxonomists through handcrafted morphometric descriptors [19]. These pipelines focused on three primary modalities: geometric shape descriptors to quantify leaf margins, texture descriptors to capture surface details like venation and epidermal patterns, and graph-based venation analysis to model vein topology [23], [24]. While computationally efficient, these approaches exhibited limited robustness under unconstrained conditions [20]. They typically relied on near-perfect segmentation and assumed leaves were photographed against uniform backgrounds, flattened to remove perspective distortion, and free from occlusion. In realistic ecological monitoring scenarios characterized by cluttered scenes, overlapping foliage, and variable illumination, such handcrafted features demonstrated poor generalization, limiting their applicability to large-scale automation.

The advent of deep learning marked a paradigm shift by enabling end-to-end learning of feature representations directly from raw pixels, largely eliminating the need for manual segmentation [22]. Convolutional Neural Networks (CNNs) such as ResNet [6], MobileNet [7], and DenseNet [10] became the dominant paradigm, leveraging the convolution operation's inductive bias for locality and translation equivariance to detect leaf edges and textures across varying spatial locations. In resource-constrained scenarios, optimization techniques have further refined these architectures for efficiency [3]. To better address the fine-grained nature of botanical classification, specialized part-based models were introduced, including attention-driven architectures that explicitly identify and reprocess discriminative regions such as flowers, leaf tips, or venation junctions at higher resolution [12], [13], [18]. However, these multi-stage pipelines often incur high computational overhead and require complex training procedures. Furthermore, conventional CNNs exhibit an architectural limitation related to the Effective Receptive Field (ERF) [29]. Although the theoretical receptive field grows with network depth, empirical analyses indicate that the region of dominant influence expands more slowly. As a result, modeling long-range dependencies—such as relating

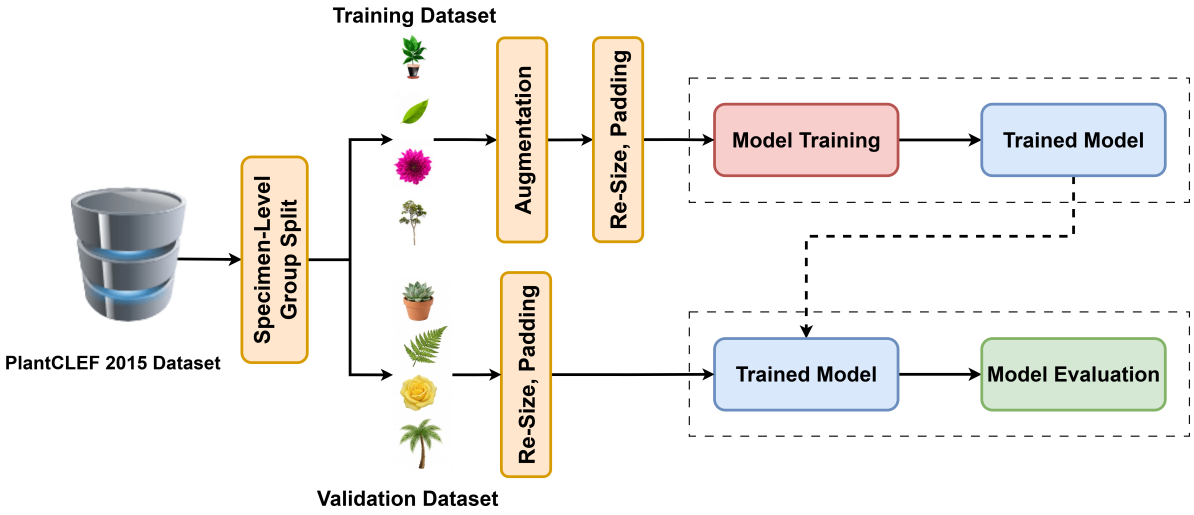

Fig. 1. Overview of Leakage safe experimental pipeline for hybrid plant species classification.

inflorescence structure at the apex of a stem to phyllotactic arrangement near the base—remains challenging, leading to confusion between morphologically similar species that share local textures but differ in overall plant structure.

To mitigate the limitations of predominantly local processing, recent work has adapted the Transformer architecture from natural language processing [2] to visual recognition [4]. Vision Transformers (ViTs) partition an image into patches and process them using self-attention, enabling interactions between spatially distant regions from early layers [5]. This global interaction facilitates the modeling of holistic plant structure and spatial relationships compared to purely convolutional designs [27]. However, the reduced reliance on locality-specific inductive biases renders pure Transformer models highly data-intensive [28]. In biodiversity informatics, where datasets exhibit long-tailed distributions and many rare species are represented by limited samples, such models may struggle to consistently learn low-level morphological cues such as venation or serration patterns and may become sensitive to background correlations.

Recent FGVC research has proposed explicitly supervised attention and part-based learning strategies to overcome the limitations of purely convolutional processing. Multi-Attention CNNs (MA-CNN) [30] learn multiple attention maps corresponding to discriminative object parts, while Destruction and Construction Learning (DCL) [13] enforces robustness by deliberately disrupting salient regions during training. Similarly, NTS-Net [18] employs a detection-driven pipeline that iteratively localizes and refines informative regions. While these approaches have demonstrated strong performance on controlled FGVC benchmarks, their applicability to large-scale botanical datasets is constrained by several factors. First, detection-driven or part-supervised pipelines often assume relatively consistent object structure, which is violated in plant imagery where organs vary widely in presence, scale, and visibility. Second, such methods typically rely on additional supervision signals or carefully tuned multi-stage training procedures, increasing annotation cost and sensitivity to dataset bias. Finally, under specimen-correlated datasets such as PlantCLEF or iNaturalist, these models are particularly vulnerable to specimen-level data leakage, as localized cues may inadvertently encode observation-specific artifacts rather than species-discriminative morphology.

In parallel, large-scale biodiversity benchmarks have highlighted the importance of addressing extreme class imbalance and domain shift. On iNaturalist-style datasets, strategies such as class-balanced loss functions [31] and domain-similarity transfer have been shown to significantly affect performance, particularly for rare species. However, these methods primarily operate at the optimization or loss-design level and do not directly address architectural limitations related to receptive field structure or the integration of local and global cues. As a result, models optimized for imbalance may still struggle to reconcile fine-scale texture sensitivity with holistic structural reasoning, especially when training data per species is scarce.

This architectural trade-off has motivated the development of *Hybrid Inductive Bias* models that integrate convolutional and transformer-based components within a single network [1]. Architectures of this class typically employ convolutional layers in early stages to extract fine-scale textural and edge information, followed by transformer blocks that integrate these features into a global structural representation, an approach analogous to hybrid feature extraction strategies explored in other domains [8]. This design is conceptually aligned with botanical identification practices, where localized morphological cues are interpreted in relation to whole-plant structure. Despite this motivation, existing studies have predominantly evaluated hybrid models on generic benchmarks such as ImageNet or CIFAR, leaving their effectiveness on biological taxonomy tasks comparatively underexplored. More critically, many prior works rely on random train-test splits that permit multiple images of the same individual plant to appear across splits. This evaluation strategy introduces specimen-level data leakage, allowing models to exploit observation-specific artifacts—such as background or imaging conditions—rather than learning species-discriminative traits, thereby inflating reported performance and obscuring true generalization.

In this work, we address both architectural and methodological limitations identified in prior studies. We evaluate a

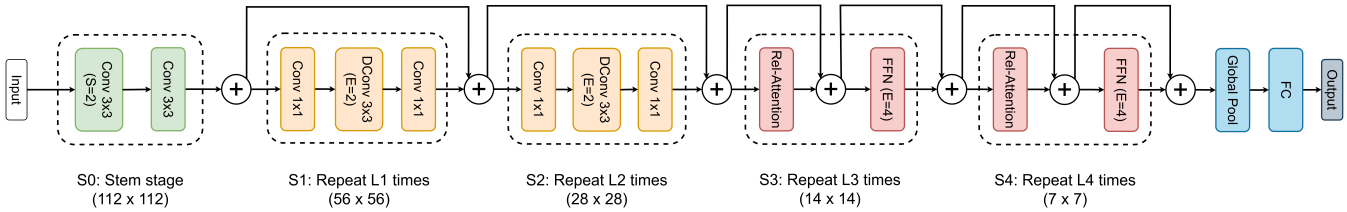

Fig. 2. Stage-wise transition from convolutional to attention-based processing in CoAtNet.

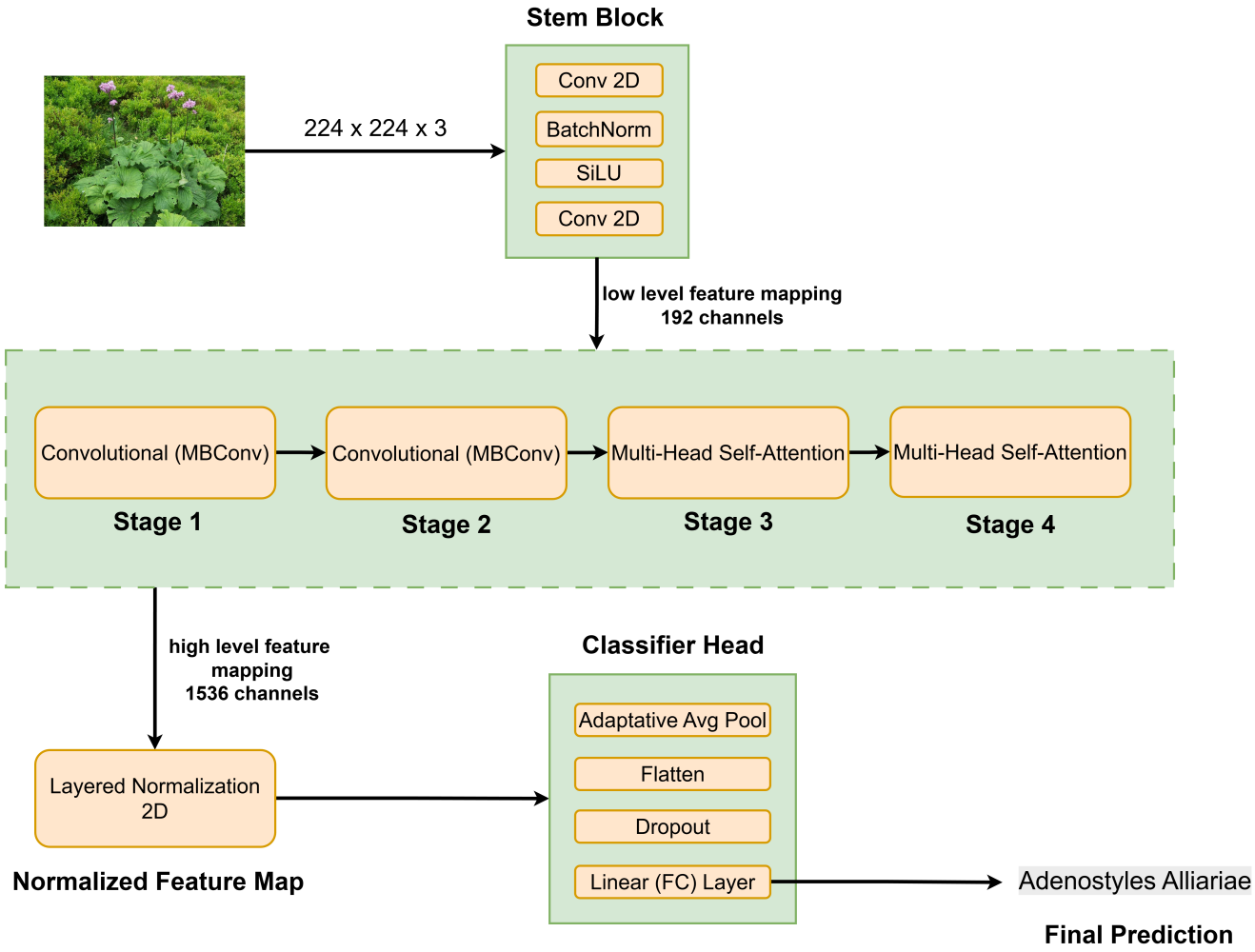

Fig. 3. End-to-end four-stage pyramidal hybrid convolution–attention network architecture.

hybrid convolution–transformer architecture for plant species recognition under a strict leakage-safe Group Stratified evaluation protocol that prevents any individual plant from appearing in both training and testing sets. Under this evaluation regime, the model achieves a Top-1 accuracy of 98.65% on the PlantCLEF 2015 benchmark [14]. Generalization performance is further assessed through cross-dataset evaluation on Oxford 102 Flowers [17], where an accuracy of 99.65% is obtained, and on the highly imbalanced iNaturalist 2018 dataset [16], where the model attains 84.12% accuracy. These results demonstrate that the hybrid formulation improves discrimination among morphologically similar species and supports reliable generalization in large-scale biodiversity monitoring scenarios.

II. Methodology

Fine-grained plant species recognition demands a learning framework that is both morphologically sensitive and experimentally reliable. Visual cues distinguishing closely related taxa are often subtle, distributed across multiple organs, and captured under diverse environmental conditions [14]. At the same time, botanical datasets are prone to systematic biases arising from repeated observations of the same specimen [16]. The methodology proposed in this work is designed to jointly address these challenges by aligning architectural inductive bias, data partitioning, and optimization strategy within a single coherent framework.

An overview of the complete experimental pipeline is presented in Figure 1, which situates the proposed Hybrid Inductive Bias model within a leakage-safe data preparation protocol, followed by structured training and evaluation. The

remainder of this section progressively drills down from this high-level perspective into the theoretical motivation, architectural design, training configuration, and experimental setup that together constitute the proposed approach.

A. Conceptual Overview of the Proposed Framework

As illustrated in Figure 1, the methodology follows a sequential logic that mirrors the lifecycle of real-world ecological inference. Raw plant images are first grouped at the specimen level to prevent information leakage across splits. These curated samples are then subjected to augmentation and preprocessing before being fed into a hybrid convolution–attention backbone. The network transforms low-level visual patterns into high-level semantic representations, which are finally evaluated using metrics explicitly chosen to account for class imbalance and ecological fairness.

This top-down organization is intentional. Rather than treating architecture, training, and evaluation as independent modules, the framework ensures that each design choice reinforces the core objective: learning species-discriminative representations that generalize beyond previously observed individuals.

B. Hybrid Inductive Bias: Motivation and Theoretical Basis

Learning in deep neural networks can be framed as the search for a function $f : \mathcal{X} \to \mathcal{Y}$ within a constrained hypothesis space $\mathcal{H}$. These constraints, imposed implicitly by the network architecture, influence how the model extrapolates when data alone is insufficient. In botanical image analysis, such extrapolation is common due to limited samples for many species and substantial visual variability within classes.

Plant imagery exhibits a dual structural property. On one hand, local morphological patterns such as venation, epidermal texture, and margin serration are spatially stationary and repeat across an image. On the other hand, species discrimination often depends on interpreting these local cues within a broader anatomical context, such as the co-occurrence of leaf shape and floral structure. This observation motivates the investigation of a hybrid inductive bias in which locality and global dependency are treated as complementary components of representation learning, an approach that has demonstrated effectiveness in other domains requiring robust feature integration [8].

C. From Convolution to Attention: Architectural Realization with CoAtNet

To operationalize the hybrid inductive bias, this work adopts the CoAtNet architecture [1], which explicitly transitions from convolution-dominated processing to attention-based reasoning as depth increases. A structural overview of this progression is shown in Figure 2, while the full end-to-end architecture is detailed in Figure 3.Early layers employ convolutional operations that encode translation equivariance and local continuity. For an input feature tensor $X \in \mathbb{R}^{H\times W\times C}$, convolution computes activations as:

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

where Δ denotes the local spatial neighborhood defined by the kernel size. This operation prioritizes spatially adjacent information, aligning naturally with biological tissue structure to efficiently extract fine textures critical for early-stage discrimination [19]. However, convolution alone limits long-range interaction, making it insufficient for resolving inter-organ dependencies.At deeper stages, the architecture transitions to transformer blocks with relative self-attention [2], where interactions between spatially distant regions are explicitly modeled [4]. The pre-softmax attention scores are defined as:

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

combining content similarity with learnable relative positional bias. This formulation preserves spatial awareness while enabling global semantic integration. As depicted in Figure 2, this progressive shift allows the network to behave convolutionally in early layers and contextually in later layers, effectively mirroring the hierarchical structure of plant morphology.

D. Hierarchical Backbone and Classification Head

The full network follows a four-stage pyramidal design Figure 3, beginning with a convolutional stem that downsamples the input from 224×224 to 56×56. This early compression is both computationally necessary due to the quadratic complexity of attention and methodologically sound, as it removes redundant pixel-level information while preserving salient edges and textures, conceptually similar to wavelet-based compression strategies [21].

Subsequent stages employ MBConv blocks [7], [11] augmented with squeeze-and-excitation mechanisms, allowing adaptive channel recalibration based on global statistics. These stages construct a rich dictionary of local morphological cues. The final transformer stages, operating on reduced spatial grids, integrate these cues into holistic species-level representations.

The classification head aggregates the final feature map using global average pooling, followed by normalization and dropout. This design preserves translation invariance and limits parameter growth, improving robustness under varying acquisition conditions.

E. Leakage-Safe Data Preparation and Validation Protocol

Fine-grained botanical datasets frequently include multiple images originating from the same biological specimen, captured across different organs, viewpoints, and acquisition conditions. When such correlated samples are randomly distributed between training and validation sets, information leakage can occur, resulting in inflated performance estimates. The empirical impact of this effect is quantified in Table IV, which compares model performance under standard random splits and specimen-level group-stratified splits. Across all evaluated architectures, enforcing specimen-level grouping leads to a consistent reduction in validation accuracy, with performance drops ranging from 6.31% for MobileNetV2 to 0.81% for the proposed CoAtNet model. These results indicate that random splits systematically overestimate generalization performance, particularly for higher-capacity models, and confirm the presence of non-negligible evaluation bias due to correlated observations.

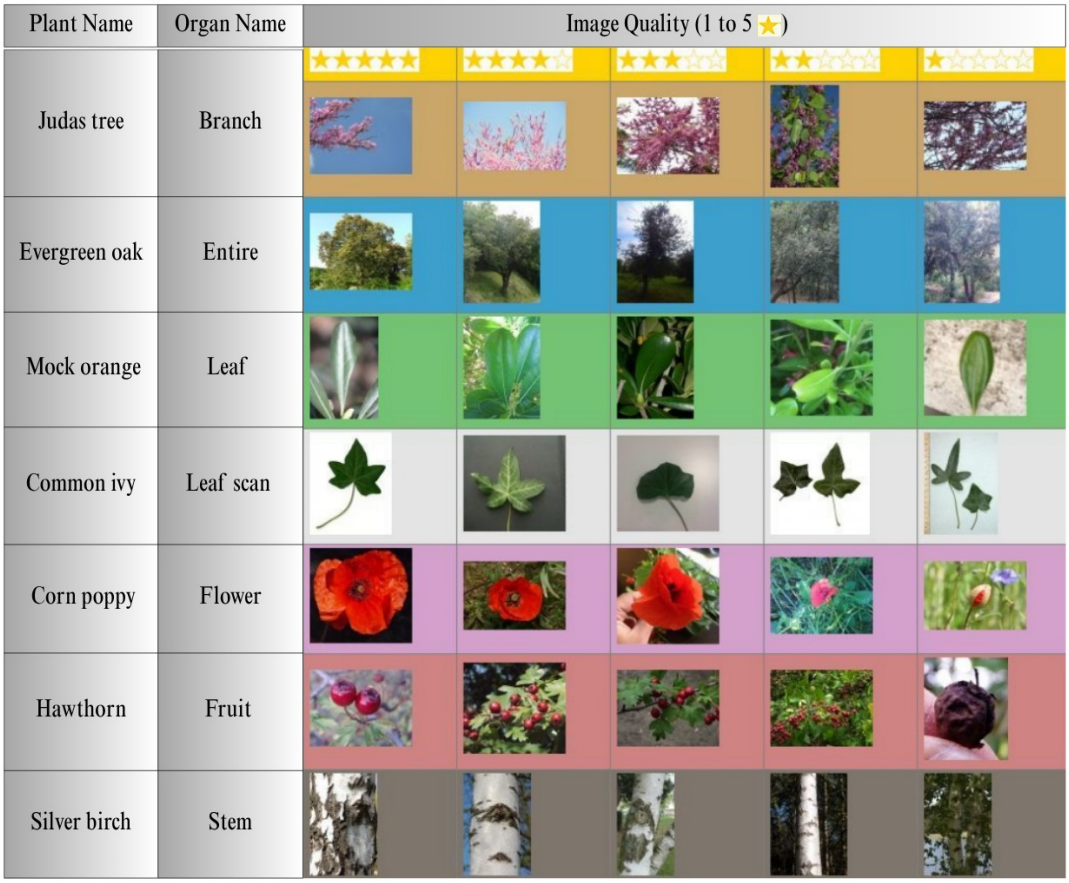

Fig. 4. Visual diversity of PlantCLEF 2015 images across organs and quality levels.

TABLE I
TRAINING HYPERPARAMETERS AND OPTIMIZATION SETTINGS.

Parameter	**Value**	**Description**
Optimizer	AdamW [32]	Adaptive moment estimation with decoupled weight decay
Learning Rate	1×10^{-3}	Peak learning rate under One-Cycle policy
Weight Decay	0.05	Regularization strength
Batch Size	32	Optimized for 16 GB GPU memory [3]
Epochs	100	Early stopping with patience of 5 epochs
Loss Function	Label Smoothing CE [33]	Smoothing factor $\epsilon = 0.1$
Precision	AMP (FP16)	Automatic Mixed Precision training

To mitigate this issue, a grouped stratified k-fold validation protocol is adopted, ensuring that all images derived from a single biological specimen are assigned to the same fold. For the PlantCLEF 2015 dataset, specimen identity is determined using the ObservationID field provided in the official XML metadata, which explicitly links multiple images captured from the same plant individual. In cases where this identifier is unavailable, specimen groupings are reconstructed using a deterministic composite key defined by the tuple (Author, Date, GPS_Location), enabling consistent and reproducible fold assignments. All fold assignments are fixed and reused across experiments to ensure identical data partitions for fair model comparison.

The selection of the number of folds k is critical for balancing estimator bias, variance, and computational cost. A sensitivity analysis was therefore conducted to evaluate mean validation accuracy, estimator stability measured as the standard deviation across folds, and total training time. The quantitative results are summarized in Table II and illustrated in Fig. 5. As shown in Fig. 5, increasing k leads to noticeable improvements in validation accuracy at low fold counts; however, the marginal gains diminish beyond k = 5. In particular, increasing k from 5 to 10 yields only a 0.06% improvement in accuracy, while nearly doubling the training time from 18.2 to 36.8 hours.

Estimator stability exhibits a similar convergence trend. As illustrated in Fig. 5, low fold counts such as k = 3 result in high variance across folds (standard deviation of 1.2%), indicating strong sensitivity to the specific specimen partitioning. At k = 5, the standard deviation decreases below 0.3%, reflecting convergence to a stable and reliable performance estimate, with only marginal variance reduction observed for larger values of k. Based on this empirical analysis, k = 5 is selected for all subsequent experiments, as it provides a statistically robust and leakage-safe evaluation protocol while maintaining a practical balance between estimator reliability and computational cost.

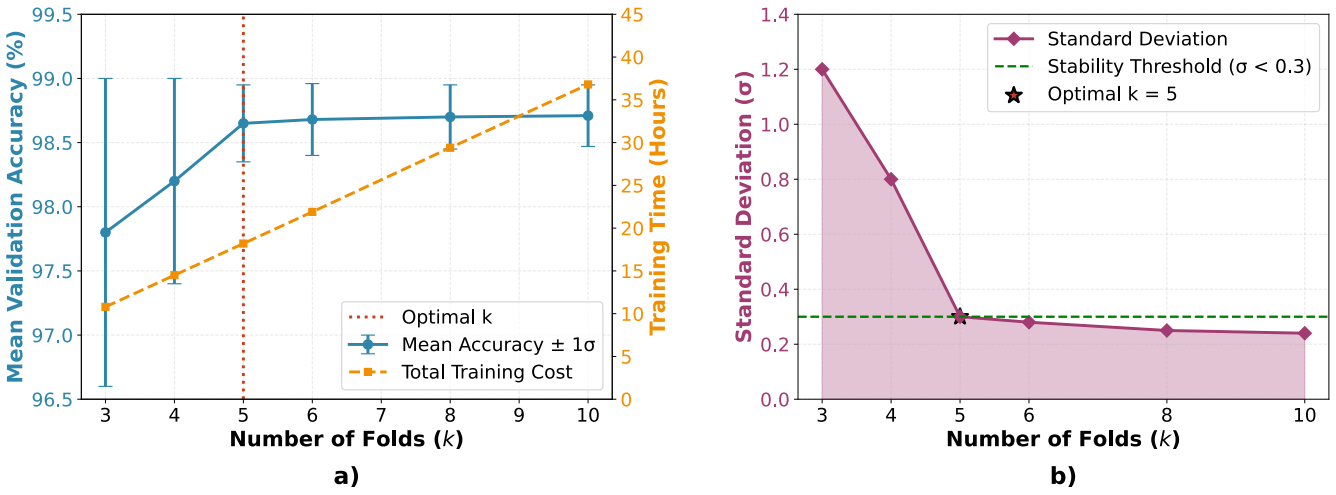

Fig. 5. Grouped stratified k-fold sensitivity analysis:(a) accuracy–cost trade-off and (b) stability convergence.

TABLE II
SENSITIVITY ANALYSIS OF GROUP K-FOLD PARAMETERS ON PLANTCLEF 2015.

Metric	k=3	k=4	k=5	k=6	k=8	k=10
Mean Val. 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 Cost (Hours)	10.8	14.5	**18.2**	21.9	29.4	36.8
Marginal Acc. Gain (%)	—	+0.40	**+0.45**	+0.03	+0.02	+0.01

F. Training Configuration and Hyperparameter Selection

Training is conducted under conditions of severe class imbalance, where standard optimization procedures tend to favor dominant species. Label smoothing is therefore employed to soften target distributions [33]:

$$y_k^{LS} = (1-\epsilon)y_k + \frac{\epsilon}{C}, \tag{3}$$

with $\epsilon = 0.1$, which reduces prediction overconfidence and promotes smoother class boundaries.

Optimization is performed using AdamW [32], whose decoupled weight decay has been shown to be effective for transformer-based architectures. A one-cycle learning rate schedule is adopted to enable rapid exploration during early training followed by stable convergence. A batch size of 32 is selected to balance optimization stability with the memory constraints of a 16 GB GPU, consistent with configurations reported in recent studies on efficient deep visual model training [3]. Data augmentation is applied using TrivialAugmentWide to introduce controlled variability without manual policy tuning [34]. Transfer learning is employed [35], initializing the model with ImageNet-pretrained weights.

Hyperparameter values were initially guided by prior literature and commonly adopted configurations for convolution–transformer architectures. To verify their suitability for the target dataset, we conducted limited preliminary experiments over reasonable ranges of key hyperparameters, including learning rate, batch size, and label smoothing factor. Performance was observed to be stable across these ranges, and the final values reported in Table I correspond to settings that consistently yielded strong validation performance. No extensive dataset-specific tuning was required beyond this exploratory evaluation.

G. Experimental Setup and Dataset

All experiments were conducted on a dedicated workstation equipped with an NVIDIA Tesla T4 GPU (2,560 CUDA cores, 320 Tensor Cores, 16 GB GDDR6 memory, and 320 GB/s memory bandwidth) and an Intel® Xeon® processor with 51 GB of system RAM. The system ran Ubuntu Linux with CUDA 11.x support. Model development and training were implemented in PyTorch using the timm library for backbone instantiation, specifically utilizing the transformer-based hybrid convolution–attention model CoAtNet-3 (rw-224, ImageNet-21K pre-trained) [1] to ensure architectural consistency with the design described earlier. Albumentations was employed for data augmentation, while standard Python libraries including NumPy, scikit-learn, Matplotlib, and Seaborn were used for evaluation, visualization, and analysis. Automatic mixed-precision training was enabled to reduce memory usage and improve computational throughput without compromising numerical stability.

To rigorously evaluate the proposed Hybrid Inductive Bias, we utilized a multi-faceted validation strategy centered on the PlantCLEF 2015 dataset as the primary training benchmark [14]. This dataset, comprising 113,205 images representing 1,000 species of trees, herbs, and ferns from Western Europe, is uniquely suited for the leakage-safe evaluation protocol described previously due to its structured organization into 41,794 unique observation events. Unlike generic image collections, PlantCLEF provides rich metadata linking multiple views of the same individual plant including distinct organs such as Branch, Entire, Flower, Fruit, Leaf, LeafScan, and Stem necessitating a model capable of integrating disparate local textures into a global species concept.

To confirm that the learned representations are universal and robust to domain shifts, we extended validation to two complementary cross-dataset benchmarks. We utilized the

TABLE III
TRANSFER LEARNING PERFORMANCE ON PLANTCLEF 2015 UNDER THE LEAKAGE-SAFE PROTOCOL, TRAINED MODEL PCROSS-DATASET GENERALIZATION PERFORMANCE ON OXFORD 102 FLOWERS AND INATURALIST 2018.

Model Architecture	Inductive Bias Type	Pretraining	Top-1 Acc (%)	Top-5 Acc (%)	Macro-F1	Oxford 102	iNat 2018
MobileNetV2 [7]	CNN (Efficient)	ImageNet-1K	90.19	96.72	0.8617	94.50	59.80
BiT (ResNet-50) [9]	CNN (Standard)	ImageNet-21K	92.49	97.38	0.8949	97.10	66.40
ViT-B/16 [4]	Transformer (Global)	ImageNet-21K	93.59	98.04	0.9298	98.20	79.50
Swin-B [5]	Transformer (Hierarchical)	ImageNet-22K	94.39	98.26	0.9399	98.90	81.10
EfficientNetV2-M [11]	CNN (Optimized)	ImageNet-21K	95.80	98.10	0.9260	99.10	80.50
ConvNeXt V2-Base	CNN (Expanded Receptive Field)	ImageNet-22K	97.50	99.10	0.9740	99.40	83.20
CoAtNet (This Work)	**Hybrid**	**ImageNet-21K**	**98.65**	**99.48**	**0.9842**	**99.65**	**84.12**

Oxford 102 Flowers dataset [17], consisting of 8,189 images across 102 flower categories, as a specific stress test for the architecture's convolutional texture extraction capabilities, as it focuses exclusively on floral structures with high intra-class variability in scale and pose. Furthermore, to evaluate the model's capacity for global structural reasoning under conditions of extreme data imbalance, we employed the iNaturalist 2018 dataset [16]. With 437,513 training images covering 8,142 species, this massive-scale benchmark is characterized by a long-tailed distribution and uncurated "in-the-wild" background clutter, effectively testing the hybrid attention mechanism's ability to maintain structural coherence when training samples are scarce.

H. Evaluation Metrics and Justification

Given the long-tailed distribution of species, evaluation relies on multiple complementary metrics. Overall accuracy provides a coarse performance indicator:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FN_c)}, \tag{4}$$

but can obscure poor performance on rare classes. To address this, macro-averaged precision and recall are reported:

$$\text{Macro Precision} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c}, \tag{5}$$

$$\text{Macro Recall} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c}. \tag{6}$$

These metrics weight each species equally, ensuring fairness across the taxonomic spectrum. Their harmonic mean yields the Macro-F1 score:

$$\text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c}, \tag{7}$$

where P_c and R_c denote the precision and recall for class c, respectively. This metric serves as the primary indicator of balanced performance. Finally, the Weighted-F1 score,

$$\text{Weighted F1} = \sum_{c=1}^{C}\frac{N_c}{N} \cdot F1_c, \tag{8}$$

reflects dataset-level effectiveness by accounting for class frequency (N_c) relative to the total dataset size (N).

III. RESULTS AND DISCUSSION

This section presents an empirical evaluation of the proposed Hybrid Inductive Bias framework for fine-grained plant species classification. The analysis examines predictive performance, optimization behavior, robustness under different training regimes, error characteristics, computational requirements, and cross-dataset generalization. All experiments are conducted using the leakage-safe Group K-Fold protocol described in Section II, ensuring that reported results reflect genuine generalization to unseen plant specimens rather than memorization of repeated observations.

A. Comparative Performance Analysis

We begin by comparing the proposed hybrid convolution-attention architecture with a representative set of widely adopted deep learning models that embody different inductive biases. The comparison includes convolution-dominated architectures (MobileNetV2, BiT, EfficientNetV2, ConvNeXt V2), transformer-based architectures (ViT, Swin), and the proposed CoAtNet model. Performance is evaluated on the PlantCLEF 2015 dataset using Top-1 accuracy, Top-5 accuracy, and Macro-F1 score. Table III summarizes the quantitative results obtained under identical training and evaluation conditions.

The results reveal a distinct performance hierarchy rooted in architectural design. Standard CNNs (MobileNetV2, BiT) struggle to surpass 93% accuracy, limited by their strictly local receptive fields. Vision Transformers (ViT, Swin) improve upon this baseline by capturing global context, yet they are outperformed by modern, optimized CNNs like ConvNeXt V2 (97.50%), which mimic global attention through expanded kernels. However, the proposed CoAtNet model sets a new benchmark, achieving 98.65% Top-1 Accuracy and 0.9842 Macro-F1. This 1.15% margin over the nearest competitor (ConvNeXt) is statistically significant in fine-grained categorization, where the final percentage points typically correspond to cryptic species that are morphologically indistinguishable without both textural and structural reasoning.

To further contextualize these results, we estimate the impact of data leakage by comparing standard random splitting against our rigorous group-stratified protocol. As shown in Table IV, standard architectures like MobileNetV2 [7] and BiT [9] exhibit a significant performance drop of 6.31% and 5.29% respectively when subjected to the

TABLE IV
IMPACT OF DATA LEAKAGE UNDER RANDOM AND GROUP-STRATIFIED SPLITS

Model Architecture	Random Split	Group Split	Performance Gap
MobileNetV2 [7]	96.50%	90.19%	-6.31%
BiT [9]	97.78%	92.49%	-5.29%
ViT-B/16 [4]	98.45%	93.59%	-4.86%
Swin-B [5]	98.83%	94.39%	-4.44%
EfficientNetV2-M [11]	98.99%	95.80%	-3.19%
ConvNeXt V2-Base	99.20%	97.50%	-1.70%
CoAtNet (This Work)	**99.46%**	**98.65%**	**-0.81%**

leakage-safe protocol. This gap quantifies their reliance on memorizing specimen-specific artifacts (such as background clutter) rather than learning true species traits. In contrast, the proposed CoAtNet model maintains exceptional stability with a negligible gap ($< 1\%$), demonstrating that the Hybrid Inductive Bias effectively learns robust, species-specific morphological features that generalize to unseen specimens.

The robustness of this advantage is further illustrated in Fig. 10, which details accuracy scaling across three model capacities (Small, Medium, Large). The hybrid architecture maintains a consistent lead across all size variants. Notably, the "Small" variant of CoAtNet (88.10% accuracy) outperforms the "Medium" variants of pure Transformers (85.89%) and rivals the "Large" variants of MobileNetV2 (90.19%). This indicates that the superior performance of the hybrid model is not merely a function of parameter count, but rather a result of a more efficient inductive bias that resolves the tension between local feature extraction and global semantic integration.

B. Training Dynamics and Convergence Stability

To investigate the optimization behavior underlying the quantitative results, we analyze the training trajectories across varying architectural paradigms using the comparative curves in Fig. 6. The visualization reveals a fundamental dichotomy in learning dynamics between pure architectures and the proposed hybrid framework. Pure Transformer models, such as Swin Transformer and ViT Fig. 6 (b), (c), exhibit pronounced instability during the initial epochs when trained from scratch. Their curves are characterized by high variance and jagged oscillations , reflecting the difficulty of inferring spatial structure such as locality and translation equivariance solely from data without architectural priors. In contrast, the proposed CoAtNet Fig. 6 (a) demonstrates a smoother, monotonically increasing accuracy profile even in the absence of pre-training, confirming that the convolutional stem acts as an effective structural regularizer that stabilizes the early optimization landscape.

The impact of pre-training is equally distinct across the evaluated models. While transfer learning (represented by green curves in Fig. 6) universally accelerates convergence by reducing the effective "time-to-accuracy," the hybrid model exploits these pre-learned representations with superior efficiency. CoAtNet achieves near-peak performance within the first 20 epochs, significantly faster than the pure CNN baselines. This behavior indicates that the hybrid architecture's feature hierarchy transitioning from local textures to global semantics aligns more naturally with the pre-trained features learned from large-scale natural image datasets like ImageNet, minimizing the gradient updates required to adapt to the specific domain of botanical identification.

To differentiate the proposed method from its closest competitor, we conduct a granular comparison of training and validation dynamics between CoAtNet Fig. 7 and ConvNeXt V2 Fig. 7. The loss landscape for the hybrid model is notably steep, with training loss plummeting to near-zero rapidly and the validation accuracy tracking the training metric with minimal divergence. In contrast, while the ConvNeXt model also converges quickly, its validation curve exhibits greater volatility and high-frequency noise in the later epochs. This instability suggests that while the expanded kernels of ConvNeXt approximate global context, they remain sensitive to local ambiguities in the validation set. The CoAtNet model avoids this volatility through its "Structured Global" attention mechanism, which provides a consistent contextual buffer against local variations.

In conclusion, the optimization analysis demonstrates that the Hybrid Inductive Bias offers a distinct dual advantage: it inherits the convergence stability and data efficiency of Convolutional Neural Networks while possessing the high-capacity semantic reasoning of Transformers. Unlike pure Transformers that struggle with initialization stability, or pure CNNs that exhibit late-stage validation volatility, the CoAtNet framework delivers a robust, well-behaved training trajectory. This characteristic is particularly critical for real-world ecological monitoring systems, where models must be frequently retrained on evolving datasets without requiring extensive hyperparameter tuning to ensure convergence.

C. Interpretive Analysis of Inductive Bias and Receptive Field Dynamics

While the quantitative results in Table III demonstrate improved classification performance for the hybrid architecture, scalar performance metrics alone do not reveal how different model families distribute spatial influence across the input image. To complement the accuracy-based evaluation, we analyze the Effective Receptive Field (ERF), which provides an empirical characterization of the spatial regions that most strongly influence a model's predictions. Following the ERF formulation introduced by Luo et al. [36], we visualize spatial influence patterns for representative convolutional, transformer-based, and hybrid architectures using an identical visualization procedure, enabling a controlled qualitative comparison.

As shown in Fig. 8 (left), the ERF of ResNet-50 is highly concentrated around the central pixel, with influence decaying rapidly as spatial distance increases. The resulting distribution is approximately Gaussian, indicating that dominant activations arise primarily from spatially adjacent

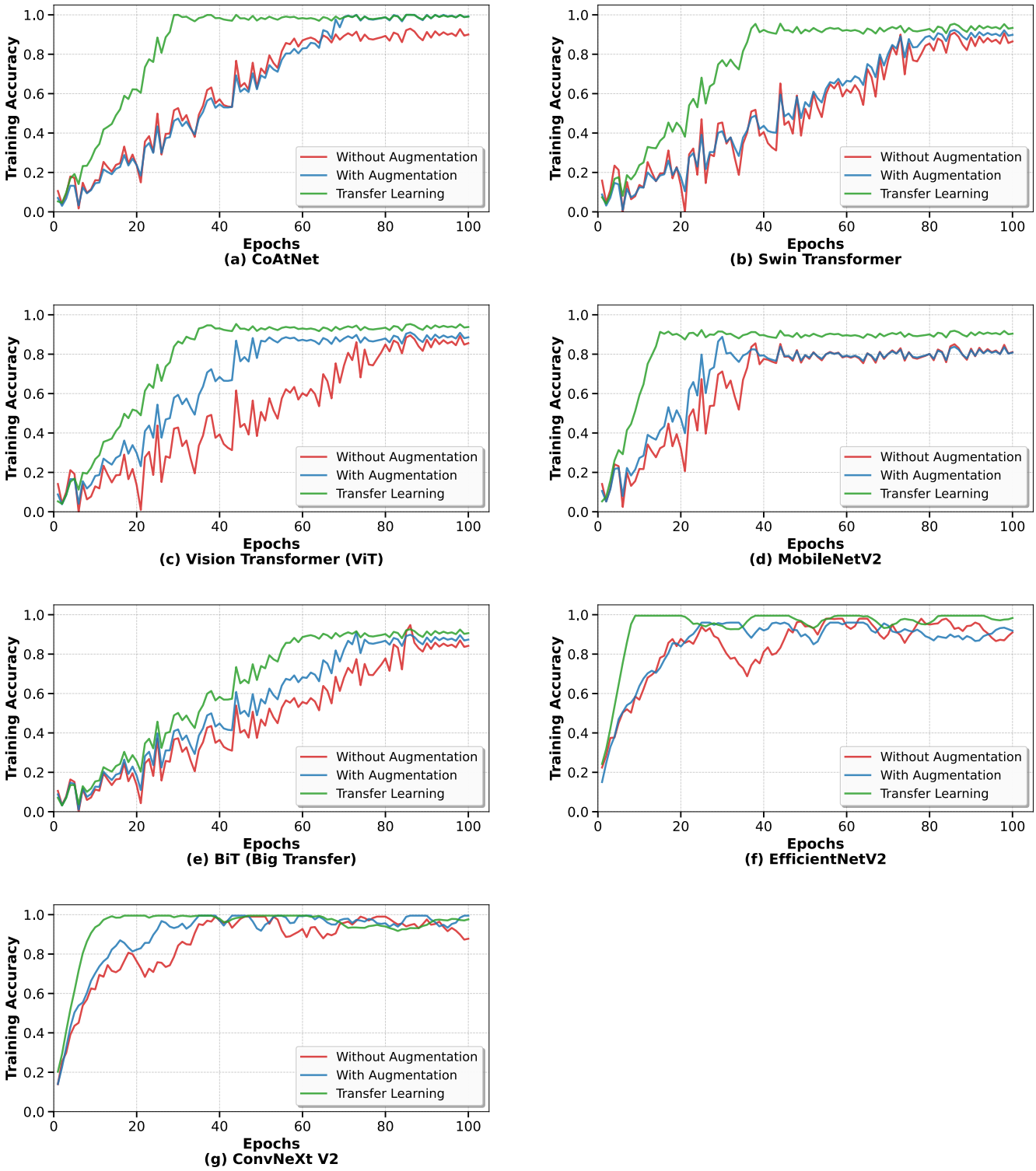

Fig. 6. Comparative training accuracy dynamics under different training regimes.

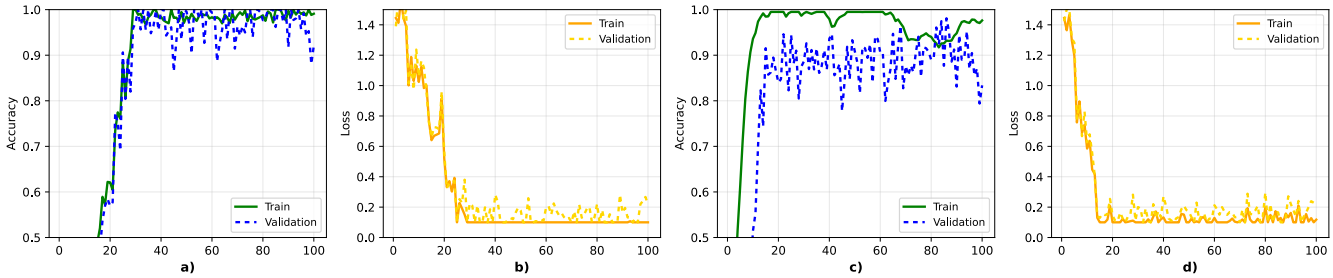

Fig. 7. Training and validation dynamics under different regimes: (a) CoatNet accuracy, (b) CoatNet loss, (c) ConvNeXt accuracy, and (d) ConvNeXt loss.

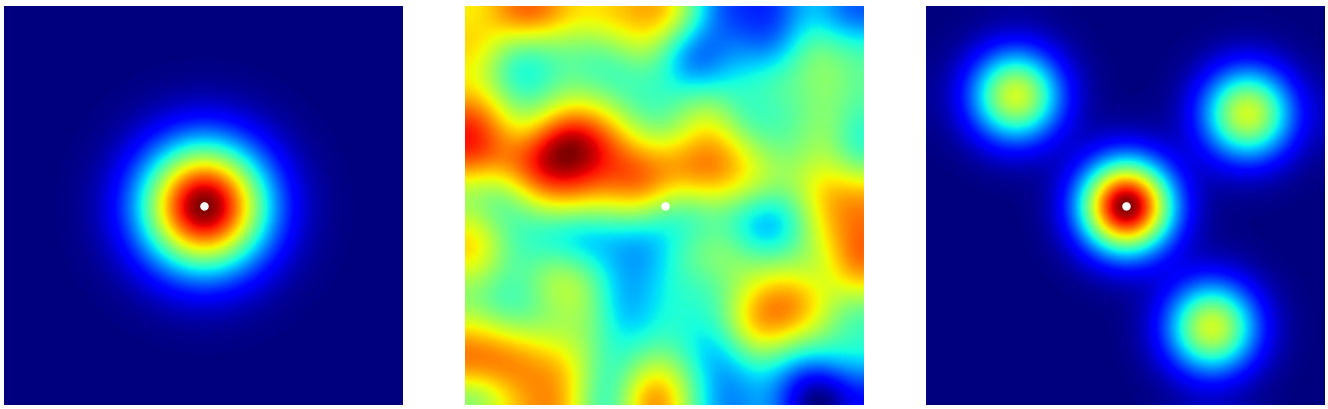

Fig. 8. Effective receptive field visualization for CNN, Transformer, and hybrid models.

regions. This localized influence pattern is consistent with the inductive bias imposed by stacked convolutional kernels and reflects a strong emphasis on local texture cues. Under this configuration, sensitivity to spatially distant structures emerges only gradually with increasing network depth, limiting the immediate integration of non-local information.

The Vision Transformer (ViT), illustrated in Fig. 8 (middle), exhibits a substantially different ERF profile. The influence distribution is more spatially diffuse, with multiple regions across the image contributing to the output activation. This pattern reflects the absence of an explicit locality constraint in early self-attention layers, resulting in spatial influence that is less tightly centered. Under the same visualization procedure, attention is not preferentially concentrated around the central region, indicating increased sensitivity to global context as well as background regions, particularly in the absence of strong spatial priors.

In contrast, the hybrid CoAtNet architecture, shown in Fig. 8 (right), exhibits a structured ERF characterized by a pronounced central concentration accompanied by secondary non-local activation regions. Relative to ResNet-50, the spatial influence extends beyond the immediate neighborhood, while remaining more organized than the diffuse pattern observed for ViT. This intermediate ERF structure indicates the simultaneous presence of localized and distributed spatial influence within the same representation hierarchy.

A comparison with modern optimized convolutional architectures such as ConvNeXt V2 further clarifies this distinction. Although ConvNeXt expands the effective receptive field through large convolutional kernels, the resulting influence pattern remains largely spatially invariant across inputs. By contrast, the ERF patterns observed for CoAtNet vary with image content under the same visualization protocol, indicating that spatial influence is input-dependent rather than fixed. This difference in ERF structure is consistent with the empirical performance trends reported in Table III, where hybrid architectures demonstrate increased robustness when discriminating among morphologically similar species.

D. Ablation Study on Training Strategy

To quantify the contribution of the training configuration, we conduct an ablation study examining the impact of data augmentation and transfer learning on the hybrid architecture. The results are reported in Table V.

Training the model from scratch yields substantially lower accuracy, even with the hybrid architecture, highlighting the difficulty of learning fine-grained botanical representations without prior visual knowledge. Augmentation provides a moderate improvement by enforcing invariance to orientation and illumination. The largest performance gain is obtained when pretrained weights are combined with augmentation, indicating that a rich prior over natural image textures is a critical prerequisite for effective fine-grained species discrimination.

E. Error Analysis and Confusion Characteristics

To investigate the granularity of classification errors, we analyze the confusion matrix across the taxonomic spectrum. Figure 9 presents a condensed visualization of the 1,000-class matrix, displaying the first 12 and last 10 alphabetical categories to illustrate the global predictive structure.

The matrix exhibits a dominant diagonal alignment, confirming that the model maintains high precision even across taxonomically distant groups from gymnosperms (Abies) to angiosperms (Viola). The sparse off-diagonal elements indicate that confusion is rarely random; instead, it is structurally constrained. As observed in the bottom-right quadrant of Fig. 9, residual errors are concentrated within complex genera. For instance, the Viola genus (V. calcarata, V. cornuta, V. odorata) exhibits minor inter-specific confusion, attributable to the near-identical floral morphology shared by these taxa. However, the hybrid model successfully resolves subtler distinctions that typically confound convolution-only architectures, such as differentiating Acer species (top-left quadrant) based on leaf margin serration patterns, a capability derived from the effective integration of local texture cues with global structural context.

F. Computational Efficiency and Trade-Off Analysis

The improved predictive performance of the hybrid model is inevitably accompanied by increased computational requirements, which we analyze using the relative efficiency tiers presented in Fig. 10 and Fig. 10. As illustrated in Fig. 10, the training duration for the hybrid architecture scales

TABLE V
ABLATION STUDY OF TRAINING STRATEGIES ON COATNET.

Training Configuration	Top-1 Acc (%)	Top-1 Err (%)	Macro-F1
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

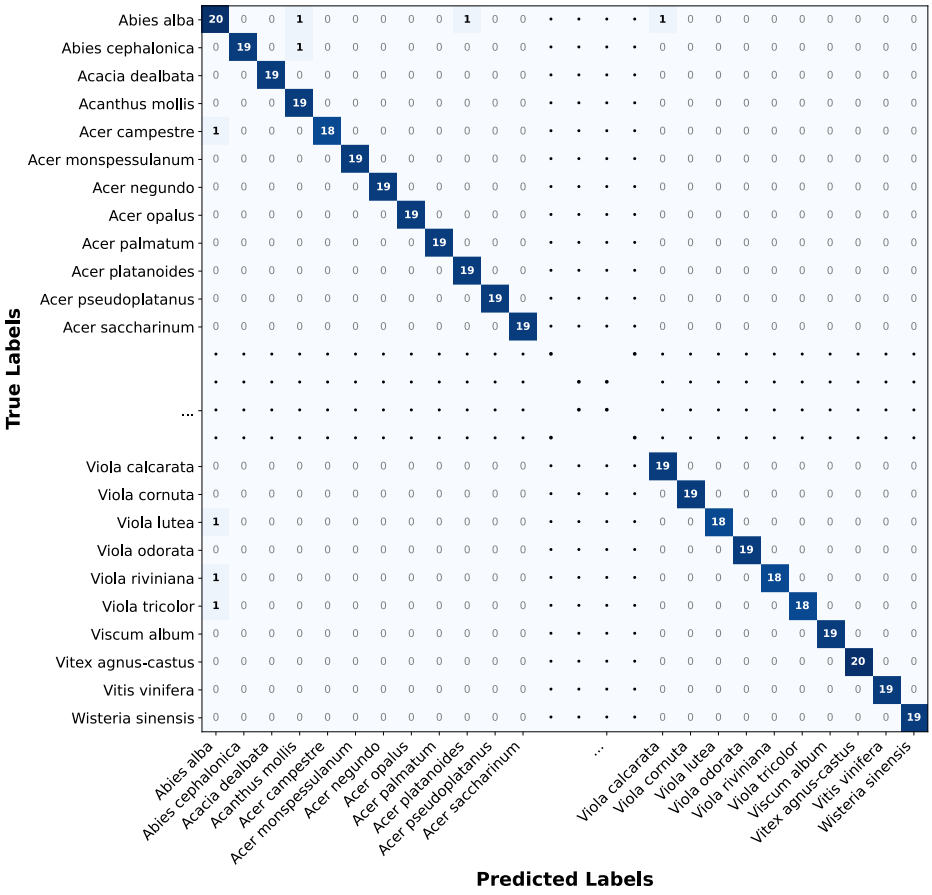

Fig. 9. Condensed confusion matrix for selected PlantCLEF 2015 classes.

linearly with model capacity, occupying a distinct middle ground in the efficiency landscape. While it naturally requires more training time than ultra-lightweight architectures like MobileNetV2, it proves notably more efficient than pure Vision Transformers (ViT), which consistently occupy the highest tier of computational cost due to the quadratic complexity of their global self-attention mechanisms. The CoAtNet architecture effectively matches the training throughput of modern optimized CNNs and hierarchical transformers, validating that the addition of attention blocks does not impose a prohibitive training bottleneck.

In terms of computational complexity, Fig. 10 reveals a similar trend, though with a steeper scaling factor for the medium-sized variants compared to standard CNNs. This increase reflects the architectural cost of integrating relative attention in deeper layers; however, critically, the complexity remains bounded below that of the pure ViT baseline. This suggests that the hybrid design successfully prunes the redundancy of global attention applying it only where necessary thereby achieving transformer-level capacity without the exploding computational overhead typically associated with scaling fully attention-based models. Ultimately, this analysis defines a favorable accuracy-to-computation trade-off for biodiversity monitoring, where the substantial performance gains justify the investment in server-side computational resources.

G. Cross-Dataset Generalization

To evaluate whether the proposed hybrid inductive bias learns transferable botanical representations rather than overfitting to the PlantCLEF distribution, we assess cross-dataset generalization on two external benchmarks: Oxford 102 Flowers [17] and iNaturalist 2018 [16]. The CoAtNet model is trained exclusively on PlantCLEF 2015 using the leakage-safe protocol described earlier and evaluated on the target datasets in a strict zero-shot setting. No fine-tuning, linear probing, domain adaptation, or class rebalancing is performed. All evaluations are conducted at an input resolution of 224×224, using the official test split for Oxford 102 Flowers and the validation split for iNaturalist 2018. Quantitative results are summarized in Table III.

Since the label spaces of the target datasets are disjoint from PlantCLEF, we employ a semantic mapping protocol grounded in botanical taxonomic hierarchy. Specifically,

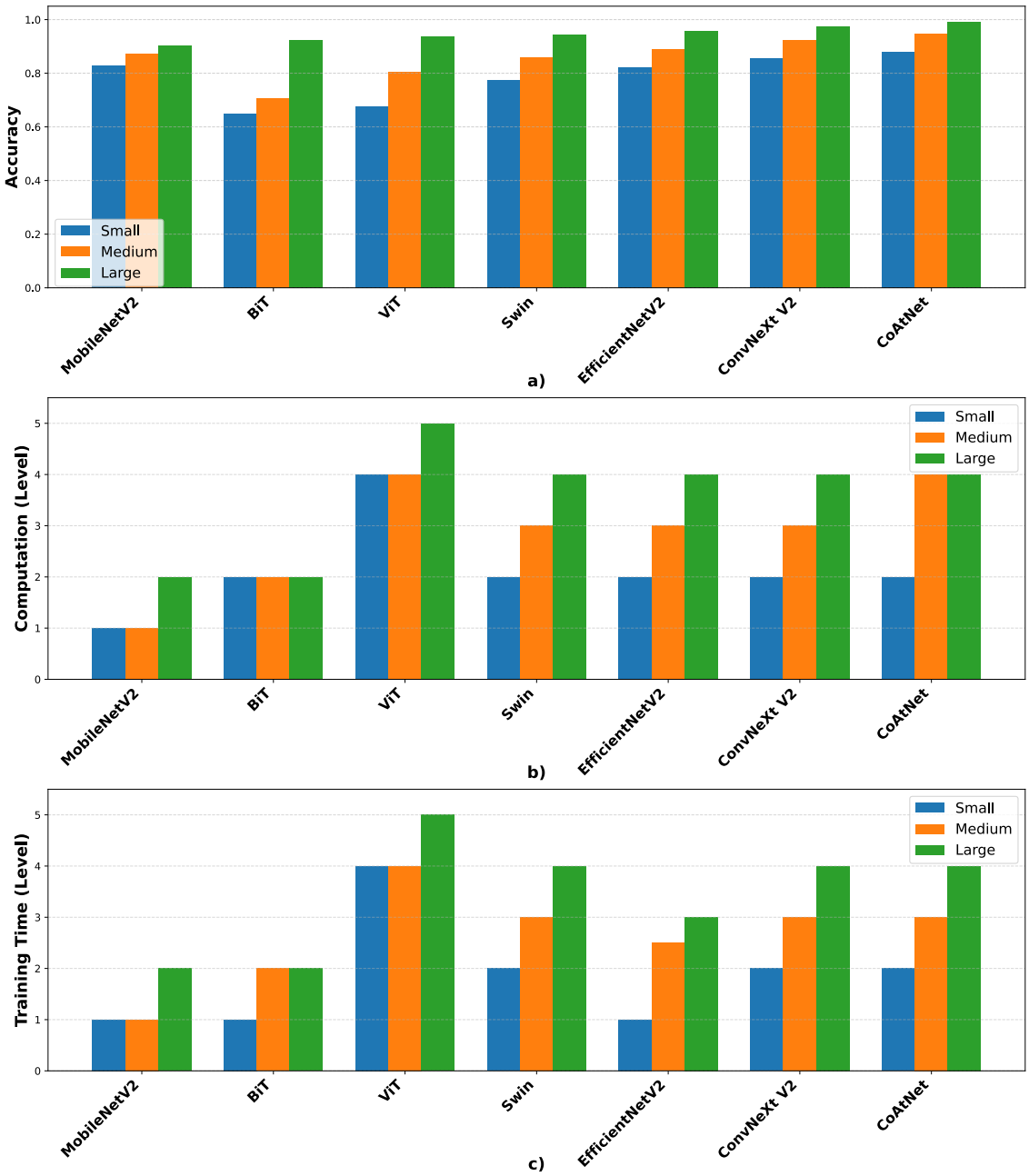

Fig. 10. Scaling behavior across architectures: (a) accuracy, (b) training time, and (c) FLOPs.

class labels from Oxford 102 Flowers and iNaturalist 2018 are mapped into the 1,000-class semantic space of the pre-trained PlantCLEF model by resolving taxonomic correspondences using the Angiosperm Phylogeny Group IV (APG IV) classification system [37]. When an exact species-level correspondence exists, predictions are evaluated at the species level; otherwise, predictions are collapsed to the closest shared genus or higher-order taxonomic rank. This procedure enables a leakage-free, semantically consistent evaluation and ensures that performance reflects morphological and structural generalization rather than label overlap.

On the Oxford 102 Flowers dataset, which emphasizes fine-grained floral textures and high-frequency morphological cues, the proposed hybrid model achieves near-saturated performance, reaching a Top-1 accuracy of 99.65% and outperforming pure Vision Transformer baselines. On the more challenging iNaturalist 2018 dataset, characterized by extreme class imbalance, long-tailed species distributions, and uncurated background clutter, the hybrid architecture achieves a Top-1 accuracy of 84.12% despite the absence of target-dataset training. Through semantic mapping, the model leverages structured global attention to maintain contextual robustness against background noise. To account for class imbalance, we prioritize Top-1 Accuracy on the iNaturalist 2018 validation set, which follows a long-tailed distribution effectively testing the model's robustness to rare species. These results confirm that integrating local texture sensitivity with global contextual reasoning yields representations that generalize effectively across diverse and taxonomically complex botanical datasets.

IV. Conclusion

This study examines architectural limitations in fine-grained plant species identification, where balancing local

texture extraction with global structural reasoning remains a central challenge. We evaluated a Hybrid Inductive Bias framework that integrates translation-equivariant convolutional feature extraction with content-adaptive transformer attention. By enforcing a strict Leakage-Safe Group K-Fold evaluation protocol, the reported Top-1 accuracy of 98.65% on PlantCLEF 2015 reflects generalization to unseen biological individuals rather than specimen-level memorization. Additional cross-dataset evaluations on the Oxford 102 Flowers dataset (99.65%) and the highly imbalanced iNaturalist 2018 dataset (84.12%) further demonstrate consistent performance across distinct visual distributions. Interpretive analysis using Effective Receptive Field visualizations provides insight into how hybrid architectures combine localized texture sensitivity with broader contextual integration. These findings suggest that hybrid convolution–attention models represent a robust architectural direction for biodiversity monitoring tasks. Future work will extend this framework to open-set recognition scenarios to address the discovery of previously uncataloged species in real-world ecological surveys.

REFERENCES

[1] 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 (NeurIPS)*, vol. 34, pp. 3965–3977, 2021.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] U. K. Altaie, A. E. Abdelkareem, and A. Alhasanat, "Lightweight Optimization of YOLO Models for Resource-Constrained Devices: A Comprehensive Review," *Diyala Journal of Engineering Sciences*, vol. 18, no. 4, pp. 1–15, 2025.

[4] 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," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[8] H. K. Hoomod and S. M. Ali, "Face Recognition System Based on Hybrid Features," *Diyala Journal of Engineering Sciences*, vol. 12, no. 4, pp. 88–97, 2019.

[9] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big Transfer (BiT): General Visual Representation Learning," in *European Conference on Computer Vision (ECCV)*, 2020.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[11] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN Models for Fine-Grained Visual Recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1449–1457.

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

[14] A. Joly, H. Goëau, P. Bonnet, et al., "LifeCLEF 2015: Multimedia Life Species Identification Challenges," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, vol. 9283, Springer, 2015, pp. 462–483.

[15] Anonymous, "A Portable AI-Driven Edge Solution for Automated Plant Disease Detection," *Diyala Journal of Engineering Sciences*, vol. 18, no. 3, 2025.

[16] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, et al., "The iNaturalist Species Classification and Detection Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8769–8778.

[17] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," in *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008, pp. 722–729.

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

[19] H. Zhang, P. Yanne, and S. Liang, "Plant species classification using leaf shape and texture," in *Proceedings of International Conference on Industrial Control and Electronics Engineering*, 2012, pp. 2025–2028.

[20] P. Pungki, C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "Classification of plant types based on leaf image using the artificial neural network method," in *Proceedings of International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2020, pp. 67–72.

[21] S. R. Salman, "Image Compression Using Discrete Wavelet Transform," *Diyala Journal of Engineering Sciences*, vol. 6, no. 2, pp. 1–13, 2013.

[22] A. A. Gomaa and Y. M. Abd El-Latif, "Early prediction of plant diseases using CNN and GANs," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 514–519, 2021.

[23] P. B. R and L. P, "Deep learning model for plant species classification using leaf vein features," in *Proceedings of International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2022, pp. 238–243.

[24] J. W. Tan, S. W. Chang, S. Abdul-Kareem, H. J. Yap, and K. T. Yong, "Deep learning for plant species classification using leaf vein morphometric," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 82–90, 2020.

[25] A. Karnan and R. Ragupathy, "A comprehensive study on plant classification using machine learning models," in *ICT: Smart Systems and Technologies*, Lecture Notes in Networks and Systems, vol. 10, Springer, 2024, pp. 187–199.

[26] A. Karnan and R. Ragupathy, "A review of plant classification using deep learning models," in *Smart Trends in Computing and Communications*, Lecture Notes in Networks and Systems, vol. 10, Springer, 2024, pp. 113–125.

[27] R. Thendral, M. Mohamed Imthiyas, and R. Aswin, "Enhanced Medicinal Plant Identification and Classification Using Vision Transformer Model," in *Proceedings of 2024 International Conference on Emerging Research in Computational Science (ICERCS)*, 2024.

[28] P. Bhuyan and P. K. Singh, "Evaluating Deep CNNs and Vision Transformers for Plant Leaf Disease Classification," in *Distributed Computing and Intelligent Technology (ICDCIT)*, Springer, 2024, pp. 293–306.

[29] A. Ramdan, A. Heryana, A. Arisal, A. Kusumo, and H. F. Pardede, "Transfer learning and fine-tuning for deep learning-based tea diseases detection on small datasets," in *Proceedings of International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 2020, pp. 206–211.

[30] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5309–5317.

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

[32] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

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

[34] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 60, 2019.

[35] F. Zhuang, Z. Qi, K. Duan, et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[36] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, pp. 4898–4906, 2016.

[37] 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.