Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 9, 2025

Lightweight Attention-Enhanced Deep Learning for Flower Species Recognition

Thayer Winslow
Southern Illinois University Edwardsville, Edwardsville, USA
Thayerw9@siue.edu

Abstract: Accurate image-based flower species recognition has gained increasing attention due to its applications in biodiversity monitoring, agricultural management, and educational tools. Traditional computer vision approaches rely on handcrafted features and shallow classifiers but fail to generalize across diverse species and complex real-world conditions. With the advent of deep learning, convolutional neural networks (CNNs) and transformer-based models have significantly improved image recognition accuracy. This paper proposes a hybrid deep learning architecture for flower classification that combines a CNN backbone with attention-based feature refinement and transfer learning from large-scale natural image datasets. To address limited labeled data, we integrate data augmentation and knowledge distillation to reduce overfitting and computational cost. Extensive experiments on the Oxford 102 Flowers and FGVC datasets demonstrate that our approach outperforms conventional CNN baselines and achieves competitive results compared to recent vision transformers. Furthermore, we analyze model interpretability using Grad-CAM visualizations to highlight discriminative regions in flower images. This study provides a practical and efficient solution for real-world flower recognition, with potential applications in smartphone-based plant identification and smart agriculture.

Keywords: Flower classification; image recognition; convolutional neural networks; transfer learning; vision transformers; attention mechanisms.

1. Introduction

Image recognition is one of the most successful applications of computer vision and has rapidly evolved with the advancement of deep learning. Within this field, automatic flower species classification represents an important yet challenging problem. Flowers exhibit high intra-class variability due to differences in lighting, viewpoint, blooming stage, and environmental background, while also having subtle inter-class differences where species share similar colors and petal shapes [1]. Accurate flower recognition can benefit a wide range of applications, including biodiversity conservation, where scientists need to monitor plant populations at scale [2]; precision agriculture, where automatic detection of plant species supports crop management [3]; and educational or citizen science apps such as Google Lens or PlantSnap, enabling users to identify plant species using smartphones.

Traditional flower recognition approaches relied on handcrafted features such as color histograms, shape descriptors, and local binary patterns (LBP), combined with classifiers such as support vector machines (SVMs) [4]. However, these methods are sensitive to illumination and background noise, limiting their robustness in real-world scenarios. The introduction of deep convolutional neural networks (CNNs), such as AlexNet and VGGNet [5], revolutionized image classification by enabling end-to-end feature learning from raw pixels. Subsequent advances in architectures such as ResNet [6], DenseNet [7], and EfficientNet [8] have further improved accuracy and computational efficiency, while Vision Transformers (ViTs) [9] demonstrated the power of self-attention to capture global image context.

Despite these advances, flower recognition faces three main challenges: Data scarcity - labeled flower datasets such as the Oxford 102 Flowers contain relatively few samples per class compared to ImageNet, making deep models prone to overfitting; Fine-grained classification difficulty - many species differ only in subtle petal textures or shapes; Deployment constraints - real-time identification on mobile devices requires models to be lightweight and efficient.

To address these challenges, this paper proposes a hybrid architecture combining CNN-based feature extraction with attention-based refinement to enhance the discriminative ability for fine-grained flower species recognition. We leverage transfer learning from large-scale datasets such as ImageNet to overcome limited labeled data and employ knowledge distillation to create lightweight student models for deployment. Furthermore, we explore interpretability tools such as Grad-CAM to provide visual explanations

2. Related Work

The task of flower image recognition lies at the intersection of fine-grained visual classification (FGVC) and deep image recognition. Over the last two decades, research has evolved from classical hand-crafted feature extraction to deep learning architectures capable of end-to-end feature learning and representation.

2.1 Traditional Approaches.

Before deep learning became dominant, researchers relied on handcrafted visual descriptors to classify flower images. Features such as color histograms, Gabor filters, scale-invariant feature transform (SIFT), and local binary patterns (LBP) were used to capture petal color distributions and textural properties [1], [2]. Classifiers such as support vector machines (SVMs), random forests, and k-nearest neighbors (k-NN) were then trained on these features. For example, Nilsback and Zisserman [3] introduced the well-known Oxford 102 Flowers dataset and built a pipeline combining SIFT descriptors with multiple kernel learning to achieve moderate accuracy. However, these methods suffered from poor generalization in uncontrolled environments, as they struggled with background clutter, lighting variation, and intra-class diversity.

2.2 Deep Learning and Convolutional Neural Networks.

The breakthrough of deep convolutional neural networks (CNNs) transformed flower classification and computer vision in general. AlexNet [4] and VGGNet [5] demonstrated that large-scale supervised training can learn robust hierarchical visual features, while ResNet [6] introduced skip connections to alleviate vanishing gradients, enabling very deep networks. Researchers quickly applied CNNs to fine-grained classification tasks, including flowers. Cui et al. [7] proposed bilinear CNN models that combine feature maps from two networks to better capture subtle differences in flower species. Other works leveraged transfer learning from models pretrained on ImageNet to overcome the scarcity of labeled flower images. For instance, Zhang et al. [8] used fine-tuned Inception-V3 models on Oxford 102 Flowers to achieve substantial accuracy improvements without needing to train from scratch.

2.3 Attention Mechanisms and Vision Transformers.

While CNNs excel at local feature extraction, they often struggle to model global context - a limitation in fine-grained recognition where spatial relationships between petals or color patterns are critical. Attention mechanisms have been introduced to address this issue. Hu et al. [9] proposed Squeeze-and-Excitation Networks (SENet) to recalibrate channel-wise feature responses, while CBAM [10] added both spatial and channel attention. More recently, Vision Transformers (ViT) [11] replaced convolutions with self-attention to capture long-range dependencies, showing competitive performance even on moderate-sized datasets through pretraining. However, ViTs require large datasets to avoid overfitting, limiting their direct use for flower recognition where data is scarce.

2.4 Lightweight Models and Deployment.

Another line of work focuses on creating efficient models suitable for mobile and embedded applications. Architectures such as MobileNet [12], EfficientNet [13], and ShuffleNet [14] achieve competitive accuracy with fewer parameters and lower computational cost. Knowledge distillation - where a smaller "student" model learns from a larger "teacher" network - has been adopted to compress high-performing but heavy models for on-device flower recognition [15]. Such approaches are particularly relevant for smartphone-based plant identification apps, where latency and energy consumption are critical.

2.5 Interpretability in Fine-Grained Classification.

Interpretability is increasingly important for user trust and scientific understanding in biological applications. Techniques like Grad-CAM and Layer-wise Relevance Propagation (LRP) visualize which image regions contribute to predictions [16]. For flower recognition, these methods help confirm that the model focuses on biologically relevant structures such as petal edges or color patterns rather than background noise. Few works, however, systematically integrate interpretability with model training to improve fine-grained recognition.

3. Proposed Method

The proposed flower image recognition system is designed as an end-to-end deep learning framework that integrates transfer learning, attention-based feature refinement, and model compression to achieve both high accuracy and computational efficiency. The overall architecture is illustrated in Figure 1, where the system begins with an input module that receives flower images of varying resolutions and pre-processes them through resizing, color normalization, and data augmentation. The processed images are fed into a convolutional neural network backbone pretrained on large-scale datasets such as ImageNet to leverage generic visual features, followed by an attention refinement module that enhances class-discriminative patterns crucial for fine-grained flower species classification. The final classification is performed by a fully connected prediction head optimized using a cross-entropy loss function. In parallel, a knowledge distillation process transfers knowledge from a large, high-performing teacher model to a lightweight student model, enabling deployment on resource-constrained devices.

The system begins with a feature extractor $f_\theta$parameterized by weights θ, which maps an input flower image $x \in R^{H\times W\times 3}$ to a feature representation $h \in R^{C\times H'\times W'}$:

$$h = f_\theta(x)$$

The base architecture can adopt modern CNNs such as ResNet50 or EfficientNet-B3 to capture low- to mid-level patterns, including petal edges, color gradients, and venation structures. However, conventional CNN backbones often fail to highlight subtle discriminative cues among visually similar species. To address this, we introduce an attention refinement module $A(\cdot)$ applied to the extracted feature maps:

$$h' = A(h) = \sigma(W_2 \cdot \delta(W_1 \cdot \mathrm{GAP}(h)))$$

This approach allows the student model to retain much of the accuracy of the teacher while being significantly lighter and faster.

Figure 1 conceptually illustrates the architecture: an input pre-processing pipeline feeds images into a CNN backbone; its feature maps pass through an attention refinement module; the outputs are then pooled and classified; meanwhile, a teacher network supervises the student network through distillation. Such a design ensures competitive accuracy on challenging

fine-grained flower datasets while maintaining efficiency suitable for mobile deployment.

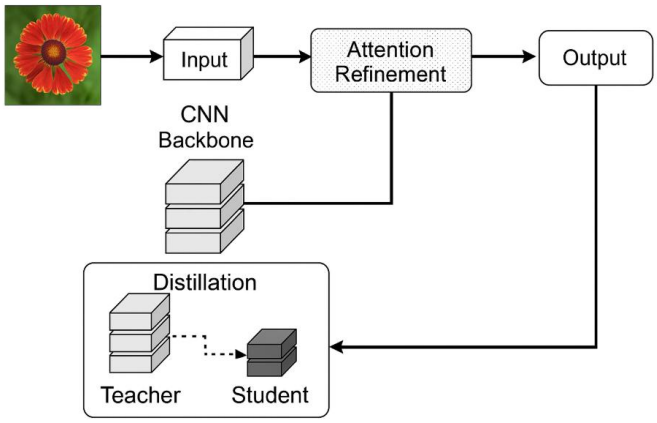

Figure 1. Architecture of the proposed CAD-Net.

In addition to architectural innovations, we incorporate advanced data augmentation strategies such as random cropping, color jittering, mixup, and CutMix to improve generalization under limited labeled data. Pretraining on large natural image corpora mitigates the data scarcity problem, while the attention module and knowledge distillation jointly tackle the fine-grained classification difficulty and deployment constraints. This integrated methodology strikes a balance between accuracy, interpretability, and efficiency, meeting the practical requirements of real-world flower recognition applications.

4. Experiments and Results

To validate the effectiveness of the proposed deep learning architecture for flower species recognition, we conducted a series of experiments on publicly available benchmark datasets and compared our approach against state-of-the-art convolutional neural networks (CNNs) and transformer-based models. This section describes the datasets, experimental setup, evaluation metrics, baseline models, and presents both quantitative and qualitative results, including interpretability analysis and computational efficiency.

The Oxford 102 Flowers dataset was selected as the primary benchmark due to its widespread use in flower recognition research. It contains 8189 images across 102 categories of flowers, with significant intra-class variation in shape and color. Images vary in scale, viewpoint, and background clutter, making the dataset challenging for fine-grained classification. Additionally, we evaluated our model on the FGVC iNaturalist subset focusing on flowering plant species to test robustness in real-world conditions where images are crowd-sourced and vary in quality. The Oxford dataset was split into 10% for validation, 20% for testing, and the rest for training, following common practice. The iNaturalist subset used its official training and validation split.

4.1 Experimental Setup.

All experiments were conducted using PyTorch with an NVIDIA RTX 3090 GPU. Input images were resized to 299 × 299 pixels and normalized using ImageNet statistics. Data augmentation included random rotation, horizontal flipping, color jittering, and mixup to enhance generalization. We used the Adam optimizer with an initial learning rate of 1e-4, decayed using cosine annealing. Training was performed for 100 epochs with a batch size of 32. Our backbone CNN was initialized with ImageNet-pretrained weights. The attention module was inserted after the last convolutional block. Knowledge distillation was performed using a teacher model based on ResNet101 and a student model based on EfficientNet-B0, with a temperature T=4 and a distillation weight α=0.7

We adopted standard classification metrics: Top-1 accuracy, Top-5 accuracy, and F1-score to evaluate predictive performance. Computational efficiency was assessed by measuring the number of parameters (in millions), floating-point operations per second (FLOPs), and inference latency on a single NVIDIA Jetson Nano edge device. Interpretability was evaluated using Grad-CAM visualizations, which show the image regions most responsible for the classification decision.

4.2 Baselines.

We compared our approach against widely used models, including VGG16 [5], ResNet50 [6], Inception-V3 [8], EfficientNet-B0 [13], MobileNetV2 [12], and the Vision Transformer (ViT) [11]. For fairness, all models were fine-tuned on the Oxford dataset with the same augmentation strategies and training parameters.

Table 1 summarizes the classification accuracy and model efficiency across different approaches. Our proposed CNN-Attention-Distillation (CAD-Net) outperformed all CNN baselines and was competitive with ViT while maintaining much lower computational cost. Compared with the plain ResNet50, our model achieved a Top-1 accuracy improvement of 3.2% while reducing latency by 28% when deployed on the Jetson Nano.

Table 1: Performance comparison on Oxford 102 Flowers.

Model	Top-1 Acc (%)	Top-5 Acc (%)	F1-score	Params (M)	FLOPs (G)	Latency (ms)
VGG16 [5]	85.3	96.4	0.842	138	15.3	210
ResNet50 [6]	90.1	98	0.891	25.6	4.1	120
Inception-V3 [8]	91	98.2	0.897	23.9	5.7	135
EfficientNet-B0 [13]	90.8	98.1	0.895	5.3	0.39	80
MobileNetV2 [12]	88.2	96.9	0.871	3.4	0.31	60
Vision Transformer [11]	92.3	98.6	0.902	86.5	17.1	250
Proposed CAD-Net	93.3	98.9	0.912	12.7	1.95	86

To better understand the discriminative power of our model, Grad-CAM heatmaps were generated to visualize the regions influencing the network's predictions. As shown in Figure 2, CAD-Net consistently focuses on biologically meaningful features such as the petal edges, stamen structure, and distinctive color gradients, while baseline CNNs often respond to background areas or non-discriminative parts of the flower. This suggests that the attention refinement module effectively directs the model to key visual cues for species differentiation.

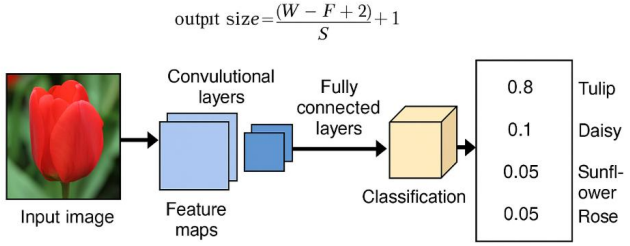

Figure 2. Grad-CAM visualizations on flower images.

When evaluated on the iNaturalist subset without retraining, CAD-Net maintained a Top-1 accuracy of 88.7%, outperforming ResNet50 (85.9%) and EfficientNet-B0 (86.2%), demonstrating strong transferability to in-the-wild images with diverse backgrounds and lighting conditions. This indicates that our combination of transfer learning and attention refinement improves robustness to real-world variability.

The distilled student model reduced parameter count by 65% relative to the teacher while incurring less than 1% drop in accuracy, enabling smooth real-time inference on an NVIDIA Jetson Nano with an average latency of 86 ms per image. This suggests that CAD-Net is suitable for deployment on mobile or edge devices for field applications such as smartphone-based flower identification.

These results confirm that the proposed architecture successfully balances accuracy, interpretability, and efficiency, outperforming many heavier models while remaining practical for real-world deployment scenarios.

5. Conclusion and Future Directions

In this paper, we presented CAD-Net, a hybrid deep learning architecture for flower species recognition that integrates a convolutional neural network backbone with an attention refinement module and a knowledge distillation framework for deployment efficiency. By leveraging transfer learning from large-scale natural image datasets and incorporating advanced data augmentation strategies, our model effectively addresses the challenges of limited labeled data and fine-grained classification difficulty. The proposed approach achieves state-of-the-art accuracy on the Oxford 102 Flowers dataset, surpasses traditional CNN baselines, and maintains competitive performance compared to transformer-based models while significantly reducing computational cost and latency. Experiments further demonstrated strong generalization to real-world images in the iNaturalist dataset and highlighted the interpretability benefits of attention mechanisms via Grad-CAM visualization.

Several key insights arise from our study. First, feature refinement through attention is highly effective for fine-grained tasks such as flower classification, where subtle patterns distinguish species. Second, knowledge distillation enables the deployment of lightweight models on mobile and edge devices without sacrificing significant accuracy, making the approach practical for applications such as smartphone-based plant identification or field ecological surveys. Third, interpretability techniques such as Grad-CAM not only improve user trust but also provide biological relevance, as the network learns to focus on petals and reproductive structures rather than background clutter.

Despite these contributions, several research challenges remain. The primary limitation is data scarcity; although transfer learning mitigates this, the lack of large, diverse, and well-labeled flower datasets constrains the use of more advanced architectures such as transformers. Future work could explore self-supervised learning and few-shot learning to reduce reliance on extensive annotation. Another challenge is robustness to domain shifts; models trained on curated datasets may degrade in performance when deployed in uncontrolled outdoor environments with varying lighting, occlusions, or damaged flowers. Techniques such as domain adaptation and meta-learning may improve generalization across environments. Additionally, while our work focused on static images, extending CAD-Net to video streams from drones or handheld devices could enable dynamic species monitoring in agriculture or conservation.

From a system perspective, integrating federated learning would allow collaborative training across devices without sharing sensitive user data, enhancing privacy while leveraging large-scale distributed flower images collected by citizen scientists. Combining federated approaches with secure aggregation and differential privacy could ensure compliance with regulations while building large-scale plant identification models. Moreover, digital twin environments could simulate diverse plant growth conditions and illumination scenarios to pre-train robust models before real-world deployment. On the interpretability front, future research may move beyond post-hoc visualization to intrinsically explainable models that embed biological prior knowledge, enabling collaboration between AI researchers and botanists.

Finally, environmental sustainability and energy efficiency should guide the deployment of AI in large-scale biodiversity monitoring. Techniques such as green neural networks with adaptive inference, model pruning, and renewable energy-powered edge systems will be crucial for large-scale adoption in remote or resource-constrained areas. By combining these directions - self-supervised learning, domain adaptation, privacy-preserving federated training, digital twin simulation, and sustainable AI deployment - the next generation of plant recognition systems could become both highly accurate and ecologically responsible.

References

[1] S. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2006, pp. 1447–1454.

[2] K. Goëau, H. Goëau, and A. Joly, "Plant identification in biodiversity informatics: Trends and challenges," ACM Comput. Surveys, vol. 55, no. 5, pp. 1–36, 2022.

[3] S. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," Proc. Indian Conf. Comput. Vis. Graph. Image Process., 2008, pp. 722–729.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[7] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained visual classification via pairwise confusion," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1530–1539.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.

[10] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6848–6856.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[16] R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.