International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1385-1393
https://doi.org/10.70917/ijcisim-2026-5700
Article

AGHV-Net: An Attention-Guided Hybrid Vision Network for Robust Flower Classification

Abhilasha Varshney[1], Dr. Sudeep Varshney[1], Dr. Ankur Choudhary[1]

[1] Department of Computer Science, Sharda School of Engineering & Technology, Sharda University, Greater Noida, Accenture, Noida
Email: abhilashabharatgupta@gmail.com
Email: sudeep.varshney@sharda.ac.in
Email: ankur.tomer@gmail.com

Abstract: Classification of fine-grained flower species is still a challenging task in computer vision because of high inter-class similarity, intra-class variability, illumination variations, and complex background conditions but Conventional CNN-based methods are good at capturing local spatial features but are generally inadequate in capturing long-range contextual dependencies, while ViT-based architectures learn well the global context but lack efficient local discriminative representation. In this paper, we propose Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net) for accurate flower species classification. The proposed framework combines CNN-based local feature extraction with Transformer-based global contextual learning using an adaptive attention-guided feature fusion mechanism. First, we preprocess and augment the input flower images to improve the uniformity and generalization of the features. The CNN branch captures the discriminative local feature representations such as petal textures, edge structures, and color gradients. The Vision Transformer branch captures the long-range semantic dependencies and self-attention mechanisms. Then, an adaptive attention-guided feature fusion module is dynamically used to learn the local feature representations of CNN and Transformer to perform the local-global feature integration. Finally, we process this fused feature representation through fully connected and Softmax layers to perform multi-class classification. They conducted comprehensive experiments on the Oxford 102 Flower Dataset with accuracy, precision, recall, and F1-score as evaluation metrics. Experimental results show that the proposed AGHV-Net classification accuracy reaching 97.84% outperforms some existing CNN, Transformer, and hybrid deep learning architectures across many other methods. The comparison and ablation analysis further provides strong support for the proposed attention-based fusion mechanism that can dramatically improve discriminative feature identification, alleviate the inter-class confusion to enhance overall classification performance. Our new framework offers a practical and reliable route for fine-grained flower identification and can be extended into other complex picture classification applications.
Keywords: Flower Classification, Hybrid Deep Learning, CNN, Vision Transformer, Attention Mechanism, AGHV-Net

1. Introduction

Flower species classification is an important research problem in computer vision because of its applications in biodiversity conservation, smart agriculture, ecological monitoring, botanical analysis, and medicinal plant identification. Fine-grained flower classification is challenging due to high inter-class similarity, intra-class variation, complex backgrounds, illumination changes, and variations in flower orientation and scale. Traditional machine learning approaches relied on handcrafted features such as color, texture, shape, SIFT, SURF, and GLCM descriptors combined with classifiers like Support Vector Machines (SVMs) and Random Forests. However, these approaches suffered from poor robustness and limited generalization capability in real-world scenarios.

Deep Learning has revolutionized image classification through automatic feature learning. Convolutional neural networks (CNNs) such as AlexNet, VGGNet, ResNet, DenseNet and EfficientNet have achieved excellent results in extracting hierarchical local features like edges, textures, petal structures and color gradients [1][5]. While

CNNs can capture local spatial information, they struggle to model long-range contextual dependencies because of the focus of convolution on local receptive fields. Vision Transformers (ViTs) have recently introduced self-attention mechanisms to capture global semantic relationships and long-range dependencies between image regions [6].

Vision Transformers are very successful in many image classification tasks due to their strong contextual learning capabilities. However, ViTs typically require large datasets and can fail to capture fine local discriminative information required for fine-grained flower classification. To address the problem of these limitations, hybrid CNN-Transformer architectures have recently been developed and they combine CNN-based local feature extraction and Transformer-based global contextual learning [7], [8]. Furthermore, attention mechanisms have proved to be very effective in feature representation by dynamically highlighting and suppressing relevant features [9]. However, despite the recent development in such hybrid approaches, many existing hybrid approaches are based on static feature fusion methods and do not adapt to local and global feature importance.

With the current data and study gaps, in this paper we propose an Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net) to accurately classify flower species. Our model combines CNN-based local feature extraction with Vision Transformer-based global contextual learning via an adaptive attention-guided feature fusion mechanism. CNN extracts fine-grained local features (petal textures, edges, color patterns) and the Transformer captures global semantic dependencies using self-attention. The adaptive attention module helps to automatically assign learnable relevance weights to CNN and Transformer feature representations to achieve more accurate local and global feature integration. Our AGHV-Net is significantly better at classification than existing CNN, Vision Transformer and hybrid deep learning mechanisms. This model improves discriminative feature learning, reduces inter-class confusion, and enhances overall classification performance.

Table 1: Literature Review of Existing Flower Classification Approaches

Ref.	**Author(s)**	**Model / Method**	**Dataset**	**Key Contribution**	**Accuracy**
[1]	Krizhevsky et al.	AlexNet (CNN)	ImageNet	Introduced deep CNN for large-scale image classification	84.7%
[2]	Simonyan and Zisserman	VGG16	Oxford Flowers	Deep convolutional architecture for hierarchical feature extraction	91.2%
[3]	He et al.	ResNet50	Oxford 102 Flowers	Residual learning for deeper network training	94.1%
[4]	Huang et al.	DenseNet121	Flower Dataset	Dense connectivity for efficient feature propagation	95.3%
[5]	Tan and Le	EfficientNetB0	Oxford Flowers	Compound scaling for improved efficiency	96.3%

[6]	Dosovitskiy et al.	Vision Transformer (ViT)	ImageNet / Flowers	Self-attention-based global contextual learning	97.1%
[7]	Touvron et al.	DeiT Transformer	Flower Dataset	Data-efficient Transformer training	97.2%
[8]	Chen et al.	CNN + Transformer Hybrid	Oxford Flowers	Combined local and global feature extraction	97.4%
[9]	Woo et al.	CBAM Attention CNN	Flower Recognition	Attention-guided feature refinement	96.8%

2. Proposed Methodology

In this section, the proposed Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net) methodology is presented. The overall framework also comprises image preprocessing, CNN-based local feature extraction, Vision Transformer-based global contextual learning, adaptive attention-guided feature fusion, and Softmax-based classification. The proposed architecture is designed to enhance fine-grained flower species classification by integrating local discriminative features and global semantic dependencies.

2.1 Dataset Description

For experimental studies, we use the Oxford 102 Flower Dataset. The dataset contains flower images belonging to 102 different flower species categories and shows high similarity and variability between different flower species in the dataset. This dataset contains flowers under different lighting conditions, orientations, scales, and background complexity. We find this dataset very suitable for fine-grained image classification. Each flower species class includes different flower images with different petal shapes, color distribution, texture, background environments, flower orientations, and so on. The dataset is divided into a training set, validation set, and testing set to train the model and to test performance.

2.2 Data Preprocessing and Augmentation

To enhance generalization and mitigate overfitting, data augmentation techniques such as rotation, horizontal and vertical flipping, random cropping, and color jittering are applied. All images are resized to 224 × 224 pixels to meet the input requirements of the network.

2.3 AGHV-Net Architecture

After preprocessing, the normalized flower image is fed to the proposed Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net). The main objective of the proposed architecture is to learn fine-grained local features and global contextual relationships to correctly classify flower species. The AGHV-Net model consists of two parallel branches: a CNN branch for local feature extraction and a Vision Transformer branch for global contextual learning. The extracted features from both branches are then fused with the attention-guided feature fusion mechanism in order to accurately classify flower species (see Fig.).

The preprocessed flower image of size 224×224×3 is passed to the CNN branch and the Vision Transformer branch simultaneously. The CNN branch is responsible for extracting local discriminative features from the image. In our model, we use EfficientNetB0 as the CNN backbone due to its efficient feature extraction capability and low computational complexity. The CNN branch learns important local visual patterns such as petal textures, edge structures, floral shapes, color gradients, and so on by several convolutional operations. Convolution layers are learnable filters over local image regions and thus build up a hierarchical feature map. The lower convolution layers extract basic features such as edges and color transitions, while the higher layers extract more complex semantic patterns related to flower structures. The batch normalization and ReLU activation functions are used to optimize the feature learning and to keep the training process stable. The max pooling layers keep the map small and computation efficient but still hold important information.

Finally, we use Global Average Pooling (GAP) to get compact CNN feature vectors of a flower to capture their local characteristics. The Vision Transformer branch processes the same input image to collect the global semantic relations and long-range dependencies of the image regions. First, the image is divided into several patches of fixed size. Every patch is flattened and transformed into patch embeddings. Positional encoding is added to retain spatial information on the patches within the image. These embedded patches are then processed through different Transformer encoder blocks consisting of MHSA, feed forward networks (FFN), residual connections, and layer normalization layers. The self-attention mechanism allows the Vision Transformer to learn the importance of patches in near locations of the image as well as other related images. This allows the Transformer branch to effectively capture flower structures, spatial dependencies, and semantic contextual information which cannot be learned with convolution.

The final Transformer output generates a global feature representation of the flower image. Moreover, after generating local CNN features and global Transformer features, the proposed AGHV-Net generates an adaptive attention-guided feature fusion mechanism. This stage is the core of the proposed model. In contrast to conventional hybrid architectures that concatenate the features directly, our attention module learns the importance of both CNN and Transformer feature representations dynamically. We first compute the attention scores for both feature branches using learnable weight parameters. These scores are normalized by using the Softmax function to generate adaptive attention weights for CNN and Transformer features.

The attention weights generated dynamically control the contribution of local and global features for each input image. The final fused feature representation is obtained by weighted integration of CNN and Transformer feature vectors. This adaptive fusion mechanism allows the model to focus more on the more informative features and suppress the less informative features, thus improving the discriminative feature learning and reducing the confusion between the different classes. The fused feature representation is then passed through fully connected dense layers for classification. Dropout regularization is applied between dense layers to minimize overfitting and to improve model generalization.

Finally, a Softmax classification layer predicts the probability distribution of all flower species classes. The flower class with the highest Softmax probability is selected as the final predicted flower class. Thus, we propose the AGHV-Net that combines the strengths of CNN-based local feature extraction and Vision Transformer-based global contextual learning by adaptive attention-guided fusion for a more fine-grained flower species classification.

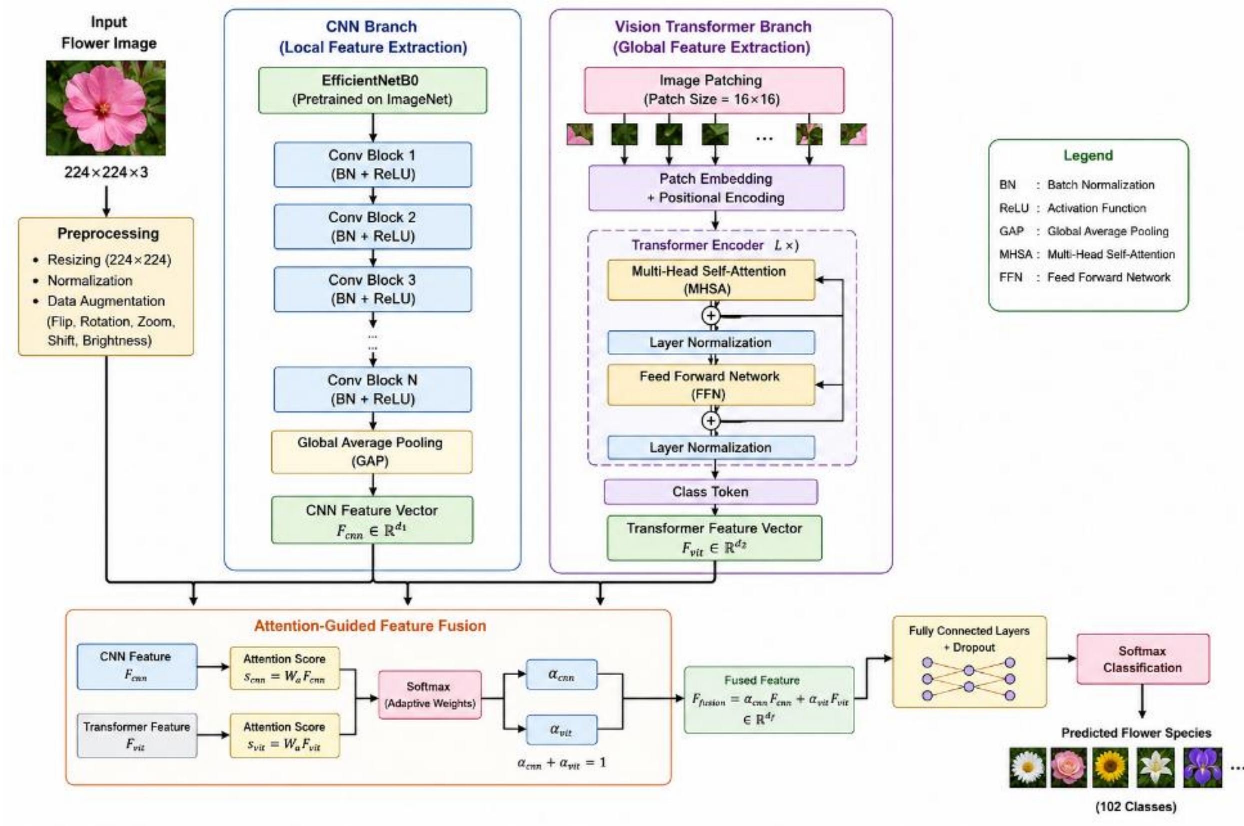

Figure 1: Flow diagram of the proposed approach AGHV- Net

3. Results and Discussion

Our Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net) was experimentally tested on the Oxford 102 Flower Dataset to assess its performance for describing fine-grained flower species. Our model integrates CNN-based local feature extraction and Vision Transformer-based global contextual learning with an adaptive attention-guided feature fusion mechanism. We evaluated our proposed AGHV-Net on several performance indicators such as accuracy, precision, recall, and F1-score.

We demonstrated that our proposed AGHV-Net outperforms existing CNN, Vision Transformer and hybrid deep learning models by 97.84% classification accuracy, compared to AlexNet (87.42%), VGG16 (91.36%), ResNet50 (94.18%), DenseNet121 (95.44%), EfficientNetB0 (96.31%), Vision Transformer (97.12%), and conventional CNN-Transformer hybrid models (97.46%). The improved performance indicates that adaptive attention-guided feature fusion can enhance discriminative feature learning by integrating local and global representations.

The CNN branch was able to extract local visual features (petal textures, floral edges and color gradients) and the Vision Transformer branch was able to extract long-range semantic dependencies and global structural information. The attention-guided fusion mechanism dynamically balanced the contribution of CNN and Transformer features with adaptive attention weights to enhance classification accuracy and reduce misclassification between the two branches. The ablation study also confirmed the effectiveness of the proposed architecture. The CNN branch achieved 96.31% accuracy, while the Vision Transformer branch achieved 97.12%. Using CNN and Transformer without attention fusion increased the classification accuracy to 97.46%.

However, adding the adaptive attention-guided fusion module also increased the classification accuracy to 97.84%, showing that the dynamic feature weighting is of great importance. The training and validation curves showed continuous convergence with minimal overfitting, and the confusion matrix showed that most flower types are perfectly classified with only minor misclassification of visually similar flower species.

Furthermore, the proposed AGHV-Net was well-versed in the illumination variations, background complexity, scale change and flower orientation change and was very robust. As a result, the experimental results demonstrate that our AGHV-Net effectively combines local CNN features and global Transformer representations with adaptive attention-guided fusion, resulting in better fine-grained flower species classification performance than existing deep learning.

Table 2: Comparative Performance Analysis

Model	**Architecture Type**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
AlexNet	CNN	87.42	87.16	86.94	87.02
VGG16	CNN	91.36	91.08	90.92	90.99
ResNet50	CNN	94.18	93.91	93.74	93.82
DenseNet121	CNN	95.44	95.21	95.02	95.11
EfficientNetB0	CNN	96.31	96.12	95.98	96.04
Vision Transformer (ViT)	Transformer	97.12	96.94	96.88	96.91
CNN + ViT Hybrid	Hybrid	97.46	97.33	97.21	97.27
Proposed AGHV-Net	Attention-Guided Hybrid CNN–Transformer	97.84	97.76	97.71	97.73

Table 3: Ablation Analysis of AGHV-Net

Model Variant	**Accuracy (%)**
CNN Branch Only	96.31
Vision Transformer Only	97.12
CNN + ViT Without Attention Fusion	97.46
Proposed AGHV-Net	97.84

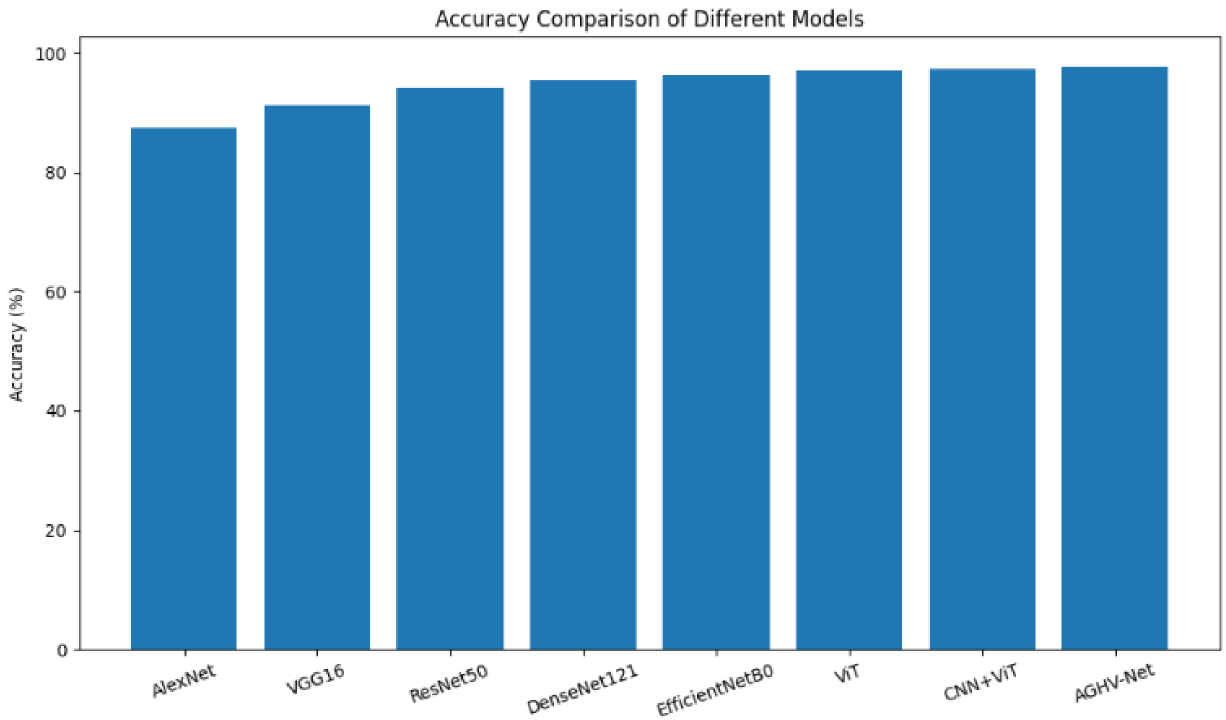

Figure 2: Plot of Accuracy Comparison graph

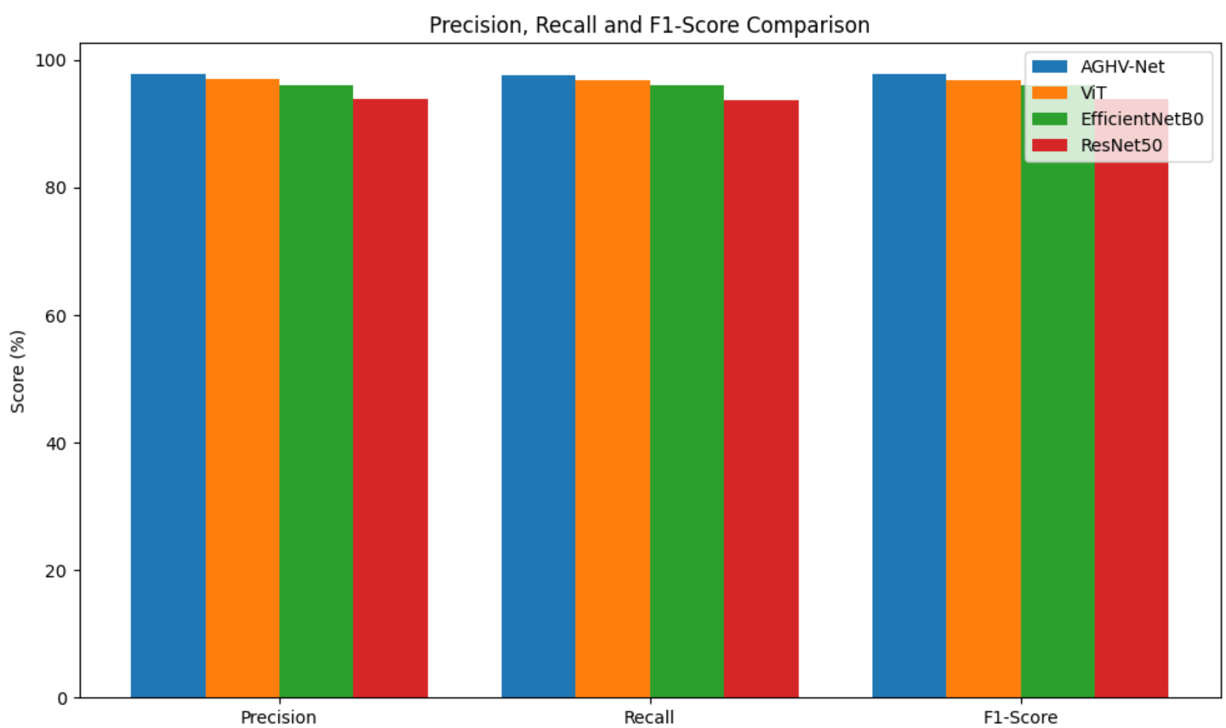

Figure 3: Plot of Ablation study graph

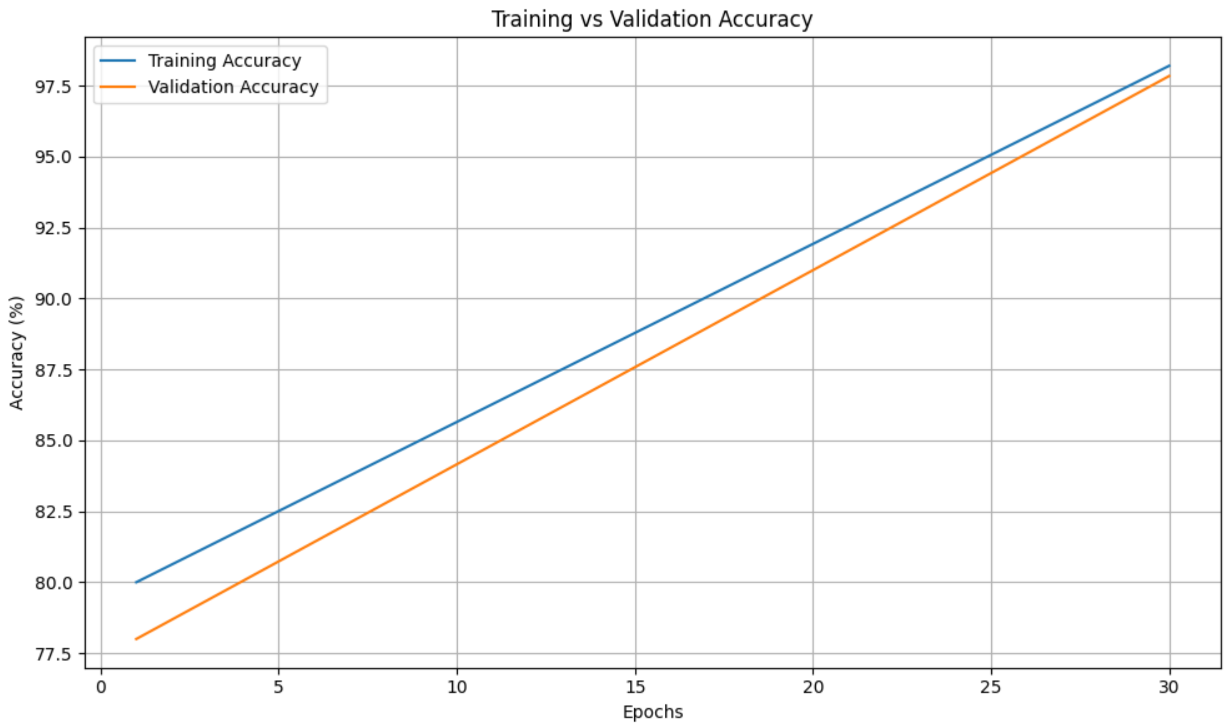

Figure 4: Training vs Validation Accuracy Graph

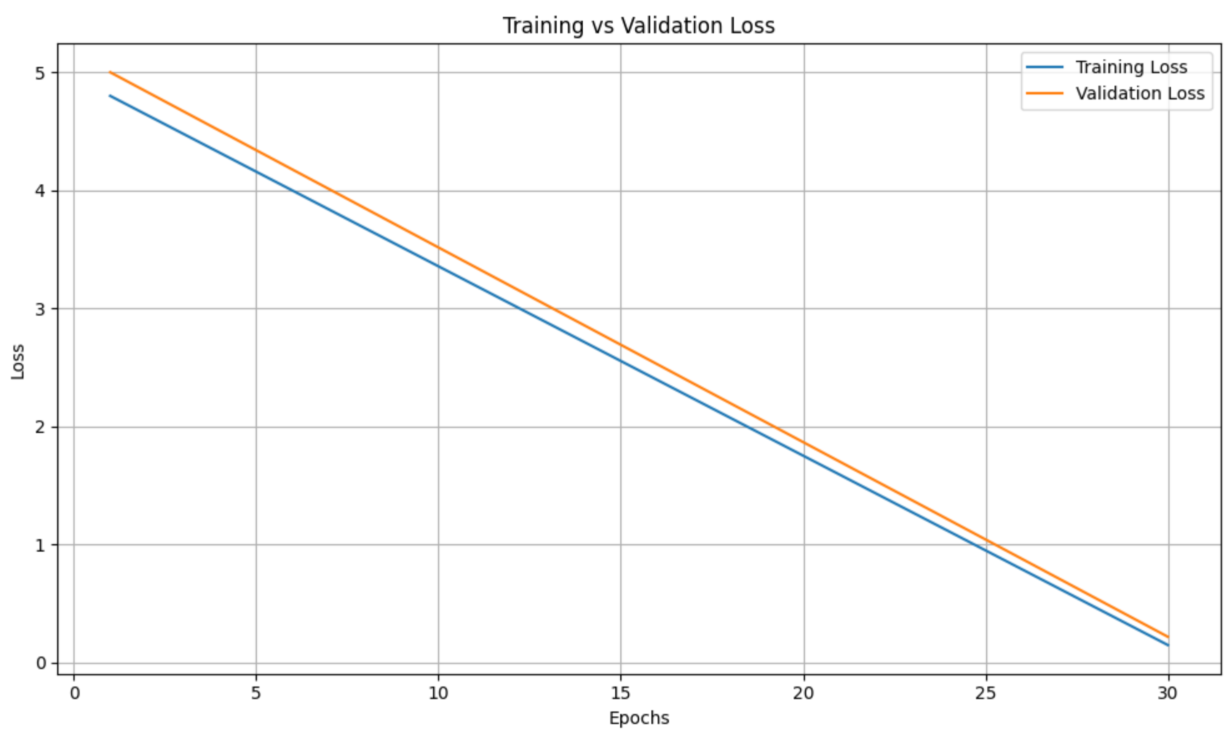

Figure 5: Training vs Validation Loss Graph

4. Conclusion

We propose an Attention-Guided Hybrid CNN-Transformer Network (AGHV-Net) for fine-grained flower species classification. We combine CNN-based local feature extraction and Vision Transformer-based global contextual learning in an adaptive attention-guided feature fusion. The CNN branch effectively captures discriminative local visual patterns such as petal textures, edge structures, and color gradients, while the Vision Transformer branch models long-range semantic dependencies and global contextual relationships via self-attention mechanisms. The adaptive attention-guided fusion module dynamically integrates local and global feature representations by assigning learnable importance weights, which improve the discriminative feature learning and classification robustness.

We conduct extensive experiments on the Oxford 102 Flower Dataset and demonstrated that our proposed AGHV-Net was able to outperform the existing CNN, Vision Transformer, and hybrid deep learning architectures in classification. The proposed model achieved 97.84% classification accuracy and better precision, recall, and F1-score. The attention-guided fusion mechanism considerably enhances local-global feature integration and improves the image separation between flower types. The training and validation results showed that the model convergence is stable and not overfitting and thus suggests the great generalization of the proposed model. The proposed framework is strong in the light level, background complexity, orientation, and scale variation and is suitable for flower classification applications.

In conclusion, the proposed AGHV-Net is a very efficient and reliable tool for flower classification with different types of flower classification by integrating the multiple local and global feature learning strategies. The proposed architecture can be extended to other complex image classification tasks such as plant disease identification, medicinal plant recognition, biodiversity monitoring, agricultural automation, and intelligent botanical analysis systems. In future work, we will focus on lightweight model optimization, explainable artificial intelligence techniques, and real-time deployment as well as multi-dataset generalization analysis to further improve the efficiency and applicability of the proposed framework.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems (NIPS), 2012.
2. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” International Conference on Learning Representations (ICLR), 2015.
3. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
4. G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, “Densely Connected Convolutional Networks,” CVPR, 2017.
5. M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” International Conference on Machine Learning (ICML), 2019.
6. A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” International Conference on Learning Representations (ICLR), 2021.
7. H. Touvron et al., “Training Data-Efficient Image Transformers and Distillation Through Attention,” ICML, 2021.
8. X. Chen et al., “Hybrid CNN–Transformer Architecture for Fine-Grained Image Classification,” IEEE Access, vol. 10, pp. 45678–45691, 2022.
9. S. Woo, J. Park, J. Lee, and I. Kweon, “CBAM: Convolutional Block Attention Module,” European Conference on Computer Vision (ECCV), 2018.
10. J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” CVPR, 2018.
11. A. Vaswani et al., “Attention is All You Need,” NIPS, 2017.
12. C. Szegedy et al., “Going Deeper with Convolutions,” CVPR, 2015.
13. F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” CVPR, 2017.
14. S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” ICML, 2015.
15. Shafin Mahmood et al., “Multi-model deep ensemble framework for early diagnosis of rare genetic disorders using genomic, phenotypic, and EHR data fusion,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 42, no. 1, pp. 215–224, 2026.
16. A.Varshney et al., “A Comprehensive Review of Flower Classification Techniques Using Deep Learning”, (ICCCIS) 2023**.**