
A Hybrid CNN-Transformer Architecture with Bidirectional Cross-Attention Fusion for Efficient Flower Recognition**Hersh M. Hama^{*1}, Shayan I Jalal², Saman M. Omer³, Mohammed H. Ahmed⁴**hersh.hama@uor.edu.krd¹, shayan.ihsan@uor.edu.krd², saman.muhammad@uor.edu.krd³,mohammedhussein@uor.edu.krd⁴^{1,2,3,4}Department of Software Engineering, College of Engineering University of Raparin, Ranya, Iraq

Article Information

Received : 19 Jun 2026

Revised : 26 Jun 2026

Accepted : 30 Jun 2026

KeywordsFine-grained flower classification,
Bidirectional cross-attention, Hybrid CNN Model, Floral Image Dataset, Visual Similarity in Flowers

Abstract

Automated flower recognition is crucial to the fields of agriculture and biodiversity monitoring, but deep learning models are extremely high in computing demand for resource-constrained devices. In this paper, a compact and efficient hybrid model based on EfficientNet-B4 and ViT-Small/16 is introduced. The design uses a bidirectional cross-attention fusion mechanism that uses both local edge-level features and global context to learn highly discriminative representations for fine-grained classification. The model was tested using 3,500 images from 35 species of flowers from a curated, augmented database, using 5-fold cross-validation. It outperformed the state-of-the-art architectures such as Swin Transformer and ResNet50, with 98.71% accuracy and a 98.80% F1 score with high statistical stability (95% CI:98.34%–99.09%) and with faster convergence rate. Statistically, these improvements were confirmed by pair-wise and Wilcoxon signed-rank statistical tests, which indicate the model's potential in resource-efficient real-world automated plant identification.

A. Introduction

Flowers are essential for human life as they are used in food chain, aromatics, cosmetics, ecological balance and decoration and also provide habitat for many insect pollinators. Many flowers also have medicinal importance [1] as they are abundant sources of bioactive compounds which play a significant role in human metabolism [2]. Some flowers are poisonous, however, proper identification is necessary to assure their safe use in food and medicine.

Beyond its practical uses, an in-depth knowledge of the species of flowers is essential for the conservation of biodiversity. Flower classification has become more important as the number of species with diverse types of flowers grows and with the increasing importance of horticulture in the society due to the improvement in the standard of living. The traditional method of floral species identification is done visually, but because of the similarity in floral shape and color among species, this method is often inefficient and inaccurate. Furthermore, manual identification is time-consuming and is still difficult even for professional botanists.

Most of the traditional machine learning methods used to classify flowers need to represent the raw images in a particular format before the handcrafted features like color, shape and texture can be extracted from the images [3]. The methods rely heavily on input image properties, making them sensitive to the properties of the input image sets and sensitive to the generalizability of the methods across data sets.

In contrast, deep learning approaches, especially Convolutional Neural Networks (CNNs), have great potential to learn the hierarchical representations of visual patterns automatically, such as the texture and edge-level features [4]. In recent years, Vision Transformers (ViTs) have also become a promising paradigm that uses global multi-head self-attention mechanisms to represent long-range contextual dependence over an entire image.

These advances have not solved all the problems encountered in fine-grained flower classification in the real world, however, as in some cases the distinct species are so similar that they share virtually identical color tones and layered petal structures, referred to as extreme inter-class similarity, and in others, there is variation in the flowers of the same species. Purely convolutional architectures tend to have a narrow receptive field, not being able to capture the global spatial relationships, while standalone Vision Transformers, which are difficult to train without massive datasets, lack local inductive biases. In addition, most of the publicly available flower data base is geographically restricted, and very few of them include the vast floral diversity found in certain ecological regions such as the Kurdistan Region of Iraq, where there are lots of flowers with unique cultural, ecological and medicinal values.

To overcome these limitations, this paper introduces a new, symmetrical dual-stream hybrid architecture that combines the advantages of CNNs and Vision Transformers. The framework consists of a pre-trained EfficientNet-B4 backbone for extracting the texture and color gradient at localized locations, and a ViT-Small stream for extracting global structural patterns. Instead of the naive feature concatenation, we propose a Bidirectional Cross-Attention Fusion mechanism that allows structured and continuous information to flow between the two streams,

enabling each stream to selectively attend to and enhance the other's representations.

This work makes the following major contributions:

- **Novel Hybrid Dual-Stream Architecture:** We propose a symmetric architecture which well connects the local inductive bias and the global contextual modelling, namely, combining the EfficientNet-B4 and ViT-Small.
- **Bidirectional Cross-Attention Fusion:** We propose a sophisticated multi-head cross-attention module which enables parallel, symmetric feature query between the convolutional stream and the transformer stream, ensuring the best feature interaction before classification.
- **Regional Floral Data Set Construction:** We introduce a carefully curated and class balanced data set of 3500 images of 35 species dedicated to Kurdistan Region of Iraq flora for regional ecological monitoring and smart agriculture.
- **Superior Performance and Convergence:** Proposed model achieves state of the art mean accuracy of 98.71% which is better than Swin Transformer, ResNet50, InceptionV3, DenseNet201, MobileNetV2, with significantly lesser training convergence time (average of just 12.8 epochs).

Related works in flower classification using machine learning and deep learning methods are reviewed in Section 2. Section 3 outlines the procedure for constructing the dataset, the data augmentation pipeline, and the mathematical implementation of the proposed hybrid dual-stream architecture. Section 3 explains the procedure followed in creating the dataset, data augmentation pipeline, and the mathematical implementation of the proposed hybrid dual-stream architecture. The experimental setup, evaluation metrics, statistical verification and empirical results are presented in Section 4. In Section 5, a detailed discussion of the results is provided and the methods are compared with current state-of-the-art approaches. Finally, Section 6 summarizes the paper and outlines directions for future research.

B. Literature Review

The classification of flowers has become an important research field in computer vision for several applications, such as, biodiversification conservation, agricultural use, and plant phenotyping. It is an interesting task due to the delicate differences in colour, texture and form between different flower species. Initial strategies were based on manually designed features and on traditional classifiers like Support Vector Machines (SVM) and Decision Trees. These strategies successfully had some degree of success, but they struggled to generalize to complex data sets.

Recent improvements in deep learning techniques have made the CNN the most popular method for flower classification. CNN-based models like InceptionV3, DenseNet121, VGG16 and MobileNetV2 have proven to be effective because they are able to learn local spatial information like edges, textures and color distributions. For example, DenseNet121 gives an accuracy of 98.6% on Oxford 102 Flower Dataset [8] and high results have been achieved by [10] using VGG16 with transfer learning. Likewise, [11] evaluated several architectures and showed competitiveness based on the optimizer. The studies validate the suitability of CNNs for learning discriminative local features in fine-grained classification tasks.

There are also a few works that have expanded CNN-based methods with hybridization and optimization. In [5] InceptionV3 was fused with SVM and evolutionary algorithms for feature selection and the classification accuracy was improved. Furthermore, [6] introduced an optimized DL architecture for improvement in efficiency and performance. Multimodal data is also explored with other studies combining visual and textual information [9] or combining RGB and depth images [13] which showed that adding complementary information sources greatly improves classification accuracy.

Although these advances have been made, CNN-based models have their inherent limitations of modelling long-range dependency and global relationships in images. To overcome this limitation, a new paradigm of network has been recently proposed: Vision Transformers (ViTs). ViT models are different from CNNs in their use of self-attention mechanisms that capture global relationships among image patches, allowing them to better capture the structure of the image. This makes them very well suitable for fine-grained classification problems that require a strong global context. On the other hand, ViTs tend to need huge numbers of images and are not effective at discerning local details as well as CNNs.

Recent studies have been directed to the hybrid CNN-Transformer architectures to take advantage of both methods. These models are a fusion of the local feature extraction power of CNNs and global context modelling power of Transformers. These hybrid methods have been found to be effective in multiple image classification scenarios, as they offer a combination of local and global features. There are, however, many current hybrid approaches that use basic feature fusion techniques, such as concatenation or averaging, which may not be effective at fully utilizing the complementary information provided by the two feature types.

In order to address this, attention-based fusion mechanisms, such as cross-attention, have been investigated. Cross-attention allows for interaction between different feature representations since one feature space can pay attention to another feature space. This makes for more efficient fusion of features than conventional fusion. Bidirectional cross-attention enables the CNN and Transformer features to share information from each other, which helps the model to learn both local and global features well.

Besides architectural improvement, different data sets have been employed to test flower classification systems. The Oxford 102 Flower Dataset has been one of the most predominantly used datasets, for instance, [7] attained high detection and classification performance with NAS-FPN and Faster R-CNN. The deep learning models can also be applied in different environments, as indicated by other datasets such as the Kaggle flower dataset [10] and region-specific datasets like the Kurdistan flower dataset [14]. Furthermore, these methods are not limited to just flower detection; they have been adapted for other special applications like the detection of chili flowers [15] and tracking cotton flowering [16], which further demonstrates their wide applicability in agricultural automation and plant monitoring.

In general, the literature shows a clear evolution from the basic machine learning techniques to complex deep learning architectures. CNN-based models have achieved great performance in learning local features, while Transformers

have shown good performance in learning global representations, but current methods are not able to integrate both well. Thus, it is necessary to develop models that are able to integrate these complementary strengths using more complex fusion mechanisms. Given these challenges, a hybrid architecture with a bidirectional cross-attention module between CNN features extracted by EfficientNet and the representations from Vision Transformer is proposed to facilitate the interaction of features and enhance the flower classification accuracy.

C. Methods and Materials

C.1. Dataset Preparation

To facilitate flower recognition research with a direct focus on the flora of the Kurdistan Region of Iraq, a specialized dataset was constructed. This comprehensive collection totals 3,500 images, spanning 35 distinct flower classes, all of which are found in the region's natural environment.

The dataset was compiled by combining samples from multiple sources. The process began with the creation of a Custom Flower Image Dataset (CFID) [14], which comprised 23 classes of native and regionally significant flowers. To ensure a clean and consistent starting point for machine learning, the CFID contained 50 high-resolution images per class, all photographed from above with the flower isolated against a plain white background.

To expand this initial collection and standardize each class to 100 images, the dataset was supplemented with images from the Oxford 17 [17] and Oxford 102 datasets [18]. However, a key challenge with these external sources was the variability in image composition, which often included whole plants, complex backgrounds, or multiple flowers. To maintain data quality, a meticulous cleaning process was undertaken to remove these unsuitable images.

When a class still had insufficient samples, additional images were collected via Google Search. This method was also used to introduce several entirely new classes, ensuring the final dataset covered all 35 flower species found in the region.

The final dataset, with its 100 images per class, effectively mitigates class imbalance issues. It includes a variety of compositions, from isolated flowers to full plants, captured under different lighting and angles. This curated collection is specifically designed to support robust deep learning models for flower recognition, with applications in smart farming, biodiversity monitoring, and regional ecological studies within the Kurdistan Region of Iraq.

The final dataset was meticulously compiled from several sources to represent the flora of the Kurdistan Region of Iraq. For a detailed breakdown of the number of classes and images contributed by each source, see Table 1. This compilation resulted in a total of 3,500 images spanning 35 classes, with each class standardized to 100 images to mitigate class imbalance.

Figure 1 Several flower classes in the dataset exhibit strong visual similarity, which likely contributes to misclassification during image recognition. For instance, Rose, Garden rose, Camella jap, and Shrub rose all share similar red or pink tones and layered petal structures, making them difficult to distinguish based solely on shape and color. Likewise, Marigold, Chrysanthemum, and Dandelion feature dense, rounded yellow blossoms that can confuse models relying primarily on texture and hue. Additionally, Homestead purple, Zrizana, and Osteospermum

possess overlapping symmetrical purple petals, increasing the likelihood of cross class predictions. These visual similarities are reflected in the confusion matrix and highlight the need for fine grained feature extraction in floral classification tasks.

The dataset integrates classes and images from the Oxford 17 and Oxford 102 Flower Datasets, a Custom Flower Image Dataset (CFID) representing regionally significant species, and additional images collected via Google Search. The final balanced dataset contains 3,500 images distributed evenly across 35 classes to support robust model training and evaluation.

Table 1. Composition of the combined flower image dataset used in this study. The dataset integrates classes and images from the Oxford 17 and Oxford 102 Flower Datasets, a Custom Flower Image Dataset (CFID) representing regionally significant species, and additional images collected via Google Search. The final balanced dataset contains 3,500 images distributed evenly across 35 classes to support robust model training and evaluation.

Dataset name	Number of classes to from	Number of images
Oxford 17	4	288
Oxford 102	3	992
Custom Flower Image Dataset (CFID)	23	1150
Google	5	1070
Total	35 class	3500 image (100 image per class)



Figure 1. Sample images of the 35 flower classes used in the dataset, illustrating the diversity in color, shape, and petal structure. Each flower is labeled with its common name for reference

C.2. Data Augmentation Strategy

In order to diversify the representative power of the proposed hybrid model that consists of EfficientNet-B4 and ViT/small architecture and help mitigate the risk of overfitting on the dataset, a comprehensive multi-stage data augmentation pipeline was adopted.

All training images were scaled down to 256×256 pixels and then randomly cropped to the desired resolution of 224×224 . Random horizontal flipping ($p = 0.5$), random vertical flipping ($p = 0.2$), random rotation within the range of -20 to $+20$ and random affine transformation that consisted of translation of up to ± 5 percent of image dimension and isotropic scaling of image dimension within the range of $[0.9, 1.1]$. Photometric perturbations included colour jitter of brightness and contrast of 0.4, saturation of 0.3 and hue of 0.1. Random erasing [19] followed the process of normalisation of tensors with an occurrence probability of 0.25 and a ratio of erased area sampled in $[0.02, 0.15]$. All images were normalised using ImageNet channel statistics (mean = $[0.485, 0.456, 0.406]$; standard deviation = $[0.229, 0.224, 0.225]$). The validation images were downsampled to 224×224 directly without stochastic modification.

To further enhance smoothness of decision boundaries and reduce overconfident predictions, stochastic sample-pair augmentation was further applied at the batch level as it has been previously done. At the start of each training epoch, one of three augmentation modes was picked randomly: Mixup [20] with probability 0.35, CutMix [21] with probability 0.35, or no augmentation with probability 0.30. Under Mixup, two training images are mixed by a convex linear interpolation using mixing coefficient λ sampled under a Beta distribution ($\alpha = 0.4$), resulting in a soft composite label that is proportional to $1/\lambda$. In CutMix, a rectangular block of the image is replaced with the corresponding block on the second image ($\alpha = 1.0$), with the label mixing ratio adjusted to reflect the actual pixel-level proportion. The training loss is calculated in both situations as a convex combination of the cross-entropy losses with respect to the two source labels, weighted by λ .

C.3. Proposed Architecture

The model proposed implements a dual-stream design that simultaneously uses convolutional and self-attention-based representations. Convolutional neural networks are good at capturing local texture, edge-level features, through spatial filtering at a hierarchical level [22], whereas vision transformers (ViTs) leverage global multi-head self-attention to model long-range contextual dependencies across the whole image [23]. One of the fundamental weaknesses of purely convolutional architectures is that, when not pretrained, they inherently favor locality, meaning that they can only learn useful spatial inductive biases when very deep stacks of layers are used; conversely, ViTs require large amounts of training data to learn useful spatial inductive biases when trained from scratch [24]. The proposed hybrid architecture mitigates both limitations and integrates the complementary advantages of the two paradigms using a cross-attention fusion mechanism as shown in Figure 2.

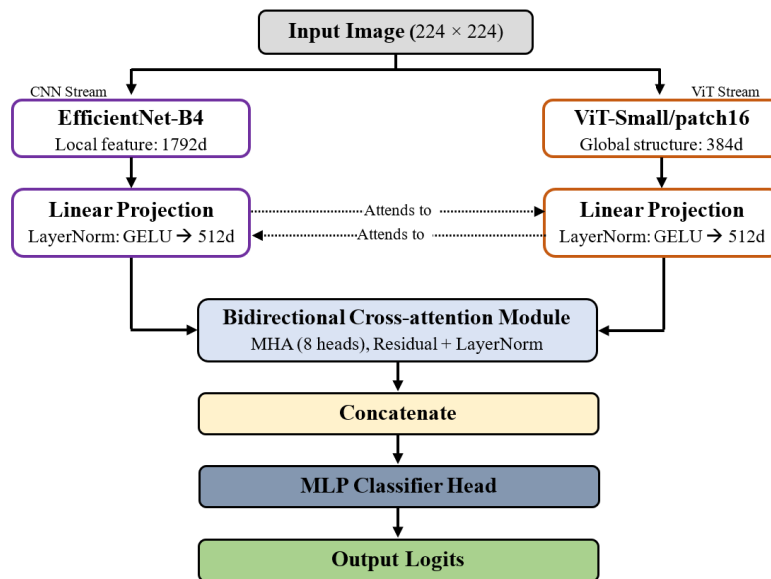


Figure 2. Schematic architecture of the proposed hybrid dual-stream model featuring bidirectional cross-attention fusion between EfficientNet-B4 and ViT-Small

C.3.1. Convolutional Stream: EfficientNet-B4

The convolutional backbone used was EfficientNet-B4 [25] that was trained on ImageNet-1K. EfficientNet uses a principled compound scaling strategy, which uniformly scales network depth, width and input resolution based on a fixed set of coefficients obtained via neural architecture search, and resulting in better accuracy-to-parameter-count trade-offs than do architectures scale along a single dimension [25]. The B4 variant was chosen because it provides a good tradeoff between representational and computational cost at the 224×224 input resolution. The average global pooled output of EfficientNet-B4 results in a 1,792-dimensional feature vector representing information about spatially localised texture and colour gradient information.

C.3.2 Transformer Stream: ViT-Small/16

The second stream used ViT-Small patch size 16 (ViT-S/16) [24], also pre-trained on ImageNet-1K. The input image is tiled with a series of non-overlapping 16×16 pixel patches linearly projected to a 384-dimensional embedding. A learnable classification token is added to the patch sequence and processed together through twelve layers of transformer encoder through multi-head self-attention and feed-forward sub-networks. The extracted final classification token representation, which is obtained through global pooling of the classification and token representations, gives a 384-dimensional global context vector that encodes structural and relational information in spatially distant regions of images. ViT-Small was selected instead of the larger version ViT-Base because the risk of overfitting is high due to the moderate size of the training set.

C.3.3 Projection to Shared Embedding Space

To allow cross-modal interaction, both stream outputs were independently projected to a common embedding space of dimension 512 through learned linear

transformations and Layer normalisation [26] and GELU activation [27]. This projection guarantees that both representations are in a common metric space of equal dimensionality, a condition necessary to meaningfully compute cross-stream attention between the two representations. The projection weights are learnable in full, and are trained at the full base learning rate together with the classification head.

C.3.4. Bidirectional Cross-Attention Fusion

To allow the exchange of structured information between the two streams before fusion, a bidirectional cross-attention module was proposed. Instead of combining representations by simply concatenating or adding the elements of the two modalities, which treats the two modalities as totally independent, cross-attention allows each stream to selectively attend the other, weighting its contribution based on learned query-key compatibility scores. Such a design is inspired by the fact that local convolutional features and global transformer features carry complementary information that can not be summed out by either stream; explicit cross-modal attention provides a principled process of selective feature sharing before fusion. Theoretically, the module is based on scaled dot-product attention [28]. Assuming a query matrix Q and a key matrix K , as well as a value matrix V , the attention weighted output is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V \quad (1)$$

where d_k is the per-head key dimensionality and the factor $1/\sqrt{d_k}$ is to stabilise the gradients by ensuring that dot products do not become large in magnitude [29]. In the suggested implementation, the shared embedding dimension is $d = 512$ and the number of attention heads is $h = 8$, yielding $d_h = d / h = 64$ dimensions per head. A dropout of attention of 0.1 is implemented in each of the heads at any given time during training in order to regularise the learned attention weights. The number of heads $h = 8$ follows the standard multi-head configuration established in the literature [28]) that ensures adequate distribution of attention across the feature space without any unnecessary computational overhead.

Two different multi-head attention (MHA) processes are calculated simultaneously. In the former the query is the projected EfficientNet-B4 representation $e_e \in \mathbb{R}^{512}$ as key and value:

$$\hat{e}_e = \text{MHA}(Q = e_e, K = e_v, V = e_v) \quad (2)$$

This step adds local convolutional details of EfficientNet-B4 with global structural context supplied by the ViT stream. The roles are swapped in the second operation: e_v serves as the query, taking care of e_e as the key and value:

$$\hat{e}_v = \text{MHA}(Q = e_v, K = e_e, V = e_e) \quad (3)$$

This latter operation provides a concrete local texture evidence of the convolutional stream to provide the global representation of the transformer. Each

attended output undergoes residual connections and Layer Normalisation, resulting in the final cross-attended representations:

$$\begin{aligned}\tilde{E}_e &= \text{LayerNorm}(\hat{e}_e + e_e) \\ \tilde{E}_v &= \text{LayerNorm}(\hat{e}_v + e_v)\end{aligned}\tag{4}$$

The residual connections maintain the original stream representations and avoid the degradation of gradient during the back-propagation, and Layer Normalisation stabilizes the distribution of activations along the feature dimension. Resulting attended representations $\tilde{E}_e, \tilde{E}_v \in \mathbb{R}^{512}$ now contain enriched cross-modal information on the verge of fusion. Its two-way, symmetric design is important because it means that both streams equally benefit as a result of cross-modal context [30] which is not guaranteed by unidirectional cross-attention schemes in which only one stream is allowed to query the other [31].

C.3.5. Classification Head

The fused features are concatenated into a 1024-dimensional vector and fed into a multi-layer perceptron (MLP) head. This head is made up of two fully connected layers with GELU activation as well as a Dropout rate of 0.4 to further reduce overfitting. The last layer uses a Softmax to give the probability distribution over the 35 flower classes.

C.4. Training Strategy and Experimental Setup

We adopted a stratified 5-fold cross-validation scheme to ensure a solid assessment of the suggested hybrid model as well as to limit bias. The dataset was divided into five different subsets, in which each fold retained the original distribution of classes. In each iteration, four folds were used to train and the rest was used as the validation set. This was done five times and the final performance measurements were computed as the mean and standard deviation of the performance in all the folds.

C.4.1. Optimization and Learning Rate Schedule

The AdamW optimizer with a learning rate of 2×10^{-4} and a batch size of 16 was used to optimize the model. The training epochs were made stable with a linear warmup phase in the first 10% of the training epochs (warmup ratio = 0.1), followed by a cosine annealing schedule to decay the learning rate. This means that the model will converge smoothly with the learning rate slowly increasing at the beginning and decreasing as the model approaches an optimal solution.

C.4.2. Regularization and Convergence

Since it utilizes pretrained backbones (ImageNet-1k) the maximum number of epochs was set to a limit of 30 epochs because in a transfer learning setup the number of epochs needed to reach convergence is usually significantly fewer than 30 epochs. To further address overfitting and maximize the use of computational resources, an early stopping mechanism was combined with a patience of 6 epochs; training was stopped in case the validation accuracy did not improve in 6 consecutive cycles. This was to make sure that the final weights were the highest

generalization state of the model before some possible performance degradation. All the experiments were carried out using a high-performance desktop workstation with the Windows 11 Pro operating system. The hardware setup consisted of an Intel Core i9-13900KF processor, 64 GB of RAM, and an NVIDIA RTX 4080 processor with 16 GB of dedicated video memory. The suggested hybrid architecture was developed in Python 3.10 on an environment made of Anaconda3. To speed up the process of training and inference, support of CUDA was used and thus efficient parallel processing occurred on the GPU.

D. Experimental Results

D.1. Performance of the Proposed Hybrid Model

A 5-fold cross-validation was performed on the proposed hybrid architecture on 35 flower classes. As showed in Table 2, the model achieved accuracy values of 99.00%, 99.00%, 98.29%, 98.57%, and 98.71% for folds 1 through 5 respectively, yielding a mean accuracy of 98.71%. Precision, recall and F1 score were both high, at 99% in four of the five folds and 98% in fold 3. The results show strong generalization performance for each of the data partitions without any significant differences between them.

The early stopping mechanism stopped training at epochs 12, 11, 18, 12, and 11 in the five folds with a mean of 12.8, which is well below 30 epochs implying rapid and stable convergence.

Table 2. The classification performance of the proposed hybrid model and its individual streams per fold.

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Hybrid (EfficientNet-B4 + ViT-Small)	Accuracy	99.00%	99.00%	98.29%	98.57%	98.71%	98.71%
	Precision	99%	99%	98%	99%	99%	98.80%
	Recall	99%	99%	98%	99%	99%	98.80%
	F1-Score	99%	99%	98%	99%	99%	98.80%
ViT-Small (standalone)	Accuracy	99.00%	99.00%	98.29%	97.29%	97.71%	98.26%
	Precision	99%	99%	98%	97%	98%	98.20%
	Recall	99%	99%	98%	97%	98%	98.20%
	F1-Score	99%	99%	98%	97%	98%	98.20%
EfficientNet-B4 (standalone)	Accuracy	94.00%	94.00%	94.14%	94.43%	93.67%	94.05%
	Precision	94%	94%	94%	95%	94%	94.20%
	Recall	94%	94%	94%	94%	94%	94.00%
	F1-Score	94%	94%	94%	94%	93%	93.80%

The proposed Hybrid EfficientNet-B4 + ViT-Small model also achieved the best mean accuracy of 98.71%, which is higher than the mean accuracy of the ViT-Small model (98.26%) and the EfficientNet-B4 model (94.05%). This indicates that the improvement is not solely due to the strong backbone, but also to the combination of CNN and transformer features.

D.2. Ablation Study: Individual Stream Performance

The training and evaluation protocol is the same as in training and evaluating both the ViT-Small and EfficientNet-B4. The standalone ViT-Small, demonstrated a mean accuracy of 98.26%, with good performance in folds 1-3 and a slight drop in folds 4 and 5 (97.29% and 97.71% respectively). Standalone EfficientNet-B4

achieved a mean accuracy of 94.05% with F1-scores remaining stable at 93-94% for all folds. EfficientNet-B4 also needed significantly more training time, with three out of five folds (mean stopping epoch: 25.0) stopping at the 30-epoch target.

D.3. Comparison with State-of-the-Art Architectures

The accuracy of the proposed model and five baseline architectures are shown in Table 3 when they are tested on the same protocol. Among the baselines, Swin Transformer achieved the highest mean accuracy (97.17%), followed by ResNet50 (96.63%), InceptionV3 (96.60%), DenseNet201 (96.11%), and MobileNetV2 (94.00%). The proposed hybrid model achieved the highest mean accuracy (98.71%) and F1-score (98.80%) on all folds compared to all the baseline models.

Table 3. Per-fold accuracy and mean performance comparison against state-of-the-art baseline architectures.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	average Acc.	average F1
Hybrid (proposed)	99.00%	99.00%	98.29%	98.57%	98.71%	98.71%	98.80%
Swin Transformer	97.86%	96.71%	97.57%	97.14%	96.57%	97.17%	97.40%
InceptionV3	96.71%	96.57%	96.14%	96.00%	97.57%	96.60%	96.80%
ResNet50	96.43%	97.14%	96.71%	96.14%	96.71%	96.63%	96.60%
DenseNet201	95.71%	95.57%	96.29%	96.86%	96.14%	96.11%	96.40%
MobileNetV2	94.14%	93.57%	94.43%	94.57%	93.29%	94.00%	93.80%

D.4. Statistical Comparison of Model Performance

A new fold-level statistical comparison was performed with the five cv accuracy values obtained for the proposed model and another model to further examine the statistical significance of the performance improvement obtained by the proposed hybrid model. The same five data folds were employed in all models, thus paired statistical tests were appropriate. The scores obtained with the fold-wise accuracy were used for comparison of the proposed hybrid model with each baseline model with the help of paired t-test. Moreover, a non-parametric test, Wilcoxon signed-rank test, was also added to verify the results when the paired sample size is small. The hybrid model proposed was tested on five folds with a mean accuracy of 98.71% and standard deviation of 0.30. The 95% confidence interval was 98.34% – 99.09%, which shows a low variation between the folds and stable performance.

Table 4. Fold-level statistical comparison between the proposed hybrid model and baseline architectures using paired t-test and Wilcoxon signed-rank test.

Comparison	Average Accuracy of Baseline (%)	Average Accuracy Difference (%)	Paired t-test p-value	Wilcoxon p-value	Interpretation
Hybrid vs ViT-Small	98.26	+0.46	0.1821	0.5000	Positive but not statistically significant
Hybrid vs EfficientNet-B4	94.05	+4.67	0.00003	0.0625	Significant by t-test; Wilcoxon borderline
Hybrid vs Swin Transformer	97.17	+1.54	0.0065	0.0625	Significant by t-test; Wilcoxon borderline

Comparison	Average Accuracy of Baseline (%)	Average Accuracy Difference (%)	Paired t-test p-value	Wilcoxon p-value	Interpretation
Hybrid vs InceptionV3	96.60	+2.12	0.0011	0.0625	Significant by t-test; Wilcoxon borderline
Hybrid vs ResNet50	96.63	+2.09	0.0003	0.0625	Significant by t-test; Wilcoxon borderline
Hybrid vs DenseNet201	96.11	+2.60	0.0016	0.0625	Significant by t-test; Wilcoxon borderline
Hybrid vs MobileNetV2	94.00	+4.71	0.0002	0.0625	Significant by t-test; Wilcoxon borderline

Based on the paired t-test results, the proposed hybrid model showed statistically significant improvements over EfficientNet-B4, Swin Transformer, InceptionV3, ResNet50, DenseNet201, and MobileNetV2 with p values <0.01. This reinforces the effectiveness of the proposed bidirectional cross-attention fusion strategy over the CNN-based and transformer-based baselines.

The improvement compared to the standalone ViT-Small was not significant at the 0.05 level, however. The difference between the mean accuracy of the proposed model and ViT-Small was very small, around 0.46 percentage points. Thus, this finding is considered a small, but reliable, improvement in performance, not a significant performance gain.

Borderline p-values were obtained for most comparisons using the Wilcoxon signed-ranked test. This is primarily because of the limited number of paired observations, since a five-fold cross-validation only yields a set of five paired accuracy values for each comparison. The Wilcoxon results thus are less definitive, but are a conservative complementary analysis. Overall the statistical results indicate that the proposed hybrid model is able to demonstrate reliable performance improvements over most of the baseline models, and the improvements over ViT-Small need to be further validated using larger datasets, cross validation or per image prediction level statistical analysis.

D.5. Convergence Efficiency

The early stopping epoch for each model in all 5 folds is reported in Table 5. The proposed hybrid model converged the earliest with a mean stopping epoch of 12.8, whereas, the MobileNetV2 consistently achieved a maximum of 30 epochs (mean: 29.6). The mean stopping epochs for ResNet50, inceptionV3 and DenseNet201 were greater than 24. The hybrid model outperforms the purely convolutional baselines while ViT-Small performs at a mean of 17.6, later than the hybrid model.

Table 5. Early stopping epochs across five cross-validation folds for all evaluated architectures.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Hybrid (proposed)	12	11	18	12	11	12.8
ViT-Small	12	17	22	13	24	17.6
EfficientNet-B4	18	30	28	29	20	25.0

Swin Transformer	24	20	24	24	27	23.8
InceptionV3	23	30	25	21	30	25.8
ResNet50	21	30	25	15	30	24.2
DenseNet201	19	25	27	30	20	24.2
MobileNetV2	30	30	30	28	30	29.6

D.6. Per-Class Accuracy

The accuracy of the proposed hybrid model is presented in per class accuracy averaged over all 5 folds in figure 3. The model is able to recognize the majority of the 35 classes with a mean accuracy of 0.987 per class, with perfect or near perfect recognition of the majority of the classes. There are a few classes that are under the mean accuracy: CHRYSANTHEMUM (about 0.90), Luminosa_zinnia (about 0.94), garden_rose (about 0.94). These are the same class groups that were identified in Section 3.1 from a visual similarity. Several classes (Iris, Orchid, Poppy, Sunflower) have a perfect accuracy on all folds.

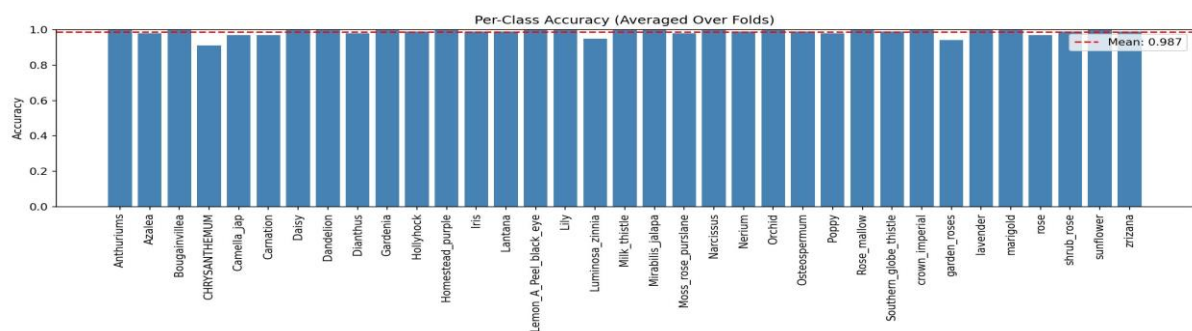


Figure 3. Per-class accuracy of the proposed hybrid model averaged over five cross-validation folds. The red dashed line indicates the overall mean accuracy of 0.987.

D.7. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves of five representative flower classes are shown in Figure 4. The Area Under the Curve (AUC) value is 1.00 for all 5 classes: Anthurium, Azalea, Bougainvillea, CHRYSANTHEMUM and Camelia_jap. All curves have true positive rates of 1.00 at very low false positive rates, which suggests very high confidence and good calibration in class-vs-rest discrimination at all classification thresholds. Even if the AUC is not as high as the other methods in the same class, as seen in Figure 3, the model's probabilistic discriminability is not affected by the misclassifications that occur in visually similar classes, as in CHRYSANTHEMUM.

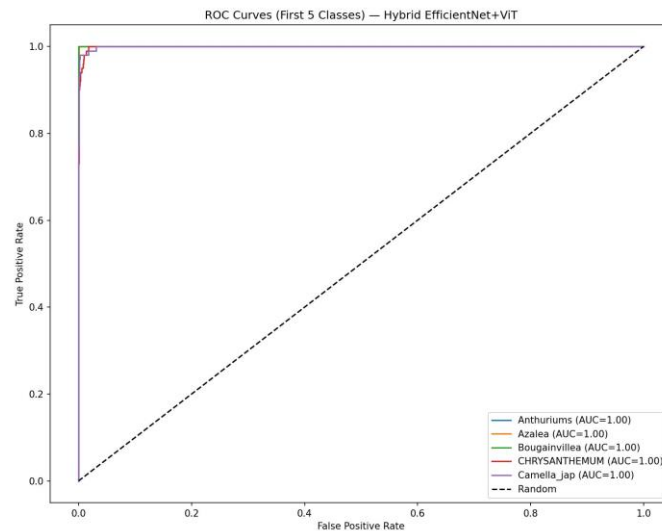


Figure 4. ROC curves for five representative flower classes produced by the proposed hybrid model.

D. Discussion

The experimental results indicate that the proposed bidirectional cross-attention fusion based on hybrid EfficientNet-B4 and ViT-Small model achieves a good performance in fine-grained flower recognition. The proposed model was evaluated with 5-fold cross validation, and it gave the highest mean accuracy of 98.71% and the mean F1-score of 98.80%, outperforming the individual constituent streams and all the evaluated baseline architectures. Results indicate the utilization of both convolutional and transformer-based representations may be useful for flower classification tasks especially when there are several flower classes with similar colour, texture, and petal structure in a flower dataset.

This was significant improvement over its standalone counterpart, EfficientNet-B4. The proposed hybrid model improved the efficiency model (EfficientNet-B4) by ~4.67 percentage points with a mean accuracy of 94.05%. This means that local convolutional features do not necessarily represent a useful approach to differentiation of visually similar flower species. EfficientNet-B4 can perform well at capturing local texture, colour gradients, and patterns at the edge level, but it might not be as effective at modeling global spatial relationships between flower petals, centers, and overall form. The proposed architecture leverages the global self-attention introduced by ViT-Small, enabling the model to handle dependencies and structural relationships in the entire image.

The improvement was less remarkable compared to standalone ViT-Small and the hybrid model gained about 0.46 percentage points on average. This indicates that the representation from ViT-Small is also quite good for this task. The hybrid model, however, was more stable across the folds later on and performed the best across all the folds on average. There was no statistically significant difference between the hybrid model and ViT-Small, suggesting that the improvement is a relatively small but consistent performance difference. This is crucial, because the proposed architecture provides not only increased accuracy, but also combines complementary local and global features in a single model.

It is important to note that the value of the proposed hybrid model over standalone ViT-Small should not be evaluated on the basis of raw accuracy alone. From a practical deployment perspective, the hybrid architecture demonstrated a meaningfully faster convergence, reaching its best validation performance at a mean of 12.8 epochs compared to 17.6 epochs for standalone ViT-Small — a reduction of approximately 27% in training time. This faster convergence is not merely a computational convenience; it reflects the ability of the bidirectional cross-attention mechanism to accelerate the learning of discriminative representations by allowing the convolutional and transformer streams to mutually refine each other from the earliest training stages, rather than relying on a single stream to develop all representational capacity independently. Furthermore, the hybrid model exhibited superior cross-fold stability, with a standard deviation of 0.30% and a 95% confidence interval of 98.34%–99.09%, compared to standalone ViT-Small which showed a more pronounced performance drop in folds 4 and 5 (97.29% and 97.71% respectively), suggesting that the convolutional stream provides a stabilising local inductive bias that reduces sensitivity to data partition variability. Taken together, the faster convergence and improved stability across folds represent meaningful practical advantages of the hybrid design, particularly in resource-constrained training environments and real-world applications where consistent, reliable performance across unseen data partitions is a primary concern.

The proposed approach clearly outperformed other state-of-the-art models. It was the top performer on all five folds compared to Swin Transformer, InceptionV3, ResNet50, DenseNet201 and MobileNetV2. The paired t test statistics revealed that these models improved significantly when compared to the proposed fusion strategy, confirming the effectiveness of the fusion strategy. However, Wilcoxon's signed rank test was not as definitive because of the small number of folds. This also brings a major problem with the statistics since it only offers 5 pairs of observations to compare the model. The results are promising, although a more compelling evidence of generalisation would be achieved with per-image prediction-level statistical testing or external testing sets.

The convergence behaviour also validates the effectiveness of the proposed model. The hybrid model achieved an average of 12.8 epochs till it reached its best validation performance, while ViT-Small and EfficientNet-B4 needed 17.6 and 25.0 epochs respectively. CNN baselines needed more than 24 epochs on average, with MobileNetV2 almost hitting the limit of epochs. It indicates that the cross-attention fusion not only contributes to improve classification performance but also can accelerate the learning of discriminative representation. The use of pre-trained backbones and the complementary relationship between the convolutional and transformer streams, in which both local and global information is used to refine each other before being classified, have contributed to the convergence.

The accuracy per class shows additional information about the model. The majority of the flower classes were classified with near perfect accuracy, and only a few groups of flowers, which are only slightly different visually, had lower accuracy. Chrysanthemum, Luminosa_zinnia, and garden_roses had lower accuracy per class than the average accuracy. This is to be expected as there are some flower species in the data set that have similar visual attributes, such as flower petal

density, colour distribution, and rounded flower shape. The remaining classification errors are primarily due to fine-grained visual similarity, as suggested by these results.

Based on results from the ROC analysis, the proposed model has a high discriminative capacity. The selected representative classes yielded AUC values of 1.00, which suggests that the model is able to discriminate between the above-mentioned classes in a one vs rest classification framework. Interestingly, while Chrysanthemum had the lowest per-class accuracy it never managed to get an AUC lower than 1.0. This indicates that model is able to separate Chrysanthemum from other classes with a relatively high probability, but with occasional hard-label misclassifications among other visually similar classes. Thus, a normalised confusion matrix and the example images misclassified needs to be analysed in the future to better understand the errors made.

While this was a good performance, there are a number of points to be taken into account. First, the dataset is balanced having 100 images per class, but the overall number of images is still quite small for deep learning, particularly transformer based models. Secondly, the data set was assembled from various sources such as the Oxford data set, a custom flower data set, and data from Google images, which can vary in quality, background complexity, and acquisition conditions. Third, the performance was estimated by cross validation within the full combined database, so the model's ability to generalise in the real world needs to be further evaluated by using completely independent external databases or images collected in the field in the Kurdistan Region. Last, hybrid architecture might be more complex than lightweight CNN models, which should also be taken into consideration if the model is going to be deployed on edge or mobile devices.

In general the results show that the bidirectional cross-attention fusion technique is an effective approach for flower classification at a finer level. The proposed model is able to obtain high accuracy, strong cross-fold stability and efficient convergence by enabling complementary information to be shared between the convolutional stream and the transformer stream. The results demonstrate hybrid CNN–Transformer models for use in biodiversity monitoring, smart farming, and applications for region-specific plant recognition.

E. Conclusion

This paper suggested a hybrid dual-stream deep learning architecture to recognize flowers in the Kurdistan Region of Iraq. The model is developed based on EfficientNet-B4 and ViT-Small/16 with a bidirectional cross-attention fusion mechanism, which allows for the fusion of local features from the convolutional layer with global features from the transformer. This design overcomes the disadvantage of using a CNN model alone or a transformer model alone, and gives a more discriminative representation for the classification of flowers with high resolution.

The proposed model has been evaluated on five-fold cross validation and has achieved a mean accuracy of 98.71% and a mean F1-score of 98.80%, which is better than standalone EfficientNet-B4, standalone ViT-Small, and a number of baseline architectures such as Swin Transformer, InceptionV3, ResNet50, DenseNet201, and MobileNetV2. In addition, the model was shown to have good

cross-fold stability with a 95% confidence interval of 98.34%–99.09% and to converge faster than the baselines CNN and transformer. Paired t-tests revealed significant improvements over most baseline models, and although there was a positive but not statistically significant difference between ViT-Small and the hybrid model, the hybrid model consistently outperforms the strongest single stream.

Most of the other errors were noted in classes of flowers with similar visual characteristics, such as similar colours and overlapping textures and shapes of the petals, which made classification more difficult. This emphasizes the need for further studies with bigger datasets with more diverse samples and using visualisation and explainability analysis of the confusion matrix. Future research will concentrate on adding more endemic species to the dataset, testing the model with all new and independent field-collected images, using Grad-CAM and transformer attention visualisation to understand the features learned by the model, and compressing the model to enable deployment on mobile and edge devices for smart farming and biodiversity monitoring applications.

Acknowledgements

The authors would like to express their gratitude to the University of Raparin and the Department of Software Engineering by acknowledging the technical assistance provided by the lab (Artificial Intelligence Research Lab at the University of Raparin) for supplying a powerful computer that has been used in the training phase of this study, the incentive paid by the university, and extend our appreciation to the university for supporting this study.

Funding Declarations

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no competing interests.

Data Availability Statement

The dataset supporting the findings of this study was compiled from multiple sources. A portion of the images was collected by the authors as part of the Custom Flower Image Dataset (CFID), while additional samples were obtained from publicly available datasets (Oxford 17 and Oxford 102 Flower datasets). To improve class balance and coverage across 35 flower species, supplementary images were included from publicly accessible web sources. Due to copyright and usage restrictions, the complete dataset cannot be made publicly available. However, metadata, processed subsets, and further details are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Gemini in order to improve language quality, clarity, and readability. After using this tool, the

authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

G. References

- [1] Tian, M. and Z. Liao. Research on flower image classification method based on YOLOv5. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
- [2] Huang, M.-L., Y.-X. Xu, and Y.-C. Liao, Image dataset on the Chinese medicinal blossoms for classification through convolutional neural network. *Data in Brief*, 2021. **39**: p. 107655.
- [3] Shaparia, R.H., N.M. Patel, and Z.H. Shah, Flower classification using texture and color features. *International Conference on Research and Innovation in Science, Engineering & Technology*. Kalpa Publications in Computing, 2017. **3**: p. 113-118.
- [4] Pratiwi, A. and A. Fauzi, Implementation Of Deep Learning On Flower Classification Using Cnn Method. *Jurnal Teknik Informatika (JUTIF)*, 2024. **5**(2): p. 487-495.
- [5] Toğaçar, M., B. Ergen, and Z. Cömert, Classification of flower species by using features extracted from the intersection of feature selection methods in convolutional neural network models. *Measurement*, 2020. **158**: p. 107703.
- [6] Lv, Q., S. Zhang, and Y. Wang, Deep learning model of image classification using machine learning. *advances in multimedia*, 2022. **2022**(1): p. 3351256.
- [7] Patel, I. and S. Patel, An optimized deep learning model for flower classification using NAS-FPN and faster R-CNN. *International journal of scientific & technology research*, 2020. **9**(03): p. 5308-5318.
- [8] Alipour, N., et al. Flower image classification using deep convolutional neural network. in *2021 7th International conference on web research (ICWR)*. 2021. IEEE.
- [9] Bae, K.I., et al., Flower classification with modified multimodal convolutional neural networks. *Expert Systems with Applications*, 2020. **159**: p. 113455.
- [10] Giraddi, S., et al. Flower classification using deep learning models. in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. 2020. IEEE.
- [11] Duman, B. and A.A. Süzen, A study on deep learning based classification of flower images. *International Journal of Advanced Networking and Applications*, 2022. **14**(2): p. 5385-5389.
- [12] Anjani, I.A., Y.R. Pratiwi, and S.N.B. Nurhuda. Implementation of deep learning using convolutional neural network algorithm for classification rose flower. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
- [13] Sun, X., et al., Four-dimension deep learning method for flower quality grading with depth information. *Electronics*, 2021. **10**(19): p. 2353.
- [14] Omer, S.M., R.M. Hasan, and B.N. Anwer, An image dataset construction for flower recognition using convolutional neural network. *Science Journal of University of Zakho*, 2020. **8**(3): p. 112-117.
- [15] Saad, W., et al. Classification and detection of chili and its flower using deep learning approach. in *Journal of Physics: Conference Series*. 2020. IOP Publishing.

- [16] Jiang, Y., et al., DeepFlower: a deep learning-based approach to characterize flowering patterns of cotton plants in the field. *Plant methods*, 2020. **16**(1): p. 156.
- [17] Xiao, R. and R. Wang. Transfer Learning-Based Flower Image Classification: Leveraging The Pre-Trained Alexnet Model. in *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*. 2023. IEEE.
- [18] Saputra, D.B.A., C.A. Sari, and E.H. Rachmawanto, Jasmine Flower Classification with CNN Architectures: A Comparative Study of NasNetMobile, VGG16, and Xception in Agricultural Technology. *Advance Sustainable Science Engineering and Technology*, 2024. **6**(4): p. 0240409-0240409.
- [19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, Apr. 2020, pp. 13001-13008.
- [20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, arXiv:1710.09412.
- [21] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023-6032.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [23] I. O. Tolstikhin et al., "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24261-24272.
- [24] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?. *Advances in neural information processing systems*, 34, 12116-12128.
- [25] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105-6114.
- [26] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv:1607.06450.
- [27] Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2016, arXiv:1606.08415.
- [28] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [30] C. F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 347–356.
- [31] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5100–5111.