Fine-Grained Image Classification Using Vision Transformer Model

Zunaira Saleem[1], Uzma Jamil[2], Saman Iftikhar[3]
Department of Computer Science, Government College University Faisalabad, Pakistan[1, 2]
Faculty of Computer Studies, Arab Open University, Saudi Arabia[3]

***Abstract*—Fine-Grained Image Classification focuses on unique features between visually similar subclasses within a wider category, which remains a challenging task due to low inter-class variations and high intra-class similarity. Conventional Convolutional Neural Network-based methods often struggle to accurately capture these minor differences. Utilizing self-attention techniques to represent global relationships within images, Vision Transformers have recently demonstrated robust performance in image classification evaluations. To enhance classification performance on complicated visual categories, this research presents a Fine-Grained Image Classification framework utilizing the Vision Transformer Model. The CIFAR-100 dataset, which includes 100 different image classes, is used for experimental purposes. The images were up-sampled because the Vision Transformer demands higher resolution inputs. To improve training efficiency and generalization, preprocessing techniques, including normalization and data augmentation, are applied. The model is trained and evaluated using standard performance metrics, including accuracy, macro precision, macro recall, and macro F1 Score, to ensure a balanced evaluation across all classes. With an overall classification accuracy of 89.68% and good macro-level assessment scores, experimental results show that the Vision Transformer Model successfully captures subtle visual distinctions among comparable categories. Transformer-based architectures offer an effective substitute for conventional techniques in Fine-Grained Image Classification applications with better performance. This research demonstrates how the Vision Transformer Model can increase classification robustness and accuracy for a dataset with very similar item classes.**

***Keywords*—Data augmentation; fine-grained image classification; CIFAR-100; vision transformer model; deep learning**

I. Introduction

Image classification has become a fundamental task in computer vision, enabling machines to understand and classify visual data. Fine-Grained Image Classification (FGIC) is still a challenging topic, despite significant progress in general image classification. It focuses on the uniqueness between visually similar categories, where differences are often limited to minor variations in texture, shape, or color.

FGIC includes the merging of visual features between classes, in comparison to predictable classification problems, where categories are recognizable. The process is further complicated by differences within the same class depending on changes in backdrop, lighting, and attitude. Models that can capture both fine-level distinction characteristics and global context will be required to overcome such challenges. In Image Classification, deep learning techniques, in particular Convolutional Neural Networks, have shown outstanding outcomes. However, their capacity to represent long-range interdependence is restricted by dependence on locally available fields. Recently, transformer-based architectures have emerged as a powerful alternative for capturing global connections across images through self-attention-based techniques.

The classification of subclasses within a broader category is the focus of a specific category of image classification called FGIC. It is a subset of image classification that intends to classify subclasses inside a broader class [1]. The issue of overall feature similarity across closely related subclasses is addressed by FGIC. Unlike common classification challenges, where the difference between classes is clear, FGIC is a challenge of identifying subtle differences within the same high-level category. For instance, the color of a bird's feathers can only vary slightly across two species. It contains images with high inter-class unpredictability and low inter-class variability [2].

With only a few design variations, car models can be nearly similar. Flowers can be identified by the slightest changes to their shape or petal pattern. This task is more challenging than conventional image categorization since such minor details may be hidden or partially hidden by changes in lighting, background clutter, or how the item will appear in the image. According to [3], lighting, position, and context can all affect how classes seem. By reconstructing the image's local and global attributes, this method seeks to detect such minute variations. The need to distinguish subclasses with high similarity is addressed in this research. FGIC requires the discovery of slight changes between different classes, as compared to general classification issues, which call for the detection of tiny differences within a single broad category. A case study of this is provided by two bird species whose feather colors can only differ slightly.

Flowers can be recognized by slight differences in the form or features of their petals. Compared to traditional image classification, this is a difficult process since such minute features can be hidden or changed by things like lighting changes, background clutter, and the image's content. To address these limitations, recent research has introduced advanced techniques including attention mechanisms, part-based modeling, and Vision Transformer Model. These techniques have demonstrated significant advancements in the detection of both global contextual structures and fine-grained local features, which have enhanced classification performance in difficult real-world scenarios.

The lack of high-quality, labeled data is a major challenge to FGIC advancement. Large-scale fine-grained dataset development requires a high level of domain-specific knowledge, which makes data gathering and annotation expensive and time-consuming. Benchmark datasets like Stanford Cars [4], FGVC Aircraft [5], and CUB-200-2011[6] (for birds) have been generated to support research in this field by offering standardized platforms for evaluation and analysis. The problem of inadequate information remains despite these resources, delaying the development of models and their practical implementation. However, FGIC is extremely significant in a range of applications, including industrial quality control, precision agriculture, medical diagnostics, and biodiversity monitoring, highlighting its significance to the development of domain-specific AI systems.

The classification of two apparently similar woodpeckers using FGIC is shown in Fig. 1. Existing methods of classification are unable to accurately identify these species due to their high inter-class similarity and low intra-class similarity. The picture indicates how an attention-based network would automatically locate discriminative regions, particularly the head and beak areas, which include minor but significant visual signals, including color patterns and pattern differences.

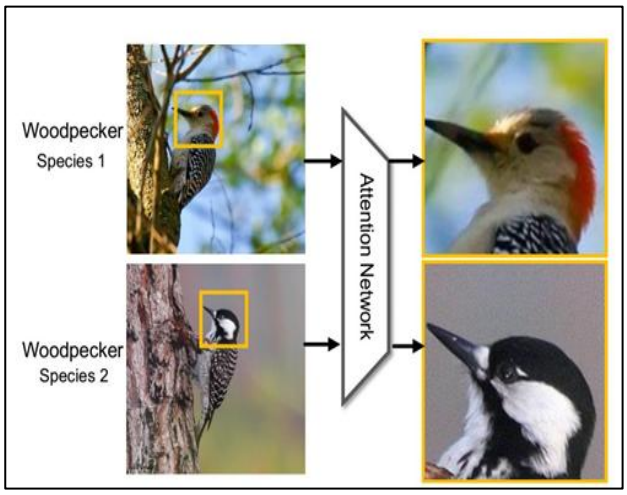

Fig. 1. Attention-based fine-grained image classification [7].

More accurate feature extraction and decision-making are made possible by the attention mechanism, which eliminates irrelevant background information and amplifies useful local characteristics. These methods indicate success in fine-grained visual classification tasks, such as the classification of bird species, where small anatomical variations are crucial [7]. The complexity and fine features of the CIFAR-100 dataset make Few-shot Learning (FSL) of the dataset difficult. Compared to large datasets like ImageNet, CIFAR-100 only has 100 classes with 600 images of each class.

It is challenging for models to develop strong and discriminative features with this limited amount of data. In a few-shot scenario, when each class has a finite number of labeled examples, the issue gets substantially worse. Many of the CIFAR-100 categories, such as various animals, cars, or household items, have visual similarities to one another.

Such a lack of examples can frequently result in models acquiring only the rough patterns that do not allow them to classify correctly. FGIC on datasets like CIFAR-100 still has a major research gap. Instead of specifically addressing small inter-class differences and high intra-class similarity, which are major issues in fine-grained tasks, the majority of current methods emphasize increasing overall classification accuracy. Furthermore, the global contextual modeling power required to capture subtle discriminative features across related categories is insufficient in many CNN-based designs. This gap emphasizes the need for more advanced designs, like Vision Transformers, that are better able to represent fine-grained interaction between features and long-range dependencies. To increase feature sensitivity, improve generalization under situations of limited data, and acquire improved discriminative performance across visually comparable classes in CIFAR-100, FGIC research is necessary. The main issue with traditional image classification is the imbalance between intra-class and inter-class variability. When several classes have similar visual features, the model is unable to differentiate between them, leading to category misclassification. Furthermore, even within a comparable learning environment, the stances and various variations in lighting, texture, and background hinder the learning process. These problems indicate that the traditional models struggle to identify subtle features that are essential during classification.

Because these models can generate highlights over small differences across classes, they can improve classification in a few-shot scenario. FGIC makes it possible to differentiate between related classes. It is required to analyze comparable classes with the eyes, such as various bird species or automobile types. Since class lines are typically defined by subtle differences in texture, shape, or color, these are very challenging to execute. Through the use of hard sample mining and comparison-based prototype learning, this work will enhance FGIC using the Vision Transformer Model.

The implementation of the Vision Transformer Model in FGIC is addressed in this research. The goal is to find out if transformer-based models may improve category classification overall and even identify subtle variations in similar classes.

A. Research Gap

FGIC on datasets like CIFAR-100 still has a significant research deficit. Instead of particularly addressing small inter-class differences and high intra-class similarity, which are major issues in fine-grained tasks, the majority of current technologies focus on increasing overall classification accuracy. Furthermore, the global contextual modeling power required to capture subtle discriminative characteristics across related categories is lacking in many CNN-based designs. This disparity emphasizes the need for more sophisticated designs, like Vision Transformers, that are better able to represent fine-grained feature interactions and long-range dependencies.

In order to increase feature sensitivity, the Vision Transformer Model improves generalization under situations of limited data and attains improved discriminative performance across visually comparable classes in CIFAR-100. FGIC research is required.

B. Need for Research

High intra-class and low inter-class variability remain challenges for traditional models in image classification, requiring the development of advanced techniques that focus on tiny significant features. In real-world uses where tiny details are essential, FGIC research greatly improves accuracy. Although Vision Transformer Models have demonstrated good performance in broad classification, their effectiveness in FGIC has not yet been thoroughly investigated. It is essential to address these issues, emphasizing the necessity for advanced techniques that successfully resolve generalized classification problems.

C. Research Question

How can Vision Transformer Models be effectively optimized to improve the accuracy of FGIC to classify similar categories using the CIFAR-100 dataset?

D. Research Objective

The main objectives of this research are:

- To increase classification accuracy for fine-grained picture categories by refining and optimizing a Vision Transformer Model using the CIFAR-100 dataset.
- To examine the model's capacity to differentiate visually similar classes quantitatively using measures including accuracy, precision, recall, and confusion matrix analysis.
- To compare classification outcomes across related categories to examine how transformer-based design affects fine-grained categorization.

Next, the study is organized as follows. Section II is about related work on FGIC, and Section III describes our proposed system architecture. Section IV presents results and discussion. Finally, Section V discusses the conclusions of the work study and future work.

II. Related Work

This section will present an overview of the existing research that is related to Fine-Grained Image Classification and Transformer-based Vision models to determine the base of the proposed approach.

FGIC seeks to differentiate between visually similar subcategories, such as flower species or bird species, where distinguishing characteristics are frequently confined and subtle [7]. The initial approaches had limited flexibility and practicality since they mostly depended on hand-crafted features or part-based models with heavy supervision, such as bounding box and component annotations. Convolutional Neural Networks (CNNs) greatly enhanced feature representation with the development of deep learning; however, many CNN-based techniques continued to treat feature learning and discriminative domain identification as independent methods, leading to inadequate performance for fine-grained tasks.

This research proposed a system for enhancing self-supervised contrastive learning for fine-grained recognition problems called Fine-grained Adaptive Contrastive Learning (FACL) [8]. The method contributes to richer representation learning, which improves feature extraction, by including methods that provide increased sample variety. According to the experimental findings, FACL significantly outperforms conventional contrastive learning techniques. The approach demonstrated its ability to handle complicated visual differences with an accuracy of 64% in applications such as anomaly detection and fine-grained categorization. This emphasizes how important adaptive techniques are for contrastive learning, particularly when it comes to tasks that require identifying subtle differences.

Fine-grained categorization in agricultural disease detection, where categories only differ by minor images and are affected by background noise, humidity, and clarity [9]. The authors' Matrix-based Convolutional Neural Network (M-bCNN), which arranges convolutional layers in a parallel "kernel matrix", improves feature recognition. By expanding data streams, neurons, and link channels without significantly changing parameters, this setup outperforms traditional CNNs in feature representation. These methods enable the network to avoid vanishing gradients and overfitting. They produced a sizable, high-resolution dataset consisting of 16,652 photos of wheat leaves, which was expanded to 83,260 samples. One of the earliest large datasets for the classification of wheat leaf diseases is this one.

A hybrid architecture that combines the benefits of Vision Transformers (Vision Transformer Models) and Convolutional Neural Networks (CNNs) to improve FGIC [10]. The framework has three new modules. The Multi-Scale Image-to-Tokens (MIT) module extracts features at different scales to provide richer visual representations. The Mixed Convolution Feed-Forward (MCF) design enhances the network's ability to learn complex local properties. The Multi-Layer Feature Selection (MFS) method collects the most important attributes from many levels. This technique catches both broad and nuanced visual features by combining CNN's local spatial awareness with the Vision Transformer Model's global context knowledge. By enabling the model to recognize tiny differences between visually similar categories, this combination greatly increases classification accuracy in fine-grained identification tasks where tiny differences matter.

The current study addresses the problem of small inter-class changes in CIFAR-100, based on previous FGIC research that focuses on attention mechanisms and multi-scale feature learning. The proposed approach improves the model's capacity to differentiate closely related categories by utilizing Vision Transformers to capture both local and global images.

III. Materials and Methods

The proposed solution is an automated system designed to classify 100 various image classes efficiently and accurately. To achieve the classification tasks, the suggested system consists of a number of interconnected modules that operate one after the other. The CIFAR-100 dataset was utilized in this experiment. This dataset serves as the study's basis, as the suggested classification framework is trained and analyzed using its images.

Pre-processing is the initial module, which gets the input data appropriate for the model's architectural requirements. The

last module, the model module, receives the processed images and feeds them into the Vision Transformer architecture. The model extracts important characteristics, learns discriminative patterns across classes, and performs classification. This well-structured pipeline guarantees stable training, improved feature learning, and improved overall classification performance.

Fig. 2 describes the flow of the proposed solution. The first phase of the proposed solution is preprocessing. In order to stabilize the learning process, images were enlarged to the required resolution, normalized, and enhanced utilizing augmentation techniques such as random flipping and rotation. These techniques improve the model's ability to generalize, diversify the data, and reduce overfitting.

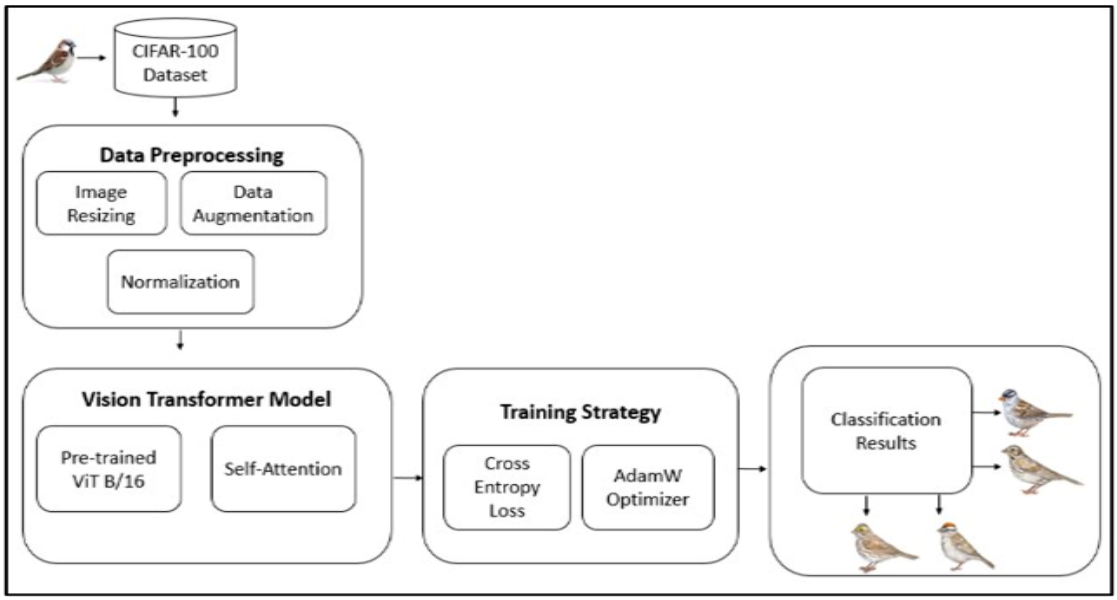

Fig. 2. An overview of the proposed methodology.

A. Image Resizing

Image resizing is one of the pre-processing stages that enhances the Vision Transformer Model by minimizing variations in image size. Vision Transformer Models with larger input sizes, which may divide pictures into meaningful patches, cannot represent the 32x32 pixel images in the CIFAR-100 dataset. This was fixed by up-sampling every image to 224 by 224 pixels. By scaling the pictures, the Vision Transformer Model may readily acquire patch embeddings that have both local and global characteristics. Learning fine details and being able to distinguish subtle visual differences across classes relies on these embedded images. The success of the classification. Fig. 3 shows the input and output images after resizing.

Eq. 1 presents a mathematical definition for this transformation, which involves sampling the original picture at scaled coordinates to determine the pixel intensity at spatial coordinates (x′, y′) in the stretched image. The scaled photos satisfy the Vision Transformer model's input size criteria while retaining structural consistency because of this method of normalizing the spatial dimensions.

Where H and W represent the image's height and breadth, and a pixel's intensity at spatial coordinates (x, y). All input photos are scaled to a fixed resolution of 224*224 in order to guarantee execution with the Vision Transformer Model architecture. The resizing process uses scaling based on the original height H and width W to transfer each pixel in the modified picture I′(x′,y′) to the appropriate place in the original image I(x',y).

$$I'(x', y') = I(x' H' . H, y' W' . W) \quad (1)$$

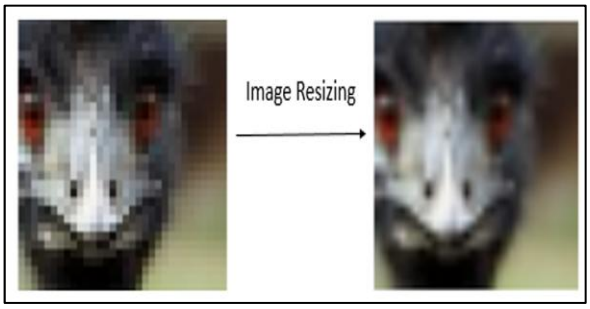

Fig. 3. Image resizing.

B. Data Augmentation

Data Augmentation in the proposed system was performed with torchvision. Data augmentation is a way to improve Vision Transformer frameworks for FGIC. The tiny images in CIFAR-100 data are barely changeable. We implement augmentation approaches to broaden the variety of the training samples. Random rotation, random cropping with padding, random horizontal flipping, and color jittering are common transformations. Image characteristics, including brightness, contrast, saturation, and hue, are affected by such changes. These modifications strengthened the model's resistance to changes in scale, orientation, and illumination.

The Vision Transformer Model was successfully trained utilizing a wider range of visual changes according to this approach. As a result, it improved overall categorization on fine-

grained categories and was better able to capture fine-grained visual distinctions. The data augmentation strategy, which enhances the model's performance on unknown test data and improves the classification rate. Fig. 4 clearly presents the Data Augmentation technique, which improves the performance of the model.

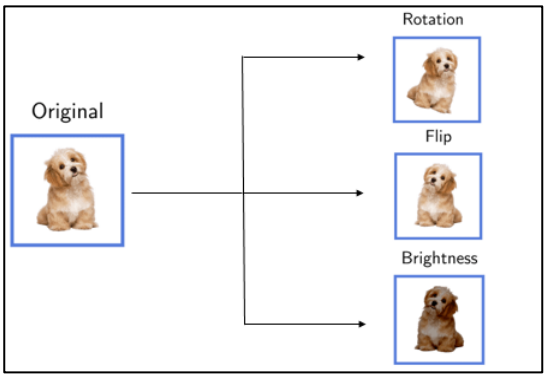

Fig. 4. Original and augmented images.

C. Random Horizontal Flip

A mirror image is produced across the vertical axis by this technique. This transformation is applied to training images at random, often with a 50% chance, in the context of Random Horizontal Flip.

In computer vision, this strategy is frequently used for data augmentation. It enhances a model's capacity for generalization by making it invariant to horizontal orientations, provided that it can accurately identify objects, regardless of whether they face left or right. To put it simply, every pixel along the horizontal axis is mirrored. The Random Flip Image is seen in Fig. 5. Each pixel at location (x', y) in the flipped image corresponds to the pixel at position (W−x−1, y) in the original image, according to Eq. 2.

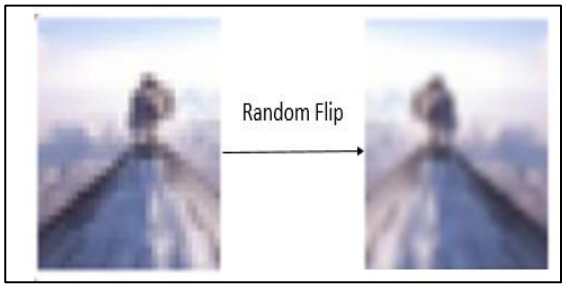

Fig. 5. Random flip image.

The expression I′ (x', y) =I (W−x−1, y) describes the mathematical operation for a horizontal flip of an image. Here, I (x', y) denotes the original image, I′ (x', y) is the transformed image after flipping, and W is the width of the image. Pixels from the left side move to the right, and pixels from the right side move to the left while the vertical position (y) stays the same.

$$I'(x,y) = I(W-x-1,y) \qquad (2)$$

D. Random Rotation

This method is frequently used to significantly enhance the range of training data in computer vision and machine learning. As a result, models become less dependent on variations in object placement and more stable. The algorithm learns to identify items despite their rotation in real-world scenarios by randomly rotating normal images during training. Eq. 3 describes a rotation transformation that is performed on the input image. The new coordinates (x', y') are obtained by rotating each pixel coordinate (x', y) by an angle θ. A typical 2D rotation matrix made up of sine and cosine terms is used to carry out the transformation. By introducing rotational invariance, this technique enhances the model's capacity to generalize to pictures with different angles. The Random Rotation is shown in Fig. 6.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (3)$$

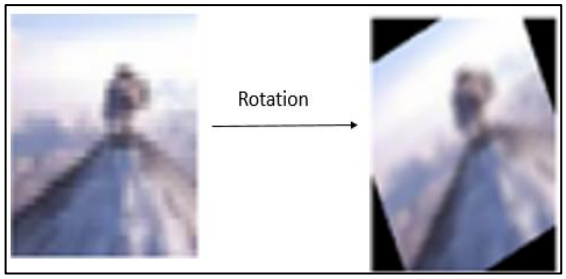

Fig. 6. Random rotation image.

E. Color Jitter Brightness Adjustment

This technique was used to enhance the robustness of image classification models by simulating various lighting conditions. Fig. 7 illustrates the Color Jittering Brightness Adjustment. Eq. 4 describes a brightness adjustment applied to the input picture, where a random brightness offset Δ (b) is added to the original image I to generate the transformed image *I*. To generate a new image, the input image's brightness is randomly changed during this procedure. This fluctuation reflects real-world situations where brightness may vary depending on the setting or time of day. The model's capacity to generalize to new data is enhanced by training on standard images with varying brightness levels, which reduces the model's sensitivity to lighting variations.

$$I' = I + \Delta_{b,} \quad \Delta_b \sim \cup(-\beta, \beta) \qquad (4)$$

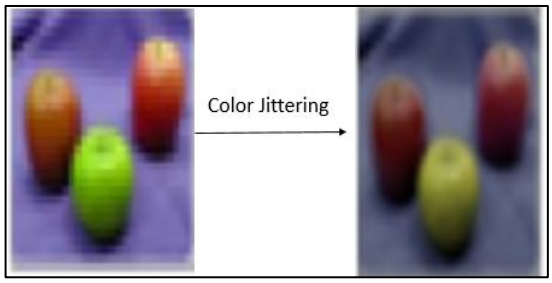

Fig. 7. Input and enhanced image with color jittering.

F. Contrast Adjustment

The operation changes the brightness difference between light and dark areas of an image. This makes features easier or harder to see. Eq. 5 describes a contrast adjustment operation applied to the input image, where the pixel intensities of the transformed image I' are scaled by a factor α around the mean intensity μ. This random contrast adjustment is often used as a data augmentation method in image classification tasks. It helps a model handle different lighting and image quality conditions better. Fig. 8 shows color illumination correction of the input image.

$$I' = \alpha \cdot I - \mu + \mu, \quad \alpha \sim \cup\ (1 - c\ , 1 + c) \quad (5)$$

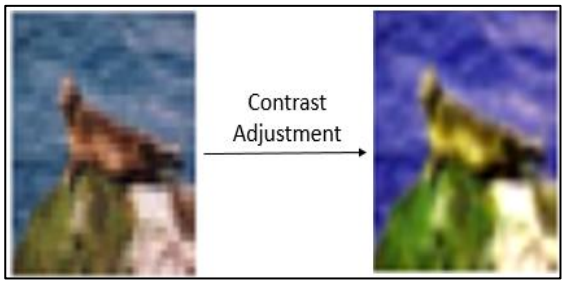

Fig. 8. Color illumination correction of the input image.

G. Normalization

This is the second pre-processing step. Images were resized, and pixel values were scaled so that each color channel's mean value and standard deviation were 0.5. Eq. 6 normalizes the input values to a fixed range, usually between -1 and 1. It helps to speed up and stabilize the training process. Normalization ensures that the pre-trained Vision Transformer model predicts the input distribution. Better convergence and performance assurance result from this. Fig. 9, which displays the Enhanced Image after Normalization, illustrates the difference gained after applying the augmentation. By eliminating internal variance shift, normalization contributes to more effective and efficient learning during training. This process is typically subject to the normalization formula of the standard score (z- score) of each channel of an image (Red, Green, Blue) [Eq. (6)].

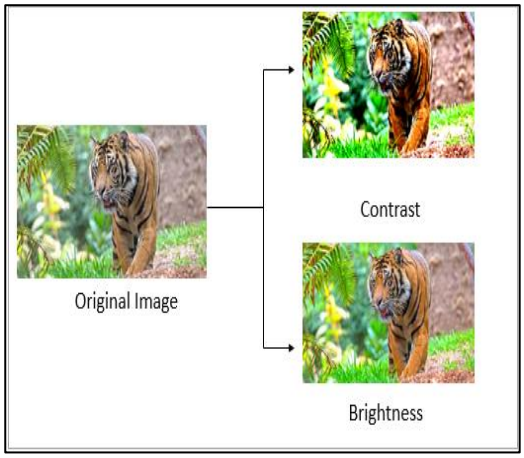

Fig. 9. Original and enhanced images.

$$X' = \frac{X - \mu}{\sigma} \quad (6)$$

where, X represents the original pixel value, μ is the mean, and σ is the standard deviation. The mean and standard deviation values used are both 0.5 for each color channel.

H. Fine-Grained Image Classification

The Fine-Grained Image Classification Model, utilizing the Vision Transformer Model, is employed in the proposed method. The strategy is based on the Vision Transformer architecture, which has shown outstanding outcomes in various types of image classification tasks. Unlike traditional convolutional neural networks, the Vision Transformer model employs self-attention processes to identify global characteristics and long-range interactions within the input image. The main architecture used in this study is the Vision Transformer Model-Base-Patch16-224 timm (PyTorch Image Models) library. This approach divides the image provided into many non-overlapping patches. These patches are then passed through a typical transformer encoder after being encoded as fixed-length vectors. Each of the 12 transformer layers in the Vision Transformer Model-Base-Patch16-224 contains 12 attention heads and 768 hidden dimensions.

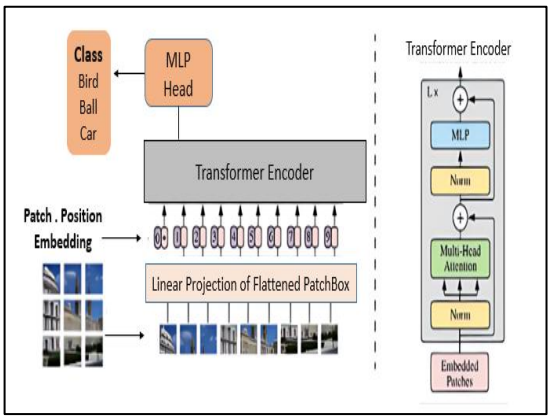

Fig. 10. Vision transformer architecture [11].

Fig. 10 illustrates that the Vision Transformer Model works with images by splitting them into patches of a fixed size. The patches are flattened and transformed into a numerical embedding representing the visual content of the patch. These embeddings are then sent into the transformer encoder together with positional encodings and a learnable class token. The encoder uses feed-forward layers with residual and normalization connections as well as multi-head self-attention. This structure helps the model recognize spatial data in images by generating significant global relationships between the patches.

A Multi-Layer Perceptron (MLP) head receives the class token's final output and is used to perform the classification task. The Vision Transformer Model can effectively capture long-range connections with this framework. The model is effective, particularly when handling recognition of images and fine-grained classification tasks, since the transformer attention mechanism allows the model to learn to focus on tiny images. Recognizing related groups of objects is essential, particularly when a collection includes sensitive classifications. The backbone may be used to save training time and enhance the

model's generalization. It makes it possible to determine tiny variations across classes in fine-grained visual datasets.

The proposed solution for FGIC:

1) The Vision Transformer (Vision Transformer Model) model uses ImageNet-pretrained weights to improve performance through transfer learning, and it was modified for the CIFAR-100 classification problem. Faster convergence and increased accuracy are made possible by this initialization, which offers a solid foundation of learnt visual features, especially in fine-grained classification settings with less training data. The technique successfully addresses the problem of data scarcity by transferring rich visual information to the target domain. Furthermore, the framework uses PyTorch to provide effective computing in both CPU and GPU settings, dynamically assigning model parameters to the available hardware.

2) The pre-processed and augmented CIFAR-100 dataset is used for training after the Vision Transformer (Vision Transformer Model) has been initialized and configured. Using a cross-entropy loss function appropriate for multi-class tasks, model parameters are optimized to minimize classification errors across 100 object categories. For transformer-based architectures, the optimization process uses the AdamW optimizer, which offers better weight regularization than standard Adam. A learning rate of 3e-5 guarantees gradual convergence, while a weight decay of 0.01 aids in reducing overfitting and enhancing generalization. Furthermore, the learning rate is gradually reduced using a cosine annealing learning rate scheduler, which facilitates smoother convergence and lowers the possibility of becoming stranded in local minima.

3) Monitor loss and accuracy metrics during training. This helps the model develop useful visual characteristics. To ensure effective learning, hyperparameters were adjusted as required. Once the training is completed, the Vision Transformer will be capable of generating good feature representations that allow the system to perform correct classification in the fine-grained disjointed categories in the CIFAR-100 dataset.

4) After training, the Vision Transformer Model's ability to generalize is determined using the test split of the CIFAR-100 dataset, with classification accuracy functioning as the primary measure. A confusion matrix is examined to provide a deeper understanding of performance. This analysis highlights misclassifications and reveals class-wise prediction patterns, particularly among visually similar categories with overlapping attributes like texture, color, and shape. This analysis helps in highlighting the model's problems in fine-grained parameters and offers recommendations for additional optimization, such as enhancements to data augmentation, class balancing, and architectural modifications, which will ultimately improve the model's capacity to differentiate closely related classes and boost overall performance.

IV. Results and Evaluation

The proposed Vision Transformer-based method for FGIC is evaluated on the CIFAR-100 dataset, which shows how well it learns discriminative visual characteristics and achieves consistent generalization performance. To achieve a thorough evaluation, the model was evaluated using a variety of performance indicators, including test accuracy, precision, recall, F1-score, and confusion matrix analysis. Accuracy, macro precision, macro recall, and macro F1-score were used to assess the recommended Vision Transformer Model's performance on the CIFAR-100 dataset. The model demonstrated good prediction abilities across 100 item categories, with an overall classification accuracy of 89.68% on the test set. The findings utilizing Vision Transformer Model B/16 on CIFAR-100 are shown in Fig. 11.

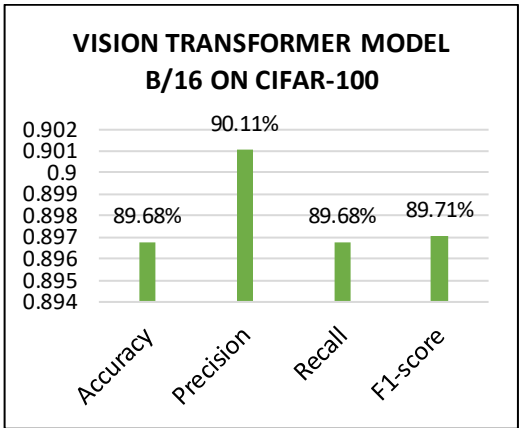

Fig. 11. Evaluation of proposed solution using vision transformer model B/16 on CIFAR-100.

By evaluating both class-wise performance and overall prediction accuracy, these metrics offer a comprehensive picture of the model's performance. Combining accuracy with precision, recall, and F1-score provides a broader overview of the model's performance in recognizing certain categories over the whole dataset.

A. Analysis of Test Accuracy

The Vision Transformer Model's generalization performance on the CIFAR-100 dataset can be enhanced by analyzing test accuracy over epochs. The model shows a sharp increase in test accuracy in the early epochs. This sharp rise indicates that the pretrained Vision Transformer Model backbone quickly adapts to the CIFAR-100 classification task and learns meaningful feature representations within a few training iterations. In transfer learning circumstances, where pretrained weights speed up learning, early convergence is frequently seen. Overall, the test accuracy trend shows that the model reaches its best generalization performance in the early epochs and continues to provide reliable outcomes beyond that. The model test accuracy results are displayed in Fig. 12. The model achieves about 89.68% test accuracy on CIFAR-100, demonstrating continuous generalization ability across epochs.

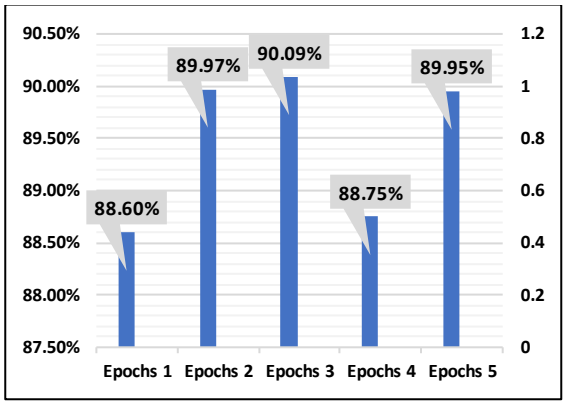

Fig. 12. Model test accuracy results.

B. CIFAR-100 Per-Class Precision Results

The proposed Vision Transformer Model's per-class accuracy in all 100 CIFAR-100 dataset classes. Most classes have high accuracy values, which often fall between 0.85 and 1.00, meaning that the model is typically accurate when predicting a certain class. This indicates that the model has a high capacity to reduce false positive predictions across various item types. For the majority of classes, consistent predictive reliability is shown by the very small range in bar heights. However, other classes show significantly better accuracy, with values lying closer to 0.70–0.80.

This consistency supports the effectiveness of the Vision Transformer architecture in handling multi- class image recognition on CIFAR-100. Fig. 13 shows that the per-class results typically show variability across different categories, indicating that some classes achieve higher precision.

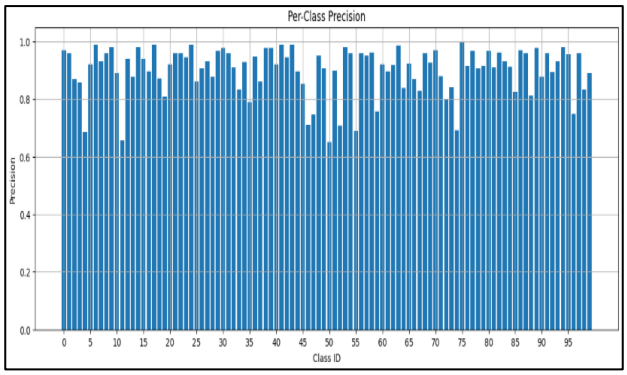

Fig. 13. Per-class precision results.

C. CIFAR-100 Per-Class Recall Results

The proposed Vision Transformer model's recall performance for each of the 100 CIFAR-100 classes is shown in the per-class recall graph. The majority of classes have high recall values, usually between 0.85 and 1.00, which suggests that the model correctly identifies a significant portion of real examples that correspond to each class. This shows how well the model can reduce false negatives and accurately gather relevant data across a variety of object classifications. Stable detection performance across the dataset is shown in the bars' overall consistency. Recall levels for a few classes are significantly lower, falling closer to around 0.70–0.80. The per-class recall is shown in Fig. 14.

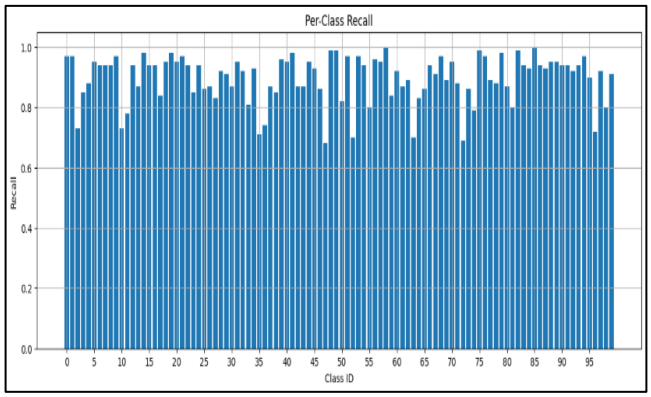

Fig. 14. Per-class recall results.

Overall, the distribution demonstrates how effectively the Vision Transformer-based architecture handles challenging multi-class image classification tasks on the CIFAR-100 dataset.

D. Confusion Matrix Analysis

The confusion matrix displays a strong diagonal pattern, indicating that the majority of samples are properly identified, providing a deeper understanding of class-wise performance beyond total accuracy. This illustrates how well the model captures significant visual patterns and how resilient the taught features are.

Despite this, the model performs well across the majority of classes, and the small number of mistakes indicates that misclassifications are not common. All things considered, the attention-based Vision Transformer (Vision Transformer Model) successfully captures both local and global features, producing dependable and balanced classification performance with significant room for improvement through improved feature refinement. Fig. 15 provides a class-by-class evaluation of the proposed model's prediction performance.

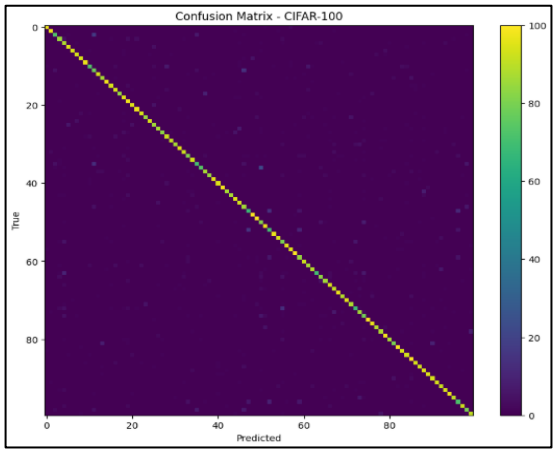

Fig. 15. Confusion matrix of predicted classes.

A major problem in fine-grained picture classification, because of minute variations in form, texture, and color, is that, although some misclassifications occur in off-diagonal regions, they are primarily limited to classes with similar visual features.

In datasets like CIFAR-100, where low resolution and closely related categories make classification challenging, this type of confusion is especially visible.

V. Comparative Analysis with Existing Research

Using the CIFAR-100 dataset, three sample image classification and few-shot learning techniques are compared in this section. Model architecture, learning methodology, and applicability for the CIFAR-100 dataset FGIC difficulties are the main areas of comparison. On the CIFAR-100 dataset, existing methods have shown excellent performance on challenging picture classification tasks. In order to increase recognition when training data is limited, several techniques concentrate on improving feature representation by utilizing attention processes and multi-branch structures to capture subtle correlations among visual characteristics. Other strategies focus on enhancing model generalization by using structured regularization techniques that hide spatially correlated areas in feature maps to avoid overfitting and enhance classification performance on multi-class datasets. Fig. 16 shows the graphical representation of the comparative analysis of the proposed solution with existing research on the CIFAR-100 Dataset.

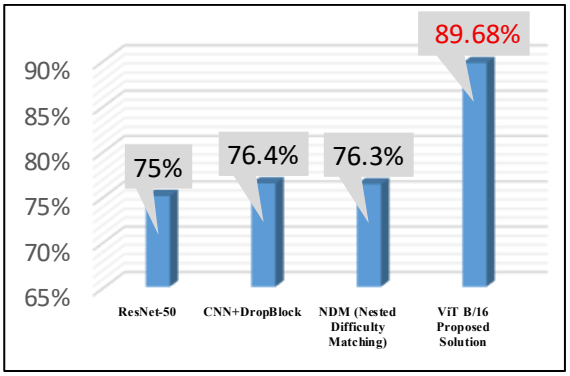

Fig. 16. Comparative analysis with existing research on CIFAR-100.

Additionally, dataset distillation techniques have been proposed to generate compact yet informative synthetic datasets by prioritizing difficult training samples and aligning them with different stages of the learning process, enabling models to achieve performance close to full-dataset training. However, many of these approaches rely on convolutional neural network backbones and require additional modules to enhance feature learning. In comparison, transformer-based architectures such as the Vision Transformer Model naturally capture global relationships between image patches through self-attention mechanisms, making them particularly suitable for FGIC, where accurate recognition depends on identifying subtle differences between visually similar categories.

VI. Conclusion and Future Work

Using the CIFAR-100 dataset, this study effectively developed and examined a Vision Transformer-based framework for FGIC. The proposed framework performed well in macro-level precision, recall, and F1-score, yielding stable and reliable results across all 100 categories, with an accuracy of 89.68%. The overall results highlight the potential of the Vision Transformer Model for FGIC despite obstacles like the dataset's low resolution and high computing needs. The findings imply that transformer topologies can be a reliable and effective method for applications requiring complex and subtle image distinction.

Future studies will expand the assessment to benchmark fine-grained datasets, including FGVC-Aircraft, Stanford Cars, and CUB-200-2011, in order to validate the suggested model more thoroughly in fine-grained image classification, even though the Vision Transformer showed strong performance on the CIFAR-100 dataset. In order to take advantage of both local and global feature representations, future research may potentially look at hybrid architectures that mix convolutional neural networks with vision transformers. Additionally, using few-shot or self-supervised learning techniques might lessen the need for sizable labeled datasets and enhance the model's capacity for generalization in practical situations.

Acknowledgment

The authors extend their appreciation to the Arab Open University for funding this work.

References

[1] Y. Liu *et al.*, "Deep learning for fine-grained image analysis: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] W. Ge, X. Lin, and Y. Yu, "Weakly supervised fine-grained classification with part selection from CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014. doi: 10.5244/C.28.1.

[4] J. Krause, M. Stark, J. Deng, and Li Fei-Fei, "3D object representations for fine-grained categorization," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops), 2013.

[5] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," arXiv preprint arXiv:1306.5151, 2013.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Institute of Technology, 2011.

[7] Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] T. Yin, J. Wang, Y. Zhao, H. Wang, Y. Ma, and M. Liu, "Fine-grained adaptive contrastive learning for unsupervised feature extraction," *Neurocomputing*, vol. 618, p. 129014, 2025, doi: 10.1016/j.neucom.2024.129014.

[9] Lin, Q. Chen, Y. Huang, and X. Wang, "Matrix-based convolutional neural network for fine-grained wheat disease recognition," *IEEE Access*, vol. 7, pp. 123450–123462, 2019.

[10] R. Shao, X.-J. Bi, and Z. Chen, "Hybrid Vision Transformer Model-CNN network for Fine-Grained Image Classification," *IEEE Signal Processing Letters*, vol. 31, pp. 1109–1113, 2024, doi: 10.1109/LSP.2024.3386112.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint* arXiv:2010.11929, 2021.