Emotion Recognition Using Vision Transformers (ViT)

Di Iorio Matteo
Politecnico di Torino
Corso Duca degli Abruzzi, 24 (TO)
s316606@studenti.polito.it

Abstract

Emotion recognition from facial expressions is critical across domains ranging from human–computer interaction to mental health. However, conventional convolutional neural network (CNN)–based models, such as ResNet and VGG, struggle in complex scenarios characterized by ethnic variation, non-uniform lighting, and noisy data. This project investigates Vision Transformers (ViT), an innovative architecture that partitions images into fixed-size patches and treats them as token sequences, leveraging self-attention to model global relationships within the image. This capability enables the capture of subtle details—such as micro-expressions—thereby improving emotion recognition over CNNs. The model will be trained on the FER-2013 dataset following careful pre-processing; images are labeled with seven emotions. We apply data augmentation and regularization strategies for optimization. The ViT will be compared against standard CNN architectures (ResNet and VGG) to assess gains in accuracy and generalization. Finally, model interpretability will be examined via attention maps to highlight facial regions most relevant to classification. Results will be evaluated using standard metrics—accuracy, F1-score, and confusion matrix—demonstrating the potential of ViTs for emotion recognition.

1. Introduction

Emotion recognition from facial expressions is a key component in the development of human–computer interaction systems, with applications in mental health, marketing, and intelligent surveillance. This technology—often termed Emotion AI—has advanced substantially with deep learning. Although convolutional neural networks (CNNs) have achieved notable results in this area, they still exhibit limitations in complex settings, especially in the presence of ethnic variability, non-uniform illumination, and data noise.

To address these challenges, this work explores Vision Transformers (ViT), introduced by Dosovitskiy et al. (2021) in An Image is Worth 16×16 Words [1]. Unlike convolutional networks that operate locally on image regions, ViTs divide an image into fixed-size patches and treat them as token sequences, preserving spatial structure through positional embeddings. This allows the model to capture global relationships among different facial regions—particularly beneficial for emotion recognition, where fine-grained cues such as facial muscle tension and lip curvature are decisive.

Through a self-attention–based architecture, ViTs can identify the most informative areas of an image for classification, improving not only accuracy but also model interpretability. In emotion recognition, this provides a critical advantage over traditional CNNs, which tend to focus on local features without fully accounting for broader relationships among facial parts.

This paper presents the fine-tuning of a Vision Transformer (ViT) on the FER-2013 dataset and analyzes its performance relative to traditional CNNs. To enhance generalization, we apply regularization and data-augmentation techniques, together with pre-processing strategies to mitigate class imbalance. We further implement optimization procedures with particular attention to imbalance handling through an adaptive loss designed to promote more balanced learning.

Lastly, we conduct an attention-map analysis to identify the facial regions most influential in model decisions, thereby improving interpretability and deepening our understanding of the classification process. The remainder of the paper is organized as follows: Section 2 details the methodology; Section 3 presents the experimental analysis; Section 4 reviews the results; Section 5 concludes with a final discussion and directions for future work..

2. Materials and Methods

This section describes the dataset, pre-processing operations, and balancing strategies, as well as the architecture and configuration of the Vision Transformer (ViT) employed for facial-emotion recognition. We also outline the optimization strategies used to improve performance.

2.1. Dataset

We adopt the public FER-2013 dataset (available at:

dataset_FER-2013), widely used for facial-expression–based emotion recognition. Introduced in the "Challenges in Representation Learning: Facial Expression Recognition Challenge" (ICML 2013), FER-2013 is a de facto benchmark containing ~30,000 RGB images. Images are down-sampled to 48×48 pixels and labeled into seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The label distribution is notably skewed: some classes (e.g., happiness, sadness, anger) have nearly 5,000 samples, whereas others (e.g., disgust) are under-represented with ~600 images. Image quality also varies considerably due to contrast differences and occlusions—factors that can hinder learning. Consequently, the dataset undergoes careful pre-processing before training.

2.2. Data pre-processing

During pre-processing, images are resized and normalized to match ViT requirements, which split images into fixed-size patches. A central goal is to mitigate issues arising from low image quality and class imbalance. In particular, the "Disgust" class is augmented via enhanced oversampling using diverse geometric and intensity transformations to synthesize new instances. Randomized transformations markedly increase the number of "Disgust" samples, reducing imbalance and improving model generalization to real-world conditions. These operations are combined with limited under-sampling of majority classes. Guided by findings in Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets [2], sampling bounds were tuned to avoid overfitting, bias, and excessive loss of data diversity (kept between 40% and 60%). The dataset is split into training, validation, and test subsets. Approximately 80% of samples labeled "Training" are used for training, whereas "PublicTest" and "PrivateTest" (about 10% each) serve for validation and final evaluation, respectively. This partitioning ensures a large training base while keeping validation and test sets fully independent for a reliable assessment of generalization.

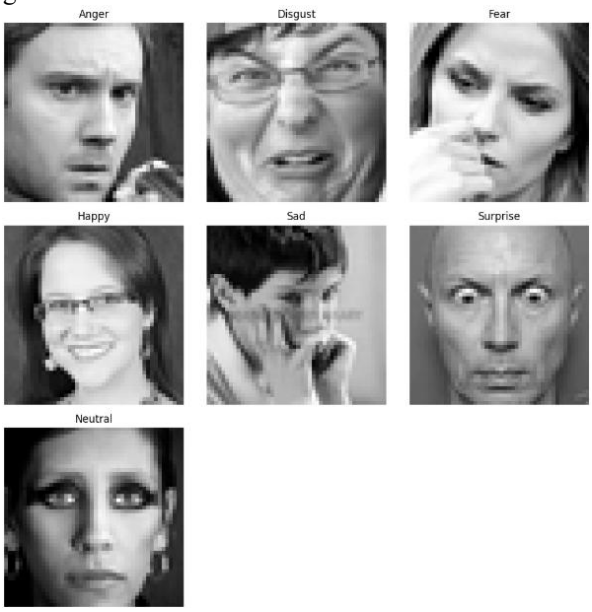

Figure 1: one example per class from FER-2013 dataset

2.3. Model

Our model is based on Vision Transformers, introduced by Dosovitskiy et al., An Image is Worth 16×16 Words (ICLR 2021) [1]. In contrast to traditional CNNs, ViTs split an image into fixed-size patches, treat them as token sequences, and use self-attention to capture global relationships and complex interactions among image regions. A special [CLS] token is prepended; its final embedding is used for classification. Positional encodings are added to each embedding to retain spatial information.

The embedding sequence (patches + [CLS]) passes through a stack of Transformer blocks (self-attention and feed-forward layers), enabling the capture of global

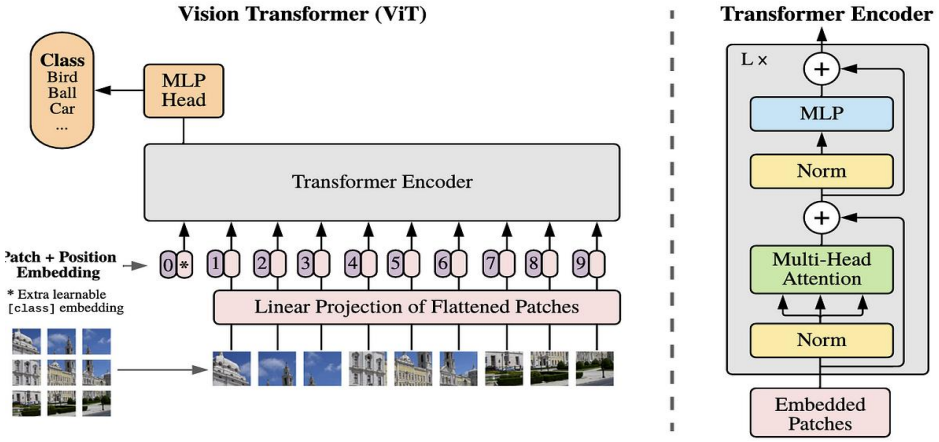

Figure 2: Vision Transformer (ViT) model architecture

relations. This is particularly advantageous for emotion recognition, where subtle cues such as lip curvature variations or periocular muscle tension matter.

For transfer learning, we select the pre-trained vit-base-patch16-224-in21k. This choice leverages powerful feature representations learned from large-scale pre-training and an architecture that models spatial relations via positional embeddings and self-attention. The model was pre-trained on ImageNet-21k (~14 M images, ~21k classes), providing rich visual priors and improved generalization to downstream tasks. To mitigate overfitting, we modify the final classifier: the dimensionality is reduced from 768 to 256 neurons, followed by dropout and a final linear layer producing logits for the seven emotion classes. This configuration balances learning capacity and generalization, yielding competitive performance on training, validation, and test data. followed by dropout and a final linear layer producing logits for the seven emotion classes. This configuration balances learning capacity and generalization, yielding competitive performance on training, validation, and test data.

3. Experiments

We present an experimental analysis to evaluate the effectiveness of our optimization strategies and hyperparameter configurations. We compare four approaches: a baseline, two fine-tuning modes (partial and full), and a warm-up strategy with a learning-rate scheduler. Each configuration is assessed using quantitative metrics (accuracy and loss).

3.1. Optimization Strategies

We adopt a set of targeted optimization strategies to address dataset limitations and improve model generalization. Data augmentation, including rotations, translations, and color perturbations, is applied during pre-processing to increase dataset diversity, reduce overfitting, and enhance robustness to real-world variability. For optimization, we evaluate different algorithms (SGD, Adam, AdamW) and select AdamW as the most suitable for Transformer architectures, ensuring stability through the decoupling of weight decay from gradient updates. Learning-rate scheduling combines Step Decay, Cosine Annealing (particularly effective during warm-up), and ReduceLROnPlateau to support controlled initialization and fine convergence.

Regularization plays a central role: weight decay, dropout within the classifier, and early stopping are applied to mitigate overfitting and improve generalization. To handle class imbalance, we employ complementary techniques such as class-weighted losses and moderate sampling, discarding SMOTE due to limited effectiveness and data-integrity concerns. Loss functions are also carefully chosen, starting from Cross-Entropy and extending to Focal Loss, which prioritizes harder-to-classify examples.

Together, these strategies provide a robust and generalizable model capable of coping with the heterogeneity and imbalance of the FER-2013 dataset.

3.2. Hyperparameter Selection

Hyperparameter selection was central to robust optimization and generalization. For dataset rebalancing, we applied under- and over-sampling: majority classes were reduced using a max_class_limit between 35% and 60% (preferably ~40%) to preserve salient information, while minority classes were oversampled with a factor between 1.5 and 3.0 (commonly 2×), providing controlled increases without undue overfitting. Data augmentation included horizontal flips with probability 0.5, rotations up to ±15°, brightness/contrast adjustments (factors up to 0.25), and Gaussian blur with $\sigma \in [0.1, 1.0]$, enhancing diversity and robustness.

For optimization, we use AdamW, adopting a fine-tuning learning rate in the range 3e-5 to 0.01; overly large values risk divergence, whereas overly small values slow learning. Weight decay is set to 0.02; the scheduler uses a step size between 5 and 10 epochs with gamma = 0.1 to progressively reduce the learning rate for fine convergence. Regularization includes dropout at 30% and early stopping with patience between 2 and 5 epochs. Focal Loss ($\gamma = 2.0$, $\alpha \in [0.25, 1]$) is considered to further address minority classes, though balancing strategies reduce the need for heavy re-weighting.

We compared batch sizes of 32 and 64. Despite the theoretical stability of larger batches, batch 64 tended to converge to less favorable local minima, increased memory usage, and lengthened training without clear gains. We therefore adopt batch 32, which balances stability, efficiency, and generalization.

We tested 10–15 total epochs (subject to early stopping), balancing GPU usage and learning stability while ensuring reproducibility and periodic checkpointing. This careful hyperparameter search produced a model that balances accuracy, robustness, and generalization, effectively addressing dataset imbalance and task complexity.

3.3. Experiments Setup

Within the project, four experiments were conducted, each contributing in a specific way to adapting the model to the FER-2013 dataset:

Baseline: This configuration employs a pre-trained model in which the final classification layer is replaced with one tailored to FER-2013. During training, only this new layer is updated while the weights of all other layers remain frozen. This approach leverages the features already learned by the model without altering its internal structure,

providing an initial benchmark for evaluating the effectiveness of subsequent strategies.

Partial Fine-tuning: In this experiment, only certain portions of the network (typically the higher layers) are updated, while the representations learned in earlier stages remain frozen. This approach exploits the model's prior knowledge and reduces the risk of overfitting when the new task is closely related to the original pre-training task.

Full Fine-tuning: Unlike partial fine-tuning, here all model weights are updated. This strategy allows for deeper adaptation to the specific characteristics of the dataset, improving performance in scenarios where merely adjusting the final layers is insufficient to capture data complexity.

Warm-up Stage with Learning Rate Scheduler: This experiment introduces an initial warm-up phase, during which the learning rate is gradually increased, followed by a controlled decay using a scheduler. This approach enables more stable convergence, avoiding instabilities in the early stages of training and refining the model's ability to reach optimal minima.

These experiments were designed to identify the most effective strategy for balancing robustness and accuracy, providing a comprehensive overview of the impact of different adaptation methods.

Modello	Accuracy
ViT	**64.02%**
ResNet18	44.73%
VGG16	**33.87%**

Table 1: Mean Accuracy across models

4. Results

Facial-expression–based emotion recognition is an important challenge in domains such as human–computer interaction and mental health. We explored Vision Transformers (ViT) as an alternative to traditional CNNs (ResNet, VGG), showing how ViTs can overcome several CNN limitations in complex contexts marked by ethnic variability, uneven lighting, and noisy data. By leveraging self-attention, ViTs capture global relationships within images and substantially improve recognition of subtle cues such as micro-expressions.

4.1. Evaluation Metrics

We adopt standard metrics: accuracy, F1-score, and confusion matrix. The F1-score—the harmonic mean of precision and recall—is particularly appropriate for the class imbalance in FER-2013. Precision quantifies the fraction of predicted positives that are correct, while recall measures the fraction of actual positives captured. For each class, these values are derived from the confusion matrix, which details correct predictions and errors and reveals correlations between true and predicted labels. We also report macro and weighted averages for an overall perspective: macro averages treat classes equally, whereas weighted averages account for class distribution. Finally, attention maps identify facial regions that most influence decisions, adding an interpretability layer to our analysis.

Classe	ViT	ResNet18	VGG16
Angry	**53.8%**	16.3%	5.6%
Disgust	**69.5%**	48.6%	41.0%
Fear	**33.6%**	15.1%	14.5%
Happy	**84.1%**	77.2%	48.3%
Sad	**54.6%**	21.8%	6.8%
Surprise	**78.1%**	68.8%	72.9%
Neutral	**68.2%**	52.3%	50.3%

Table 2: Per-class accuracy of the models

4.2. Comparison Models

In this study, alongside the Vision Transformer (ViT), we employed ResNet-18 and VGG-16 as baseline models for comparison. These widely used and well-established architectures were selected for their ability to deliver competitive performance in image-classification tasks and for their role as representative examples of classical approaches: ResNet-18, with its residual connections, enables effective training even in relatively lightweight architectures, while VGG is renowned for its simplicity and robustness in visual feature extraction.

The inclusion of these models ensures a fair comparison between the innovative ViT-based approach and traditional solutions, under consistent experimental conditions. Specifically, all comparison architectures were trained on the same appropriately balanced FER-2013 dataset, using a standard configuration consisting of cross-entropy loss as the cost function, Stochastic Gradient Descent (SGD) as the optimizer, and StepLR as the learning rate scheduler. This uniformity in settings guarantees a direct performance comparison and enables a thorough analysis of the impact of different architectures on emotion recognition.

4.3. Final Model Configuration

In the final configuration, the classifier is a sequence comprising: dropout ($p = 0.3$), a linear layer reducing the pre-trained output dimensionality (typically 768) to 256 to curb overfitting, a ReLU activation, another dropout ($p = 0.3$), and a final linear layer mapping 256 units to the 7 classes. Image pre-processing employs the

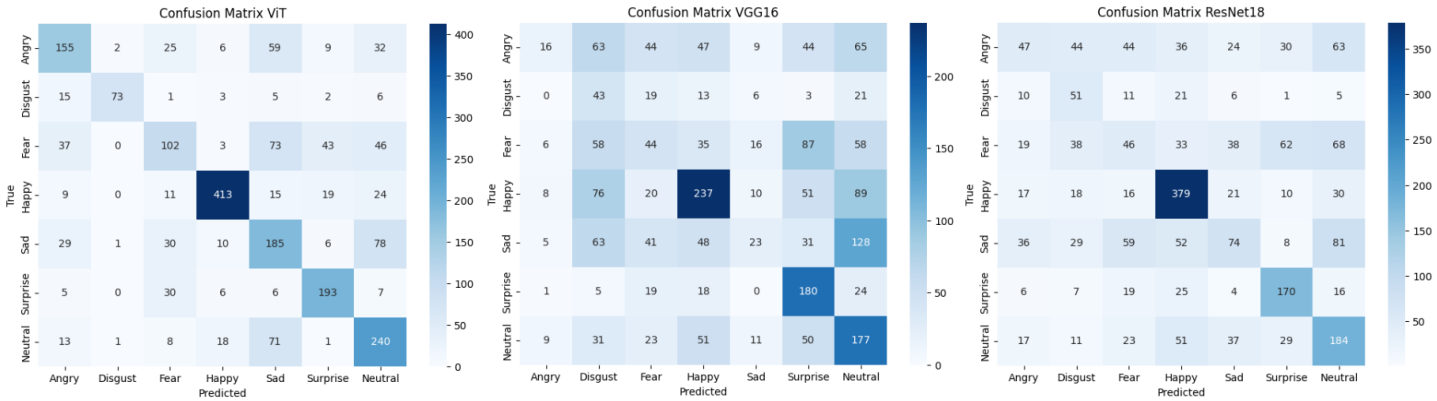

Figure 3: Confusion matrices of the models under study

ViTImageProcessor associated with the pre-trained model. For FER-2013, we set the oversampling limit to 2× and under-sampling to 40%. Class weights are computed with the balanced method and integrated into the loss. We use Focal Loss ($\alpha = 0.25$, $\gamma = 2$) to emphasize difficult examples and counter class imbalance. Optimization uses AdamW with an initial learning rate of 1e-5 and weight decay of 1e-4 (L2 regularization). Learning-rate scheduling adopts a 3-epoch warm-up (multiplier = 10) followed by ReduceLROnPlateau monitoring validation loss. Training runs for up to 15 epochs with early stopping (patience = 3), yielding stable convergence and a good balance of accuracy and generalization.

4.4. Result Analysis

During training, models used different losses aligned with their architectures: Cross-Entropy for VGG-16 and ViT, and Focal Loss for ResNet-18 to handle imbalance. For fair comparison, all models were evaluated with Cross-Entropy at test time so that accuracy, F1-score, and confusion matrices are computed under uniform conditions.

Experimental results show that ViT consistently outperforms CNN-based approaches such as ResNet-18 and VGG-16, achieving an overall accuracy of 64.02% vs 44.73% and 33.87%, respectively (Table 1). Per-class analysis (Table 2) indicates that ViT is superior in almost all categories, notably "Happy" (84.1%) and "Surprise" (78.1%), suggesting a greater ability to capture expressive nuances typical of positive emotions. By contrast, ResNet-18 and VGG-16 show larger variability and pronounced difficulties especially in "Angry" and "Fear." These findings indicate that the Transformer's global self-attention models spatially distributed complex features more effectively than CNNs, which appear less suited to capture such dynamics in these data.

Confusion-matrix analysis (Figure 4) corroborates these observations: ViT reduces misclassifications between semantically close categories, whereas ResNet-18 and VGG-16 exhibit more dispersed errors, with frequent confusions between, for instance, "Fear" and "Neutral" or "Disgust" and "Angry." While results are promising, some failure cases remain: ViT underperforms on "Disgust" and "Fear," likely due to limited training instances and resulting generalization issues. Despite balancing efforts, oversampling "Disgust" may still induce overfitting. Future improvements could include stronger augmentation, ensemble architectures, or using Generative Adversarial Networks (GANs) to synthesize additional images.

4.5. Attention Maps

Within the project, attention map generation was carried out using the Rollout method. Attention Rollout is a technique that produces a global attention map, highlighting the relative importance of different patches in the image. Specifically, during the forward pass of the Vision Transformer (ViT), attention matrices from each layer are extracted, to which residual connections are added in order to preserve the original information. These matrices are then normalized and cumulatively propagated, making it possible to visualize how information spreads and aggregates across the various layers of the model. The final result-optionally upsampled to match the dimensions of the original image-yields a map that emphasizes the contribution of each patch to the final decision. This technique is essential, as it provides a clear visual interpretation of how the model processes and integrates facial information, thereby enhancing understanding of the model's internal functioning and helping to identify potential biases in its decisions.

As shown in Figure 4, the analysis reveals that the model not only offers an interpretable view of its decision-making process but also confirms that it focuses primarily on the key facial regions relevant to emotion analysis. In particular, the model exhibits strong attention to the mouth and eye areas, which are crucial for decoding emotional expressions. The presence of some highlights in peripheral

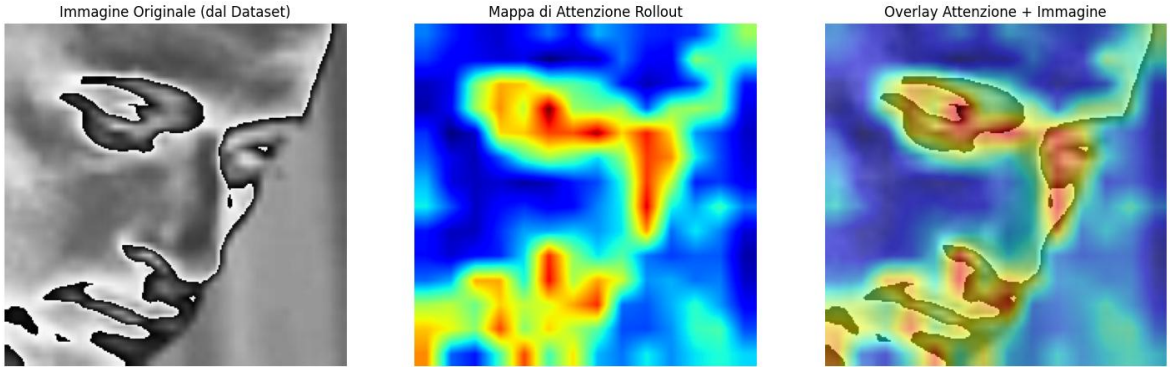

Figure 4: Attention map of the ViT model

areas may indicate slight noise or distractions, suggesting the need for further refinements in the architecture or training process. Ultimately, these attention maps not only validate the model's reliability in emotion recognition but also provide valuable insights for identifying and addressing potential areas for improvement.

5. Conclusions and Future Work

Facial expression recognition is a key challenge in domains such as human–computer interaction and mental health. This study explored Vision Transformers (ViT) as an alternative to CNNs (ResNet, VGG), demonstrating how self-attention enables ViTs to capture global image relationships and better recognize subtle cues such as micro-expressions, even under challenging conditions of variability and noise.

5.1. Results

The results obtained in this study clearly highlight the advantage of Vision Transformers (ViT) over traditional CNN-based architectures, such as ResNet-18 and VGG-16, in the task of facial emotion recognition. The ViT model achieved an overall accuracy of 64.02%, significantly higher than the 44.73% of ResNet-18 and the 33.87% of VGG-16, as shown in Table 1. Per-class performance analysis (Table 2) further confirmed that ViT outperformed in almost all categories, with particularly strong results in the Happy (84.1%) and Surprise (78.1%) classes, indicating a greater ability to capture the expressive nuances of positive emotions.

Conversely, CNN architectures such as ResNet-18 and VGG-16 exhibited clear difficulties in recognizing certain emotions, particularly in the Angry and Fear classes, where their performance was significantly lower than that of ViT. These findings suggest that, owing to its global self-attention mechanism, ViT is better able to model the spatial and complex relationships among facial features, whereas CNNs tend to struggle in capturing such dynamics, making them less effective for emotion recognition.

Moreover, the analysis of the confusion matrices (Figure 4) further confirmed the effectiveness of ViT in enhancing class discrimination. By attending to global image features, the model significantly reduced misclassification errors, particularly among semantically similar classes. In contrast, the confusion matrices of ResNet-18 and VGG-16 showed greater misclassification overlap, with frequent prediction swaps between classes such as Fear and Neutral or Disgust and Angry, indicating that CNNs are more prone to confusing emotions with similar characteristics.

5.2. Future Developments

Although the proposed approach achieved promising results, several directions remain open for further improvement. First, adopting more aggressive data augmentation strategies and exploring different Vision Transformer (ViT) variants could strengthen model robustness and provide valuable insights into the impact of architectural choices on facial emotion recognition. Additionally, employing datasets with more balanced class distributions would enable a fairer evaluation of the models' ability to generalize, reducing bias toward overrepresented emotions.

Another promising avenue involves ensemble learning and the use of Generative Adversarial Networks (GANs) to generate synthetic facial expressions. Such techniques could enrich dataset diversity and enhance the model's capacity to generalize in real-world scenarios.

Moreover, adapting the model to cultural and environmental variability in facial expressions remains a key objective for ensuring broader applicability and global deployment.

References

[1] Dosovitskiy, A., & Shlens, J. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*.

[2] Gajendran, V., Somasundaram, A., & Kumar, S. (2021). Comparative analysis of Vision Transformer models for facial emotion recognition using augmented balanced datasets. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.