Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Fine - Tuning and Optimization of Live2D Facial Expression Recognition Based on Vision Transformer (ViT)

Ziyang Chen

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangzhou, China

u04zc22@abdn.ac.uk

Abstract. With the development of fields such as virtual reality, the expressiveness of virtual character facial expressions has become increasingly crucial. Traditional CNN - based facial expression recognition methods have problems such as limited local feature extraction, sensitivity to pose changes, and high computational complexity. This research is based on the Vision Transformer (ViT) model, exploring its optimization and application in the facial expression recognition task, and combining with Live2D technology to achieve real - time and efficient expression conversion. ViT can capture global information through the self - attention mechanism, better model facial expression changes, has the potential for transfer learning, and is suitable for low - resource devices. The FER - 2013 dataset was used in this study. The ViT model was fine - tuned and quantized, and compared with traditional CNN models such as ResNet50. Experiments show that the optimized ViT model has higher accuracy and real - time performance in facial expression recognition, and can also operate efficiently in low - resource environments. The quantization technology also reduces the computational overhead, making it suitable for economical consumer software. This research enhances the application value of ViT in the field of affective computing, provides support for the development of Live2D technology, and promotes the wide application of virtual characters. In the future, the potential of ViT in multimodal emotion recognition will be explored, and its performance in complex scenarios will be optimized to provide a more comprehensive solution for intelligent interactions of virtual characters.

Keywords: Vision Transformer, Facial Expression Recognition, Live2D Technology, Quantization Optimization, Low-Resource Devices.

1. Introduction

Facial Expression Recognition (FER) is an important research direction in the field of affective computing, with wide application value in human - computer interaction, virtual reality, and the entertainment industry. In recent years, with the rise of virtual YouTubers (VTubers), the demand for real - time capturing and mapping of real - person facial expressions to virtual character animations has been increasing, which puts forward higher requirements for FER technology. Traditional Convolutional Neural Networks (CNNs) have achieved remarkable results in FER tasks. For example, ResNet proposed by He et al. solved the problem of gradient disappearance in the training of deep networks through residual connections and achieved an accuracy of approximately 73% on the FER - 2013 dataset [1]. VGGNet proposed by Simonyan and Zisserman further enhanced the feature extraction ability through a deeper network structure, but its high computational cost limited its use in real - time applications [2]. In addition, Aditya achieved good performance in the FER task by fine - tuning MobileNetV2, demonstrating the potential of lightweight models in FER [3]. However, the limitations of CNN in capturing global dependencies make its performance unsatisfactory when dealing with complex facial expressions.

To overcome the limitations of CNN, the Vision Transformer (ViT), as an emerging deep - learning architecture, models global dependencies through the self - attention mechanism, providing a new solution for the FER task. ViT proposed by Dosovitskiy et al. has proven its superiority in image recognition tasks, but its high computational requirements limit its feasibility in real - time applications [4]. In addition, the Transformer architecture proposed by Vaswani et al. provides a

theoretical basis for the design of ViT, but its application in the FER task still needs further optimization [5]. Research by Nugroho et al. shows that the FER - 2013 dataset is of great value in the FER task, but its scale and data diversity still need to be expanded to improve the robustness of the model [6]. Chu and Liu improved the FER performance through a multi - task learning method combining key - point detection and emotion recognition, but the increased computational burden limited its use in real - time applications [7]. The potential of GAN proposed by Karras et al. in image synthesis provides new ideas for FER data augmentation, but the instability of its training process limits its practical application [8-10].

This study aims to optimize the facial expression recognition performance of ViT in virtual YouTuber applications and combine with Live2D technology to achieve real - time mapping from real facial expressions to virtual character animations. Through fine - tuning and quantization techniques, this study will explore the optimization methods of ViT in the FER task and compare it with traditional CNN models such as ResNet50, MobileNetV2, and VGGNet. The expected results of the study show that the optimized ViT model can achieve high accuracy in the FER task while reducing computational requirements, making it suitable for real - time VTuber application scenarios. Through this study, not only a new technical path is provided for the FER task, but also technical support is provided for the intelligent interaction of virtual characters, promoting the improvement of the expressiveness of virtual characters in virtual YouTubers, games, and social software.

2. Research Methods

The goal of this study is to explore the application of the Vision Transformer (ViT) model in the facial expression recognition task through fine - tuning and optimization and compare it with traditional Convolutional Neural Networks (CNNs).

2.1. Dataset Introduction and Preprocessing

This study used the FER - 2013 dataset, which is one of the commonly used public datasets in the field of emotion recognition. The FER - 2013 dataset is a public dataset provided by Kaggle, originally collected by Microsoft Research for emotion recognition tasks. This dataset contains 35,887 grayscale images of different facial expressions, with an image size of 48x48 pixels. The image annotations include seven different emotion labels: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral. The dataset is divided into a training set, a validation set, and a test set, as shown in Figure 1:

Training set: Approximately 80% of the data, containing 28,709 images, covering 7 emotion labels (Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral).

Validation set: Approximately 10% of the data, containing 3,589 images, used for model tuning.

Test set: Approximately 10% of the data, containing 3,589 images, used for final model evaluation.

After data preprocessing, the image pixel values were normalized to the [0, 1] interval, and data augmentation techniques (such as random rotation, flipping, etc.) were applied to improve the generalization ability of the model.

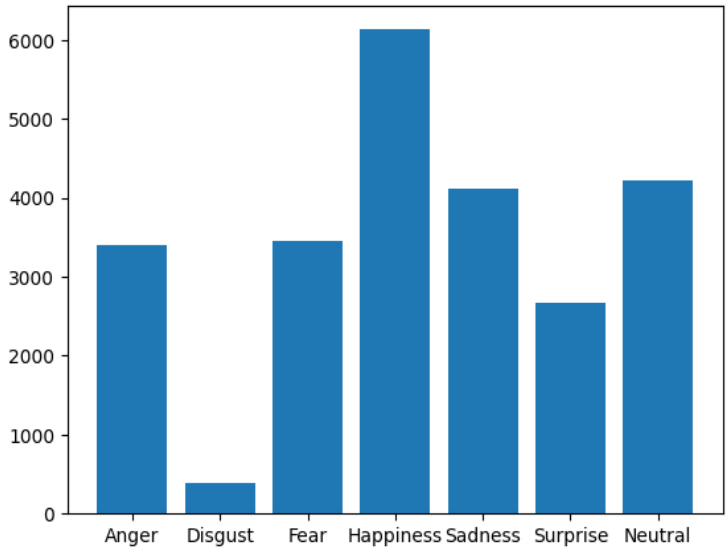

Figure 1. The number of images for each emotion in the training set (Picture credit: Original)

2.2. Method Introduction: Vision Transformer (ViT)

2.2.1. Model Overview

The Vision Transformer (ViT) is a deep - learning model based on the self - attention mechanism. Different from traditional Convolutional Neural Networks (CNNs), it does not rely on local receptive fields but models global features of the entire image through the Transformer architecture. Figure 2 shows the core idea of the ViT model. The image is divided into fixed - size Patches, and then these Patches are converted into a sequence input, similar to the Token sequence in Natural Language Processing (NLP), and feature extraction is performed through the Transformer.

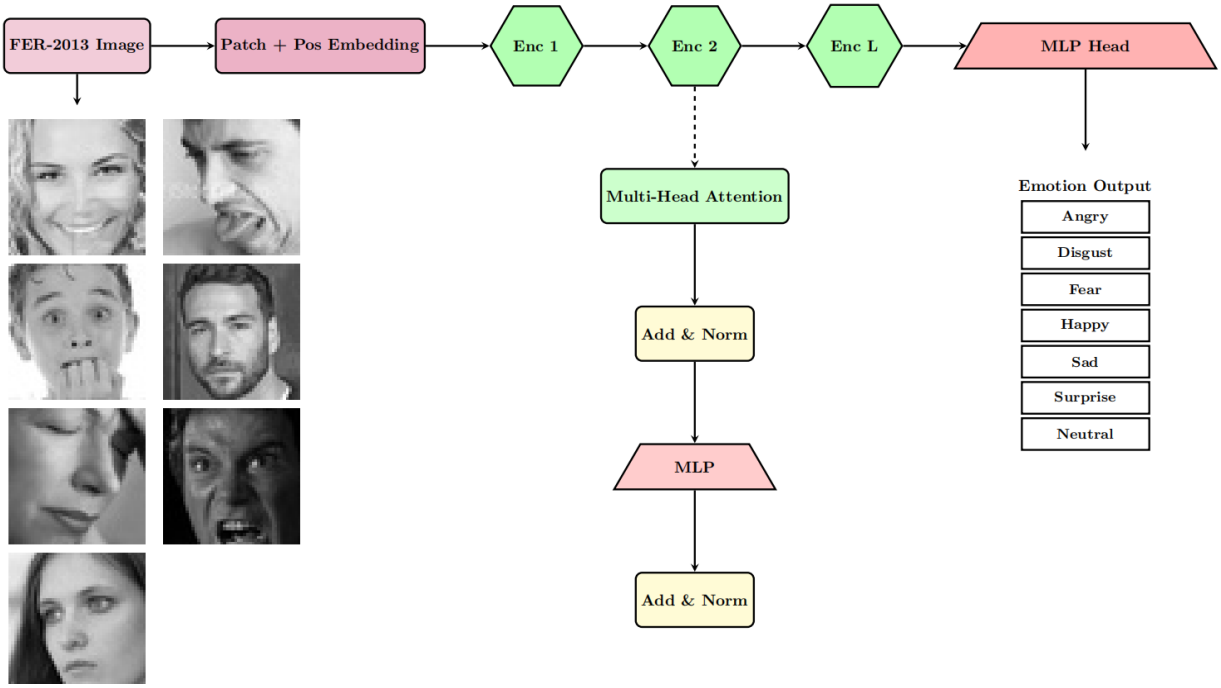

Figure 2. Overview of the ViT Model (Picture credit: Original)

2.2.2. Core Steps of ViT

Image Patch Embedding: The input image X is divided into fixed - size Patches (for example, 16×16).Each Patch is flattened and mapped to a fixed - dimension vector through a linear projection layer.

$$X_p = \text{Reshape}(X) \cdot W_p + b_p \quad (1)$$

Adding Position Embedding: Since the Transformer itself does not have spatial information, ViT adds position information to the Patch sequence through positional encoding.

$$PE(i, 2j) = \sin(i/10000^{2j/d}) \quad (2)$$

$$PE(i, 2j + 1) = \cos(i/10000^{2j/d}) \quad (3)$$

Transformer Encoder: A standard multi - head self - attention mechanism (Multi - Head Self - Attention, MSA) + feed - forward neural network (Feed - Forward Network, FFN) is adopted.

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

Classification Head: In the output sequence of the Transformer, a special classification token (CLS Token) is taken for classification:

$$y = \text{MLP}(\text{CLS Token Output}) \quad (5)$$

The final classification is performed through the MLP Head (multi - layer perceptron).

2.3. Research Process

2.3.1. Model Training

The model training used the cross - entropy loss function and was trained with the Adam optimizer (learning rate 1e - 4). To avoid overfitting, the Dropout regularization technique was adopted, and the early - stopping strategy was used during the training process to ensure that the model performed best on the validation set.

2.3.2. Model Optimization: Quantization

To improve the efficiency of the model and accelerate the inference process, this experiment used quantization technology for optimization after model training. The floating - point weights were converted into low - precision integers, reducing memory occupation and accelerating the inference process. The quantized model will operate efficiently in low - resource environments.

2.3.3. Performance Evaluation and Comparative Analysis

To verify the advantages of the ViT model in facial expression recognition, this study compared its performance with traditional Convolutional Neural Networks (CNNs) in the following indicators.

Accuracy: Measures the proportion of correct classifications of the model on the test set.

Precision: Evaluates the prediction accuracy of the model for each expression category.

Model Size: Evaluates the size of the optimized model to ensure its suitability for resource-constrained devices.

3. Experimental Results

3.1. Comparison of Test Set Accuracy

On the test set of the FER-2013 dataset, this study evaluated the classification performance of the Vision Transformer (ViT) and traditional Convolutional Neural Networks (CNNs) respectively. The experimental results show that the ViT model has a higher accuracy in the facial expression recognition task than traditional CNNs, as shown below:

Test set accuracy of the ViT model: Approximately 69.3%

Test set accuracy of the CNN model: Approximately 60.8%

As shown in Table 1 and Figure 3, the ViT model outperforms traditional CNNs in overall accuracy, indicating its advantages in feature extraction and global information modeling, making it more competitive in the facial expression recognition task.

Table 1. Accuracy of the training set and validation set

Epoch	Training Loss	Validation Loss	Accuracy
1	0.993100	0.981259	0.642519
2	0.814800	0.914304	0.669824
3	0.628500	0.916631	0.680134
4	0.470400	0.934855	0.693508
5	0.311600	1.019010	0.689886
6	0.222800	1.080805	0.686542

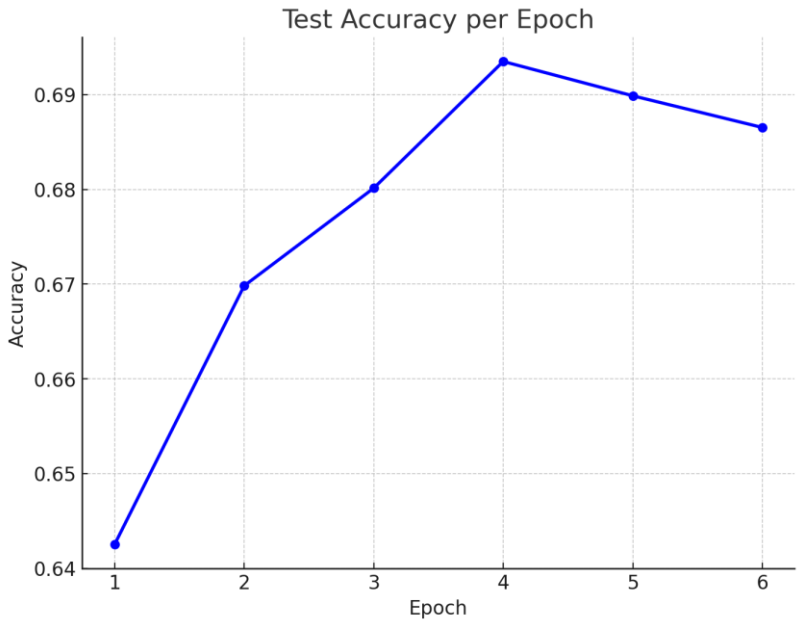

Figure 3. Test Set Accuracy Curve (Picture credit: Original)

3.2. Analysis of the Test Set Confusion Matrix

To further analyze the classification effect of the model, this study drew a confusion matrix for the test set results of ViT. Figure 4 can visually display the classification effect of the model on different types of expressions and reveal possible misclassification situations of the model.

ViT performs well on some emotions (such as happiness and neutral), but there are obvious misclassifications in some similar emotions (such as fear and sadness, anger and sadness).

The classification accuracy of disgust is the lowest (only 21 correct classifications), indicating that the model's recognition of this category is weak.

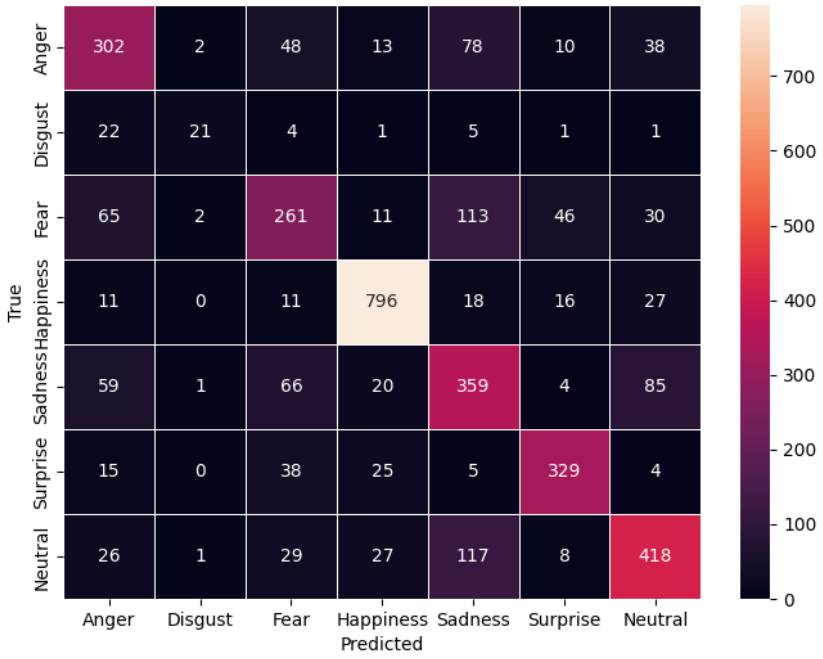

Figure 4. Confusion Matrix of the Test Set (Picture credit: Original)

3.3. Comparative Analysis of Model Sizes

In this experiment, 4-bit quantization was performed using bitsandbytes to reduce the storage requirements and inference computational load of the model. Through quantization, the size of the model was compressed from 327.38 MB to 84.43 MB, achieving a compression ratio of 3.9 times, greatly reducing the storage overhead.

From the results:

Reduced storage requirements: Compared with the unquantized ViT, the storage space requirements of the 4 - bit quantized model are reduced by approximately 75%, making it suitable for storage - constrained devices (such as mobile devices and embedded devices).

Inference acceleration: Since 4 - bit quantization reduces the computational complexity, the inference speed can be further increased.

Suitable for lightweight deployment: The compressed ViT is more suitable for economical consumer software and can still perform efficient inference under limited computing resources.

3.4. Result Analysis

ViT can better learn the global features of expression images through the self-attention mechanism, while CNN mainly relies on local feature extraction, which may lead to low discrimination between some similar categories.

The experimental results show that the ViT model outperforms traditional CNNs in the facial expression recognition task, with better performance in terms of accuracy, classification ability, and model size, further verifying the application potential of ViT in the field of facial expression recognition.

4. Conclusion

This study focuses on optimizing the performance of the Vision Transformer (ViT) in the Facial Expression Recognition (FER) task and applying it to the virtual YouTuber (VTuber) system. Through fine - tuning and quantization techniques, the ViT model shows high recognition accuracy on the FER - 2013 dataset, and significant improvements in computational efficiency and real - time

performance. Compared with traditional Convolutional Neural Network (CNN) models (such as ResNet50 and MobileNetV2), ViT has obvious advantages in capturing global dependencies, especially when dealing with complex facial expressions. The experimental results show that the optimized ViT model can not only meet the requirements of real - time applications but also operate efficiently on low - resource devices, providing a feasible solution for the practical implementation of affective computing technology.

There are several promising directions following the findings of this research. First, exploring hybrid models could yield even better results – for instance, combining CNN-based feature extractors for low-level details with transformer layers for global reasoning might capture the best of both worlds. Some researchers have reported that such hybrid CNN-Transformer models improve the accuracy and stability of FER. Implementing a hybrid on FER2013 could potentially push accuracy beyond what either alone achieves. Second, addressing data limitations of FER2013 is important. Data augmentation strategies or training on a curated, balanced subset (as done in some works) could help the CNN perform better and also benefit the ViT.

Future research can also further explore virtual YouTuber (VTuber) applications and technical optimizations, including improving the facial expression recognition ability of the ViT model in complex lighting, pose change, and occlusion scenarios, optimizing the real - time mapping algorithm to achieve smoother virtual character animations, further reducing the model's computational overhead by combining lightweight technologies (such as pruning and quantization), exploring distributed computing strategies to support large - scale concurrent processing, and developing more efficient training and inference frameworks, so as to provide more accurate, efficient, and real - time facial expression recognition and animation generation solutions for VTuber systems.

References

[1] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778 (2016).

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR) (2015).

[3] Q. Aditya. Fer-2013 mobilenet: Fine-tuning mobilenetv2 for facial expression recognition. https://github.com/qwerty-aditya/FER2013-MobileNet (2023).

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR) (2021).

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998–6008 (2017).

[6] A. A. Nugroho et al. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), pages 277–283. IEEE (2020).

[7] W. J. Chu and Y. B. Liu. Thermal facial landmark detection by deep multi-task learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4401–4410 (2019).

[9] Z. T. Shen. A comparative study of hybrid cnn and vision transformer models for facial emotion recognition. In 2024 11th International Conference on Dependable Systems and Their Applications (DSA), pages 401–408 (2024).

[10] S. Bobojanov, B. M. Kim, M. Arabboev, and S. Begmatov. Comparative analysis of vision transformer models for facial emotion recognition using augmented balanced datasets. Applied Sciences 13(22) (2023).