Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

Facial Expression Recognition of Students in Classroom Using Hybrid MobileNetV3-Vision Transformer with Token Downsampling

Mochamad Khaairi[1*], Rasim[2], Yaya Wihardi[3]
[1,2,3]Universitas Pendidikan Indonesia, Indonesia
[1]mokha30@upi.edu, [2]rasim@upi.edu, [3]yayawihardi@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 02-07-2025
Accepted: 18-07-2025
Published: 24-07-2025

Keywords:
Classroom; Facial Expression Recognition; Hybrid Vision Transformer; MobileNetV3; Student; Token Downsampling

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

In large classroom environments, teachers often struggle to monitor each student's facial expression throughout the learning process. Yet, facial expressions are important indicators of students' emotional states and engagement, which, when detected in real time, can support a more adaptive learning experience. Most previous research on Facial Expression Recognition (FER) has relied on Convolutional Neural Networks (CNN), which tend to be limited in capturing global relationships between facial features. Additionally, many studies focus on model accuracy without evaluating their practical effectiveness in real classroom settings. This study aims to develop a facial expression recognition model that is both accurate and efficient for use in classroom contexts. A hybrid Vision Transformer (ViT) architecture is proposed, which combines MobileNetV3 for local feature extraction and a Vision Transformer for global context modeling. To reduce the number of tokens and computational cost, a Token Downsampling method is introduced within the transformer blocks. The model is trained using the FER2013 dataset and achieves a test accuracy of 71.24%, surpassing the baseline pretrained ViT model, which reached only 70.10%. Additionally, the Token Downsampling method improves inference speed. Furthermore, the model is tested on a custom dataset collected from students in a real classroom setting to evaluate its performance in practical implementation. Although the performance on the classroom dataset is not yet optimal, the results on FER2013 demonstrate the potential of this approach for further development toward real-time and accurate facial expression recognition in educational environments.

INTRODUCTION

Emotions play a crucial role in the learning process as they influence students' motivation, interest, and academic outcomes. Positive emotions can enhance engagement, while negative emotions such as anger or anxiety may disrupt cognitive processes and hinder learning performance (Bouhlal et al., 2020). Students' facial expressions can serve as the most accurate indicators of emotional states and provide valuable feedback for educators to adjust their teaching strategies accordingly (Fang et al., 2023). However, directly monitoring the facial expressions of all students is challenging, especially in large classrooms. Therefore, a systematic and objective approach is needed to support more adaptive and responsive teaching.

The rapid advancement in image processing has driven numerous studies on the development of Facial Expression Recognition (FER). In previous research related to recognizing students' facial expressions, Convolutional Neural Network (CNN) models have consistently been the primary choice. For instance, Lawpanom et al. (2024) developed an ensemble convolutional neural network model to detect students' facial expressions in real-time during online learning. However, despite demonstrating adequate performance, CNN-based models face limitations, specifically the convolutional filters in CNNs operate only within small local regions. This hinders the model's ability to learn structural dependencies between different facial components across most neural layers. As a result, CNN-based FER models can only capture local facial features without fully understanding the global context of facial expressions (Huang et al., 2021).

With the growing popularity of transformer architectures, the Vision Transformer (ViT) has emerged as an alternative for FER tasks. ViT is more effective at capturing global relationships between image regions due to its use of the attention mechanism (Dosovitskiy et al., 2021), and several studies have shown that ViT can outperform CNN-based models in certain scenarios. However, ViT also presents some drawbacks, including high computational complexity resulting from the self-attention mechanism, which scales quadratically with the number of input tokens (Marin et al., 2021). Additionally, the fixed-size patch approach employed by ViT tends to be suboptimal in capturing local features (Indolia et al., 2023).

To address the limitations of both CNN and ViT, Hybrid Vision Transformer architecture has been introduced. This architecture combines CNN for local feature extraction in the early stages with ViT for modeling global

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

relationships through self-attention in the subsequent stages (Khan et al., 2023). One promising combination is the integration of MobileNetV3 with ViT. MobileNetV3 is known for being lightweight and efficient, while also demonstrating strong capability in extracting local features. This makes it a suitable backbone for hybrid models aimed at real-world applications where computational resources are limited.

Although various model approaches have been developed, most previous studies have primarily focused on the development of classification architectures and have yet to conduct direct evaluations in real-world settings such as classrooms. This leads to a lack of understanding regarding how models perform under actual conditions, which tend to be more complex and varied, for example in terms of lighting conditions, facial poses, and more subtle expression intensities.

This study is significant not only because it focuses on developing an accurate and efficient FER architecture, but also because it directly evaluates the model's performance in real classroom environments. The proposed model is a hybrid of Vision Transformer and MobileNetV3, enhanced with a token downsampling strategy to reduce computational complexity. This approach is expected to improve the accuracy of student facial expression recognition. Furthermore, the study utilizes a facial expression dataset collected directly from classroom settings, making the model more relevant and adaptive to real-world learning conditions.

LITERATURE REVIEW

Advancements in computer vision have steered FER research toward deep learning methods. Convolutional Neural Networks (CNN), as a deep learning architecture, have become the primary choice for FER tasks due to their ability to achieve satisfactory accuracy. For instance, Dukić & Krzic (2022) developed SVM, EM-CNN, and pretrained ResNet34 models for facial expression classification, achieving accuracy rates of 59.36%, 65.57%, and 76.45%, respectively, on a combined dataset (CK+, FER2013, and SFEW). Lawpanom et al. (2024) proposed an ensemble CNN model by combining several CNN-based architectures, such as DCNN, ResNet50, and Xception, and achieved an accuracy of 75.51% on the FER2013 dataset. Another study by Wihardi et al. (2022) developed a Gabor Convolutional Neural Network (GCN) model to detect students' moods based on facial expressions. The model was then integrated into the development of a Smart Classroom System (SCS). However, it has been noted that CNNs have limitations in capturing the global context of facial images (Huang et al., 2021).

In recent years, Vision Transformer (ViT) has been increasingly explored for FER tasks as an alternative to CNN-based models. As a transformer-based architecture, ViT is considered superior to CNN due to its ability to capture global context in images through the attention mechanism (Dosovitskiy et al., 2021). A study by Ping (2024) demonstrated that ViT outperformed CNN-based models, achieving an accuracy of 69.60% on the FER2013 dataset, compared to ResNet50 and VGG16, which obtained 65.31% and 67.01%, respectively.

One of the main challenges of ViT is its high computational complexity, particularly in the self-attention mechanism, which has quadratic time complexity with respect to the number of tokens, i.e., $O(N^2)$ (Marin et al., 2021). This poses a significant barrier to developing efficient models suitable for real-world deployment. To address this issue, Marin et al. (2021) proposed Token Pooling, a downsampling approach based on clustering (e.g., k-means or k-medoids) to reduce the number of tokens processed at each transformer block. As a result, computational complexity (in GFlops) can be reduced without compromising model accuracy.

Hybrid CNN-Vision Transformer architecture combines the strengths of CNNs in learning local features with the Vision Transformer's ability to capture global context in images. Xiao et al. (2021) integrated CNN and ViT by replacing the original patchify stem in ViT with a convolutional stem consisting of several small convolutional layers (3×3). As a result, the ViT with a convolutional stem achieved faster convergence and better accuracy compared to the ViT with a patchify stem, highlighting the advantage of using convolutional layers for initial feature extraction in ViT. Shen (2024) explored various hybrid approaches by combining different ResNet variants, such as ResNet50, ResNet101, and ResNet152, with ViT. These combinations included placing ResNet before or after the ViT, as well as parallel integrations. In the approach where ResNet was placed before ViT, it served as a backbone for feature extraction, replacing ViT's patch embedding. This combination outperformed the pure ViT model, achieving an accuracy of 54.51% on the FER2013 dataset, compared to 45.82% for the standalone ViT. However, further research is needed to optimize hybrid CNN-Transformer architectures for FER, as their potential for recognizing students' facial expressions, particularly in classroom environments, remains underexplored.

METHOD

In this study, the research design follows a methodology adapted from Santoso et al. (2020), as illustrated in Fig. 1. The design consists of five main stages: data collection, data preprocessing, model development, model evaluation, and analysis and conclusion. A detailed explanation of each stage is provided in the following sections.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

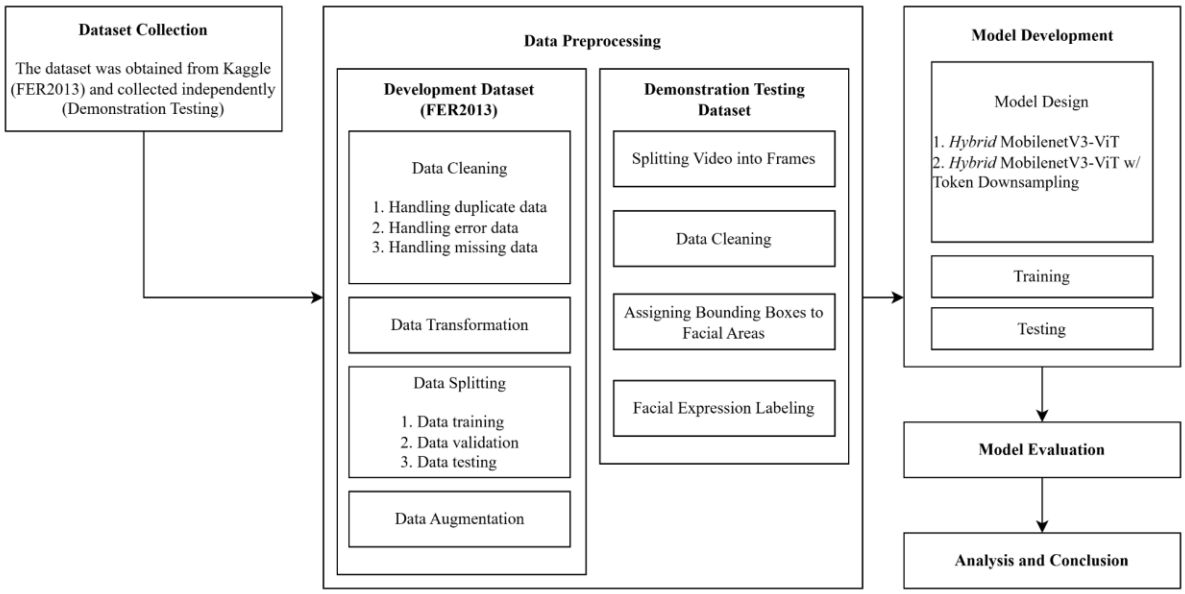

Fig. 1 Research Design

Data Collection

This study utilized two types of data: FER2013 (Goodfellow et al., 2015) as the development dataset and a dataset of students' facial expressions in classroom environments as the test dataset. FER2013 is a publicly available dataset accessible through the Kaggle platform, consisting of 35,887 facial images categorized into seven expression classes: angry, disgust, fear, happy, neutral, sad, and surprise. Each image is 48×48 pixels in size and presented in grayscale.

In the model testing phase, a dataset collected by the researchers was used. The data collection involved recording videos of various student facial expressions during classroom learning activities, with participants drawn from the Computer Science Study Program at Universitas Pendidikan Indonesia. The resulting dataset consists of video recordings divided into three sessions, with a total duration of 8 minutes and 27 seconds, recorded at 30 FPS and a resolution of 1920 × 1080 pixels. Sessions 1 and 2 involved 12 students (9 male and 3 female), while Session 3 involved 39 students (14 male and 15 female). The participants displayed typical facial characteristics of Indonesian individuals, such as medium to light brown skin tones, dark brown irises, black to dark brown hair with straight to wavy textures, and broad but not prominent nose shapes.

Data Preprocessing

Preprocessing of the Development Dataset (FER2013)

The initial stage of data preprocessing began with data cleaning, which involved identifying and handling error, duplicate, and irrelevant data. Detection of error images was carried out by analyzing the average pixel values of each image. Images with an average pixel value below 30 or above 230 were categorized as error, as they contained little to no facial information. A total of 731 images were removed, leaving 33,303 valid images. The dataset was then split into three parts with an 80:10:10 ratio for training, validation, and testing. Several transformations were applied to prepare the data for model input, including channel duplication (to convert from 1 to 3 channels), resizing images to 224×224 pixels, conversion to PyTorch tensors, and normalization of pixel values to the range [-1, 1]. These transformations were applied consistently across all data subsets to ensure compatibility during model training and evaluation.

Preprocessing of the Test Dataset (Classroom Environment)

The preprocessing stage began with splitting the video into image frames at 5 FPS to reduce data duplication caused by minimal expression changes between frames. This process yielded a total of 2,684 frames (692 from session 1, 1,026 from session 2, and 966 from session 3). However, after assessing image quality, most frames from session 3, captured from the left side angle, were deemed unsuitable. As a result, only 2,254 images were retained for further processing, including 536 usable frames from session 3. The next step involved annotating facial bounding boxes using the RetinaFace model (Deng et al., 2019). Detection was performed automatically and saved in JSON format following the COCO schema, with a single object category: face. The final step was expression labeling for each detected face in the images. Labeling was conducted using the Roboflow platform to simplify the process through its graphical user interface. The expression labeling was based on the Facial Action Coding System (FACS), which maps facial muscle movements to specific emotional expressions. The labeling results were saved again in COCO format for subsequent

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

model testing. In addition to labeling, manual refinement was carried out to add bounding boxes for undetected faces and to correct the position and size of existing bounding boxes for higher accuracy.

Data Augmentation

One of the preprocessing steps applied to the FER2013 dataset was data augmentation. This technique aims to transform the data in order to generate new variations. Several augmentation techniques were employed in this study, including: random horizontal flipping with a probability of 0.5, random rotation within a range of -10 to 10 degrees, random cropping with crop sizes ranging from 80% to 100% of the original image, brightness and contrast adjustment, and random erasing. In addition to enhancing the model's generalization capability, these techniques also aim to introduce data variations that better reflect real-world conditions. For example, in the context of this study conducted in classroom environments, techniques such as brightness adjustment and random erasing can simulate varying lighting conditions and occlusions that may occur. Data augmentation was performed on each image batch during model training, commonly known as online data augmentation. However, these augmentations were applied only to the training data, not to the validation or test sets, to ensure that model evaluation remained accurate and reflective of real-world scenarios.

Model Development

To ensure that the results of all experiments could be fairly compared, each experiment followed an identical training procedure. In this study, the applied hyperparameter configuration included a total of 1,000 epochs, with an early stopping mechanism triggered if the validation loss did not decrease for 10 consecutive epochs, in order to prevent overfitting. A large number of epochs was chosen to accommodate varying convergence times across models, allowing each model the opportunity to reach its optimal performance. The batch size was set to 64, with a learning rate of 0.00003 (3e-5). Additionally, a learning rate scheduler, ReduceLROnPlateau, was used to decrease the learning rate by a factor of 0.1 if the validation loss plateaued or did not improve for 5 epochs.

Hybrid MobileNetV3-ViT with Token Downsampling

Hybrid MobileNetV3-ViT model replaces the patch embedding component in the Vision Transformer (ViT) with MobileNetV3. MobileNetV3 is a Convolutional Neural Network (CNN) architecture designed for high efficiency on devices with limited power and computational resources (Andrew et al., 2019). There are two available variants of MobileNetV3: MobileNetV3-Small, with approximately 2.5 million parameters, and MobileNetV3-Large, with 5.4 million parameters.

The proposed hybrid architecture, as illustrated in Fig. 2, consists of two main stages. First, the input image is processed by MobileNetV3, which serves as a CNN-based feature extractor. MobileNetV3 extracts local features from the image using efficient techniques such as depthwise separable convolution and inverted residual blocks, enabling feature processing with low computational cost. The output from MobileNetV3 is then flattened and projected into a dimensional space compatible with the ViT embedding. This projection is necessary because the final output dimensions of MobileNetV3-Large and MobileNetV3-Small are 960 and 576 channels, respectively, while ViT-B requires an embedding dimension of 768. After projection, the resulting vector representation is passed into the transformer blocks of the ViT, which model the spatial relationships and global context among image patches through the self-attention mechanism. This approach allows the proposed hybrid architecture to combine the efficiency of MobileNetV3 in extracting local features with the capability of ViT to capture global structure and long-range dependencies in images.

One of the main challenges of ViT is its high computational complexity, particularly in the self-attention mechanism, which has quadratic time complexity with respect to the number of tokens, i.e., $O(N^2)$ (Marin et al., 2021). To address this challenge, a Token Downsampling technique is employed to reduce the number of input tokens processed by subsequent transformer blocks without sacrificing important information. As illustrated in Fig. 2, the Token Downsampling module is inserted after transformer blocks, aiming to significantly reduce the number of tokens processed in the following stages.

Two Token Downsampling approaches are used: clustering-based and score-based. The first approach employs a k-medoids clustering technique to reduce the number of tokens (Marin et al., 2021). The process begins by randomly selecting a set of tokens as medoids, equal to the target number of tokens to retain. The remaining tokens are then grouped into clusters based on their Euclidean distance in the embedding space relative to the medoids. Tokens with similar feature representations are grouped into the same cluster. Once clustering is complete, the tokens selected as medoids are used as the final representations and passed to the subsequent transformer blocks.

The second approach leverages a score-based technique (Goyal et al., 2020), in which each token is evaluated based on its significance score. This score is computed by summing the attention weights of the token across all self-attention heads within a transformer block. Tokens are then ranked based on their significance values, and only the top-ranked tokens are retained. This method aims to filter out tokens that contribute the most to the model's understanding of the global context. Through these two approaches, the implemented Token Downsampling method is expected to

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 5, Number 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

reduce the computational complexity of ViT without significantly compromising model performance.

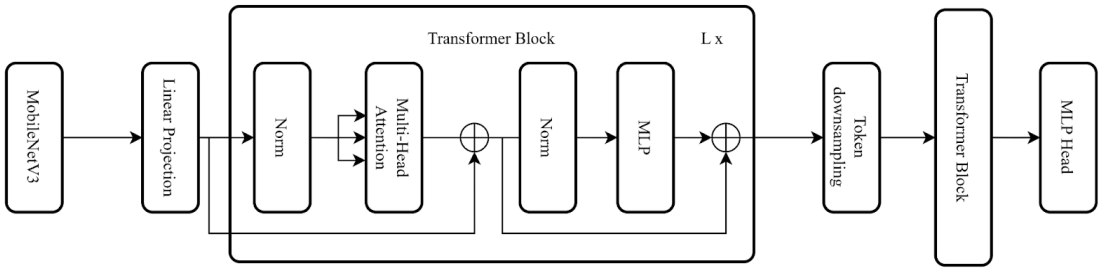

Fig. 2 Architecture of Hybrid MobileNetV3-ViT Model with Token Downsampling

Regularization

To enhance the model's generalization capability, Sharpness-Aware Minimization (SAM) regularization technique was employed. Unlike conventional methods that solely aim to find points with minimal loss values, SAM seeks regions in the parameter space where the loss is both low and relatively flat, making the model more robust to small parameter perturbations and less prone to overfitting (Foret et al., 2021). By encouraging the model to converge within flatter areas of the loss landscape, SAM can lead to more stable performance and improved generalization, especially when applied to unseen data.

Facial Expression Detection

After the facial expression classification model was successfully developed, it was integrated with a face detection model to form a complete facial expression recognition (FER) system. In this study, an existing face detection model was used, as many high-performing face detectors are already available. Specifically, RetinaFace model (Deng et al., 2019) was employed, which achieves a Average Precision (AP) of 0.914 on the WIDER FACE test set. The FER system workflow is as follows: first, an input image is processed by the face detection model to obtain regions of interest (ROIs) containing facial areas. Then, for each detected face, the expression classification model is applied to identify the facial expression. The experiment proceeded by evaluating the FER system on the preprocessed test dataset to assess the model's reliability in recognizing facial expressions in a classroom environment. However, considering that classroom learning environments exhibit different visual and expressive characteristics compared to the FER2013 dataset, additional experiments were conducted using adaptation strategies to enhance the model's generalization ability in real-world classroom contexts.

Model Evaluation

Performance of the facial expression classification model in each experiment scenario will be evaluated using accuracy and F1-score metrics, based on the components of the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Grandini et al., 2020). For the integrated model combining facial expression classification and face detection, performance on the test dataset will be evaluated using Average Precision (AP) (Zhu, 2004).

RESULT

Evaluation was conducted by comparing several facial expression classification models developed in this study. The evaluation includes an analysis of accuracy and F1-score metrics for each expression class. In addition, the facial expression recognition model was evaluated on the test dataset using the Average Precision (AP) metric.

Hybrid MobileNetV3-ViT with Token Downsampling

Model performance evaluation was carried out by calculating and comparing the accuracy and f1-score metric on the test data for each experimental scenario. Table 1 presents performance of using MobileNetV3-Large and MobileNetV3-Small as backbones in the hybrid model, along with a comparison to their respective standalone (non-hybrid) models. Additionally, the performance of the Hybrid MobileNetV3-Large–ViT model was evaluated across each facial expression class, with the results presented in Fig. 3.

Table 1. Model Performance Comparison

Model	Accuracy (%)$^{\Delta}$	F1-Score (%)$^{\Delta}$	Epoch$^{\nabla}$	Params (M)
Pretrained Vision Transformer (ViT)	69,62	69,39	17	87,4
Pretrained MobileNetV3-Large	66,62	65,89	22	4,2
Hybrid CNN-ViT	62,80	62,45	41	85,5
Hybrid MobileNetV3 Small-ViT	66,32	65,72	20	86,4
Hybrid MobileNetV3 Large-ViT	68,33	67,80	19	88,8

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

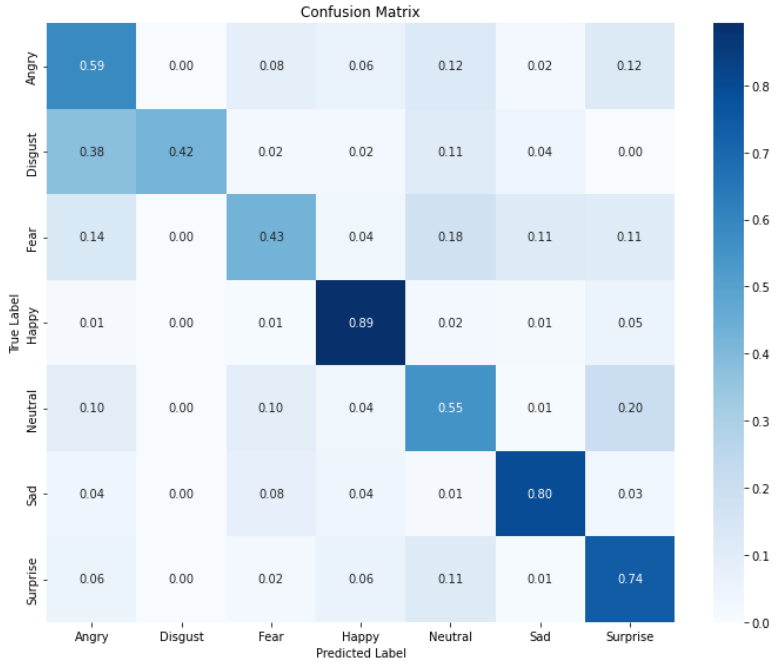

Fig. 3 Confusion Matrix of Hybrid MobileNetV3-ViT Model Performance

In this study, the number of tokens retained at each transformer layer during the token downsampling process was manually defined, using the configuration (49, 35, 35, 35, 26, 26, 20, 20, 20, 12, 12, 12) as the number of tokens to be preserved. Table 2 presents a comparison of Token Downsampling performance using the clustering-based and score-based approaches.

Table 2. Performance Comparison of Models With and Without Token Downsampling

Model	Accuracy (%)△	F1-Score (%)△	Epoch▽	Params (M)	Flops (G)▽	Inference Speed (FPS)▽
Pretrained ViT	69,62	69,39	17	87,4	8,73	123,06
Hybrid MobileNetV3-ViT	68,33	67,80	19	88,8	9,01	79,68
Hybrid MobileNetV3-ViT/S	67,40	67,08	21	53,8	5,52	86,70
Hybrid MobileNetV3-ViT w/ Token Downsampling (cluster)	68,06	67,56	19	88,8	5,48	41,75
Hybrid MobileNetV3-ViT w/ Token Downsampling (score)	68,39	67,84	19	88,8	5,48	82,73

Regularization

In this study, SAM was applied with a perturbation magnitude (rho) of 0.05. The results of using this regularization technique on the model are presented in Table x, and its effect on the model's loss landscape is illustrated in Fig. 4. The F1-score evaluation results of the Hybrid MobileNetV3-ViT model for each expression class are shown in Table 3.

Table 3. Performance Comparison of Models With and Without SAM

Model	SAM	Accuracy (%)△	F1-Score (%)△	Epoch▽
Pretrained ViT	Tidak	69,62	69,39	17
Hybrid MobileNetV3-ViT	Tidak	68,39	67,84	19
Pretrained ViT	Ya	70,10	69,59	20
Hybrid MobileNetV3-ViT	Ya	71,24	70,81	25

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

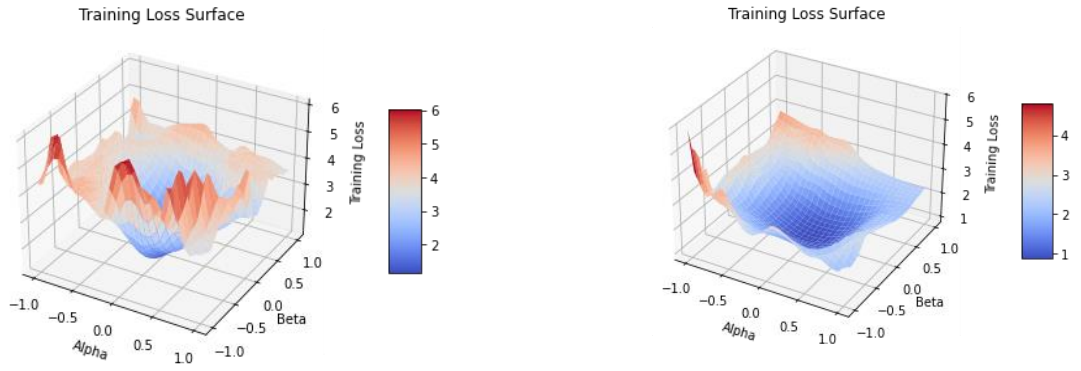

Fig. 4 Loss Landscape Plot of Hybrid MobileNetV3-ViT Model With (Right) and Without (Left) SAM

Facial Expression Detection

An evaluation was conducted by comparing the performance of integrated models combining various facial expression classification models with a face detection model on the test dataset. The evaluation metric used was Average Precision (AP), with the results presented in Table 4.

Subsequently, an experiment was conducted by training the model on a combined dataset consisting of FER2013 and the classroom dataset as part of a model adaptation strategy. A different data distribution was applied, where the combined data included sessions 1 and 2, while session 3 was used exclusively for testing. The experimental results are presented in Table 5.

Table 4. Comparison of Facial Expression Detection Model Performance

Facial Expression Classification Model	Face Detection Model	AP@.5$^{\triangle}$	AP@.75$^{\triangle}$ (PASCAL VOC)	AP@[.5:.95]$^{\triangle}$ (COCO)	Inference Speed (FPS)$^{\triangledown}$
Pretrained ViT	RetinaFace	14,5	14,3	14,1	0,58
Pretrained MobileNetV3		10,2	10,0	9,9	0,55
Hybrid MobileNetV3-ViT		16,0	15,7	15,5	0,55

Table 5. Model Performance under Adaptation Strategy

Training Dataset	AP@.5$^{\triangle}$	AP@.75$^{\triangle}$ (PASCAL VOC)	AP@[.5:.95]$^{\triangle}$ (COCO)
FER2013	16,0	15,7	15,5
Combined Dataset (FER2013 and Classroom)	19,1	18,9	18,6

DISCUSSION

This section provides a detailed analysis and discussion of the results obtained from the series of experiments conducted earlier.

Hybrid MobileNetV3-ViT with Token Downsampling

Based on Table 1, it can be observed that using MobileNetV3 as the backbone for ViT significantly improves accuracy compared to standard CNN, by 5.53% for MobileNetV3-Large and 3.52% for MobileNetV3-Small. Although integrating MobileNetV3 as a backbone in the ViT model yields better accuracy than the standalone pretrained MobileNetV3, with an improvement of 1.71%, it still falls short of outperforming the standalone pretrained ViT model.

The evaluation results of Hybrid MobileNetV3-ViT model for each expression class, as shown in Fig. 3, indicate that the model performs very well in recognizing the "happy" expression, which is the class with the largest number of samples. Interestingly, although the "surprised" expression belongs to the minority classes, the model is still able to recognize it effectively. This may be due to the fact that the surprised expression contains distinct facial features that are noticeably different from other expressions, making it easier for the model to identify.

To examine the contribution of MobileNetV3 as the backbone in the hybrid model, attention map visualizations were analyzed. As shown in Fig. 5, the attention maps of the pure ViT model appear to be more dispersed across various facial regions, with somewhat random focus, as seen in columns 3 and 6. This may be attributed to ViT's tendency to learn global relationships between image regions, leading it to attend to multiple areas simultaneously. In contrast, the attention maps in the hybrid model are more concentrated, primarily focusing on key facial features such as

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

the eyes, nose, and mouth, vertically aligned along the center of the face. This indicates the influence of convolution (MobileNetV3) in the hybrid model, enabling it to better attend to visually important local features.

The experiment results on Token Downsampling, as shown in Table 2, indicate that the cluster-based approach can reduce computational complexity from 9.01 GFLOPs to 5.48 GFLOPs, with only a 0.27% drop in accuracy, significantly lower than the accuracy drop caused by reducing the number of layers or embedding size, which led to a 0.93% decrease. However, implementing clustering poses practical challenges, as the clustering process is executed on the CPU while the main model runs on the GPU. This leads to data transfer overhead between devices, ultimately slowing down inference speed, even though fewer tokens are processed. In other words, the computational burden of clustering may offset the efficiency gains achieved by reducing model complexity.

Unlike the cluster-based approach, the score-based method only requires the computation of significance scores from attention between tokens, resulting in lower computational overhead. As a result, the score-based token downsampling model not only reduces complexity and improves inference speed, but also achieves a 0.06% increase in accuracy compared to the model without downsampling. Conversely, when high-significance tokens are removed, accuracy drops to 67.91%, indicating that the significance score is positively correlated with prediction quality.

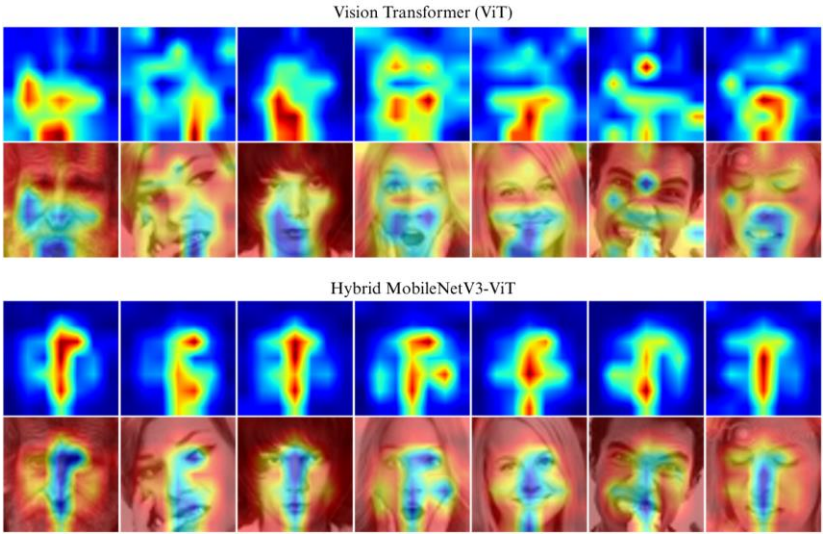

Fig. 5 Comparison of Attention Maps between ViT and Hybrid Models

Regularization

Based on the loss landscape analysis of the Hybrid MobileNetV3-ViT model, as shown in Fig. 4, the model initially converges to a relatively sharp region, indicated by a deep blue color at the optimal point and lighter blue surrounding it, which may negatively impact the model's generalization capability. After implementing SAM, the model begins to converge to a flatter region. Additionally, as shown in Table 3, there is a 2.85% increase in accuracy for the Hybrid MobileNetV3-ViT and a 0.48% increase for the Pretrained ViT. The greater accuracy improvement observed in the Hybrid MobileNetV3-ViT may be attributed to the ViT component being trained from scratch, making SAM more influential on the model. This accuracy gain highlights the advantage of achieving a flatter loss landscape through SAM, leading to better model generalization.

Facial Expression Detection

After evaluating the facial expression detection model using the classroom test dataset, the performance results are shown in Table 4. Hybrid MobileNetV3-ViT model, used as the facial expression classification model, achieved the highest Average Precision (AP) compared to other models. However, the inference speed was relatively low, primarily due to a bottleneck in the face detection model, which achieved only 0.6 FPS, in contrast to the expression classification model that reached up to 82 FPS. Additionally, the overall AP performance was not optimal, with the best AP at a 0.5 threshold reaching only 16. A further analysis was conducted to investigate the factors contributing to the low AP performance.

Hybrid MobileNetV3-ViT model demonstrated good performance on the FER2013 dataset but performed poorly on the classroom test dataset. Several factors may have contributed to the low AP results. First, the discrepancy in expression labeling formats may have played a role, while the classroom dataset uses the FACS (Facial Action Coding System), the labeling format used in FER2013 is unknown. Second, there is a notable difference in image resolution: the training dataset (FER2013) consists of uniformly sized images at 48 × 48 pixels, whereas the classroom dataset includes images with resolutions as low as 27 × 34 pixels, for example the faces of students seated at the back of the classroom. Furthermore, as shown in Fig. 6, the training data predominantly contain highly expressive facial expressions, for instance "surprise" expression typically includes wide-open eyes and a broadly open mouth. In contrast,

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

the classroom dataset presents more subtle facial expressions, such as slightly open mouths, which makes it more difficult for the model to recognize expressions accurately in the test data.

Through the domain adaptation experiment, the results shown in Table 5 indicate that training the model with a combined dataset (FER2013 and classroom data) significantly improves performance compared to training with FER2013 alone. The model trained on the combined dataset achieved an Average Precision (AP) of 19.1 at an IoU threshold of 0.5, compared to 16.0 for the model trained only on FER2013. This improvement suggests that incorporating data from the classroom context enables the model to become more adaptive and relevant to the more natural and diverse facial expression patterns exhibited by students. In other words, the model can learn expression patterns that better reflect real-world classroom environments, resulting in better detection performance than when trained solely on the standard expressions found in FER2013. These findings also highlight the importance of curating and integrating data from specific contexts to enhance the generalization capability and practical effectiveness of the model in real-world applications.

Fig. 6 Comparison of Facial Expressions in the FER2013 and Classroom Datasets Along with Their Prediction Results

Although the model has not yet achieved optimal results on the classroom test dataset, it demonstrated strong performance on the FER2013 dataset. As shown in Table 6, the Hybrid MobileNetV3-ViT model with Token Downsampling outperformed both single models and hybrid models from previous studies. Moreover, it was able to compete with ensemble models composed of multiple architectures. These results highlight the potential for further development of this approach to support real-time and context-aware facial expression recognition in classroom environments.

Table 6. Comparison of Model Performance in This Study with Previous Works

Model	Accuracy (%)[△]
Hybrid MobileNetV3-ViT with Token Downsampling (our)	71,24
Improved MobileViT (Jiang et al., 2024)	62,20
Hybrid ResNet101-ViT (Shen, 2024)	54,51
Pretrained ViT (Ping, 2024)	69,60
EM-CNN (Dukić & Krzic, 2022)	65,57
Ensamble CNN (Lawpanom et al., 2024)	75,51

CONCLUSION

This study proposes a Hybrid MobileNetV3-ViT model with Token Downsampling for recognizing students' facial expressions in classroom environments. The experiment results show that integrating MobileNetV3 as a backbone with ViT improves the model's ability to learn local facial features compared to pure CNN, with an accuracy improvement of 5.53% for MobileNetV3-ViT Large and 3.52% for MobileNetV3-ViT Small. The model also proves to be more efficient, as the score-based Token Downsampling method reduces computational complexity from 9.01 GFLOPs to 5.48 GFLOPs, while slightly increasing accuracy by 0.06%. The implementation of Sharpness-Aware Minimization (SAM) regularization also had a positive impact, especially on hybrid models trained from scratch, leading to an accuracy increase of 2.85%. Although the proposed model achieved the highest Average Precision (AP) compared to other models and showed improved AP when trained on domain-specific datasets, its performance is not yet sufficient for direct classroom deployment. The low AP score is attributed to several factors, including differences in expression labeling formats, image resolution disparities, and the subtler facial expressions in classroom test data compared to the highly expressive faces in the FER2013 training dataset. Therefore, future research should focus on developing a high-resolution classroom-specific dataset with expression labels consistent with the training scheme (e.g., FACS). Additionally, integrating a more efficient face detection method is recommended to enhance overall inference speed, thus enabling real-time deployment.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

REFERENCES

Andrew, H., Mark, S., Grace, C., Liang-Chieh, C., Bo, C., Mingxing, T., Weijun, W., Yukun, Z., Ruoming, P., & Vijay, V. (2019). Searching for MobileNetV3. Proceedings of the IEEE International Conference on Computer Vision, 1314–1324.

Bouhlal, M., Aarika, K., AitAbdelouahid, R., Elfilali, S., & Benlahmar, E. (2020). Emotions recognition as innovative tool for improving students' performance and learning approaches. Procedia Computer Science, 175, 597–602. https://doi.org/10.1016/j.procs.2020.07.086

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). RetinaFace: Single-stage Dense Face Localisation in the Wild. http://arxiv.org/abs/1905.00641

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. ICLR 2021 - 9th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

Dukić, D., & Krzic, A. S. (2022). Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment. Electronics (Switzerland), 11(8). https://doi.org/10.3390/electronics11081240

Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research from the Perspective of Machine Learning: A Systematic Review. IEEE Access, 11(August), 112060–112074. https://doi.org/10.1109/ACCESS.2023.3322454

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2021). Sharpness-Aware Minimization for Efficiently Improving Generalization. ICLR 2021 - 9th International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.01412

Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D. H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

Goyal, S., Choudhury, A. R., Raje, S. M., Chakaravarthy, V. T., Sabharwal, Y., & Verma, A. (2020). PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination. 37th International Conference on Machine Learning, ICML 2020, PartF16814, 3648–3657. https://doi.org/10.48550/arXiv.2001.08950

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. 1–17. http://arxiv.org/abs/2008.05756

Huang, Q., Huang, C., Wang, X., & Jiang, F. (2021). Facial expression recognition with grid-wise attention and visual transformer. Information Sciences, 580, 35–54. https://doi.org/10.1016/j.ins.2021.08.043

Indolia, S., Nigam, S., Singh, R., Singh, V. K., & Singh, M. K. (2023). Micro Expression Recognition Using Convolution Patch in Vision Transformer. IEEE Access, 11(August), 100495–100507. https://doi.org/10.1109/ACCESS.2023.3314797

Jiang, B., Li, N., Cui, X., Zhang, Q., Zhang, H., Li, Z., & Liu, W. (2024). Research on facial expression recognition algorithm based on improved MobileNetV3. Eurasip Journal on Image and Video Processing, 2024(1), 1–19. https://doi.org/10.1186/s13640-024-00638-z

Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. Artif. Intell. Rev., 56(Suppl 3), 2917–2970. https://doi.org/10.1007/s10462-023-10595-0

Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach. Applied Sciences (Switzerland), 14(3). https://doi.org/10.3390/app14031156

Marin, D., Chang, J.-H. R., Ranjan, A., Prabhu, A., Rastegari, M., & Tuzel, O. (2021). Token Pooling in Vision Transformers. 1–21. http://arxiv.org/abs/2110.03860

Ping, H. (2024). Advancing Facial Expression Recognition: A Comparative Study of CNNs and Transformers. 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering, ICEACE 2024, 222–226. https://doi.org/10.1109/ICEACE63551.2024.10898937

Santoso, R. R., Megasari, R., & Hambali, Y. A. (2020). Implementasi Metode Machine Learning Menggunakan Algoritma Evolving Artificial Neural Network Pada Kasus Prediksi Diagnosis Diabetes. 3(2).

Shen, Z. (2024). A Comparative Study of Hybrid CNN and Vision Transformer Models for Facial Emotion Recognition. 2024 11th International Conference on Dependable Systems and Their Applications (DSA), 401–408. https://doi.org/10.1109/DSA63982.2024.00061

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6323

Wihardi, Y., Junaeti, E., Setiawan, W., Wahyudin, W., & Erlangga, E. (2022). Smart Classroom System (SCS) Berbasis Kamera Untuk Memantau Keadaan Peserta Didik. INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS : Journal of Information System, 6(1), 67. https://doi.org/10.51211/isbi.v6i1.1771

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early Convolutions Help Transformers See Better. Advances in Neural Information Processing Systems, 36(NeurIPS), 30392–30400. https://doi.org/10.48550/arXiv.2106.14881

Zhu, M. (2004). Recall, precision and average precision. Department of Statistics and Actuarial Science, 1–11.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.