International Journal of Autonomous Robotics and Intelligent Systems
Vol. 2, No. 1, June 2026 [43-52]

A Controlled Empirical Comparison of ResNet18, ViT-Small and CNN–Transformer Hybrid for RAF-DB Facial Expression Recognition

Mohanad Qubtan[1*], Hasimah Ali[1] and Mohamed Elshiekh[2]

[1]Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, Arau, Perlis
[2]Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Arau, Perlis

Received 30 April 2026, Revised 17 May 2026, Accepted 21 June 2026

ABSTRACT

Facial expression recognition (FER) models frequently struggle with class imbalance and generalization, particularly on real-world datasets. This study provides a controlled empirical comparison under a unified training pipeline on RAF-DB. Three lightweight architectures are compared: (i) a ResNet18 CNN classifier, (ii) a pretrained Vision Transformer (ViT-Small, patch16/224) and (iii) a CNN–Transformer hybrid, that converts CNN feature maps into token sequences for transformer encoding. The same preprocessing, online augmentation, imbalance-handling strategy and evaluation protocol are applied across all models. Class-aware augmentation, weighted random sampling and class-weighted cross-entropy are used, while accuracy, macro-F1, weighted-F1 and per-class metrics are reported. On the RAF-DB test set, the ViT-Small baseline achieves the best performance with accuracy of 0.8116 and macro-F1 of 0.7353. The ResNet18 CNN obtains accuracy of 0.7386 and macro-F1 of 0.6621, while the hybrid model obtains accuracy of 0.7370 and macro-F1 of 0.6552. The results show that ViT-Small benefits from pretrained global representation learning, whereas the current hybrid configuration does not improve minority-class recognition over the CNN baseline.

Keywords: Convolutional Neural Network, Facial Expression Recognition, Hybrid Model, Vision Transformer

1. INTRODUCTION

Facial expression recognition (FER) is a critical component of affect-aware human-computer interaction, but real-world performance remains challenging since expressions are subtle, identities vary and photos are captured under unconstrained conditions (position, illumination, occlusion) rather than in lab environments [1]. Large, “in-the-wild” benchmarks such as RAF-DB have therefore become standard for evaluating how well models generalize beyond controlled data [1]. Convolutional neural networks (CNNs) are a strong baseline for FER because they learn discriminative local patterns efficiently, and residual learning (e.g., ResNet) enables deeper CNNs to train stably and scale to harder visual recognition tasks [2]. More recently, Vision Transformers (ViTs) have shown that attention-based architectures can be highly competitive in image recognition by modeling long-range dependencies, which can benefit FER when relevant cues are distributed across the face [3]. However, transformers often demand more data and compute than compact CNNs, motivating hybrid designs that combine CNN locality with transformer-style global context modeling [4]. In this study, a controlled comparison was conducted on RAF-DB using ResNet18, ViT-Small and a lightweight CNN–Transformer hybrid under the same preprocessing, training, imbalance-handling and evaluation settings. Accuracy, macro-F1, weighted-F1 and per-class metrics are reported to capture both overall performance and minority-class behavior [1].

* mhndqbtan@gmail.com

The main contributions of this study are as follows. First, it provides a controlled empirical comparison of CNN, ViT and CNN–Transformer hybrid architectures under a unified RAF-DB training pipeline. Second, it evaluates model behavior under class imbalance using macro-F1, weighted-F1 and per-class precision / recall / F1 rather than relying only on accuracy. Third, it provides empirical insight into the limitations of the tested hybrid configuration, showing that simply converting CNN feature maps into transformer tokens does not necessarily improve FER performance without careful token design and capacity tuning.

2. RELATED WORK

Early "in-the-wild" FER benchmarks such as FERPlus improved label reliability by aggregating crowd-sourced annotations and using label distributions rather than a single hard label, which helped stabilize training under noisy supervision [5]. AffectNet further scaled FER by providing a large, diverse set of facial images collected from the web with categorical and (optionally) dimensional emotion annotations, and it remains a common pretraining/transfer reference for FER systems targeting real-world variation [6]. Beyond datasets, a major line of work focuses on attention and locality to handle pose, occlusion, and background distractions. Region Attention Networks (RAN) explicitly emphasize informative facial regions and suppress less useful areas, improving robustness when expressions are partially visible or subtle [7]. Similarly, deep attentive local feature approaches learn to attend to discriminative patches (e.g., around eyes / mouth) and combine them with global cues, which is particularly relevant for confusing classes such as fear vs. surprise or sad vs. neutral [8]. More recent CNN-based designs such as Distract Your Attention Network (DAN) continue this trend by coupling feature extraction with attention mechanisms that reduce the influence of irrelevant regions and strengthen class-separable representations under typical FER confounders [9]. In parallel, transformer-based vision models gained popularity due to their global context modelling. However, training transformers from scratch is data-hungry; DeiT showed that distillation and data-efficient training recipes can make ViT-style models competitive on smaller datasets, motivating their use (often pretrained) in downstream recognition tasks including FER [10]. Overall, previous FER studies have explored CNN-based local feature learning, attention-based region selection and Transformer-based global representation learning. However, many studies differ in datasets, preprocessing, augmentation, imbalance handling, pretraining and evaluation metrics, making it difficult to isolate the effect of model architecture. A direct comparison of CNN, ViT, and CNN–Transformer hybrid models under the same RAF-DB training and evaluation pipeline is therefore useful for understanding whether global attention or hybrid tokenization provides practical benefits under class imbalance. This study addresses that gap by comparing the three model families using identical data splits, input resolution, augmentation strategy, imbalance handling, optimizer settings and macro-F1-based checkpoint selection.

3. METHODOLOGY

3.1 Dataset and Dataset Split

RAF-DB is used and organized into seven class folders, with the label mapping set to 1=Surprise, 2=Fear, 3=Disgust, 4=Happiness, 5=Sadness, 6=Anger and 7=Neutral. The dataset contains 10,183 training-directory images and 3,068 test images. From the training directory, a stratified 80/20 split is used to create 8,146 training images and 2,037 validation images. The resulting training distribution is: Surprise = 944, Fear = 211, Disgust = 549, Happiness = 2640, Sadness = 1482, Anger = 542 and Neutral = 1778. The validation distribution is: Surprise = 236, Fear = 53, Disgust = 137, Happiness = 660, Sadness = 371, Anger = 136 and Neutral = 444. The test distribution is: Surprise = 329, Fear = 74, Disgust = 160, Happiness = 1185, Sadness = 478, Anger = 162 and Neutral = 680. These distributions show a clear imbalance, especially for Fear, Disgust

and Anger. Therefore, macro-F1 and per-class recall are reported together with accuracy to provide a more reliable evaluation of class-balanced performance.

3.2 Preprocessing and Augmentation

The RAF-DB aligned images are used as provided in the dataset folders. No additional face detection or facial alignment step is applied before resizing. Each image is resized or cropped to 224×224 and normalized using ImageNet mean and standard deviation. During training, augmentation is applied online within the training dataset loader, meaning that transformed versions are generated dynamically during training rather than saved as additional offline images. A light augmentation pipeline is applied to all training samples, including random resized cropping around the target resolution, random horizontal flipping and moderate color jitter. To reduce imbalance and increase robustness for underrepresented classes, a stronger augmentation pipeline is used just on minority classes. Minority classes are determined using the median class count from the training split. In this run, the training class counts are [944, 211, 549, 2640, 1482, 542, 1778], giving a median threshold of 944. Classes with counts below 944 are treated as minority classes; therefore, Fear, Disgust and Anger receive the stronger augmentation pipeline. The stronger pipeline increases geometric and photometric variability, and includes random erasing to encourage invariance to partial occlusion.

3.3 Imbalance Handling

Two complementary techniques are used to address class imbalance during training. First, the training DataLoader employs a WeightedRandomSampler, in which each sample is weighted inversely proportional to its class frequency, increasing the likelihood of sampling minority-class samples. Second, class-weighted cross-entropy with inverse-frequency weights normalized is used so the average weight is approximately one. It includes an implementation of focal loss as an alternative objective, but it is disabled in the reported run; focal loss is commonly used to emphasize hard examples when imbalance is severe [11].

3.4 Model Configurations

Three models are trained and compared under the same pipeline. To improve fairness, all trainable backbones are initialized with ImageNet-pretrained weights where available and fine-tuned on RAF-DB using the same preprocessing, augmentation, imbalance handling, optimizer, training duration and model-selection criterion. The CNN baseline used an ImageNet-pretrained ResNet18 backbone from timm with the classifier removed and global average pooling enabled, followed by a lightweight dropout-and-linear classification head. The ViT baseline used an ImageNet-pretrained ViT-Small (patch16, 224) model fine-tuned end-to-end for seven-class classification. The hybrid model follows a step-by-step CNN–Transformer pipeline using an ImageNet-pretrained ResNet18 CNN backbone as the feature extractor. First, the CNN backbone extracts a spatial feature map from the input image. Second, a 1×1 convolution projects the feature map to an embedding dimension of 256. Third, the projected feature map is flattened into a sequence of visual tokens. Fourth, the token sequence is processed by a transformer encoder with depth 4, 8 attention heads and GELU activations. Fifth, a learnable positional embedding is initialized at the first forward pass to match the token count. Finally, attentive pooling aggregates the token representations into a single feature vector, which is passed to the classification head. Figure 1 shows the hybrid model architecture.

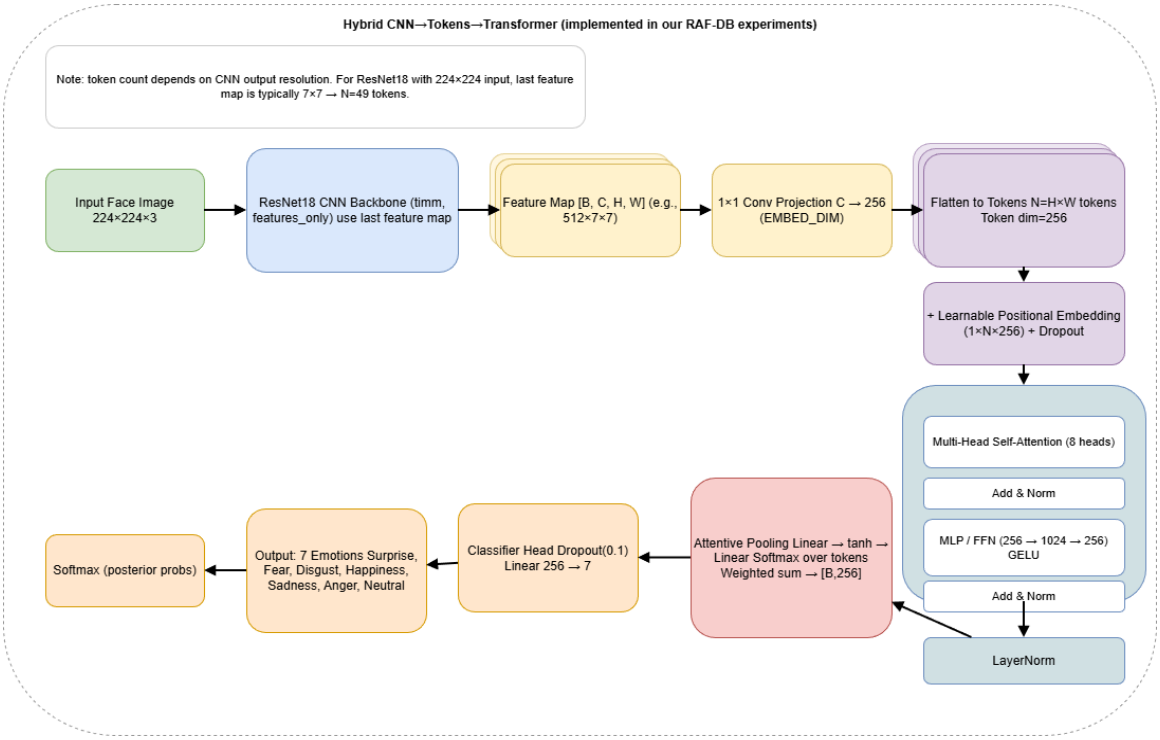


Figure 1. Hybrid model architecture

Figure 1 illustrates how the hybrid model converts CNN feature maps into token sequences before transformer encoding. This supports the methodology by showing the flow from convolutional feature extraction to token-based global modeling. This setup makes the comparison controlled at the training-protocol level: all models use the same input size, train / validation / test split, online augmentation strategy, imbalance handling, optimizer, number of epochs and checkpoint-selection criterion. However, the comparison is not intended to prove architectural superiority in all conditions because each architecture has different inductive biases and parameterization. Instead, the goal is to compare their empirical behavior under the same practical RAF-DB training pipeline.

3.5 Training Procedure and Model Selection

All models are trained for 25 epochs with a batch size of 64 using AdamW with learning rate of 3×10^{-4} and weight decay 1×10^{-4} [12]. A cosine learning-rate schedule is applied across epochs [13]. Training uses automatic mixed precision (autocast and gradient scaling) to improve throughput on GPU while maintaining stable optimization [14], and gradients are clipped to a maximum norm of 1.0. For each model, the checkpoint that achieves the highest validation macro-F1 are saved and evaluate that checkpoint once on the held-out test set.

3.6 Test-Time Evaluation

On the test set, accuracy, macro-F1, weighted-F1 and per-class recall are derived from the confusion matrix. The full per-class precision/recall/F1 classification are reported and confusion matrix visualizations and training curves (train loss, validation accuracy, validation macro-F1) are saved as generated.

4. RESULTS AND DISCUSSION

This section reports test-set performance on RAF-DB using the fixed label mapping 1=Surprise, 2=Fear, 3=Disgust, 4=Happiness, 5=Sadness, 6=Anger, 7=Neutral, with N = 3068 test images. Macro-F1 is emphasized to reflect minority-class behaviour under imbalance, alongside accuracy and weighted-F1.

4.1 Overall Comparison Across Models

Before examining class-wise behavior, it is necessary to establish the overall ranking of the three architectures under identical preprocessing, training protocol and checkpoint selection by validation macro-F1. The Vision-Transformer baseline consistently dominates both accuracy and macro-F1, while the hybrid CNN–Transformer configuration, which converts CNN feature maps into token sequences before transformer encoding, does not surpass the CNN baseline in this run, indicating that the chosen tokenization and transformer depth / width are not yet yielding additive benefits on this split. Table 1 reports the aggregate test metrics for the CNN, ViT and hybrid models.

Table 1 Overall test-set performance on RAF-DB (N=3068)

Model	Accuracy	Macro-F1	Weighted-F1
CNN (ResNet18)	0.7386	0.6621	0.7414
ViT (ViT-Small/16)	0.8116	0.7353	0.8108
Hybrid (CNN-Transformer)	0.7370	0.6552	0.7381

Table 1 shows that ViT-Small achieves the strongest overall performance, with an accuracy improvement of 0.0730 and a macro-F1 improvement of 0.0732 over ResNet18. The hybrid model remains close to ResNet18 in accuracy but lower in macro-F1, indicating weaker class-balanced performance. The magnitude of the ViT improvement is substantial: relative to CNN, accuracy increases by +0.0730, and macro-F1 increases by +0.0732. By contrast, the hybrid model slightly underperforms CNN on macro-F1 (0.6552 vs. 0.6621), suggesting that errors concentrate more heavily in underrepresented classes. The stronger ViT performance can be attributed to two likely factors. First, the pretrained ViT-Small benefits from global self-attention, which can model relationships among distant facial regions such as the eyes, eyebrows and mouth. These long-range relationships are important for expressions where local cues alone may be ambiguous. Second, the pretrained ViT representation provides a stronger initialization for RAF-DB, reducing the need to learn global facial structure from a relatively limited and imbalanced training set. In contrast, the hybrid model depends on the quality of the CNN feature map, the tokenization strategy and the transformer capacity. In this configuration, the token sequence produced from CNN features may not preserve enough fine-grained expression information for minority classes, and the shallow transformer may not add sufficient discriminative power beyond the CNN baseline.

Because all models are fine-tuned under the same data split, augmentation, imbalance handling, optimizer and model-selection rule, the performance differences mainly reflect how each architecture uses the shared pipeline, although architecture-specific factors such as pretrained representation quality, tokenization and capacity remain important.

4.2 Per-Class Performance (Precision / Recall / F1)

Overall averages can obscure which emotions drive success or failure. In RAF-DB, minority classes, particularly Fear and Disgust, strongly affect macro-F1 because each class contributes equally to the average regardless of support size. Across all three models, the weakest recalls

occur in Fear and Disgust, while the strongest performance is generally observed for frequent and visually salient classes such as Happiness and Neutral. This pattern explains why weighted-F1 remains relatively high even when macro-F1 drops: strong majority-class performance can dominate the weighted average. To make these class-specific effects explicit, Tables 2 to 4 provide the complete per-class classification reports for CNN, ViT and hybrid models, respectively.

Table 2 CNN (ResNet18) classification report on RAF-DB test set

Class (Emotion)	Precision	Recall	F1-score	Support
1 (Surprise)	0.6863	0.8511	0.7598	329
2 (Fear)	0.6230	0.5135	0.5630	74
3 (Disgust)	0.4655	0.3375	0.3913	160
4 (Happiness)	0.9600	0.7485	0.8412	1185
5 (Sadness)	0.6145	0.7971	0.6940	478
6 (Anger)	0.6627	0.6914	0.6767	162
7 (Neutral)	0.6675	0.7559	0.7090	680
Accuracy	N/A	N/A	0.7386	3068
Macro Avg	0.6685	0.6707	0.6621	3068
Weighted Avg	0.7624	0.7386	0.7414	3068

Table 2 shows that the ResNet18 CNN baseline performs strongly on Happiness, achieving high precision of 0.9600, but its recall is lower at 0.7485. This indicates that when the CNN predicts Happiness it is usually correct, but it still misses a considerable number of Happiness samples. The weakest results appear in the minority classes, especially Disgust, with recall of 0.3375, and Fear, with recall of 0.5135. These low minority-class recalls reduce the overall macro-F1 to 0.6621, even though the weighted-F1 remains higher at 0.7414 because majority classes contribute more heavily to the weighted average.

Table 3 ViT-Small/16 classification report on RAF-DB test set

Class (Emotion)	Precision	Recall	F1-score	Support
1 (Surprise)	0.8324	0.8754	0.8533	329
2 (Fear)	0.7273	0.5405	0.6202	74
3 (Disgust)	0.5614	0.4000	0.4672	160
4 (Happiness)	0.9483	0.8506	0.8968	1185
5 (Sadness)	0.7403	0.8410	0.7875	478
6 (Anger)	0.8175	0.6914	0.7492	162
7 (Neutral)	0.7111	0.8471	0.7732	680
Accuracy	N/A	N/A	0.8116	3068
Macro Avg	0.7626	0.7209	0.7353	3068
Weighted Avg	0.8185	0.8116	0.8108	3068

Table 3 shows that ViT-Small provides the strongest class-wise performance among the three models. Compared with ResNet18, ViT-Small improves recall for Happiness from 0.7485 to 0.8506 and recall for Neutral from 0.7559 to 0.8471. It also improves the F1-scores of the difficult minority classes, including Fear from 0.5630 to 0.6202 and Disgust from 0.3913 to 0.4672. These improvements explain why ViT-Small achieves the highest macro-F1 of 0.7353 and the highest weighted-F1 of 0.8108. The result suggests that ViT-Small captures more transferable and globally informative facial features than the CNN baseline.

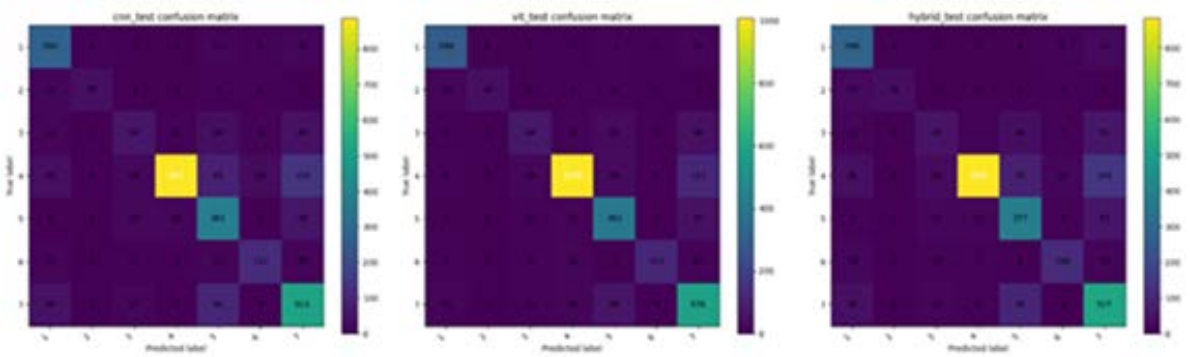
Table 4 Hybrid CNN–Transformer classification report on RAF-DB test set

Class (Emotion)	Precision	Recall	F1-score	Support
1 (Surprise)	0.6793	0.8693	0.7627	329
2 (Fear)	0.7234	0.4595	0.5620	74
3 (Disgust)	0.4091	0.2812	0.3333	160
4 (Happiness)	0.9557	0.7460	0.8379	1185
5 (Sadness)	0.6315	0.7887	0.7014	478
6 (Anger)	0.7013	0.6667	0.6835	162
7 (Neutral)	0.6474	0.7750	0.7055	680
Accuracy	N/A	N/A	0.7370	3068
Macro Avg	0.6782	0.6552	0.6552	3068
Weighted Avg	0.7597	0.7370	0.7381	3068

Table 4 shows that the hybrid CNN–Transformer model achieves strong recall for Surprise at 0.8693 and remains close to the CNN baseline in overall accuracy. However, its minority-class performance is weaker, especially for Fear and Disgust. The hybrid model obtains recall of 0.4595 for Fear and 0.2812 for Disgust, which are lower than both the CNN and ViT results for these classes. This explains why the hybrid model obtains the lowest macro-F1 of 0.6552 despite having an accuracy close to ResNet18. The result suggests that the current hybrid design does not preserve or exploit enough fine-grained expression cues for minority classes, and that the tokenization strategy, transformer depth and pooling mechanism may require further tuning. A likely reason is that the hybrid model introduces additional transformer layers after CNN feature extraction, but the token design may not provide enough discriminative local detail for subtle minority expressions. Since Fear and Disgust have fewer training samples, the hybrid model may also require stronger regularization, a different token source level, or more careful tuning of transformer depth, token count and pooling strategy. Therefore, the result should be interpreted as an empirical limitation of this specific hybrid configuration rather than evidence that CNN–Transformer hybrids are generally unsuitable for FER.

4.3 Confusion Matrix Analysis

Aggregate metrics do not show which emotions are being confused. To expose systematic error modes, confusion matrices are analyzed for each model. The strongest model should exhibit a more dominant diagonal structure (correct predictions), while weaker models typically show heavier off-diagonal leakage especially among minority negative emotions where subtle expression cues are easily confused. This behavior is visualized in Figure 2, which combines confusion matrices for the three models: (a) CNN, (b) ViT, and (c) Hybrid**.** The ViT model exhibits the cleanest separation pattern consistent with its superior macro-F1, while the hybrid model shows comparatively larger confusion mass away from the diagonal for the most difficult classes, aligning with its reduced recall for Fear and Disgust.



(a) (b) (c)

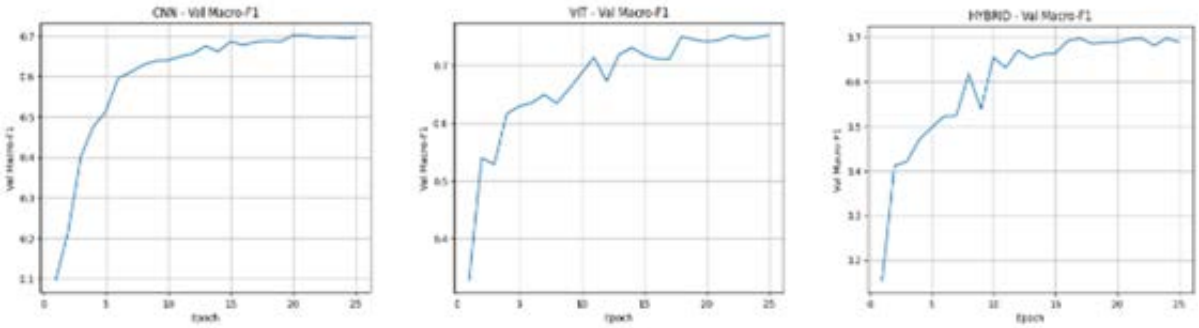
Figure 2. Confusion matrices on RAF-DB test set (N=3068): (a) CNN (ResNet18), (b) ViT-Small/16, and (c) hybrid model

Figure 2 supports the quantitative results by showing that ViT has a stronger diagonal pattern than CNN and hybrid models, indicating more correct predictions. The off-diagonal errors are more visible for minority and ambiguous expressions, especially Fear and Disgust.

The confusion patterns are also consistent with the visual similarity between some expression classes. Fear and Surprise can be confused because both may involve widened eyes, raised eyebrows and open-mouth regions, while the distinction may depend on subtle mouth shape and tension cues. Sadness and Neutral can also overlap because both may contain low-intensity facial movements, especially when mouth corners and eye regions are not strongly expressed. These ambiguities affect minority classes more strongly because fewer training samples are available to learn class-specific boundaries. The ViT model reduces some of these errors compared with CNN and hybrid models, suggesting that global context helps separate expressions with overlapping local cues.

4.4 Validation Macro-F1 Learning Dynamics

Because model selection is based on validation macro-F1, learning dynamics are critical for interpreting whether performance arises from stable convergence or from transient spikes. Validation macro-F1 curves provide a compact view of training progress under class imbalance. The ViT achieves the highest validation macro-F1 peak (matching its strongest test macro-F1), while the hybrid model peaks below both CNN and ViT, consistent with its weaker minority-class recall and overall macro-F1. Figure 3 summarizes the validation macro-F1 progression for each model: (a) CNN, (b) ViT, (c) Hybrid**.** It shows the validation macro-F1 trend used for checkpoint selection. The ViT curve reaches the strongest validation macro-F1, which is consistent with its best test-set macro-F1.



(a) (b) (c)

Figure 3. Validation macro-F1 across epochs: (a) CNN (ResNet18), (b) ViT-Small/16, and (c) hybrid model

5. CONCLUSION

This study evaluated ResNet18, ViT-Small and CNN–Transformer hybrid model for seven-class facial expression recognition on RAF-DB using a unified training pipeline with online augmentation, weighted sampling, class-weighted cross-entropy and macro-F1-based model selection. On the held-out test set, the pretrained ViT-Small achieved the best overall performance, with accuracy of 0.8116 and macro-F1 of 0.7353. The ResNet18 CNN baseline achieved accuracy of 0.7386 and macro-F1 of 0.6621, while the tested CNN–Transformer hybrid achieved accuracy of 0.7370 and macro-F1 of 0.6552. These findings show that the ViT-Small baseline benefits from pretrained global representation learning and provides stronger class-balanced performance on RAF-DB. In contrast, the tested hybrid configuration did not improve over the CNN baseline, suggesting that its tokenization strategy, transformer capacity and pooling design require further optimization. Therefore, the contribution of this work is best understood as a controlled empirical comparison and analysis of model behaviour under class imbalance, rather than as a claim of a superior hybrid architecture. Future work should investigate hybrid-model ablations, including token source level, token count, transformer depth / width, pooling strategy, stronger regularization and cross-dataset generalization.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Universiti Malaysia Perlis grant program for supporting this publication.

REFERENCES

[1] S. Li, W. Deng, 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 356–370, Jan. 2019, doi: 10.1109/TIP.2018.2868382.
[2] K. He, X. Zhang, S. Ren, J. Sun, 2016. Deep residual learning for image recognition, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778.
[3] A. Dosovitskiy *et al.*, 2021. An image is worth 16×16 words: Transformers for image recognition at scale, in Proc. Int. Conf. Learn. Representations (ICLR).
[4] H. Wu *et al*., 2021. CvT: Introducing convolutions to vision transformers, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV).

[5] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution, in Proc. ACM Int. Conf. Multimodal Interact. (ICMI).
[6] A. Mollahosseini, B. Hasani, M. H. Mahoor, 2019. AffectNet: A database for facial expression, valence, and arousal computing in the wild, IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019, doi: 10.1109/TAFFC.2017.2740923.
[7] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, 2020. Region attention networks for pose and occlusion robust facial expression recognition, IEEE Transactions on Image Processing, vol. 29, pp. 4057–4069, doi: 10.1109/TIP.2019.2956143.
[8] Y. Li, H. Liu, J. Liang, D. Jiang, 2025. Occlusion-robust facial expression recognition based on multi-angle feature extraction, Applied Sciences, vol. 15, no. 9, Art. no. 5139, doi: 10.3390/app15095139.
[9] Z. Wen, W. Lin, T. Wang, G. Xu, 2023. Distract your attention: Multi-head cross attention network for facial expression recognition, Biomimetics, vol. 8, no. 2, Art. no. 199, doi: 10.3390/biomimetics8020199.
[10] H. Touvron *et al*., 2021. Training data-efficient image transformers & distillation through attention, in Proc. Int. Conf. Mach. Learn. (ICML).
[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, 2017. Focal loss for dense object detection, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV).
[12] I. Loshchilov, F. Hutter, 2019. Decoupled weight decay regularization, in Proc. Int. Conf. Learn. Representations (ICLR).
[13] I. Loshchilov and F. Hutter, 2017. SGDR: Stochastic gradient descent with warm restarts, in Proc. Int. Conf. Learn. Representations (ICLR).
[14] P. Micikevicius *et al*., 2018. Mixed precision training, in Proc. Int. Conf. Learn. Representations (ICLR).