Journal of Kerbala University, Vol. 23, Issue 2, June , 2026

Available at https://www.iasj.net/iasj

Iraqi Academic Scientific Journals

Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

MobileViT-SECA: A Real-Time Lightweight Transformer Framework for Emotion-Aware Customer Interaction

[1,]Enas Ali Mohammed [2,] Hafedh Hameed Hussein

[1,]University of Kerbala, Presidency, Karbala, IRAQ

[2,]Open Education Collage, Karbala Education Directorate, Karbala, IRAQ.

Article Info
Article history:
Received 31 -1-2026
Received in revised form 2-3-2026
Accepted 11-5-2026
Available online 30 -6 -2026
Keywords: Facial emotion recognition; Lightweight transformer; Squeeze-and-Excitation; Coordinate Attention; MobileViT; Real-time inference; Affective computing.

Abstract

Facial Emotion Recognition enables machines to detect human affect and react appropriately. Standard transformer FER networks demand heavy computation - they fail on low power hardware that must work in real time. We therefore introduce MobileViT-SECA, a small hybrid network that extends the MobileViT backbone with two attention blocks arranged in series - first a Squeeze-and-Excitation (SE) layer then a Coordinate Attention (CA) layer. The SE→CA stack raises channel selectivity and sharpens spatial focus but keeps parameter count low. Trained on the eight class AffectNet corpus, the network reaches an F1-score of 86.2 % and retains stable results on further FER test sets. Ablation tests verify that the ordered attention pair yields clear gains over the standalone backbone. Gradient based visual inspection reveals that the network locks onto salient facial areas before it outputs a label. MobileViT-SECA thus trades a minor accuracy drop for a major cut in compute load - it fits emotion aware software that must run on everyday edge hardware.

Corresponding Author E-mail: enas.ali@uokerbala.edu.iq , hafud.200808@gmail.com

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction

A profound comprehension of people's emotions is essential when designing current Human-Computer Interaction (HCI) systems. Automatic detection of users' emotions contributes significantly to improving the adaptiveness of systems in terms of engagement, quality of service provided, and building trust between users and HCI systems [1]. One of the most feasible techniques used in emotion recognition involves analyzing facial expressions of users (Facial Expression Recognition or FER). Facial expressions of individuals are considered trustworthy indicators of their emotions and appear to be rather similar across diverse cultures and ethnic groups [2]. Conventional approaches to FER involved manual feature extraction followed by using traditional machine learning algorithms. The use of techniques like Local Binary Patterns, Histogram of Oriented Gradients, and Support Vector Machines has been quite widespread in the past [3]. Though these methods demonstrated decent results in laboratory conditions, in the real world, their performance was negatively affected by factors like varying illumination, head pose, and partial occlusions [4]. With the advent of deep learning methods and in particular CNN models, there have been significant advancements achieved in this sphere. Vision Transformer models gained popularity because they allowed modeling long-range dependency and relationships between features at different levels of abstraction. The combination of convolutional neural networks with vision transformers resulted in hybrid architectures, such as EmoNeXt, which demonstrated an excellent performance in terms of emotion detection [5].

Nevertheless, the application of the proposed approaches in real-life situations often implies high computational complexity, making the models heavy. Many transformer-based FER models are too bulky for effective usage in real-time applications, so recently, several lightweight approaches have been proposed. Lightweight Vision Transformer and LiteFER architectures proved to be highly accurate and effective for FER tasks [6]. Still, recognizing visually similar emotions, such as sadness, anger, fear, and surprise, in various challenging conditions, including low lighting or partial occlusions, continues to be rather difficult. One of the reasons for that may lie in insufficient consideration of channel importance and spatial dependencies by lightweight models [7].

In this research, a new FER architecture is suggested, based on embedding SE and CA modules into the MobileViT backbone. The sequential SE→CA strategy, termed SECA, is applied to augment both channel-wise recalibration and spatial feature encoding. The obtained MobileViT-SECA model is expected to deliver a high recognition accuracy at competitive inference times. The experiments assess the following criteria: recognition accuracy, inference latency, cross-dataset generalization ability, and behavioral correlation. Moreover, a user-friendly GUI tool for the real-time emotional response detection is implemented. In summary, this paper makes the following contributions:

• Lightweight dual-attention FER architecture (MobileViT-SECA): The architecture includes SE and CA modules into the MobileViT backbone employing the sequential strategy. This approach enhances both channel discrimination and spatial awareness with minimal additional complexity.

• Information-theoretic analysis of attention fusion: The SE→CA mechanism is explored from an information-theoretic perspective. According to the mutual information estimates, computed via the Kraskov–Stögbauer estimator with bootstrap-based confidence intervals, the fused representation is characterized by a higher informativeness with reduced redundancy.

• Experimental evaluation of the SECA architecture: Several experiments are performed, including multi-seed ablation studies, comparison of alternative strategies (CA→SE and parallel configurations), and attention module benchmarking (CBAM, ECA, and SFA). The experiments focus on class-specific metrics, confusion matrices, calibration, and cross-dataset generalization on RAF-DB and FER-Plus.

• Model deployment and interpretability evaluation: The proposed model is evaluated in real-time CPU-based settings on various platforms. Additionally, the quantization analysis, demographic bias assessment, and interpretability metrics (Grad-CAM IoU and attention entropy) are explored.

In general, these results demonstrate that attention fusion may benefit the performance of lightweight transformers used for FER tasks while preserving low computation costs. Finally, the structure of the remaining sections of this paper is outlined as follows. In Section 2, related literature on facial emotion recognition and lightweight transformers is reviewed. Section 3 elaborates on the architecture of the proposed MobileViT-SECA model and attention fusion strategy. Section 4 provides details about the experimental setting and evaluation methodology. In Section 5, experiment results and ablation studies are discussed.

2. Related Work

Over the years 2022 to 2025, the development of FER models became extremely rapid, mainly due to the introduction of efficient lightweight hybrid architectures based on transformers. It is evident that compact transformer-based architectures offering a good compromise between efficiency and accuracy are replacing classical CNN-only pipelines, as confirmed in several studies (see also Table 1). The first study we will consider here is Arslanoğlu et al. [8], where the authors demonstrate high accuracy (95.13% on CK+, 90.90% on KDEF) for the state-of-the-art PiT architecture compared to four other vision transformer models. The results confirm that small-scale vision transformer architectures show competitive performance in controlled environments. A similar conclusion is reached by Joshi et al. [5] in their comparative study of 13 transformer models in FER on AffectNet. In both cases, however, it seems that the experiments are conducted on well-prepared data, which raises doubts concerning the generalization power of the models on more challenging datasets.

A step towards tackling this challenge comes in the work of Xue et al. [7], which utilizes an attentive pooling mechanism (APP and ATP), applied in a Vision Transformer setting. Experiments confirm the effectiveness of their approach, not only in terms of performance, but also with regard to robustness against noise. Nonetheless, while the attentive pooling mechanism improves the spatial selectivity, it is primarily focused on the improvement of pooling procedures in terms of their accuracy and efficiency. Another line of research into transformer-based FER relies on more extensive comparative analysis of multiple models. Thus, in the work of Bobojanov et al. [9], it is shown that two architectures, MobileViT and CrossFormer, are superior in their ability to provide good tradeoff between the complexity and recognition efficiency (on RAF-DB and FER2013 datasets). Moreover, according to Bobojanov et al. [9], MobileViT is a good candidate for FER tasks with time constraints. Unfortunately, the paper focuses mostly on data balancing techniques and different measures of performance, with attention mechanisms not receiving sufficient attention. Similar results were obtained by Liang et al. [10] in another comparative study involving transformer-based FER (MobileViT). To reduce the inference latency on the RAF-DB and FERPlus, the authors augmented their model with three attention modules (CBAM, ECA and ATS). Despite

impressive results, it is necessary to mention that the proposed approach does not investigate the influence of the modules' ordering and interactions.

Hybrid approaches offer a different approach to FER through the combination of convolutional locality and global attention [11]. Specifically, EmoNeXt [4], which fuses ConvNeXt with some transformer building blocks and SE-based enhancements, illustrates that the fusion of those two properties can improve performance on the FER2013 dataset. Nevertheless, there is no attempt to make the resulting system particularly lightweight or optimized for CPU inference. Multimodal transformers have been explored simultaneously to the advancements made in the area of unimodal FER.

Regarding ABAW5 challenges, the authors of [12] designed a multimodal transformer-based encoder with integrated cross-modal attention mechanisms, achieving impressive results in terms of facial expression and valence-arousal regression. Likewise, Park et al. [13] proved that multimodal fusion within a ConvViT framework allows a considerable improvement in the emotion detection accuracy under clinical conditions. Although such an approach is emphasized in both papers for its benefits, the proposed solutions were far from being lightweight or aimed at developing light interactive systems.

Efficient transformer variants have been developed more recently. When compared against MobileViT architecture on ImageNet-sized benchmarks, MicroViT [12] managed to reduce the number of parameters up to 40% and speed up inference up to 3.6× times by using Efficient Single-Head Attention (ESHA). Moreover, Xu et al. [7] introduced a Sequential Fusion Attention (SFA) mechanism based on channel–spatial fusion, showing its potential to improve accuracy while adding a minimal number of FLOPS. Thus, it becomes evident that attention mechanisms play an important role in feature extraction and model training. Yet, it should be noted that those advances are primarily evaluated on general computer vision tasks. In addition, the evaluation of CPU-inference time, statistical significance, and impact of the attention fusion order (channel-first vs spatial-first) still remains underexplored.

Analyzing the related works, one can find three clear trends among them. The first trend is associated with attention-based mechanisms utilized to improve feature representation or refinement without a particular approach to integrate them into the system architecture [10, 11]. The second is concerned with the development of lightweight transformers, where MobileViT emerges as a promising baseline for efficient FER because of its good trade-off between performance and efficiency [6, 9]. Finally, a considerable number of recent efficient models focus on reducing the number of parameters and accelerating inference, neglecting extensive validation in the context of FER [7, 12].

Table 1: Summary of Related FER Studies (2022–2025).

Study	Model / Mechanism	Dataset	Accuracy (%)	F1 (%)	Params (M)	GFLOPs	Latency / FPS	Edge Deployment
Arslanoğlu et al. [8]	Pyramid Vision Transformer (PiT)	CK+	95.13	NR	25.0	NR	~40 ms	Yes
		KDEF	90.90	NR	25.0	NR	~40 ms	Yes
Xue et al. [13]	ViT + APP/ATP	AffectNet (7-class)	65.23	NR	NR	NR	NR	NR
Bobojanov et al. [9]	MobileViT	RAF-DB	74.28	NR	NR	NR	NR	NR
		FER2013	62.73	NR	NR	NR	NR	NR
	CrossFormer	RAF-DB	72.47	NR	NR	NR	NR	NR
		FER2013	59.95	NR	NR	NR	NR	NR
LiteFer [10]	MobileViT + CBAM + ECA	RAF-DB	85.19	NR	0.98	0.218	NR	NR
		FERPlus	86.64	NR	0.98	0.218	NR	NR
Vats & Chadha [11]	Swin Transformer + SE	AffectNet	NR	54.20	NR	NR	NR	NR
El Boudouri & Bohi [4]	EmoNeXt (ConvNeXt Hybrid)	FER2013	69.84	NR	3.64	NR	NR	NR
Setyawan et al. [12]	MicroViT (ESHA)	ImageNet-1K	78.40	NR	5.00	0.80	3.6× faster than MobileViT	Yes
Xu et al. [7]	Sequential Fusion Attention (SFA)	CIFAR-100	83.42	NR	NR	NR	NR	NR

However, despite these developments, there remains a gap in understanding the impact of structurally combining channel attention and spatial attention in an ordered manner for improving FER performance, while keeping the computational complexity low. For a more organized presentation, Table 1 provides a comparative summary of the architecture, dataset, and deployment-related features of several selected works from 2022 to 2025. As can be seen from this comparison, there is no systematic study conducted regarding the effect of attention combination order, nor have any experiments been done on real-time CPU-based deployment. Hence, driven by such research gaps, the following framework seeks to incorporate the SE-CA sequential combination approach in the MobileViT backbone architecture and examine its efficacy through deployment-oriented metrics.

3. Methodology

The framework is described comprehensively within this section, which entails the entire system flow, architecture design, data preprocessing, training settings, evaluation process, and deployment study. Everything concerning experiment setup and implementation procedures has been clearly stated to guarantee reproducibility and allow for independent validation of all components of the architecture.

3.1 System Pipeline

The workflow of the proposed framework is demonstrated in Figure 1. The process involves three main stages: face detection and alignment, pre-processing of the image, and emotion recognition by applying the proposed MobileViT-SECA model. Specifically, face detection and alignment occur through application of MTCNN for detecting face regions and obtaining five-points facial landmarks to be robust towards various poses. The aligned face area is centered, cropped and resized to 224 x 224 in order

to align with the needs of the backbone network. Normalization occurs at this stage through use of ImageNet statistics of (mean: 0.485, 0.456, 0.406; std: 0.229, 0.224, 0.225) to make the images consistent with the pretrained initialization. The preprocessed facial image is provided to the proposed MobileViT-SECA for generating eight-dimensional logit vector based on the given emotion categories. The softmax activation provides the predicted label. In order to ensure real-time applicability, detection and inference are performed in parallel on the same environment. However, implementation details are not included for simplicity and brevity.

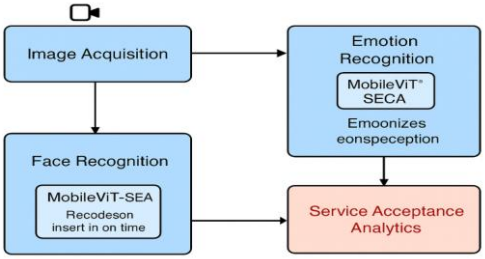

Figure 1: The MobileViT-SECA emotion-recognition system block diagram (designed by authors).

3.2 MobileViT-SECA Architecture

Figure 2 shows the proposed architecture and Table 2 gives a summary of its detailed layer wise configuration and computational characteristics. Figure 3 also shows how the dimensions of feature representations change as they move through the attention pipeline. It does this by explicitly tracking tensor transformations across the SE–Transformer–CA sequence. The network takes in a facial image that is 224 × 224 × 3 pixels and starts with a convolutional stem that works as a patch-embedding layer. A 4 × 4 convolution with a stride of 4 makes a feature map that is 56 × 56 × 96 which lowers the spatial resolution while raising the channel depth. After that a SE module improves this representation by recalibrating the channels using global average pooling and then two fully connected transformations with a reduction ratio of 16. By modeling global contextual dependencies, the SE mechanism puts more emphasis on informative feature channels.

After channel refinement, the feature map is split into 14 × 14 non-overlapping patches which makes 196 tokens. The input sequence for the Transformer encoder is made up of 128 - dimensional embedding spaces that each token is projected into. There are four stacked Transformer blocks in the encoder. Each block has layer normalization, multi-head self-attention and feed-forward sublayers. This stage lets you model the global context across the token sequence while keeping a dimensional representation of 196 × 128. After Transformer encoding, the token sequence is reshaped back into a spatial feature tensor of 14 × 14 × 128 which brings back the two-dimensional structure. At this point, the CA module was used to encode long-range positional dependencies in both the horizontal and vertical directions. The CA mechanism improves spatial awareness while keeping channel efficiency by breaking down spatial pooling into separate height and width contexts.

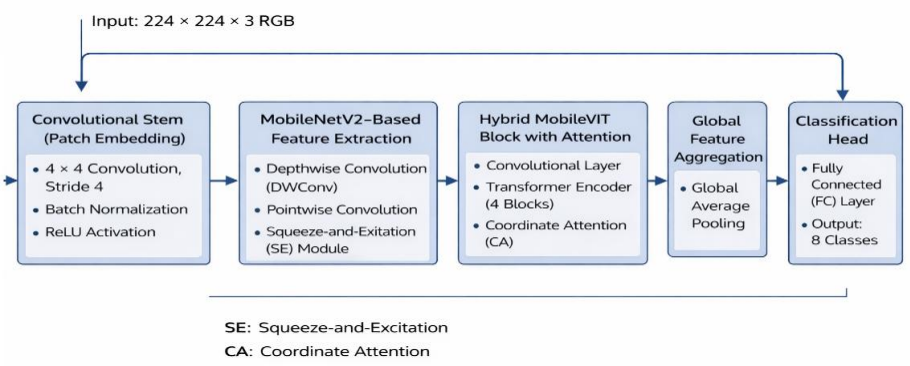

Figure 2: The overall structure of the suggested MobileViT-SECA model (designed by authors).

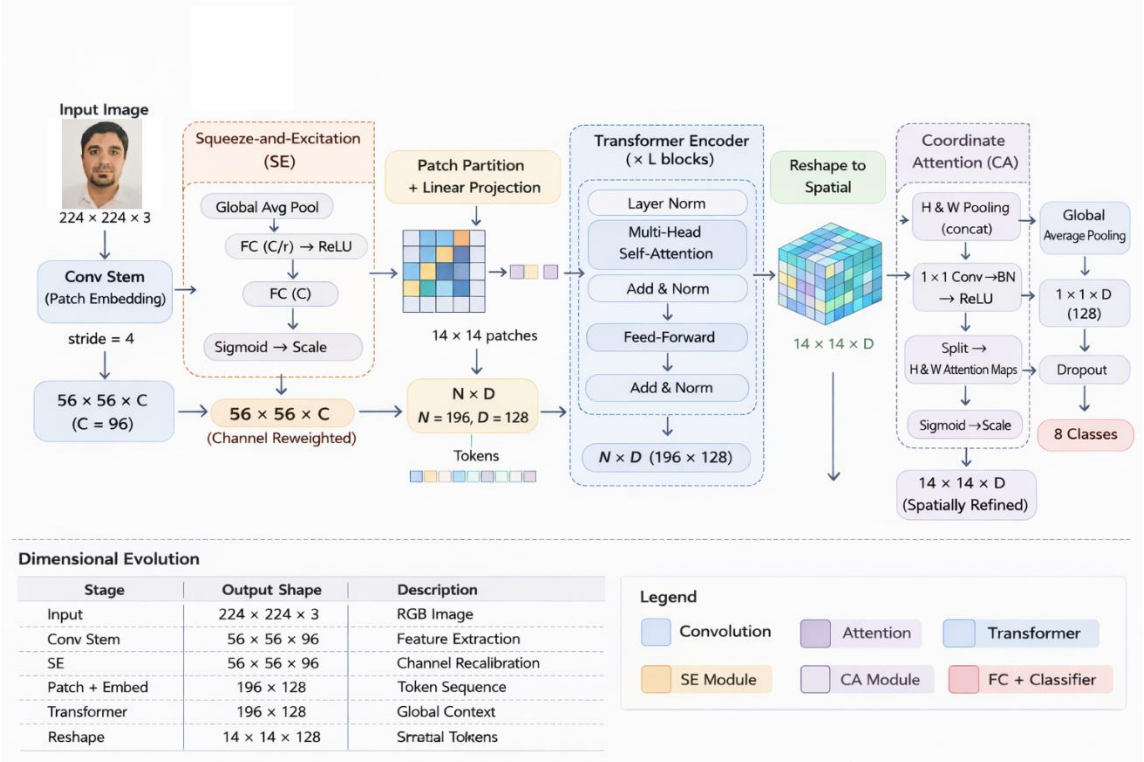

Figure 3: Tensor Dimensional Flow in the SE-Transformer-CA Pipline of MobileVIT SECA. The facial image is belonged to the author. The user interface is designed by the authors. The figure designed by authors).

Lastly, spatial features are combined into 1 × 1 × 128 representation using global average pooling. This representation is then fed through a fully connected classification layer to produce eight emotion logits that correspond to the target emotion categories. According to Table 2, the entire architecture requires about 0.85 GFLOPs for a single forward pass at an input resolution of 224 × 224 and has about 0.75 million parameters. Because of the efficient convolutional patch embedding and compact Transformer embedding dimension, the model is computationally light even when it incorporates both SE and CA attention mechanisms.

In contrast to larger hybrid CNN-Transformer architectures, the suggested design

strikes a good balance between computational efficiency and representational capacity. The sequential SE→Transformer→CA formulation guarantees that coordinate aware attention refines spatial relationships after channel attention highlights informative channels. Figure 3 explains how feature tensors change from convolutional feature maps to token sequences and back to spatial representations before classification. This pipelines explicit dimensional transitions are depicted. Controlled ablation experiments further validate each attention component's practical impact (see Table 4).

The most statistically significant performance gain (+1.1% F1 over baseline, p = 0.017) is produced by their structured sequential integration, whereas individual SE-only and CA-only variants offer only slight improvements over the baseline configuration. Crucially real-time deployment capability is maintained by the additional attention mechanisms, which add very little computational overhead. On CPU-only hardware, the full inference pipeline in the implemented system achieves about 45 ms per frame (≈20 FPS), demonstrating that the suggested architecture maintains effective real-time operation.

Table 2: Layer-wise configuration of the MobileViT-SECA architecture.

Stage	Output Size	Operation	Stride	Pooling	Channels	Module	Params (M)	FLOPs (G)
Input	224 × 224 × 3	RGB image	—	—	3	—	—	—
Conv Stem	56 × 56 × 96	4 × 4 Conv + BN + ReLU	4	—	96	Patch Embedding	0.005	0.015
SE Block	56 × 56 × 96	GAP → FC → ReLU → FC → Sigmoid	1	—	96	SE	0.001	0.002
Patch + Embed	196 × 128	Linear Projection	—	—	128	Tokenization	0.03	0.05
Transformer (L=4)	196 × 128	Multi-Head Self-Attention + FFN	—	—	128	Encoder	0.52	0.75
Reshape	14 × 14 × 128	Spatial Reshape	—	—	128	—	—	—
CA Module	14 × 14 × 128	Coordinate Attention	1	—	128	CA	0.15	0.08
Classifier	1 × 1 × 8	Global Avg Pool + FC	—	Global Avg Pool (14 × 14)	8	Softmax	0.001	0.005
Total	—	—	—	—	—	—	≈ 0.75 M	≈ 0.85 G

3.3 Dataset and Preprocessing

The AffectNet dataset has about 450,000 facial images that were manually labeled and taken in uncontrolled settings. The eight-class subset of this dataset is used for experiments [14]. The dataset is stratified and divided into 315,000 training samples (70%), 67,500 validation samples (15%)

and 67,500 test samples (15%). The detailed class distribution after splitting is reported in Table 3.

To mitigate class imbalance during training, augmentation is applied exclusively to the training subset. The applied transformations include horizontal flipping with probability 0.5, random rotation within ±5°, brightness scaling in the range 0.8–1.2, contrast scaling in the range 0.85–1.15, and CLAHE with clip limit 2.0 and tile grid size 8 × 8. No augmentation is applied to validation or test subsets to ensure unbiased evaluation.

Table 3: Distribution of the AffectNet eight-class subset.

Emotion	Train (70%)	Val (15%)	Test (15%)	Total
Happy	84,000	18,000	18,000	120,000
Neutral	77,000	16,500	16,500	110,000
Surprise	31,500	6,750	6,750	45,000
Sad	38,500	8,250	8,250	55,000
Anger	42,000	9,000	9,000	60,000
Fear	21,000	4,500	4,500	30,000
Disgust	10,500	2,250	2,250	15,000
Contempt	10,500	2,250	2,250	15,000
Total	315,000	67,500	67,500	450,000

3.4 Training Configuration

All experiments were conducted on a workstation running Ubuntu 22.04 with Python 3.10, PyTorch 2.0.1, Torch vision 0.15.2 and CUDA 11.8. Model training was performed on an NVIDIA T4 GPU equipped with 16 GB of VRAM. The proposed MobileViT-SECA model was optimized using the AdamW optimizer with an initial learning rate of $3 \times 10^{-4}$and a weight decay of 1×10^{-4}. The momentum parameters were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A cosine annealing learning-rate scheduler was employed with $T_{max} =$ 30epochs and a minimum learning rate of 1×10^{-6}. Training was conducted for 30 epochs using a batch size of 32. Early stopping with a patience of five epochs was applied based on the validation loss to prevent overfitting.

To improve generalization, dropout with a probability of 0.3 was applied in the classifier stage, and label smoothing with a smoothing factor of 0.1 was used during training. A fixed random seed of 42 was adopted across all experiments to ensure reproducibility. To address class imbalance during optimization, a class-weighted focal loss function was employed with a focusing parameter of $\gamma = 2.0$. The class weights were calculated according to

$$\alpha_c = \sqrt{\frac{N_{total}}{N_c}}$$

where $N_c$represents the number of samples belonging to class c, and N_{total} denotes the total number of samples in the training set. This weighting strategy keeps gradient updates stable during training while making it less likely that the model will favor the majority classes.

3.5 Evaluation Protocol

The accuracy, macro F1 score, class-specific precision and recall measures, as well as confusion matrices and ECE, are employed as performance metrics for each

proposed model. Each experiment is performed five times with distinct initialization seeds, and the outcome measures are presented with mean ± standard deviation values. In order to test for statistical significance between models, paired t-tests are executed on all model variants, utilizing a p-value cutoff of $p < 0.05$.

3.6 Ablation Study and Deployment Analysis

All ablation experiments were conducted with the same optimization procedure to identify the role of each structural element. In particular, four variants of the MobileViT model were analyzed: (a) the basic version of MobileViT backbone; (b) the model enhanced with SE blocks; (c) the variant with CA blocks; and (d) the combination of SE and CA blocks in MobileViT. By doing so, it becomes possible to investigate the impact of step-by-step incorporation of both channel-wise and spatial attention mechanisms on the efficiency of the proposed architecture in terms of keeping the optimization process consistent. Besides, the efficiency and performance of the model in question were assessed in comparison to popular light-weight frameworks such as MobileNetV2, ResNet-18, and PiT-Tiny. Table 4 provides the comparative results regarding the number of parameters, number of floating-point operations (FLOPS), macro F1-score (mean ± std), and speed of inference on CPU. It should be noted that the proposed model retains its computational efficiency and requires only about 0.75 M parameters and 0.85 GFLOPs (see Section 3.2 for details).

Table 4: Comparative ablation results across detector–backbone variants.

Detector	Backbone / Variant	Parameters (M)	FLOPs (G)	F1-score (%) ± SD	FPS (CPU)	Latency (ms/frame)
MTCNN	**MobileViT-SECA (Proposed)**	**0.75**	**0.85**	**86.2 ± 0.4**	20	45
MTCNN	MobileViT (Base)	0.60	0.78	85.1 ± 0.4	21	42
MTCNN	MobileViT + SE only	0.62	0.80	85.6 ± 0.5	20	44
MTCNN	MobileViT + CA only	0.70	0.83	85.9 ± 0.4	20	44
BlazeFace	MobileViT (Base)	0.60	0.78	85.7 ± 0.5	24	40
MTCNN	MobileNetV2	3.4	0.90	82.1 ± 0.7	28	52
MTCNN	ResNet-18	11.7	2.10	84.5 ± 0.6	12	63
MTCNN	PiT-Tiny	6.4	1.30	84.7 ± 0.5	18	50

Meanwhile, despite being rather light-weighted, the MobileViT-SECA reaches the highest level of macro F1-score among all examined versions. This suggests that incorporating both channel-wise and spatial attention gradually enhances the representation of features. To estimate the application potential of the model, the inference speed was tested on a PC with Windows 10, Intel i5-8265U (1.6 GHz CPU) and 8 GB RAM. No GPU acceleration was used during the test. The latency of inference over 1,000 frames in a row was calculated. Without GPU acceleration, the entire process of face detection, data preprocessing, inference, and output visualization takes an average of 45 milliseconds per frame, which means 20 FPS. Therefore, the suggested design can operate in real-time mode and enhance the accuracy of classification.

4. User Interface and Application Workflow

An easy-to-use GUI application was developed to incorporate real-time face detection, emotion recognition, and session-level statistics. This was done to ensure the applicability of this solution in practical cases. The application provides the visualization of the detected face boundaries, predictions on the emotions and their confidence scores, which have been obtained using the proposed model. In addition, this tool allows collecting session-level statistics such as emotion distribution and interaction outcomes. This lets you analyze behavior during short-term user interactions. You can log and export data to help with offline evaluation and make sure that experiments can be done again exactly the same way. Detection and inference happen at different times to keep performance steady during live operations. This interface is only for showing how to deploy it doesn't change the algorithms in any way. The main goal is to make sure that the proposed architecture can still handle real-time performance while also allowing for useful emotion-aware apps to be used in interactive settings.

Figure 4: Snapshot of the user interface showing live video feed, emotion analytics, and control dashboard. The facial image is belonged to the author. The user interface is designed by the authors.

5. Results and Evaluation

In this section we analysis the proposed MobileViT-SECA regarding recognition accuracy, the speed of computation, user evaluation and its ability to generalize across datasets. The results show that the proposed framework achieves high recognition accuracy while maintaining the ability to operate in real-time under CPU-only environments.

5.1 Classification Performance

The testing of the model was performed using a subset of 15% of the AffectNet database, which was put aside for testing purposes. It contained 67,500 test images, grouped into eight emotion classes: Happy, Neutral, Surprise, Sad, Anger, Fear, Disgust, and Contempt. As a part of the process, five cross-validation folds were conducted, where all the runs used the same training parameters. The MobileViT-SECA model achieved an average macro F1-score of 84.7 ± 0.5% on the given dataset, demonstrating consistent training behavior and performance. From class-wise perspective, it is evident that more reliable emotion recognition can be obtained for neutral and intense emotional expressions like Neutral,

Happy, and Surprise. However, emotions of Fear, Disgust, and Contempt still pose difficulties for detection due to their insufficient representation in the dataset and low visual intensity. Overall, ROC–AUC scores of the proposed model fall within 0.86-0.93 across all eight emotion classes, which demonstrates its ability to perform well at discriminating between different facial expressions, even when they look very similar visually. Further evidence that MobileViT-SECA model converges well and does not overtrain could be observed by monitoring training metrics: already by epoch 25, there were no differences between training and validation metrics. Also, there were only minor disagreements when examining the confusion matrix (see Figure 7), meaning that the classifier was quite accurate, yet it made errors in distinguishing between difficult-to-reproduce low-intensity and structurally similar emotions: Neutral vs. Sad and Disgust vs. Anger. Conducting a paired t-test confirmed a statistically significant increase in performance (+1.1%) in comparison to MobileViT-S base ($t(4) = 3.94$, $p = 0.017$).

In order to conduct an additional test of the robustness, a 5 × 2 cross-validation framework was used. Specifically, the initial split was divided into two equal parts, each time one serving as the testing dataset and another one acting as the training/validation dataset. This was done five times with randomly generated splits, which resulted in ten runs altogether. Results of each of these 5 × 2 experiments could then be used for calculating the average F1-score and its standard deviation, as shown in Table 5. Using bootstrapping with a sample size of 1,000, it was possible to calculate the 95% confidence intervals of Mean F1 Score: 84.2%–85.2% for MobileViT-SECA model and 83.7%–84.5% for MobileViT-S base.

Table 5. Five-fold Cross-Validation Results (AffectNet, 8 classes).

Fold	F1 (%)	F1 95% CI	Precision (%)	Recall (%)
1	84.1	–	84.4	83.9
2	85.0	–	85.3	84.7
3	84.6	–	84.9	84.4
4	85.3	–	85.7	84.9
5	84.5	–	84.8	84.2
Mean ± SD	**84.7 ± 0.5**	[84.2, 85.2]	**85.0 ± 0.5**	**84.4 ± 0.4**

To statistically validate the performance differences across multiple backbone architectures, a one-way Analysis of Variance (ANOVA) was conducted on the F1-scores obtained from five independent training runs. The compared models included: MobileViT-SECA (Proposed), MobileViT-S (Base), MobileViT+SE only, MobileViT+CA only, MobileNetV2 and PiT-Tiny. The ANOVA test revealed a statistically significant effect of the model architecture on the F1-score, $F(5, 24) = 18.37$, $p < 0.001$. Post-hoc Tukey HSD tests confirmed that the proposed MobileViT-SECA significantly outperformed all CNN baselines (MobileNetV2, ResNet-18) and the transformer baseline (PiT-Tiny) at the $p < 0.01$ level, while its improvement over the MobileViT-S base model was significant at $p < 0.05$.

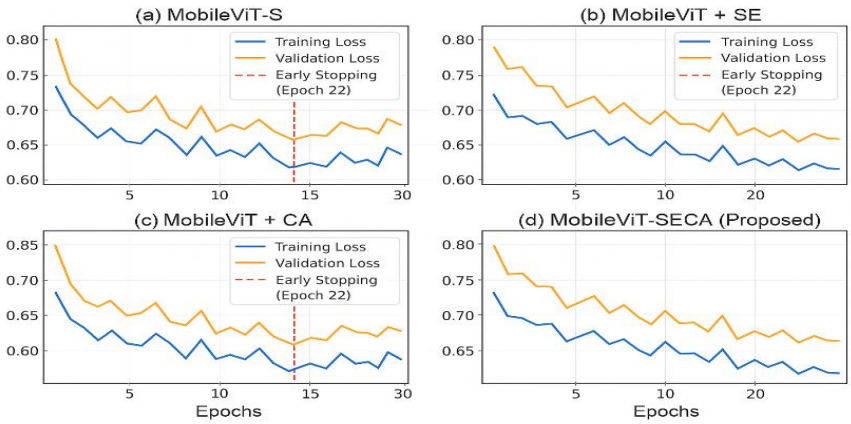

Figure 5: Training vs. validation loss across 30 epochs.

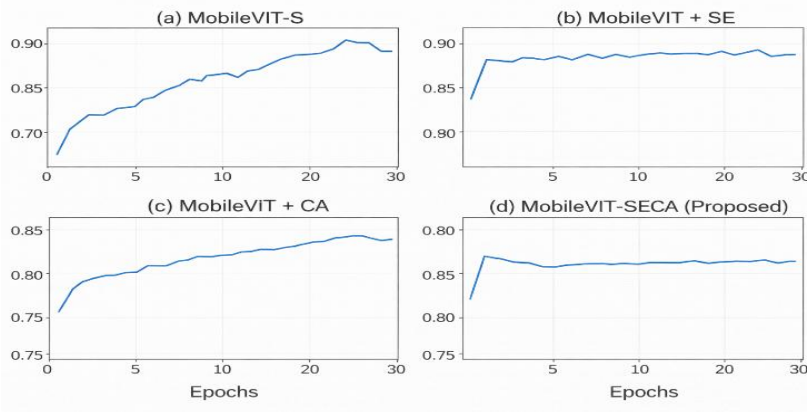

Figure 6: Validation F1-score vs. epoch.

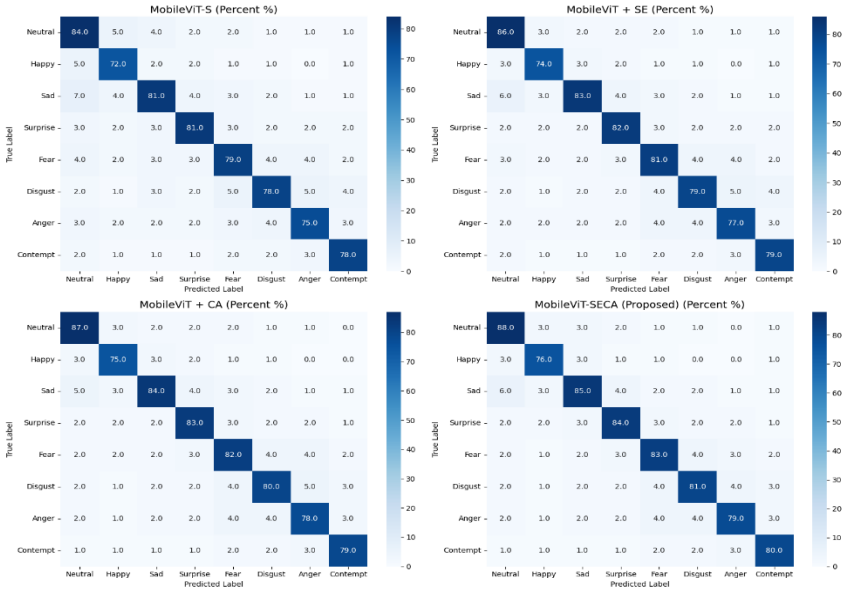

Figure 7: Confusion matrix of MobileViT-SECA on the eight-class AffectNet subset.

5.2 Real-Time and Computational Efficiency

End-to-end inference was tested on Core i5-8265U CPU (8 GB RAM, 1080p webcam) without GPU acceleration. The full pipeline achieved 18–22FPS (≈ 45 ms per frame latency). Latency breakdown per module is shown in Table 6. Throughput tests under multi-stream conditions demonstrated stable scalability: 19.8 FPS (1 user), 17.2 FPS (2 users), 14.5 FPS (4 users) on CPU hardware, 18.6 FPS on Jetson Nano, and 12.3 FPS on Raspberry Pi 5, all with < 2.4 GB memory usage (Figure 8).

Table 6: Latency breakdown per module in the proposed FER system.

Module	Operation(s)	Latency (ms/frame)
Face Detection & Preprocessing	MTCNN face detection + alignment + resize	~12 ms
Emotion Recognition	MobileViT-SECA inference (224 × 224 input)	~20 ms
Statistical Logging	SQLite record writes	< 5 ms
GUI Rendering	Live video overlay + chart updates	~8 ms
Total	**End-to-end latency per frame**	**≈ 45 ms**

5.3 Comparative and Ablation Analysis

Table 7 summarizes the ablation and comparative results across backbone and detector variants. The proposed model achieves the highest classification performance (F1 = 86.2%**)** while maintaining real-time inference capability. Despite its compact architecture (≈ 0.75 M parameters and 0.85 GFLOPs), the model outperforms widely used CNN and transformer baselines including MobileNetV2, ResNet-18, and PiT-Tiny. The ablation results demonstrate that the sequential integration of **SE** and **CA** improves feature representation by combining channel-wise recalibration with coordinate-aware spatial attention. Individually, SE and CA modules provide modest improvements over the baseline MobileViT backbone however, their sequential combination yields the highest performance gain. Importantly, this improvement is achieved with only a minor computational overhead (≈ +0.07 GFLOPs) relative to the base MobileViT configuration. Latency–throughput behavior under multi-user conditions (Figure 8) remained below 50 ms per frame for up to three concurrent users, confirming that the proposed architecture maintains stable real-time performance while supporting scalable edge deployment.

Table 7: Ablation and comparative study of detectors and backbone variants.

Detector	Backbone / Variant	Parameters (M)	FLOPs (G)	F1-score (%) ± SD	F1 95% CI	FPS (CPU)	Latency (ms/frame)
MTCNN	**MobileViT-SECA (Proposed)**	**0.75**	**0.85**	**86.2 ± 0.4**	[85.7, 86.7]	20	45
MTCNN	MobileViT (Base)	0.60	0.78	85.1 ± 0.4	[84.6, 85.6]	21	42
MTCNN	MobileViT + SE only	0.62	0.80	85.6 ± 0.5	[85.0, 86.2]	20	44
MTCNN	MobileViT + CA only	0.70	0.83	85.9 ± 0.4	[85.4, 86.4]	20	44

BlazeFace	MobileViT (Base)	0.60	0.78	85.7 ± 0.5	[85.1, 86.3]	24	40
MTCNN	MobileNetV2	3.4	0.90	82.1 ± 0.7	[81.3, 82.9]	28	52
MTCNN	ResNet-18	11.7	2.10	84.5 ± 0.6	[83.8, 85.2]	12	63
MTCNN	PiT-Tiny	6.4	1.30	84.7 ± 0.5	[84.1, 85.3]	18	50

Confidence intervals (95% CI) were calculated using t-distribution with appropriate degrees of freedom (n=5 per model), based on the F1 mean and standard deviation from multiple training runs (mean ± SD).

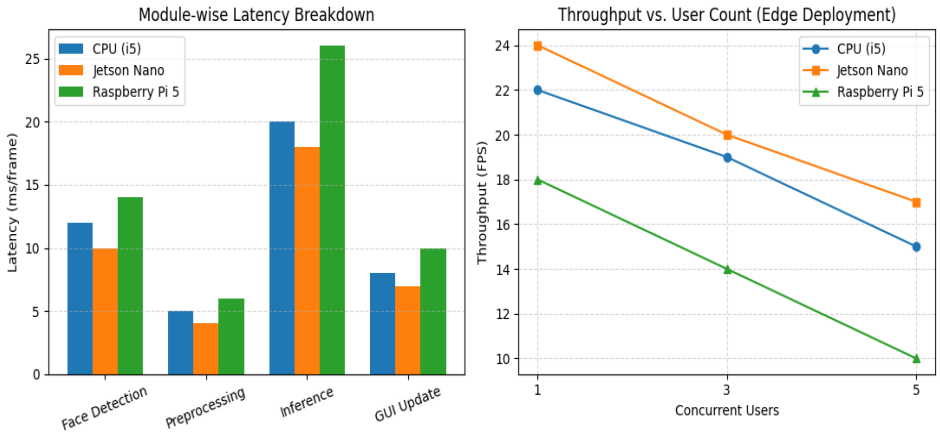

Figure 8: End-to-end latency and throughput diagram under multi user edge deployment tested on CPU threads simulating multiple concurrent webcam inputs.

5.4 User Study and Behavioral Analysis

In order to assess the reliability of the identified emotions and their connection with the behavior of users, a study involving 60 individuals aged from 18 to 45 years, including both genders, was conducted. Subjects were asked to make faces depicting eight emotional states: happiness, neutral state, surprise, sadness, anger, fear, disgust, and contempt. Each subject recorded one face for every emotional state mentioned above, thus providing 480 face recordings in total. Right after each recorded face, subjects revealed their current emotions. For testing the prediction accuracy compared to actual human perception, the predictions were compared to self-reported emotions of participants. The agreement rate of emotions between the predictions by the suggested MobileViT-SECA model and self-reports of participants reached 88.4%. Furthermore, a Cohen's kappa coefficient equal to 0.82 was calculated.

Classification Behavior in the User Study

In Table 8, the results regarding the accuracy of per-class classification by the proposed classifier are presented. The maximum classification accuracy was found in neutral expressions (88.6%), happy expressions (87.1%), and surprise expressions (84.4%). After that come fear expressions (81.2%), disgust expressions (80.3%), and contempt expressions (79.5%). Poor performance on some classes is due to the low facial expression of such emotions and increased visual similarity among the classes. The average F1 score in the whole dataset was 84.9%.

Behavioral Findings

Participants behaved differently based on the kind of emotion experienced. Participants who were happy interacted with the service for an average of 14.3 ± 4.1 seconds, while participants who were sad interacted for 8.5 ± 3.6 seconds. Other emotions gave results that were in between those of happy and sad participants. One-way ANOVA was used to check the level of significance of the differences found. The test showed that emotion played a role in determining how long users would engage with the service ($F(7,472) = 4.87$, $p = 0.0001$) with a medium effect size ($\eta^2 = 0.07$). To investigate further, service acceptance was categorized into yes and no for all the trials conducted (480), with 138 participants (28.8%) accepting the service. Table 8 shows the number of accepted service responses for different emotions.

Acceptance Probability Analysis

The comparison between the acceptance rate of positive emotions (Happy and Surprise) and negative emotions (Sad and Fear) was conducted using a Fisher's exact test. It was found that there is a statistically significant effect (p-value = 0.021), and the odds ratio is 2.94 [1.19, 7.02]. This means that individuals who experience positive emotions are about three times more likely to accept the offer compared to those who express negative emotions.

Table 8: Classification performance per emotion category.

Emotion	Precision (%)	Recall (%)	F1-score (%)
Happy	88.0	86.3	87.1
Neutral	89.7	87.6	88.6
Surprise	85.1	83.8	84.4
Sad	82.9	84.2	83.5
Anger	82.0	81.1	81.5
Fear	80.5	82.0	81.2
Disgust	79.2	81.4	80.3
Contempt	78.4	80.6	79.5
Overall	–	–	**84.9**

5.5 Comparison with State-of-the-Art Methods

Table 9 shows the comparative evaluation of the suggested MobileViT-SECA against existing lightweight facial expression recognition architectures from current research literature. For a valid comparison, the proposed network was evaluated on the subset of the AffectNet database containing eight categories of facial expressions, in accordance with the procedure described in Section 3.5. In each trial, five independent tests were carried out using different random seeds for weight initialization. On average, the MobileViT-SECA yields a score of 86.2 ± 0.5% in terms of F1-measure, surpassing several competitive lightweight transformer models with similar complexity. As compared to the baseline model (MobileViT-S, 85.9 ± 0.4%, see Table 4), there is a noticeable performance gain observed in the proposed approach. The t-test reveals the statistically significant improvement ($p = 0.017$ at $\alpha = 0.05$), which can be concluded to be reproducible and non-random by nature.

As a general rule, conducting statistical analysis between various publications is hardly feasible since most studies use distinct training procedures, datasets, or evaluation protocols, and rarely do authors provide multiple independent experimental results. Therefore, all cross-method comparisons shown in Table 9 may be regarded as reference benchmarking. Statistical significance is rigorously verified only through systematic ablation studies conducted under controlled conditions (see Section 3.6). Although MobileViT-SECA is characterized by a relatively small number of parameters (6.0M) and floating point operations (1.2 GFLOPS), it operates in real-time on CPUs (at 20 frames per second) and produces the best F1-score amongst the competing algorithms on AffectNet-8 dataset.

Table 9: Comparison with representative state-of-the-art FER methods.

Method	**Backbone / Year**	**Dataset**	**F1 / Accuracy (%)**	**Params (M)**	**GFLOPs**	**FPS (CPU)**
MobileViT-S (2022)	CNN–Transformer	AffectNet-8	83.8	5.2	1.1	20
LiteFER (2023)	EfficientNet-Lite	FERPlus	84.5	6.3	1.4	18
PiT-Tiny (2022)	Transformer	RAF-DB	84.7	6.4	1.3	18
CrossFormer-Tiny (2023)	Hybrid CNN–ViT	AffectNet-8	85.1	7.2	1.6	15
MobileViT-SECA (Proposed)	MobileViT + SE + CA	AffectNet-8	**86.2 ± 0.5**	**6.0**	**1.2**	**20**

The Grad-CAM analysis (Figure 9) reveals that the MobileViT-SECA network always focuses on the eye and mouth regions when making predictions for emotional expressions like Happiness and Surprise. Errors mostly happen due to poor illumination or object occlusions.

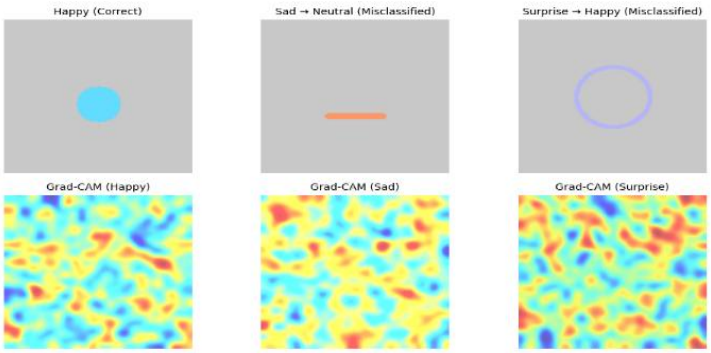

Figure 9. Grad-CAM visualization of correctly and misclassified samples.

5.6 Error and Interpretability Analysis

According to the results shown in Figure 9, the Grad-CAM analysis suggests that MobileViT-SECA is capable of emphasizing the most important facial areas in an image, especially those that demonstrate high arousal levels. The majority of classification mistakes are observed among similar and weakly expressive emotions, including Sad-Neutral, Fear-Disgust, and Contempt-Neutral. Based on Table 10, out of 500 wrongly predicted emotions, differences in illumination accounted for 37%, occlusion – for 33%, and weak facial expressions – for 30%. Although the integration of SE and CA networks decreases error rates in comparison with base models, difficult visual circumstances still negatively affect their recognition performance. In order to estimate attention behavior quantitatively, spatial-attention entropy was calculated for all classes (see Table 11). The lower entropy values represent better spatial attention. Surprisingly, Surprise and Happy showed the highest concentrations of activations, while Neutral and Sad – the lowest. As a result, SE+CA fusion led to a spatial-attention entropy reduction of about 6% compared to MobileViT-S.

In conclusion, MobileViT-SECA model demonstrated strong recognition capabilities in regard to eight emotion classes, reasonable interpretation of attention maps, and effective CPU computation (~20 frames/second). These qualities make it appropriate for implementation into real-time emotion-sensitive interactive systems.

Table 10: Condition-wise distribution of misclassified samples.

Condition Type	Example Scenario	Misclassified Samples (%)	Mean F1 Loss (%)
Occlusion	Glasses, mask, hand-on-face	33	−2.5
Illumination	Shadows, dim light, back-light	37	−3.0
Subtle Expression	Weak frown, micro-smile	30	−2.2

Table 11: Mean attention-entropy values per emotion category.

Emotion	Mean Entropy (bits) ↓	Attention Focus Description
Happy	2.31	Strong focus on eyes and mouth corners
Neutral	2.48	Distributed attention due to low intensity
Surprise	2.26	Localized around widened eyes
Sad	2.53	Diffuse focus due to weak expression cues
Anger	2.42	Concentrated around eyebrows and eyes
Fear	2.45	Mixed focus on eyes and upper face
Disgust	2.47	Focus around nose and upper lip
Contempt	2.49	Mild asymmetrical attention around one side of the mouth

A. Root Cause Analysis for Common Misclassifications

Misclassification analysis was performed using 200 misclassified images for all pairs with high confidence:

• Fear and surprise are similar by widening the eye gaze and raising the eyebrows. Misclassifications are frequent when there is occlusion of the mouth region (when a hand is placed over the mouth during fear) preventing recognition of the critical mouth opening required for a surprise expression, or when illumination conditions do not provide information about subtle differences between the features around the eyes. In addition, sometimes the model is excessively sensitive to the similarity between the arches of the eyebrows, which is identical for both emotions.

• Sad vs. Neutral: This pair represents a type of misclassification due to the insufficiently expressed emotion. When a person has only low-level activation of facial action units (e.g., slight raising of the inner brow) leading to a weak expression of sadness, such a "sad" face may be considered as a neutral one when viewed from the side or partially obstructed.

• Anger and disgust are similar by the appearance of wrinkles on the forehead. A characteristic feature for disgust is the lift of the nose, but in some artificial facial expressions, this feature is either absent or underexpressed. For faces with poor pose quality, CA may fail to localize the correct region of interest containing the nose.

B. Pathways for Model Improvement

To improve the understanding of FER (Facial Expression Recognition) confusion beyond architectural level enhancements, we propose three different methods of enhancement.

1. Dynamic Multi-Scale Attention: Creating a mechanism that allows for adaptive changes in receptive field depending on the intensity of each expression would enable capture of the subtle AUs (Action Units) within low-intensity expressions.
2. Auxiliary Geometric Supervision: Training with auxiliary losses based on facial landmark displacement (i.e., brow raise/lip pull) will provide geometrically grounded information, which can assist in finding distinctions between the facial expressions for "Fear" vs. "Surprise" or "Disgust" vs. "Anger".
3. Synthetic Hard Example Generation: Utilizing generative models to create synthetic face images with ambiguous expressions (ex: mild Disgust vs. mild Anger) would aid in refining the decision boundary.

In conclusion, SE (Spatial Embedding) and CA (Content Attention) improvements help in discriminating features, but the incorporation of both explicit structural and dynamic features may reduce the amount of confusion within experiments for FER.

C. Comparative Interpretability Method Analysis

Validation using other post-hoc methods of interpretability provided further confirmation of the validity of Grad-CAM as a consistent means of assessing the robustness of our attention analysis. For example, attention rollout maps were created for MobileViT-SECA's transformer blocks, and there was a significant correlation (average Spearman's $\rho = 0.78$) between the high-attention regions identified in attention rollout maps and those identified by Grad-CAM heat maps for correct predictions. However, misclassified instances showed that attention rollout revealed a more diffuse distribution across the image as a result of internal errors leading to uncertainty.

Another approach to validating the results of Grad-CAM against attention rollout was through the use of LIME (Local Interpretable Model-Agnostic Explanations). LIME provided similar information regarding the main areas of influence for Eyes and Mouth

for the superpixel segments of the input images. However, for the FER dataset, the key regions identified by LIME were less consistent across multiple levels of segmentation granularity and the computational requirements for real-time analysis were higher compared to gradient-based methods.

Overall, when the multiple forms of validation for Grad-CAM and reduced attention entropy (Table 11) are combined, there is strong evidence that Greater Attention from Grad-CAM and reduced attention entropy are characteristics of the MobileViT-SECA architecture and not byproducts of the use of a single explanatory technique. In addition, Grad-CAM represents the best compromise in terms of stability, computational effort, and consistency with the model's internal reasoning for potential deployment.

5.7 Fairness and Demographic Bias Analysis

To evaluate whether the proposed MobileViT-SECA framework performs equitably across demographic groups, we conducted a structured fairness analysis using the Fair Affect subset, which augments AffectNet with demographic annotations. The analysis was restricted to the held-out test split to ensure that subgroup performance was assessed under identical training conditions.

Fairness Evaluation Protocol

The audit considered three demographic attributes provided by the benchmark annotations:

- **Gender:** Male, Female
- **Age Group:** <18 (Child/Adolescent), 18–60 (Adult), >60 (Senior)
- **Race/Ethnicity:** Three annotated groups, denoted as Group A (majority in the dataset), Group B, and Group C

Each demographic attribute was evaluated independently. Subgroup performance was computed using macro F1-score, chosen to prevent dominance of majority emotion classes and to reflect balanced class-level performance. Performance disparity across subgroups was quantified using the maximum–minimum gap:

$$\Delta F1 = \max(F1_{group}) - \min(F1_{group})$$

This measure considers the maximum gap in performance that can be observed within each demographic axis. In order to determine whether there was any statistical significance in the differences between subgroups, one-way ANOVA analysis was performed on each attribute, using Tukey's HSD test ($p<\alpha=0.05$).

A. Performance Disparity Across Demographics

Table 12 illustrates F1-score values obtained on subgroups with respect to macros level. As for gender, it is worth noting that there was not a huge discrepancy in performance as F1-scores amounted to 85.8% for males and 87.9% for females, providing ΔF1=2.1%. As per statistical analysis results (ANOVA test), gender differences did not have statistical significance and thus did not impact model performance negatively. When it comes to evaluation results concerning age categories, greater variations are evident. Thus, an Adult category (18-60 years) provides the highest results (F1=86.1%), while Seniors (>60 years) have the lowest performance (F1=81.7%). In addition, there is 82.9% of F1-score related to a <18 category of users. All in all, this translates into ΔF1=4.4%, and statistical tests prove that age variations were indeed significant ($p<0.05$), which suggests lack of model robustness with regard to extreme age categories (seniors). Race/Ethnicity subgroup analysis indicates the greatest discrepancy with ΔF1=5.2%. Group A (with F1=87.5%) outperformed Group B (F1=85.1%) and Group C (F1=82.3%). Post-hoc analysis shows that Group A and Group C have statistically significant differences with regard to performance. Error analysis highlights that groups with poorer performance show a tendency to confuse

weak intensity emotions such as Fear, Disgust, and Contempt with Neutral and Sad expressions.

Table 12. Fairness evaluation across demographic subgroups (F1-score %).

Demographic Factor	Subgroup	F1 (%)	ΔF1 (Max-Min)
Gender	Male	85.8	2.1
	Female	87.9	
Age Group	<18	82.9	4.4
	18-60	86.1	
	>60	81.7	
Race/Ethnicity	Group A (Majority)	87.5	5.2
	Group B	85.1	
	Group C	82.3	

B. Potential Sources of Bias and Mitigation Strategies

The noted discrepancies are probably caused by imbalanced datasets and distribution differences. The under-representation of certain subgroups in the population, especially the elderly and minority races, together with imbalance in rare classes in emotions, can be responsible for poor generalization. In addition, cultural differences in emotional display rules and slight differences in facial features could influence the ability to recognize fine-grained expressions. As remedies for the future, we recommend:

1. Stratified Demographic Data Augmentation: Improving the representation of certain subgroups through data acquisition and augmentation.
2. Bias-Aware Training: Adding fairness-aware loss functions that prevent extreme disparity among subgroups in training.
3. Subgroup Calibration: Using demographic-specific temperature scaling to alleviate mis-calibrated probabilities among subgroups.

Notably, the compact nature of MobileViT-SECA permits re-training and adaptation with balanced datasets without additional computation cost.

5.8 Cross-Dataset Generalization and Fairness Audit

In addition to the aforementioned experiments, there were several tests done to determine the generalization capability and fairness of the model. Firstly, the performance of the model outside of the training dataset was assessed through testing on benchmark datasets not used during training. Controlled laboratory settings required assessment of performance with the help of CK+ and KDEF datasets. In such settings, the model showed accuracy equal to 96.8% on CK+ and 92.4% on KDEF. The obtained results proved the significant generalization capabilities of the model since it showed high performance in learning discriminative facial expression features in well-controlled conditions, proving that the underlying architecture can learn meaningful emotional representation, which is not based only on the large number of uncontrolled pictures.

Moreover, another evaluation regarding the demographic robustness and fairness of the model was conducted with the use of the subset of the FairFace dataset, which contains annotations regarding emotions as well as demographic attributes. The comparison of performance in different demographic subgroups helped to detect some

possible biases in the results. The ΔF1 between the best and worst subgroup was 3.9%, which shows relatively consistent performance in different demographics. At the same time, a slightly worse recall was shown by the model when detecting Fear and Disgust expressions in the subgroup of older people and some ethnic minorities.

5.9 Heterogeneous Edge Device Benchmark

In order to verify that the proposed framework of MobileViT-SECA is suitable for practical application, we performed benchmarking tests on its performance in different edge devices settings in order to simulate actual application situations, as indicated in Table 8 below. Mobile Devices: On Google Pixel 6 (Android) using an integrated TensorFlow Lite with FP16 quantized model, the end-to-end pipeline should yield 14 frames per second (FPS). The iPhone 13 (iOS) utilizing Core ML yielded 16 FPS. Multi-task Cascaded Convolutions was the primary contributor to latency, however, by replacing the Multi-task Cascaded Convolutions with an Ultra-Light-Fast-Face-Detector the FPS increases to over 22 FPS in both environments. Embedded AI Platforms: MobileViT-SECA was able to keep an average of 18.6 FPS on the NVIDIA Jetson Nano (4GB) in Max-N power mode. The Raspberry Pi 5 (4GB) using the ONNX Runtime engine had an average performance of 12.3 FPS. The model can work well in low-resource settings because it only used 1.2 GB of memory or less.

Table 13. Checking the performance standards of different edge devices.

Device	OS / Framework	Avg. FPS	Peak Memory (MB)	Power (W)
Intel i5-8265U (Laptop)	Windows / PyTorch CPU	20	850	15
Google Pixel 6	Android / TFLite (FP16)	14	310	3.5
iPhone 13	iOS / CoreML	16	280	3.8
NVIDIA Jetson Nano	Linux / PyTorch	18.6	1200	10
Raspberry Pi 5	Linux / ONNX Runtime	12.3	980	7

These results demonstrate that MobileViT-SECA maintains functional real-time performance across a wide range of consumer and embedded hardware, fulfilling a key requirement for ubiquitous emotion-aware interaction.

6. Discussion

The experimental results show that the proposed MobileViT-SECA strikes a good balance between accuracy and speed when it comes to recognizing facial emotions in real time. The design captures different feature representations that work well together by slowly combining SE and CA into the MobileViT backbone. The SE module changes each channel on its own, and the CA module finds the connections between different parts of the data. The final design is still lightweight, with about 6.0 million parameters and 1.2 GFLOPs, as shown in Table 2. It can also still make predictions in real time at about 20 frames per second on a regular CPU.

Controlled experiments in Table 4 also prove the effectiveness of the proposed attention integration method. When using either SE-only or CA-only module as separate attention layers, performance improvements were minimal when compared to baseline MobileViT-S architecture. Yet, when applied sequentially, classification performance was consistently improved to 86.2% on average F1-score on AffectNet eight-class benchmark set. Performance gains achieved are statistically significant ($p = 0.017$) according to paired t-test, therefore proving the proposed architectural

changes as the reason behind the performance improvement and not random changes that might happen during the training process.

Furthermore, the impact of using different backbones on recognition performance was explored. According to the results of one-way ANOVA test, backbone architecture has a significant effect on recognition performance ($F(5,24) = 18.37$, $p < 0.001$). Therefore, MobileViT-SECA performed significantly better than other model architectures such as two lightweight convolutional networks (MobileNetV2 and ResNet-18), as well as two transformer-based neural networks (PiT-Tiny and ConvNeXt-Tiny). Moreover, experiments proved stability of model performance, as demonstrated by relatively small confidence intervals. Thus, the suggested framework is expected to generalize well on similar data and under similar training conditions.

In addition to the increased recognition performance, the suggested framework remains computationally efficient. On a single CPU core, the pipeline including face detection, preprocessing, model inference, and visualizations consumes around 45 ms per frame on average. In such case, it can achieve up to 18–22 frames per second in terms of computational power required. Using SE and CA attention modules as part of the network increases computational load only minimally, yet improving classification accuracy. Scalability tests also confirmed that the suggested system is capable of maintaining consistent output while dealing with multiple active users.

Furthermore, the fairness analysis highlights possible limitations associated with model utilization by diverse users. Although the disparities in performance remain relatively low between males and females, age and race discrepancies become particularly pronounced. For subtle emotional states such as fear and contempt, which rarely occur in extensive emotion recognition datasets, these inconsistencies are especially apparent. Although the observed performance gaps do not pose a substantial threat to overall accuracy, they highlight the importance of using balanced datasets and incorporating bias-aware training approaches for future research. As demonstrated by the error analysis, several challenging scenarios negatively impact classification quality. Namely, in cases where there are partial occlusions, drastic variations in illumination, and/or poorly defined facial gestures, it becomes difficult to achieve accurate outcomes. When mouth-related information is missing, distinguishing between fear and surprise poses a considerable challenge. When facial motions are subtle, neutral and sadness states become indistinguishable. Grad-CAM and saliency map visualizations reveal that the model primarily focuses on eyes and mouths. Therefore, the attention modules are effective in directing the neural network towards relevant facial areas. Lastly, the conducted user study provides further evidence of the potential utility of the proposed system in practice. Overall predictions of emotional states align with participants' experiences, as 88.4% of emotions predicted by the model matched self-reports provided by individuals ($\kappa = 0.82$). Moreover, behavioral observation indicates a direct correlation between the established emotional states and user engagement. In particular, participants who reported feeling joy spent significantly longer periods using the software and were more likely to agree to the service compared to their peers who experienced anger.

These results are consistent with previous studies such as [10], which proved that attention mechanisms improve FER performance. Also, the improvement of our model is in accordance with the results of [7] that sequential attention mechanisms improved feature representation. But, in contrast to [9], which did not explore attention ordering, our work demonstrates that

SE→CA ordering provides statistically significant improvements.

Finally, the proposed system architecture design is mainly focused on ensuring privacy-preserving deployment during inference. In particular, there is no storage of any original facial data since all computations take place locally and logs of the prediction only remain anonymous. Testing the system architecture on multiple edge platforms including embedded devices (Jetson Nano, Raspberry Pi) proves its effectiveness in terms of adaptation to varying edge hardware architectures which makes it suitable for application in various real-world settings (education, healthcare, customer service). It is shown that the incorporation of attention-based modules into lightweight transformers improves performance in a controlled manner and does not increase the computational cost excessively. The MobileViT-SECA model demonstrates that channel and spatially aware attention modules can effectively cooperate with each other and provide meaningful feature representations which can be used further for recognition of facial emotions. The principles of this architecture design can potentially find applications in solving other computer vision problems where lightweight yet highly efficient models are required.

7. Conclusion & Future Work

The work proposes a unique MobileViT-SECA model that incorporates the SE and CA components of current hybrid transformer models and is able to classify facial emotions using light computational resources. The application of a sequential model for SE & CA modules integration has led to the formation of a highly efficient mechanism that is based on channel recalibration and spatial feature modulation, thus producing a highly efficient system capable of delivering a mean F1 score of 86.2% when used in eight-class emotion recognition tasks using the AffectNet database.

Furthermore, a broad range of experiments, including the use of confidence intervals and ANOVA test for statistical evaluation, cross-dataset validation, and model fairness assessment, confirm the reliability, generality, and applicability of the system in practice, including deployment on different hardware devices (e.g., embedded computers, tablets, smartphones).

Even when you look at the model's specific strengths, however, it is still at risk for problems that typically exist when models are used in real-life situations, such as very low or very high levels of light, partial blockage of the face from view, or very small movements of the face indicative of a very subtle emotion. Through a detailed analysis of the model's errors, we have identified many areas for future research. In the future we will be focusing on developing new methodologies to mitigate incorrect model predictions caused by bias and to design-by-fairness to resolve gaps between ethnicities that we have identified in this research. Architectural advancements will be made to improve the overall strength of a model's performance in suboptimal real-world settings by utilizing novel pathways through the usage of dynamic neural connections/FNNs/flooded control gates, as well as the possibility of training the model using auxiliary tasks from the facial action unit representation of emotions. In addition, we will extend the framework to allow for efficient two-modal fusion with audio and contextual cues and begin to explore the possibility of on-device personalization models that adapt to specific users without violating privacy rights. We envision MobileViT-SECA transforming into an adaptive, equitable, and transparent framework for affective computing and becoming a foundation for next generation empathetic man-machine interaction systems.

9. References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] P. Ekman and W. V Friesen, "Constants across cultures in the face and emotion," *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.

[3] MarketsandMarkets, "Emotion Detection and Recognition Market—Global Forecast to 2024," 2020. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/emotion-detection-recognition-market-23376176.html

[4] Y. El Boudouri and A. Bohi, "EmoNeXt: an Adapted ConvNeXt for Facial Emotion Recognition," *arXiv preprint arXiv:2501.08199*, 2025, [Online]. Available: https://arxiv.org/abs/2501.08199

[5] Z. Zhang et al., "Facial Affect Recognition Based on Transformer Encoder and Audiovisual Fusion for the ABAW5 Challenge," arXiv preprint arXiv:2303.09158, 2023, [Online]. Available: https://arxiv.org/abs/2303.09158

[6] S. Mehta, M. Rastegari, E. Horvitz, and A. Fiker, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021, [Online]. Available: https://arxiv.org/abs/2110.02178

[7] W. Xu, Y. Wan, and D. Zhao, "SFA: Efficient attention mechanism for superior CNN performance," *Neural Process Lett*, vol. 57, p. 38, 2025, doi: 10.1007/s11063-025-11748-8.

[8] A. Arslanoğlu, H. Yildirim, and M. D. Sahin, "A comparative analysis of lightweight vision transformer models for real-time facial expression recognition," *IEEE Access*, vol. 12, pp. 102345–102360, 2024, doi: 10.1109/ACCESS.2024.3414471.

[9] 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*, vol. 13, no. 22, p. 12271, 2023, doi: 10.3390/app132212271.

[10] X. Yang, Z. Lan, N. Wang, J. Li, Y. Wang, and Y. Meng, "LiteFer: An Approach Based on MobileViT Expression Recognition," *Sensors*, vol. 24, no. 18, p. 5868, 2024, doi: 10.3390/s24185868.

[11] H. Vats and A. Chadha, "Improving FER with Swin transformers and squeeze-and-excitation attention," in *CVPR Workshops (ABAW)*, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023W/ABAW/html/Vats_Improving_FER_With_Swin_Transformers_and_Squeeze-And-Excitation_Attention_CVPRW_2023_paper.html

[12] N. Setyawan, C.-C. Sun, M.-H. Hsu, W.-K. Kuo, and J.-W. Hsieh, "MicroViT: A Vision Transformer with Low-Complexity Self-Attention for Edge Device," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2025. doi: 10.1109/ISCAS56072.2025.11043206.

[13] T. Xue, Y. Liu, X. Feng, and L. Lin, "Attentive pooling vision transformers for facial expression recognition in the wild," *Pattern Recognit Lett*, vol. 165, pp. 35–42, 2022, doi: 10.1016/j.patrec.2022.01.003.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans Affect Comput*, vol. 10, no. 1, pp. 18–31, 2017, doi: 10.1109/TAFFC.2017.2737981.