Attention-Enhanced EfficientNet for Facial Expression Recognition on Imbalanced Datasets

Wasan Alhababheh, Nedal Abushadoof, Cynthea Qaqish
Mohammed Abulathoa, Khalil Alhileh, Mohammad Hassanat
Artificial Intelligence Department
University of Jordan
Amman, Jordan

Instructor: Tamam AlSarhan

***Abstract*—Facial expression recognition (FER) is a challenging pattern recognition task due to subtle inter-class variations, high intra-class diversity, and the presence of imbalanced data. In this paper, we propose a robust FER framework based on an EfficientNet-B0 backbone enhanced with Squeeze-and-Excitation (SE) attention mechanisms to improve discriminative feature learning. To address the class imbalance inherent in the collected dataset, we employ a weighted sampling strategy combined with Focal Loss, enabling the model to focus on hard-to-classify and underrepresented emotion classes. The proposed approach is evaluated on a custom facial expression dataset collected by the authors. Experimental results demonstrate stable training behavior and improved class-wise performance, achieving an overall accuracy of 95% and a Macro-F1 score of 0.93. These findings highlight the effectiveness of attention mechanisms and imbalance-aware optimization for practical facial expression recognition tasks.**

***Index Terms*—Facial Expression Recognition, EfficientNet, Squeeze-and-Excitation, Class Imbalance, Focal Loss**

I. Introduction

Facial Expression Recognition (FER) is a fundamental problem in affective computing and computer vision, aiming to enable machines to interpret human emotional states from facial cues. FER has been widely applied in human–computer interaction, intelligent tutoring systems, healthcare monitoring, surveillance, and driver assistance systems [4], [5], [17]. Psychological studies suggest that facial expressions convey universal emotional information across cultures, making them a reliable modality for emotion analysis [10].

Despite recent advances driven by deep learning, FER remains a challenging task due to subtle inter-class variations, high intra-class diversity, and sensitivity to environmental factors such as illumination changes, pose variations, and occlusions [6], [7]. In addition, facial expressions such as *Sad* and *Surprise* often exhibit similar visual patterns, increasing the likelihood of misclassification. These challenges are further exacerbated in real-world settings where datasets are limited in size and lack controlled acquisition conditions.

Another critical challenge in FER is the inherent class imbalance present in most datasets. Emotions such as *Happy* are typically overrepresented, while expressions like *Fear* and *Surprise* occur less frequently [8]. Models trained using standard cross-entropy loss tend to be biased toward majority classes, resulting in degraded performance on minority emotions. Consequently, addressing data imbalance is essential for building reliable and fair FER systems.

Recent studies have demonstrated that attention mechanisms significantly enhance FER performance by enabling models to focus on discriminative facial regions and informative feature channels [9], [15], [16]. Channel-wise attention, in particular, has proven effective in emphasizing salient features while suppressing irrelevant information. Motivated by these findings, this paper proposes an enhanced FER framework based on an EfficientNet-B0 backbone integrated with a Squeeze-and-Excitation (SE) attention module. Furthermore, to explicitly address class imbalance, Focal Loss and weighted sampling are employed. The proposed approach aims to achieve robust and balanced performance across all emotion categories, as measured by Macro-F1 score.

II. Related Work

A. Deep Learning for Facial Expression Recognition

Early FER approaches relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). While effective under controlled conditions, these methods struggled to generalize to unconstrained environments. The emergence of deep learning shifted FER research toward Convolutional Neural Networks (CNNs), which automatically learn hierarchical facial representations [6], [13]. Deep CNN-based models have achieved significant performance improvements by leveraging large-scale datasets and deeper architectures; however, they often suffer from overfitting and high computational cost when applied to small or imbalanced datasets [4], [7].

B. Efficient Architectures for FER

To address efficiency and scalability challenges, lightweight and parameter-efficient architectures have gained attention in FER research. Popular deep models such as VGG, Inception, and ResNet provide strong feature extraction capabilities but are computationally expensive [11], [12]. EfficientNet introduces a compound scaling strategy that jointly optimizes network depth, width, and resolution, achieving state-of-the-art performance with significantly fewer parameters [1]. Due to its favorable accuracy-to-complexity trade-off, EfficientNet-B0 is particularly suitable for FER tasks involving limited data and constrained computational resources.

C. Attention Mechanisms in FER

Attention mechanisms have been widely adopted to improve FER by enabling networks to focus on informative facial regions or feature channels. Spatial attention emphasizes critical facial areas such as the eyes and mouth, while channel-wise attention selectively amplifies discriminative feature maps [9]. Squeeze-and-Excitation (SE) networks explicitly model channel interdependencies and recalibrate feature responses based on global contextual information [2]. Recent studies confirm that attention-enhanced FER models exhibit improved robustness against noise, occlusion, and subtle expression variations [15], [16].

D. Learning from Imbalanced Data

Class imbalance is a persistent challenge in FER datasets and has a significant impact on classification performance. Standard cross-entropy loss treats all samples equally, leading to biased predictions toward majority classes [8]. Focal Loss mitigates this issue by down-weighting easy examples and focusing learning on hard and minority-class samples [3]. In addition to loss-level strategies, data-level approaches such as weighted sampling have been shown to further improve class-wise performance and stability during training [8].

III. Methodology

A. Dataset Acquisition and Preprocessing

The dataset utilized in this study consists of facial images collected by students at the University of Jordan under unconstrained conditions. The dataset is categorized into five emotion classes: **Angry, Fear, Happy, Sad, and Surprise**. Similar to many real-world FER datasets, the collected data exhibits class imbalance, with the *Happy* class significantly outnumbering other expressions.

To ensure data consistency, an extensive preprocessing pipeline was implemented. Raw images included High Efficiency Image File (HEIF/HEIC) formats, which were converted to standard JPEG format using the `pillow-heif` library. Directory structures were sanitized to merge inconsistent labels, and all images were converted to three-channel RGB format. Each image was resized to 224×224 pixels and normalized using a mean and standard deviation of 0.5 across all channels.

The dataset was partitioned using a stratified split strategy to maintain the original class distribution across subsets. The data was divided into **70% training, 15% validation, and 15% testing** sets. To address class imbalance (e.g., the "Happy" class being significantly larger than "Surprise"), we employed a **Weighted Random Sampler** during training. Samples were weighted inversely proportional to their class frequency, ensuring the model viewed underrepresented classes more frequently within each batch.

All images were resized to 224×224 pixels and normalized with a mean and standard deviation of 0.5 across all channels prior to entering the network.

B. Network Architecture

The proposed FER model employs a hybrid architecture combining a robust convolutional backbone with a channel-wise attention mechanism.

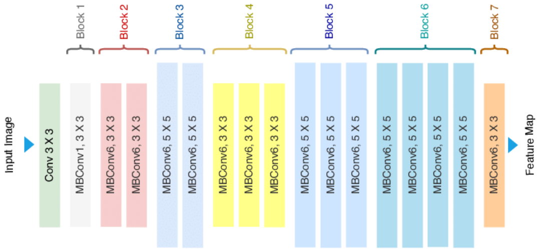

Fig. 1. EfficientNet-B0 Architecture utilized as the feature extractor.

1) Backbone Feature Extractor: We utilized **EfficientNet-B0** pretrained on ImageNet as the feature extractor. This backbone was selected for its balance between parameter efficiency and accuracy. The network was initialized with pretrained weights to leverage transfer learning, and features were extracted from the final convolutional stage.

2) Squeeze-and-Excitation (SE) Attention Block: To enhance the model's sensitivity to subtle facial features, a custom **Squeeze-and-Excitation (SE)** block was inserted after the backbone. This mechanism recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels.

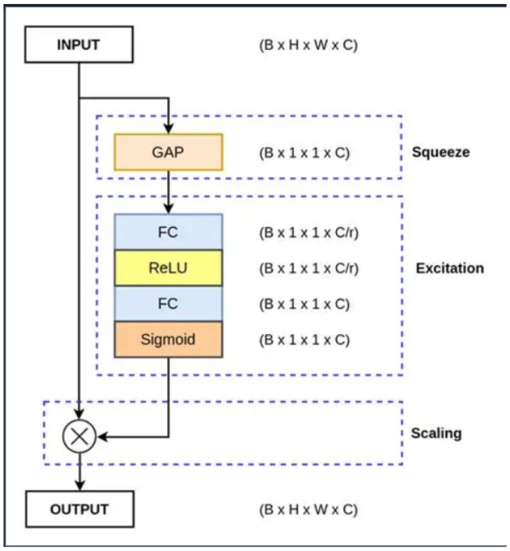

Fig. 2. Architecture of the SE block showing Squeeze, Excitation, and Scaling stages.

The SE mechanism operates in two distinct stages as illustrated in Fig. 2:

1) **Squeeze:** Global spatial information is compressed into a channel descriptor via global average pooling (GAP).
2) **Excitation:** A self-gating mechanism is employed using a bottleneck structure with two fully connected (FC) layers, producing rescaling weights for the original feature maps.

The output X_{out} is computed as:

$$X_{out} = F_{scale}(X_{in}, W) = X_{in} \cdot \sigma(W_2 \delta(W_1 \text{AvgPool}(X_{in}))) \tag{1}$$

where δ denotes the ReLU activation function and σ denotes the Sigmoid function.

3) Classification Head: The feature map from the SE block is passed through an Adaptive Average Pooling layer to reduce spatial dimensions to 1×1. The classification head consists of a dense layer with 512 units (ReLU activation), a **Dropout layer with a probability of 0.4** to prevent overfitting, and a final fully connected layer mapping to the 5 emotion classes.

C. Loss Function and Optimization

Standard Cross-Entropy loss often struggles with class imbalance and easy negatives. To mitigate this, we employed **Focal Loss**, defined as:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

where $\gamma = 2$ is the focusing parameter that down-weights easy examples and focuses training on hard negatives. Additionally, we incorporated inverse class frequency weights (w) into the loss calculation to further penalize misclassifications of minority classes.

The model was optimized using the **AdamW optimizer** with a learning rate of 1×10^{-4}. Training was conducted with a batch size of 16. We monitored the Macro F1-score on the validation set to track performance and employed early stopping where appropriate.

IV. Experimental Results

A. Quantitative Performance

The proposed EfficientNet-B0 model with Squeeze-and-Excitation (SE) attention and Focal Loss was evaluated on a test set of 112 facial images across the five emotion classes. The model achieved an **overall accuracy of 95%**, with a **Macro-F1 score of 0.93** and a Weighted-F1 score of 0.95. These metrics indicate a strong and balanced classification performance, validating the effectiveness of the proposed architecture.

The class-wise F1-scores were as follows:

- **Angry:** 0.96
- **Fear:** 0.98
- **Happy:** 1.00
- **Sad:** 0.87
- **Surprise:** 0.85

Emotions with distinct facial patterns, such as Happy, Fear, and Angry, were recognized with very high reliability. While the more subtle classes, Sad and Surprise, exhibited slightly lower performance compared to the others, the scores remain robust (> 0.85), demonstrating the model's capability to handle difficult expressions. This performance can be attributed to the channel-wise recalibration provided by the SE attention mechanism and the imbalance-aware optimization introduced by Focal Loss, which together enhance discrimination between visually similar emotion categories.

B. Training Behavior and Generalization

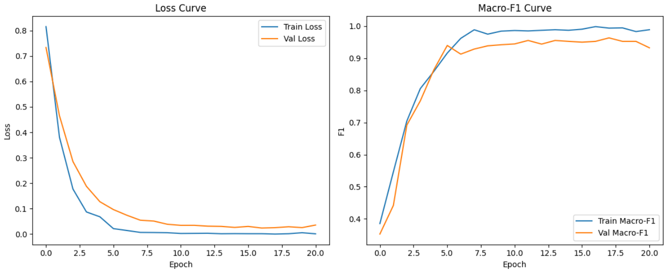

Fig. 3. Training and validation Macro-F1 scores showing stable convergence. The validation Macro-F1 peaked at 0.963 at epoch 17.

As illustrated in Fig. 3, the validation Macro-F1 reached a peak of **0.963 at epoch 17**, while the training Macro-F1 approached 0.99. The curves demonstrate stable convergence throughout the learning process. The consistent gap observed between training and validation performance indicates effective generalization, despite the limited size and inherent class imbalance of the custom dataset.

C. Confusion Matrix Analysis

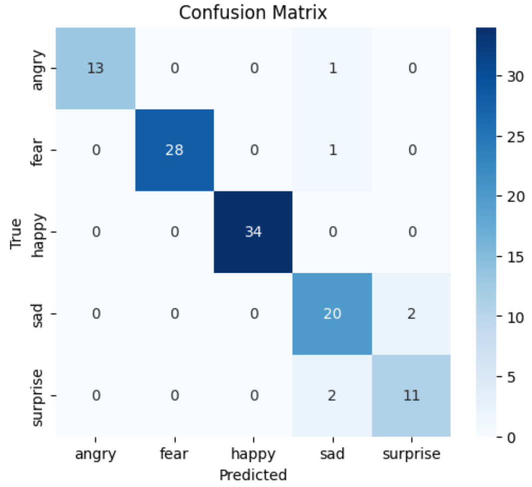

Fig. 4. Confusion Matrix of the test set predictions. High diagonal values indicate accurate classification.

The confusion matrix (Fig. 4) exhibits strong diagonal dominance, confirming reliable classification across most expressions. The analysis reveals that the majority of errors occur between the **Sad** and **Surprise** classes. This misclassification is attributed to the similar visual characteristics shared by these emotions in certain subjects, which is a well-known challenge in facial expression recognition tasks.

V. Discussion

The experimental results demonstrate that integrating channel-wise attention with imbalance-aware optimization significantly enhances FER performance. The Squeeze-and-Excitation module improves feature discrimination by emphasizing informative channels associated with facial muscle movements, leading to more robust representation learning. Additionally, the combination of Focal Loss and weighted sampling effectively mitigates bias toward majority classes, resulting in balanced performance across emotions.

The strong Macro-F1 score indicates that the proposed model generalizes well despite the limited size and imbalance of the custom dataset. Remaining misclassifications primarily occur between visually similar expressions such as *Sad* and *Surprise*, which aligns with observations reported in prior FER studies [4], [9]. These results suggest that the proposed approach is well-suited for real-world FER scenarios where data collection is constrained.

Despite the strong performance, the proposed approach was evaluated on a relatively small, custom-collected dataset. Future evaluations on larger, publicly available FER benchmarks are necessary to further validate the generalizability of the proposed framework.

VI. Conclusion

This paper presented a facial emotion recognition system based on EfficientNet-B0 enhanced with SE attention and Focal Loss. The model achieved 95% accuracy and a Macro-F1 of 0.93, demonstrating high reliability and balanced performance across all emotion classes. The SE attention mechanism improved feature discrimination, while Focal Loss reduced the impact of class imbalance. Together, they enabled strong recognition even for challenging emotions such as Surprise and Sad. Future work will explore the integration of temporal modeling and multimodal inputs to further enhance recognition accuracy.

References

[1] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[5] Z. Zhao, Q. Liu, S. Yang, F. Chen, and D. Metaxas, "Facial expression recognition: A comprehensive survey," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–40, 2021.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. WACV*, 2016.

[7] I. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[9] Y. Wang, J. Li, and Y. Zhao, "Attention-based convolutional neural network for facial expression recognition," *Pattern Recognition Letters*, vol. 131, pp. 366–374, 2020.

[10] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015.

[15] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, 2021.

[16] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[17] R. Ranjan *et al.*, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, 2018.