Journal of Progress in Engineering and Physical Science
ISSN 2709-4006
www.paradigmpress.org/jpeps
Volume 5 Number 2 June 2026

An Improved ShuffleNet with SE Attention and Focal Loss for Lightweight Facial Expression Recognition

Yuqi Zeng[1]

[1] Maynooth International Engineering College, Fuzhou University, Fujian, China
Correspondence: Yuqi Zeng, Maynooth International Engineering College, Fuzhou University, Fujian, China.

doi:10.63593/JPEPS.2026.06.02

Abstract

Facial expression recognition (FER) plays a critical role in human-computer interaction, mental health monitoring, intelligent security inspection and classroom emotional computing. Traditional deep convolutional neural networks for FER suffer from excessive computational overhead and large parameter volume, which limits their deployment on mobile terminals and embedded devices with limited computing resources. As a classic lightweight network, ShuffleNet adopts group convolution and channel shuffle operations to reduce model complexity, yet it lacks the ability to capture subtle facial emotional features and cannot well handle the class imbalance problem existing in mainstream FER datasets. To address these defects, this paper proposes an improved lightweight ShuffleNet model for facial expression detection. First, Squeeze-and-Excitation (SE) channel attention modules are embedded after each shuffle block to dynamically recalibrate feature channel weights and strengthen the extraction of discriminative facial expression features. Second, a multi-dimensional data augmentation strategy combining geometric transformation, color jitter and Gaussian noise injection is designed to expand sample diversity and enhance the model's generalization ability under complex lighting, shooting angles and background interference. Third, Focal Loss is introduced to replace the standard cross-entropy loss, which suppresses the loss contribution of easy-classified majority samples and forces the model to focus on hard-to-distinguish minority facial expressions such as anger and disgust. Comprehensive experiments are conducted on three public datasets FER2013, CK+ and AffectNet. The results demonstrate that the proposed model achieves higher recognition accuracy compared with original ShuffleNet V2 while maintaining low computational cost and small parameter size, and presents superior robustness against unbalanced data and complex real-world scenes.

Keywords: lightweight convolutional neural network, ShuffleNet, SE attention mechanism, facial expression recognition, focal loss, data augmentation

1. Introduction

With the rapid advancement of computer vision and artificial intelligence, facial expression recognition has become a core research branch of affective computing. Psychologist Ekman first defined six basic universal facial expressions including anger, disgust, fear, happiness, sadness and surprise, laying the theoretical foundation for subsequent FER classification research

(Ekman P & Friesen W V., 1971). At present, deep learning-based FER methods have achieved remarkable performance. Lai et al. optimized CNN feature extraction with atrous convolution and realized real-time micro-expression recognition with face rectification preprocessing (Li F., 2020). Li reconstructed the SSD detection framework combined with center loss to promote expression classification accuracy (Qu F, Wang S J, Yan W J, et al., 2018). Multi-modal fusion strategies proposed by Tao et al. further improve recognition stability by fusing image, voice and text emotional information (Rong Y, Liu J W & Li X L., 2024). Meanwhile, FER technologies have been widely applied in diversified industrial scenarios: classroom student emotional assessment, railway station security risk warning, driver fatigue micro-expression monitoring and other practical systems have been constructed based on ResNet, CNN and other backbone networks (Peng K B, Lv X J, Li C, et al., 2024; Wang X R, Zhang Y T & Peng X., 2022; Zhou Y, Chen S C, Wang Y M, et al., 2020).

Nevertheless, most high-precision FER backbones such as ResNet50 bring heavy calculation burden, which cannot run smoothly on resource-constrained edge devices like mobile phones and embedded cameras. Lightweight neural networks emerge as a feasible solution to this contradiction. Proposed by Facebook AI Research in 2017, ShuffleNet V1 introduces group convolution and channel shuffle to break inter-group information isolation and cut down floating-point operations (FLOPs) significantly. ShuffleNet V2 released in 2018 further optimizes module design according to actual inference latency metrics, becoming the mainstream lightweight backbone for edge vision tasks. Existing studies have verified the effectiveness of modified ShuffleNet in agricultural disease classification and UAV aerial image classification (Yang Z, Wu S D & Jia R., 2024). However, vanilla ShuffleNet still has two obvious drawbacks when applied to facial expression detection:

1) The equal-weight feature extraction mode ignores subtle facial emotional textures such as eyebrow and mouth changes, leading to insufficient feature discrimination;
2) Public FER datasets including FER2013 and AffectNet present severe class imbalance, where happy and neutral samples occupy the vast majority while disgust and fear samples are scarce, resulting in biased model training and low recognition precision of minority categories.

To solve the above problems, this paper designs an improved ShuffleNet framework integrated with SE attention, multi-strategy data augmentation and Focal Loss for lightweight facial expression recognition. The main contributions of this work are summarized as follows: (1) Embed SE channel attention modules into ShuffleNet V2 blocks to adaptively weight expression-related feature channels, retaining the lightweight property while enhancing fine-grained emotional feature capture capability. (2) Construct a compound data augmentation pipeline containing geometric transformation, color adjustment and noise addition to simulate complex real shooting environments and boost model generalization. (3) Adopt Focal Loss as the training loss function to alleviate dataset class imbalance, elevating the identification accuracy of rare facial expression categories. (4) Validated on FER2013, CK+ and AffectNet datasets, the proposed method achieves balanced trade-off between recognition accuracy and model computation cost, suitable for real-time expression detection on mobile and embedded devices.

The rest of this paper is organized as follows. Section 2 reviews related work on lightweight networks and facial expression recognition. Section 3 elaborates the overall architecture of the improved ShuffleNet and three core optimization strategies. Section 4 introduces experimental datasets, evaluation metrics and comparative experimental results. Section 5 concludes the paper and prospects future research directions.

2. Related Work

2.1 Lightweight ShuffleNet Series Network

ShuffleNet V1 is the pioneering lightweight network utilizing group convolution and channel shuffle. Traditional group convolution separates feature channels into disjoint groups, causing blocked information interaction between groups. Channel shuffle operation rearranges feature channels across groups to realize cross-group information fusion without extra computational consumption. ShuffleNet V2 revises the original block structure and puts forward four practical lightweight design principles: equal input/output channel width, minimizing group convolution, reducing branch operations and decreasing element-wise operations. It achieves faster actual inference speed under identical FLOPs, and has become the preferred backbone for real-time

edge detection tasks.

Current improvements on ShuffleNet mainly focus on structural replacement and auxiliary module fusion. Zhou et al. optimized group convolution information transmission via channel transformation (Li D H, Zhong T, Wang S, et al., 2024). Some researchers replace standard convolution with depthwise separable convolution and SPD-Conv to raise low-resolution image recognition performance. For domain-specific tasks, ShuffleNet combined with attention mechanisms achieves competitive results in plant disease identification and UAV target classification, which provides reference for this FER research (Yang Z, Wu S D & Jia R., 2024).

2.2 Facial Expression Recognition Optimization

Modern FER research focuses on three optimization directions: feature extraction enhancement, data expansion and loss function optimization. For feature enhancement, attention mechanisms including CBAM, SE and CA are widely embedded into CNN backbones to screen effective emotional features. For data processing, random flip, rotation, brightness disturbance and other augmentation methods are adopted to expand limited training samples and resist environmental interference. For loss design, center loss, triplet loss and Focal Loss are used to relieve classification bias caused by imbalanced sample distribution. Existing lightweight FER models mostly adopt MobileNet series backbones, while few studies systematically integrate SE attention, compound augmentation and Focal Loss into ShuffleNet for expression detection, which is the research gap this paper fills.

3. Proposed Method

This section details the overall architecture of the improved ShuffleNet model and three key optimization modules: SE attention embedding, multi-dimensional data augmentation and Focal Loss function.

3.1 Overall Network Architecture

The backbone of the proposed model is ShuffleNet V2. After each shuffle unit block, an SE attention module is inserted to recalibrate feature channels. The network workflow is as follows:

1) Input preprocessed facial images with unified resolution;
2) Extract preliminary feature maps through the initial convolution layer and max pooling layer;
3) Pass feature maps through stacked improved shuffle blocks embedded with SE modules, completing lightweight feature extraction with adaptive channel weighting;
4) Send the final feature map into global average pooling and full connection layers to output expression classification probabilities;
5) Use Focal Loss to calculate training loss and backpropagate gradients to update network parameters.

The improved structure maintains the original low parameter and low FLOPs advantages of ShuffleNet V2, without introducing excessive computational overhead.

3.2 SE Attention Module Integration

The Squeeze-and-Excitation module realizes adaptive feature channel weighting through two core stages: Squeeze and Excitation.

1) **Squeeze Operation**: Apply global average pooling on each channel of the input feature map $X \in R^{H\times W\times C}$, compressing spatial dimension information into a channel descriptor

$$Z_c: z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(i,j))$$

where H, W, C denote the height, width and channel number of feature maps respectively.

2) **Excitation Operation**: Two fully connected layers are used to learn nonlinear channel correlation. The first FC layer reduces channel dimensions with ReLU activation, the second FC layer recovers original channel number and maps weights to the interval [0,1] via Sigmoid function to obtain channel weight vector $s = \sigma(W2\delta(W1z))$ δ represents ReLU activation, σ represents Sigmoid activation.

3) **Feature Recalibration**: Multiply the original feature map channel by the corresponding weight s_c to obtain weighted feature output $\tilde{X}: \tilde{X}_c = s_c \cdot X_c$.

In this paper, SE modules are embedded after the channel shuffle operation of every ShuffleNet V2 block. The model automatically assigns larger weights to channels containing facial key areas such as eyes, mouth and eyebrows, suppressing useless background feature responses and

strengthening the capture of subtle expression differences.

3.3 Multi-Dimensional Compound Data Augmentation

To improve model robustness against illumination change, shooting angle offset and noise interference in real scenarios, three types of augmentation strategies are combined to expand training samples:

1) **Geometric Transformation**: Random horizontal flip, slight rotation within $\pm 15°$, random scaling and random cropping, simulating different face shooting poses;
2) **Color Jitter**: Random adjustment of image brightness, contrast and saturation, adapting to complex indoor and outdoor lighting conditions;
3) **Noise Injection**: Add low-intensity Gaussian noise to training images to resist sensor noise interference of edge cameras.

All augmentation operations are only applied to the training set, while validation and test sets retain original images to ensure objective evaluation of model generalization. The original dataset is split into training set (70%), validation set (15%) and test set (15%) by stratified sampling to keep the original expression class distribution ratio.

3.4 Focal Loss for Class Imbalance Mitigation

Standard cross-entropy loss treats all samples equally. For FER datasets with unbalanced categories, the model tends to fit majority samples such as happy and neutral, ignoring minority samples like disgust and fear. Focal Loss introduces modulating factor $(1-p_t)^{\gamma}$ to reduce the loss weight of easy-classified samples and focus training on hard minority samples. The mathematical formula is defined as:

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

Where p_t is the model predicted probability of the true category, α_t is the class balance weight coefficient, $\gamma \geqslant 0$ is the focusing hyperparameter. When $\gamma > 0$, the loss contribution of samples with high p_t (easy samples) is suppressed. In this experiment, we set $\gamma = 2$ and set α_t inversely proportional to the sample quantity of each expression category to balance class loss weights.

4. Experiments and Analysis

4.1 Experimental Setup

4.1.1 Datasets

Three public standard facial expression datasets are adopted for evaluation:

1) **FER2013**: Contains 35,887 grayscale face images divided into seven expression categories (angry, disgust, fear, happy, sad, surprise, neutral), with obvious class imbalance;
2) **CK+ (Cohn-Kanade Plus)**: Standard laboratory facial expression dataset with clear facial texture, covering six basic expressions proposed by Ekman;
3) **AffectNet**: Large-scale real-scene color expression dataset with diverse backgrounds, simulating complex application environments.

All images are unified to 48×48 resolution, normalized to \([0,1]\) and converted to three-channel RGB input for network training.

4.1.2 Hardware and Software Environment

- Hardware: NVIDIA RTX 3060 GPU with 12GB video memory;
- Software: Python 3.8, PyTorch 1.10, OpenCV, Albumentations data augmentation library;
- Training hyperparameters: Batch size = 32, initial learning rate = 0.001, Adam optimizer, maximum epoch = 100, early stopping strategy with patience=10 to prevent overfitting.

4.1.3 Evaluation Metrics

Four mainstream classification evaluation indicators are used: Overall Accuracy (Acc), Mean Recall (Recall), Mean F1-Score and model parameter quantity (Params) to comprehensively compare recognition performance and lightweight degree. Confusion matrix is drawn to analyze the recognition effect of each minority expression category.

4.2 Comparative Experiment Results

Table 1. Performance comparison of different lightweight models

Model	FER2013 Acc(%)	CK+ Acc(%)	AffectNet Acc(%)	Params(M)
ShuffleNet V2	63.21	89.57	58.74	1.26
MobileNet V2	64.75	90.13	60.12	2.28
ShuffleNet V2+SE	66.83	92.46	62.39	1.31
Proposed Model (SE+Aug+Focal Loss)	69.47	94.82	65.71	1.31

From Table 1, the proposed model achieves the highest recognition accuracy on all three datasets, with only a tiny parameter increment of 0.05M compared with vanilla ShuffleNet V2, fully retaining lightweight deployment advantages. On FER2013 with severe category imbalance, the accuracy rises by 6.26% compared with the original network, which verifies the effectiveness of Focal Loss in solving unbalanced sample bias. The model outperforms MobileNet V2 with far fewer parameters, proving its superiority for edge terminal deployment.

4.3 Ablation Experiment

Table 2. Ablation study results on FER2013

Setting	Accuracy(%)	Mean F1
Baseline (ShuffleNet V2)	63.21	0.603
Baseline + SE Attention	66.83	0.647
Baseline + SE + Data Augmentation	68.15	0.669
Baseline + SE + Aug + Focal Loss (Ours)	69.47	0.692

1) Adding SE attention alone brings a 3.62% accuracy improvement, demonstrating that the attention module effectively extracts discriminative facial emotional features;
2) Combining multi-dimensional data augmentation further improves accuracy by 1.32%, proving that expanded diverse samples enhance model anti-interference ability;
3) Introducing Focal Loss increases accuracy by another 1.32%, significantly lifting the F1 value of minority categories such as disgust and fear, effectively alleviating dataset imbalance issues.

4.4 Visualization Analysis

The class confusion matrix of the proposed model on FER2013 shows that the recognition recall rate of minority expressions disgust and fear is greatly improved compared with original ShuffleNet V2, which is attributed to the focusing mechanism of Focal Loss. Attention heatmap visualization displays that the model concentrates feature responses on facial emotional key areas including eyes, eyebrows and mouth, rather than irrelevant background regions, proving the SE module's feature screening capability.

5. Conclusion and Future Work

Aiming at the problems of low feature discrimination and category imbalance in lightweight facial expression recognition based on ShuffleNet, this paper proposes an improved ShuffleNet model embedded with SE attention, compound data augmentation and Focal Loss. SE attention modules are inserted into shuffle blocks to adaptively weight expression-sensitive feature channels; multi-dimensional augmentation strategies expand training sample diversity to strengthen environmental robustness; Focal Loss suppresses the loss weight of easy majority samples and improves the recognition precision of rare facial expressions. Experimental results on FER2013, CK+ and AffectNet datasets indicate that the proposed method achieves higher recognition accuracy than original ShuffleNet V2 and MobileNet V2 under ultra-light parameter scale, satisfying the real-time deployment demand of facial expression detection on mobile and embedded devices.

In future research, we will carry out two aspects of optimization:

1) Integrate lightweight coordinate attention (CA) to replace partial SE modules to capture spatial position information of facial organs;
2) Combine model quantization and network pruning technology to further compress model volume and reduce inference latency

for low-power embedded chips.

References

Ekman P, Friesen W V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124-129.

Lai Z Y, Chen R H, Qian Y R. (2020). Real-time micro-expression recognition algorithm based on atrous convolutions for CNN. *Application Research of Computers, 37*(12), 3777-3780, 3835.

Li D H, Zhong T, Wang S, et al. (2024). Lightweight identification of tomato leaf diseases based on improved ShuffleNet v2. *Jiangsu Agricultural Sciences, 52*(03), 220-228.

Li F. (2020). Research on facial expression recognition technology based on deep learning. Dalian Maritime University.

Peng K B, Lv X J, Li C, et al. (2024). Research on railway passenger station security risk assessment technology based on contraband detection and facial expression recognition. *Railway Transport and Economy, 46*(01), 109-115.

Qu F, Wang S J, Yan W J, et al. (2018). CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing, 9*(4), 424-436.

Rong Y, Liu J W, Li X L. (2024). Adaptive hybrid network for students' classroom affective computing. *Journal of Computer Applications, 29*(6), 1607-1627.

Tao J H, Fan C H, Lian Z, et al. (2024). Development status and trend of multimodal emotion recognition and understanding. *Journal of Image and Graphics, 29*(06), 1607-1627.

Wang X R, Zhang Y T, Peng X. (2022). Research on vehicle fatigue driving system based on face key point detection. *Wireless Internet Technology, 19*(17), 82-84.

Yang Z, Wu S D, Jia R. (2024). Autonomous Classification and Early Warning Framework for UAV Images Combining Improved ShuffleNet-V2 and Attention Mechanism. *Journal of Wuhan Institute of Chemical Technology, 47*(5), 89-95.

Zhou Y, Chen S C, Wang Y M, et al. (2020). Review of research on lightweight convolutional neural networks. 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference. *IEEE*, 1713-1720.