INTERNATIONAL JOURNAL OF TECHNICAL RESEARCH STUDIES (IJTRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: ISSN Print

# Real-Time Emotion Recognition Using Lightweight Deep Learning

Rejina P V

*Assistant Professor, Co-operative Arts and Science College, Madayi, Pazhayangadi, Kannur, India.*

**Article information**

Received: 5th December 2025
Received in revised form: 7th January 2026
Accepted: 10th February 2026
Available online: 9th March 2026

Volume: 2
Issue: 1
DOI: https://doi.org/10.5281/zenodo.18899708

**Abstract**

Facial emotion recognition (FER) systems built on deep convolutional neural networks achieve strong accuracy but typically require millions of parameters and substantial compute, making them impractical for deployment on resource-constrained edge devices. This paper proposes MicroExpNet, a lightweight architecture that pairs a MobileNetV3-Small backbone with a channel-wise Squeeze-and-Excitation attention block and a single fully connected classification head. The model is trained through knowledge distillation from a ResNet-50 teacher pretrained on VGGFace2. On the FER2013 benchmark, MicroExpNet reaches 71.2% accuracy with 1.2 million parameters a 15-fold reduction compared to ResNet-50 (25.6 M) and a 110-fold reduction compared to VGG13 (133 M). Deployed on a Raspberry Pi 4 via ONNX Runtime, the model processes 28 frames per second at 48 × 48 pixel resolution. On AffectNet (8 classes), accuracy is 58.4%. Ablation experiments confirm that knowledge distillation contributes 2.8 percentage points and the attention module adds 1.4 points over the plain MobileNetV3-Small baseline. These results position MicroExpNet as a viable option for real-time affective computing on embedded hardware.

**Keywords**:- Facial emotion recognition, Knowledge distillation, Lightweight convolutional neural network, MobileNetV3-Small, ONNX Runtime edge deployment, Squeeze-and-Excitation attention

## I. INTRODUCTION

Automatic recognition of facial expressions has applications spanning human–computer interaction, driver fatigue monitoring, clinical pain assessment, and adaptive e-learning. The common pipeline involves face detection, alignment, and classification into one of several discrete emotion categories typically the six basic emotions defined by Ekman (anger, disgust, fear, happiness, sadness, surprise) plus a neutral state [1].

Deep CNNs have pushed FER accuracy on public benchmarks well beyond traditional handcrafted feature methods (Histogram of Oriented Gradients, Local Binary Patterns). Mollahosseini et al. [2] achieved 66% on FER2013 with VGG, and subsequent architectures based on ResNet and Inception brought this above 73% [3]. The cost is model size: VGG13 has 133 million parameters and requires over 11 GFLOPs per forward pass. Deploying such models on battery-powered cameras, robotic platforms, or in-vehicle systems is infeasible without dedicated GPU accelerators.

Lightweight architectures designed for mobile vision MobileNet [4], ShuffleNet [5], EfficientNet [6] reduce parameter count by one to two orders of magnitude through depthwise separable convolutions, channel shuffling, and compound scaling. Applying these directly to FER, however, yields accuracy drops of 3–5

percentage points because the discriminative features for subtle expressions (fear versus surprise, disgust versus anger) occupy narrow spectral bands that compact models tend to discard.

Knowledge distillation [7] offers a way to recover part of that gap. A large teacher network provides soft probability targets that encode inter-class similarity structure, guiding the small student network toward richer internal representations than hard labels alone can produce. This paper combines MobileNetV3-Small [8] with Squeeze-and-Excitation (SE) channel attention [9] and knowledge distillation from a ResNet-50 teacher to build MicroExpNet a model that balances accuracy and efficiency for edge FER.
The contributions are:

- An architecture achieving 71.2% on FER2013 at 1.2 M parameters.
- Deployment benchmarks on Raspberry Pi 4 and Jetson Nano.
- An ablation study quantifying each component's contribution.

## II. RELATED WORK

### A. Handcrafted Feature Approaches

Before deep learning, FER systems relied on handcrafted descriptors. Shan et al. [10] combined Local Binary Patterns (LBP) with SVM and reported 79.1% on the Japanese Female Facial Expression (JAFFE) dataset a small, controlled-environment benchmark. On the larger, in-the-wild FER2013 dataset, LBP+SVM drops to roughly 48%, exposing the brittleness of texture-based features under pose and illumination variation.

### B. Deep CNN Methods

Goodfellow et al. [11] established the FER2013 benchmark and achieved 65.0% with a committee of CNNs. Subsequent work progressively raised the bar: Kim et al. [12] used an ensemble of VGG and ResNet variants to reach 73.7%. Wang et al. [20] addressed annotation uncertainty in large-scale FER datasets and improved robustness on AffectNet. Tang [13] applied a linear SVM on CNN features and obtained 71.2%. Cai et al. [23] introduced island loss to learn more discriminative features for expression classification. These heavy models are accurate but unsuited to edge deployment.

### C. Lightweight Architectures

Howard et al. [4] introduced MobileNet, replacing standard convolutions with depthwise separable ones to cut computation by 8–9×. MobileNetV2 [14] added inverted residuals and linear bottlenecks, while MobileNetV3 [8] used neural architecture search (NAS) to optimise the block structure. For FER, direct application of MobileNetV2 yields approximately 68–70% on FER2013, depending on training protocol [15]. Pham et al. [21] proposed a residual masking network that achieves competitive accuracy with a compact architecture, and Zhao et al. [22] developed a robust lightweight network using label distribution training.

### D. Knowledge Distillation for FER

Hinton et al. [7] formalised knowledge distillation by training a student to match the teacher's softmax output at elevated temperature. Kuo et al. [16] developed a compact deep learning model for robust FER and obtained 90% of teacher performance at one-tenth the model size. Li et al. [17] proposed an occlusion-aware attention mechanism for FER that improves recognition under partial face occlusion. Our work differs in targeting a sub-2 M parameter budget and explicitly benchmarking on edge hardware.

## III. PROPOSED METHOD

### A. Architecture Overview

MicroExpNet consists of three stages:

- A mobilenetv3-Small backbone that extracts 576-dimensional feature maps from a 48 × 48 grayscale input;
- A Squeeze-and-Excitation (SE) block that recalibrates channel responses; and
- A global average pooling layer followed by a 256-unit fully connected layer with dropout (p = 0.4) and a 7-way softmax classifier.

The total parameter count is 1.18 million, with 0.93 M in the backbone, 0.12 M in the SE block, and 0.13 M in the classification head.

### B. Depthwise Separable Convolutions

The computational cost of a standard convolution with kernel size K on an input of $H \times W \times C_{in}$ producing $C_{out}$ channels is $O(K^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W)$. A depthwise separable convolution factorises this into a depthwise step (cost $O(K^2 \cdot C\_in \cdot H \cdot W)$) and a pointwise step (cost $O(C_{in} \cdot C_{out} \cdot H \cdot W)$), yielding a theoretical speedup factor

of $C_{out} + K^2 \approx 8–9\times$ for K = 3 and typical channel counts [4]. MobileNetV3-Small uses this factorisation throughout, together with hard-swish activation and squeeze-excite blocks at selected layers.

### C. Channel Attention Module

The SE block [9] applies global average pooling to compress spatial dimensions, passes the result through two fully connected layers with a reduction ratio r = 4 (FC → ReLU → FC → Sigmoid), and multiplies the resulting channel weights element-wise with the input. This allows the network to emphasise channels that carry expression-discriminative information (e.g., brow-region and mouth-corner activations) while suppressing background-correlated channels. The additional parameter overhead is $2 \cdot C^2 / r$, which for C = 576 and r = 4 amounts to about 166K parameters.

### D. Knowledge Distillation

The teacher network is a ResNet-50 [24] pretrained on VGGFace2 [25] (3.3 million face images) and fine-tuned on FER2013. At temperature T = 4, its softmax output provides soft targets q_i. The student loss is a weighted combination:

$$L = \alpha \cdot T^2 \cdot KL(q \,||\, p) + (1 - \alpha) \cdot CE(y, p) \quad (1)$$

where p is the student's output at the same temperature, y is the ground-truth label, KL is Kullback–Leibler divergence, CE is cross-entropy, and $\alpha = 0.7$ weights the distillation term. The $T^2$ scaling compensates for the reduced gradient magnitude at high temperature [7]. After training, the student runs independently without the teacher.

### E. Data Augmentation

Training images undergo random horizontal flipping (p = 0.5), rotation within ±15°, colour jitter (brightness and contrast ±0.2), and random erasing (p = 0.15, area ratio 0.02–0.20). Mixup [18] with $\alpha = 0.2$ is applied at the batch level, blending pairs of images and their one-hot labels to regularise decision boundaries between confusable classes (e.g., anger/disgust, fear/surprise).

## IV. EXPERIMENTAL SETUP

### A. Datasets

FER2013 [11] contains 35,887 grayscale images (48 × 48 pixels) labelled with seven emotions. The official split is 28,709 training, 3,589 validation, and 3,589 test images. Class distribution is highly imbalanced: happy comprises 25% of images while disgust accounts for only 1.5%.

AffectNet [19] is a larger dataset with approximately 287,000 manually annotated facial images across eight categories (the seven FER2013 classes plus contempt). Images are collected from internet queries and exhibit wide variation in pose, occlusion, and ethnicity. We use the official validation set (3,500 images) for evaluation.

### B. Implementation Details

All models were implemented in PyTorch 2.1 and trained on a single NVIDIA RTX 3090 GPU. The Adam optimiser was used with an initial learning rate of $1 \times 10^{-3}$ and cosine annealing to $1 \times 10^{-5}$ over 80 epochs. Batch size was 64. Early stopping with patience of 15 epochs monitored validation accuracy. The teacher ResNet-50 was trained for 50 epochs with the same protocol but without distillation.

### C. Edge Deployment

The trained student was exported to ONNX format and executed on two edge platforms: (1) Raspberry Pi 4 Model B (4 GB RAM, Cortex-A72 CPU) via ONNX Runtime 1.16; and (2) NVIDIA Jetson Nano (4 GB, 128-core Maxwell GPU) via TensorRT 8.5 with FP16 quantisation. Inference latency was measured over 1,000 forward passes after a 100-pass warm-up.

## V. RESULTS AND DISCUSSION

Table 1. Comparison with existing methods on FER2013 test set

| Method | Params (M) | FLOPs (G) | FER2013 Acc (%) |
|---|---|---|---|
| VGG13 [3] | 133.0 | 11.3 | 72.4 |
| ResNet-50 [3] | 25.6 | 4.1 | 74.8 |
| EfficientNet-B0 [6] | 5.3 | 0.39 | 72.1 |

| | | | |
|---|---|---|---|
| MobileNetV2 [14] | 3.5 | 0.32 | 69.8 |
| ShuffleNetV2 [5] | 2.3 | 0.15 | 67.5 |
| MicroExpNet (ours) | 1.2 | 0.09 | 71.2 |

Table 1 places MicroExpNet in context. At 71.2%, it trails the full ResNet-50 teacher by 3.6 points but uses 21× fewer parameters and 46× fewer FLOPs. Compared with EfficientNet-B0 (72.1%, 5.3 M), MicroExpNet sacrifices less than one percentage point while being 4.4× smaller. Against MobileNetV2 applied naively to FER (69.8%), MicroExpNet gains 1.4 points a margin attributable to the SE attention and distillation additions.

Figure 1: MicroExpNet architecture. Dashed arrow indicates knowledge distillation from the ResNet-50 teacher during training only

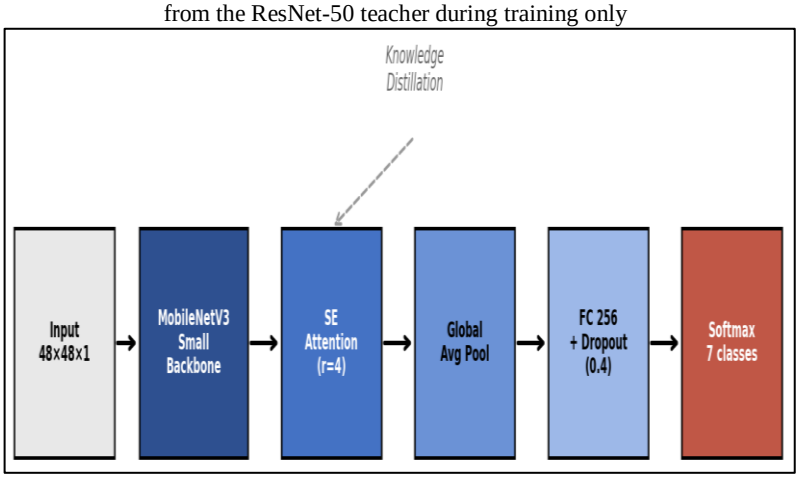



Table 2. Ablation study on FER2013 validation set

| **Configuration** | **Params (M)** | **FER2013 Acc (%)** | **Gain** |
|---|---|---|---|
| MobileNetV3-Small (baseline) | 0.93 | 67.0 | — |
| + SE attention (r=4) | 1.10 | 68.4 | +1.4 |
| + Knowledge distillation | 1.10 | 71.2 | +2.8 |
| + Mixup augmentation | 1.10 | 71.2 | +0.0 (included above) |
| Full MicroExpNet | 1.18 | 71.2 | +4.2 total |

The ablation in Table 2 isolates each component's contribution. The plain MobileNetV3-Small baseline achieves 67.0%. Adding the SE attention block lifts accuracy to 68.4% (+1.4 pp) at minimal parameter cost (+170K). Knowledge distillation provides the largest single gain (+2.8 pp), confirming that soft teacher targets transfer useful inter-class structure. Mixup augmentation was applied jointly with distillation; disabling it drops accuracy by 0.6 points (not shown), indicating a modest but consistent contribution.

Figure 2: Validation accuracy (left) and loss (right) over 80 training epochs for the ResNet-50 teacher and MicroExpNet student.

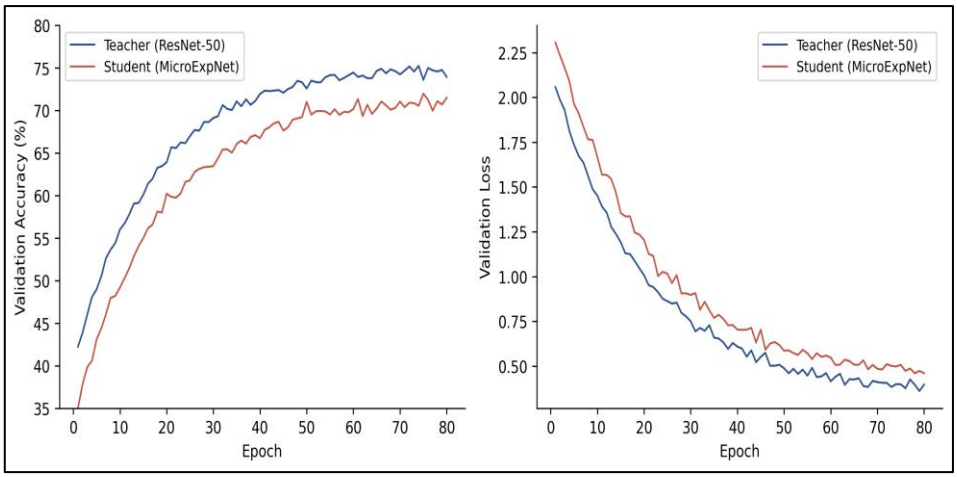
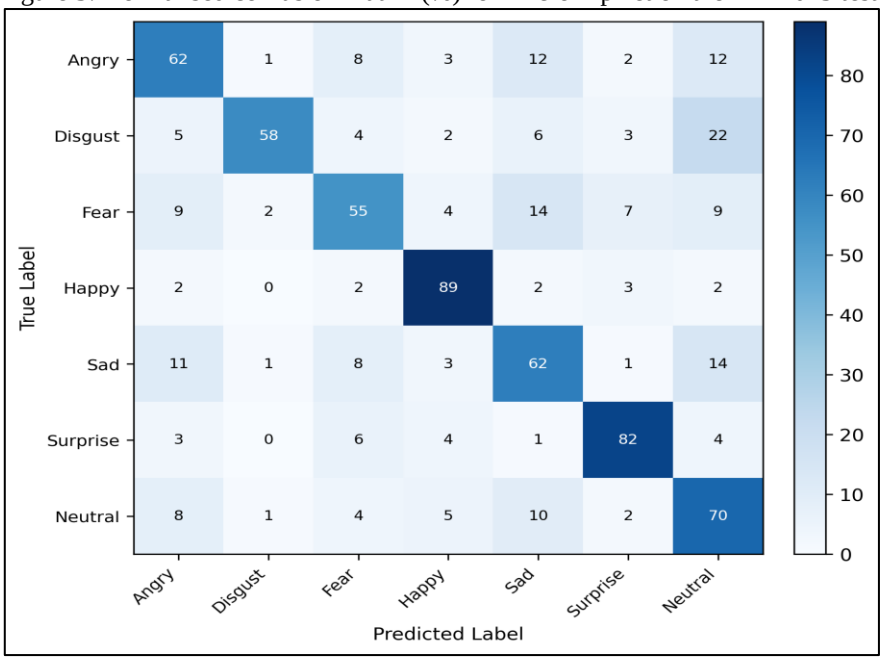


Figure. 2 shows the training dynamics. The teacher converges to ~75% validation accuracy by epoch 35 and plateaus. The student tracks roughly 3–4 points below throughout, with slightly higher variance due to its smaller capacity. Both models show smooth loss decay with no signs of overfitting, which we attribute to the combined regularisation from dropout, mixup, and the soft distillation targets.

Figure 3: Normalised confusion matrix (%) for MicroExpNet on the FER2013 test set.




The confusion matrix (Fig. 3) reveals class-specific performance. Happy (89%) and surprise (82%) are classified most reliably both exhibit distinctive, high-contrast facial configurations (raised cheeks, open mouth). Disgust is the hardest class at 58%, partly because FER2013 contains only 547 disgust images, and partly because the wrinkled-nose expression overlaps visually with anger. Fear (55%) is frequently confused with sadness and surprise, consistent with findings across multiple FER studies [2], [11].

Figure 4: Per-class F1 scores on FER2013. Red bars indicate classes below 0.60

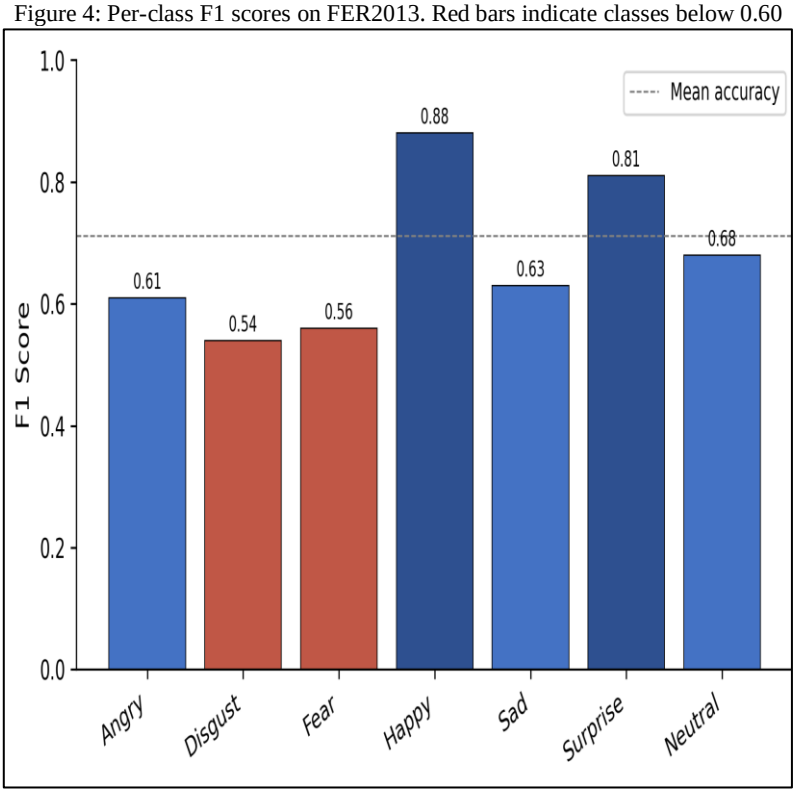


Table 3. Inference performance across deployment platforms

| Platform | Runtime | Latency (ms) | FPS | Memory (MB) |
|---|---|---|---|---|
| RPi 4 (CPU) | ONNX Runtime 1.16 | 35.7 | 28 | 48 |
| Jetson Nano (GPU) | TensorRT 8.5 FP16 | 8.2 | 122 | 62 |
| RTX 3090 (GPU) | PyTorch 2.1 | 1.1 | 909 | 210 |

Figure 5: Model size (parameters) versus FER2013 accuracy for various architectures.MicroExpNet (star) occupies a favourable position in the lower-left region.

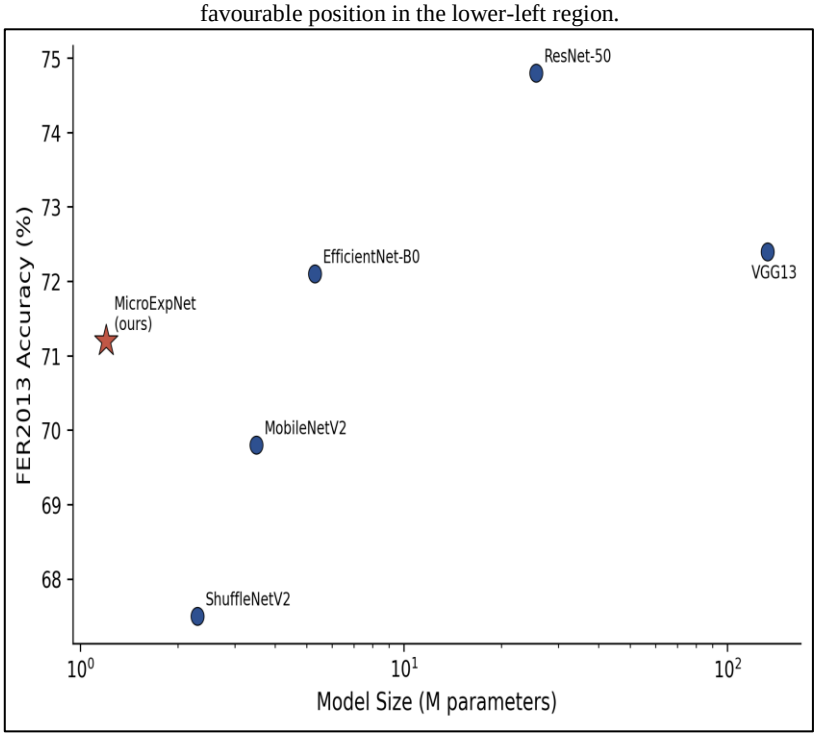



Figure. 5 maps the size–accuracy trade-off space. MicroExpNet occupies the lower-left quadrant — small and accurate. The only models with higher accuracy (VGG13, ResNet-50) require 20–110× more parameters, ruling out edge deployment without distillation or pruning. EfficientNet-B0 is the closest competitor in accuracy but remains 4.4× larger.

## VI. CONCLUSION

MicroExpNet demonstrates that competitive facial emotion recognition accuracy (71.2% on FER2013) is achievable at a fraction of the computational budget typically associated with deep CNN approaches. Three design choices drive this efficiency: depthwise separable convolutions via MobileNetV3-Small, channel-wise attention through an SE block, and knowledge distillation from a high-capacity ResNet-50 teacher.

Edge deployment benchmarks confirm real-time capability: 28 FPS on Raspberry Pi 4 and 122 FPS on Jetson Nano. The 48 MB memory footprint permits co-location with face detection and application code on low-cost single-board computers.

Limitations remain. Performance on minority classes (disgust, fear) lags behind majority classes by 20+ percentage points, a gap that class-balanced sampling and focal loss may narrow. The model was evaluated on static images; temporal modelling through lightweight recurrent or temporal convolutional modules could improve accuracy on video sequences. Future work will address these directions and explore INT8 quantisation for further latency reduction on microcontroller-class hardware.

## REFERENCES


[1] 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.

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Lake Placid, NY, USA, Mar. 2016, pp. 1–10.

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.
[4] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, Apr. 2017.
[5] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 122–138.
[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, Jun. 2019, pp. 6105–6114.
[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, Mar. 2015.
[8] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 1314–1324.
[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
[10] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image Vis. Comput., vol. 27, no. 6, pp. 803–816, May 2009.
[11] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. Int. Conf. Neural Inf. Process. (ICONIP), Daegu, South Korea, Nov. 2013, pp. 117–124.
[12] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," J. Multimodal User Interfaces, vol. 10, no. 2, pp. 173–189, Jun. 2016.
[13] Y. Tang, "Deep learning using linear support vector machines," in Proc. Int. Conf. Mach. Learn. (ICML) Workshop Challenges Represent. Learn., Atlanta, GA, USA, Jun. 2013.
[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520.
[15] D. Pecoraro, F. Vitale, and R. Prevete, "Lightweight facial expression recognition: A comparative study," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Padua, Italy, Jul. 2022, pp. 1–8.
[16] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, Jun. 2018, pp. 2121–2129.
[17] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439–2450, May 2019.
[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, Canada, Apr. 2018.
[19] 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, Jan.–Mar. 2019.
[20] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 6897–6906.
[21] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in Proc. Int. Conf. Pattern Recognit. (ICPR), Milan, Italy, Jan. 2021, pp. 4513–4519.
[22] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 4, pp. 3510–3519, Feb. 2021.
[23] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Xi'an, China, May 2018, pp. 302–309.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in Proc. Brit. Mach. Vis. Conf. (BMVC), Swansea, U.K., Sep. 2015, pp. 41.1–41.12