Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL36462

Model Capacity for Small-Sample Facial Expression Recognition

Yangxuan Xie

Software Engineering, East China Normal University, Shanghai, China
10245101534@stu.ecnu.edu.cn

Abstract. Facial expression recognition (FER) is increasingly required in classroom affect analysis, lightweight human-computer interaction, and domain-specific behavioral monitoring, where only limited labeled facial images are available. Reliable FER in such low-resource settings is critical because unstable predictions can distort downstream decisions, particularly for minority classes and visually similar negative expressions. Existing pipelines often increase depth, attention, region fusion, or transfer learning capacity, but the completed experiments show that these designs can mismatch small datasets and underperform compact alternatives. This paper presents a stability-oriented convolutional study that rethinks model capacity for seven-class small-sample FER. The framework combines stratified partitioning, grayscale normalization, compact VGG-style representation learning, global average pooling, label smoothing, dropout, L2 regularization, and momentum-based optimization to control estimator variance while preserving local expression cues. An error-aware protocol further evaluates macro-F1, weighted-F1, class-wise recall, sensitivity behavior, and confusion patterns. Experiments on a 981-image FER dataset with a stratified 70/15/15 split compare LeNet, VGG variants, ResNet-small, ROI-fusion VGG, and MobileNetV2 transfer learning. Results show that VGG-optimized achieves 97.30% accuracy and 95.64% macro-F1, while lightweight LeNet reaches 98.65% accuracy, and complex ROI/transfer alternatives drop to 56.08% and 83.78%, confirming that capacity control dominates architectural complexity.

Keywords: Facial expression recognition, Small-sample learning, Convolutional neural networks, Model capacity, Regularization

1. Introduction

Facial expression recognition (FER) aims to infer affective states from facial images and is an important component of human-centered intelligence systems [1-3]. It is widely relevant to intelligent tutoring, lightweight human-computer interaction, driver monitoring, and domain-specific behavioral analysis, where FER models are often trained with limited task-specific data rather than large-scale curated benchmarks. Such low-resource conditions are caused by privacy constraints, annotation cost, demographic imbalance, and the difficulty of collecting balanced expression samples. Without explicit treatment of small-sample instability, FER systems may produce over-confident but unreliable predictions, especially for minority classes and visually ambiguous negative

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL36462

expressions, thereby weakening downstream decisions and user trust. Existing FER methods can be grouped into four families. Classical shallow CNNs learn local appearance patterns with low computational cost [4, 5]. VGG-style and residual CNNs improve representational capacity through deeper hierarchical feature extraction [6, 7]. Attention and region-of-interest methods focus on expression-relevant facial regions such as the eyes, brows, and mouth [8, 9]. Transfer-learning methods use pretrained backbones to compensate for limited target-domain supervision [10, 11]. Although these approaches are effective in many settings, they do not fully address small-sample FER. When fewer than one thousand images are available, higher architectural complexity can increase estimation variance, ROI or attention branches may overfit local artifacts, and pretrained RGB natural-image features may not transfer cleanly to low-resolution grayscale facial expressions.

Thus, the main limitation therefore lies in the gap between model capacity and the amount of available supervision. For reliable small-sample FER, simply expanding the network architecture is often insufficient; capacity control and diagnostic evaluation also need to receive careful consideration. Based on this idea, the study examines a stability-oriented convolutional framework that combines compact representation learning, global average pooling, calibrated supervision, and conservative regularization. The framework retains the local feature-learning ability of CNNs while limiting the possibility that the model memorizes patterns that only occur in the small training set. Attention, ROI fusion, and transfer learning are often expected to improve FER performance; however, this study examines whether these mechanisms still provide advantages when the available supervision is limited and the class distribution is imbalanced.

The study provides three main contributions. First, the experimental results identify capacity mismatch as an important source of failure in small-sample FER; under the same data split, ROI-fusion and transfer-learning baselines may perform worse than lightweight CNNs. Second, a stability-oriented VGG-style framework is developed and evaluated; it combines global average pooling, label smoothing, dropout, L2 regularization, and SGD with momentum to form a low-variance training strategy. Third, an error-aware evaluation protocol is introduced to examine accuracy, macro-F1, class-wise recall, sensitivity behavior, and confusion patterns; the results indicate that ambiguity among negative expressions remains the main unresolved recognition problem. The remaining sections present the related work, methodology, experimental results, and conclusions.

Section 2 reviews related work. Section 3 formulates the small-sample FER problem. Section 4 presents the methodology. Section 5 reports comparative, ablation, sensitivity, and error analyses. Section 6 concludes the paper.

2. Related work

FER research has evolved from handcrafted feature engineering to deep representation learning [1, 12]. Early FER systems mainly described facial changes through local texture descriptors, geometric landmarks, and shallow classifiers [1]. As research progressed, convolutional networks, attention mechanisms, region-aware modeling, and transfer learning became more common. These methods provide stronger representation capabilities; however, when labeled samples remain limited, increasing model capacity does not always produce better recognition results.Classical and lightweight CNNs remain an important research direction [4, 5, 13]. LeNet-style models use convolution, pooling, and nonlinear activation to capture local facial patterns. Their relatively simple structures require fewer computational resources and can maintain a suitable bias-variance balance on small datasets [4, 5]; nevertheless, the limited network depth may prevent them from fully describing complex expression variations. VGG-style networks increase the model depth by

DOI: 10.54254/2755-2721/2026.GL36462

stacking small convolutional kernels, which allows them to learn facial features at different levels [6]. Residual networks use skip connections to simplify network optimization [7]. The performance of these deeper models still depends on whether the available dataset contains enough samples to support their larger capacity. Attention-based and region-based FER methods use the anatomical characteristics of facial expressions [8, 9, 14]. Expression-related information usually appears around the eyes, eyebrows, cheeks, and mouth; therefore, channel attention, spatial attention, and ROI fusion can help the model focus on more useful facial features. CBAM evaluates channel and spatial importance in sequence [9], whereas ROI-based models directly process selected facial regions [8]. Although these designs have clear motivations, they introduce additional parameters and make training more complicated; with only a small number of samples, the increased flexibility may enlarge sampling noise or cause the model to learn local artifacts. Transfer-learning methods attempt to reduce the demand for labeled data through large-scale natural-image pretraining [10, 11, 14]. MobileNetV2 and EfficientNet provide efficient pretrained backbones with strong accuracy-computation tradeoffs [10, 11]. Nevertheless, transfer learning is not automatically optimal for small-sample FER because the target images are low-resolution grayscale facial expressions, whereas common pretrained backbones are optimized for RGB natural images. This domain mismatch can limit adaptation when fine-tuning data are scarce. Regularization strategies, including dropout, weight decay, label smoothing, data augmentation, and learning-rate scheduling, are widely used to improve neural network generalization [15-17]. In small-sample FER, these techniques are central mechanisms for controlling variance, reducing overconfidence, and stabilizing decision boundaries. This paper differs from prior work by treating regularization and capacity control as the primary design axis and by empirically showing that complexity escalation can be counterproductive under limited supervision.

3. Methodology

3.1. Framework overview

The proposed framework follows a stability-first design. Rather than increasing the number of branches or importing a large pretrained backbone [8, 10, 11], it constructs a compact convolutional pipeline whose components are selected to control variance while retaining expression-discriminative local features.

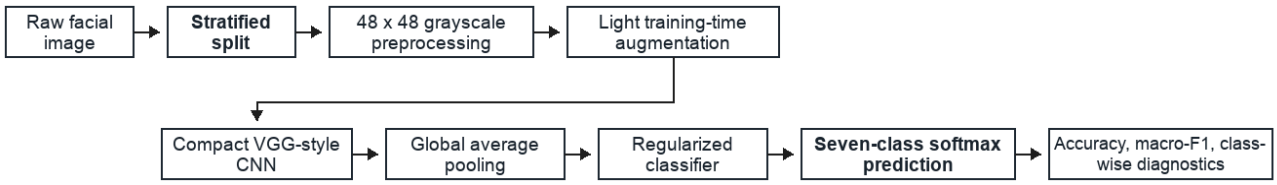


Figure 1. Stability-oriented workflow of the proposed small-sample FER framework. (picture credit: original)

As shown in Figure 1, the workflow consists of five interacting modules: stratified partitioning, input normalization, conservative augmentation, compact convolutional representation learning, and regularized classification. Each module addresses a specific failure mode of small-sample FER.

DOI: 10.54254/2755-2721/2026.GL36462

3.2. Stratified small-sample partitioning

The first challenge is evaluation reliability. In imbalanced small datasets, a random global split may leave minority classes underrepresented in validation or test subsets. To avoid this, samples are partitioned independently within each class. Let D_k denote all samples of class k. The class-specific partition is:

$$\mathrm{D_k} = \mathrm{D_k^{train}} \cup \mathrm{D_k^{val}} \cup \mathrm{D_k^{test}} \tag{1}$$

$$\mathrm{D_k^a} \cap \mathrm{D_k^b} = \emptyset, \mathrm{a} \neq \mathrm{b} \tag{2}$$

Here, D_k^{train}, D_k^{val}, and D_k^{test} refer to the training, validation, and test subsets of class k, respectively. This class-level partitioning keeps every expression category in each evaluation stage; it also improves the reliability of class-wise recall and F1, since these metrics require corresponding samples in the test set.

3.3. Input representation and normalization

Each facial image is represented as a grayscale tensor:

$$\mathrm{x} \in \mathrm{R}^{48\times48\times1} \tag{3}$$

Grayscale images are used because facial expressions mainly depend on local shapes, wrinkles, muscle movements, and the relative structure of the eyes and mouth, while color information generally has a smaller influence. A single input channel also reduces data complexity, which is suitable when labeled samples are limited. Pixel normalization is defined as:

$$\mathrm{x^{'}} = \frac{\mathrm{x}}{255} \tag{4}$$

In this equation, x represents the original pixel tensor, while x' represents the normalized tensor. This operation converts pixel values into the range ; as a result, gradient-based optimization becomes more stable, and intermediate activations become less affected by differences in the original pixel scale.

3.4. Conservative data augmentation

The training images receive several mild geometric and photometric transformations, including horizontal flipping, small-angle rotation, slight zooming, translation, and contrast adjustment. The augmentation process is expressed as:

$$x_{aug} = T\left(x;\emptyset\right) \tag{5}$$

Here, denotes a randomly selected transformation, and controls the transformation magnitude. These operations introduce reasonable visual variations while preserving the original expression category. The experimental results also indicate that augmentation does not always improve performance on this dataset; removing it produces perfect test accuracy under the current split. One possible explanation is that the original images are already well aligned, so even mild transformations may change subtle expression features or create a difference between training and

DOI: 10.54254/2755-2721/2026.GL36462

test distributions. Therefore, data augmentation serves as an adjustable regularization method rather than a method that always guarantees better performance.

3.5. Compact VGG-style representation learning

The feature extractor adopts a compact VGG-style CNN containing four convolutional blocks, following the hierarchical feature-learning principle of stacked small-kernel convolution [6]. As illustrated in Figure 2, the channel widths increase from 32 to 64, 128, and finally 256. The earlier layers mainly capture edges, textures, and small facial changes; the later layers combine these local patterns into higher-level expression features.

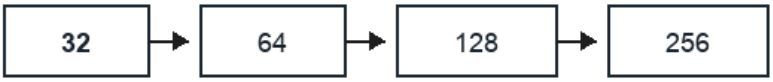


Figure 2. Channel-width progression of the compact VGG-style feature extractor. (picture credit: original)

Each convolutional block follows the operation sequence shown in Figure 3. Specifically, two convolutional transformations are combined with batch normalization and ReLU activation, followed by max pooling and dropout. This design enables local feature extraction, activation stabilization, nonlinear representation learning, spatial downsampling, and regularization within each block.



Figure 3. Operation sequence of each compact VGG-style convolutional block. (picture credit: original)

For a layer l, the convolutional transformation is:

$$F_l = \sigma\left(BN\left(W_l * F_{l-1}\right)\right) \tag{6}$$

where F_l is the feature map at layer l, W_l is the convolution kernel, * denotes convolution, BN is batch normalization, and σ is the ReLU activation. This formulation encodes the local inductive bias needed for expression recognition while avoiding excessive branching. The theoretical rationale is bias-variance control: the model is expressive enough to learn hierarchical facial features but constrained enough to reduce overfitting on small data.

3.6. Global average pooling

Instead of flattening the final feature map into a high-dimensional vector, the framework uses global average pooling:

$$z_c = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} F_{h,w,c} \tag{7}$$

Here, $F_{h,w,c}$ is the activation at spatial location (h, w) in channel c, and z_c is the resulting channel descriptor. The design purpose is parameter efficiency. A flatten classifier can memorize spatially

DOI: 10.54254/2755-2721/2026.GL36462

specific artifacts, whereas global average pooling encourages each channel to behave as a semantic detector. This is particularly important when the number of training samples is small.

3.7. Regularized classification objective

The classifier maps the pooled feature vector to seven logits s_k, which are converted into probabilities through softmax:

$$p_k = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp\left(s_j\right)} \tag{8}$$

where K is the number of expression categories. To reduce over-confident supervision, label smoothing transforms the one-hot label y_k into [16]:

$$y_k^{'} = (1-\varepsilon)y_k + \frac{\varepsilon}{K} \tag{9}$$

Here, ε controls smoothing strength. The objective function is:

$$L = -\sum_{k=1}^{K} y_k' log(p_k) + \lambda \|\theta\|_2^2 \tag{10}$$

The first term is smoothed categorical cross-entropy, which measures the distance between predicted probabilities and softened labels. The second term is L2 regularization, where θ denotes trainable parameters and λ controls weight decay. This objective simultaneously optimizes discrimination and constrains model complexity.

3.8. Training strategy

The optimization uses stochastic gradient descent with momentum, while dropout and weight decay provide complementary regularization [15]:

$$v_t = \mu v_{t-1} + \nabla_\theta L\left(\theta_t\right) \tag{11}$$

$$\theta_{t+1} = \theta_t - \eta v_t \tag{12}$$

Here, η is the learning rate and μ is the momentum coefficient. Momentum aggregates consistent gradient directions and suppresses oscillatory updates. The training process uses validation-driven checkpointing, preserving the model state with the best validation accuracy. A learning-rate scheduler reduces the learning rate when validation loss saturates, improving late-stage convergence.

Although this is a supervised deep-learning problem rather than reinforcement learning, the decision process can be interpreted analogously: the state is the input facial image and current model representation, the action is the predicted expression label, the reward is implicit and inversely related to classification loss, the optimization target is the regularized cross-entropy objective, and the training strategy is mini-batch supervised optimization with validation-based model selection.

DOI: 10.54254/2755-2721/2026.GL36462

4. Experiments

4.1. Experimental setup

The experiments are conducted on a seven-class facial expression dataset containing 981 images. The categories are anger, contempt, disgust, fear, happy, sadness, and surprise. All images are resized to 48 × 48 grayscale tensors. The dataset is split using a stratified 70/15/15 protocol, producing 686 training images, 147 validation images, and 148 test images.

The implementation uses TensorFlow/Keras with NumPy, Pandas, scikit-learn, OpenCV, PyYAML, Matplotlib, seaborn, and tqdm. The main training configuration uses SGD with learning rate 0.003, momentum 0.9, batch size 16, dropout 0.5, L2 coefficient 0.0001, label smoothing 0.05, and 120 epochs. Hardware metadata were not persisted in the result artifacts; a camera-ready version should report CPU, GPU, memory, operating system, CUDA/cuDNN status, and wall-clock time for reproducibility.

4.2. Baselines

The baselines are selected to cover the main architectural families relevant to small-sample FER, including lightweight CNNs [4], cross-connected LeNet variants [5], deeper CNNs [6, 7], ROI-aware models [8], attention-based models [9], and pretrained efficient backbones [10, 11]. LeNet represents shallow lightweight CNNs. VGG-small evaluates a simpler VGG-style model. ResNet-small tests whether residual connections improve optimization under small data. VGG-ROI-fusion tests whether explicit facial-region priors justify their additional complexity. MobileNetV2 transfer learning evaluates whether ImageNet-pretrained compact backbones transfer effectively to low-resolution grayscale expressions. These baselines are representative because they span shallow, hierarchical, residual, region-aware, and pretrained paradigms.

4.3. Main results

The comparative results are summarized in Table 1.

Table 1. Comparative performance of representative FER models

Method	Accuracy	Macro-F1	Weighted-F1	Anger Recall	Contempt Recall	Sadness Precision
LeNet	98.65	97.56	98.64	95.24	87.50	92.31
VGG-small	97.97	96.10	97.94	100.00	75.00	92.31
VGG-optimized	97.30	95.64	97.32	90.48	87.50	85.71
ResNet-small	70.27	53.12	65.66	61.90	62.50	0.00
VGG-ROI-fusion	56.08	48.97	51.04	42.86	37.50	37.50
MobileNetV2-transfer	83.78	79.49	83.50	47.62	75.00	60.00

The results reveal a non-monotonic relationship between model complexity and recognition performance. LeNet achieves the highest accuracy and macro-F1, while VGG-small and VGG-optimized remain highly competitive. In contrast, ResNet-small, ROI-fusion VGG, and MobileNetV2 transfer learning perform substantially worse. This pattern supports the central thesis of the paper: under small-sample FER, added architectural complexity does not necessarily improve generalization. The ROI-fusion model likely suffers from insufficient data to train multiple branches

DOI: 10.54254/2755-2721/2026.GL36462

reliably, while MobileNetV2 transfer learning may suffer from domain mismatch between RGB natural-image pretraining and low-resolution grayscale facial expression images.

The VGG-optimized model is therefore best interpreted not as the absolute top performer in the current split, but as a stable capacity-controlled design. Its performance remains strong while preserving a more scalable hierarchical representation than LeNet. The empirical evidence suggests that future deployment should prioritize compact architectures and class-wise diagnostics rather than assuming deeper or region-aware models are superior.

4.4. Ablation study

The comparative results are summarized in Table 2.

Table 2. Ablation study of regularization and imbalance-handling components

Variant	Accuracy	Macro-F1	Weighted-F1	Anger Recall	Contempt Recall	Sadness Precision
Full VGG-optimized	97.30	95.64	97.32	90.48	87.50	85.71
w/o augmentation	100.00	100.00	100.00	100.00	100.00	100.00
w/o label smoothing	93.92	92.04	94.17	85.71	87.50	78.57
w/o dropout	98.65	98.54	98.66	100.00	100.00	100.00
w/o L2	96.62	95.17	96.71	90.48	87.50	85.71
with class weight	93.92	90.62	94.09	90.48	62.50	70.59
focal loss	91.22	89.45	91.49	85.71	87.50	83.33

The ablation results show that different regularization methods have noticeably different effects on the model. After label smoothing is removed, both accuracy and macro-F1 decline, which indicates that calibrated supervision is helpful when the number of labels is limited. Removing L2 also causes a small performance decrease; therefore, controlling the weight norm can support more stable generalization. By comparison, class weighting and focal loss [18] lead to lower overall performance, suggesting that losses designed for class imbalance may place too much emphasis on certain categories and shift the decision boundary. This problem is particularly clear in the class-weight experiment, where the recall of contempt falls to 62.50%.

Some ablation results also need further consideration. The model achieves perfect test performance without augmentation, while removing dropout produces better results than the complete VGG-optimized configuration. These findings do not directly contradict the stability-oriented design; instead, they suggest that the images in the current training and test subsets may already have similar visual characteristics, and compact CNNs can learn their patterns without strong regularization. When the distributions of the two subsets are close, mild augmentation may alter subtle expression details, while a high dropout rate may remove useful features during training. Therefore, regularization strength should be adjusted according to the dataset rather than increased mechanically, especially for small-sample FER tasks.

4.5. Sensitivity analysis

Table 3 reports the ablation results for the main VGG-optimized configuration and its regularization variants.

DOI: 10.54254/2755-2721/2026.GL36462

Table 3. Sensitivity analysis under key hyperparameter variations

Variant	Accuracy	Macro-F1	Weighted-F1	Anger Recall	Contempt Recall	Sadness Precision
Main: dropout 0.5, LS 0.05, LR 0.003, BS 16	97.30	95.64	97.32	90.48	87.50	85.71
Batch size 8	86.49	83.12	86.65	76.19	87.50	66.67
Class-weight cap 2.0	97.97	97.42	98.00	90.48	100.00	85.71
Dropout 0.3	93.24	90.11	93.37	85.71	62.50	66.67
Learning rate 0.001	97.30	95.90	97.34	90.48	87.50	85.71
Label smoothing 0.03	95.95	92.57	95.94	85.71	62.50	70.59

The sensitivity analysis shows that several parameter settings have a clear effect on model performance, especially batch size, dropout strength, and label-smoothing magnitude. When the batch size is reduced to 8, accuracy decreases to 86.49%; the smaller batch produces greater gradient fluctuation, which may make the training process less stable. Dropout 0.3 and label smoothing 0.03 also perform worse than the main configuration, particularly in contempt recall and sadness precision. In comparison, a learning rate of 0.001 achieves a result close to that of the main learning rate, showing that the optimizer remains relatively stable within a moderate learning-rate range. When the class-weight cap is increased to 2.0, contempt recall reaches 100.00%, while overall performance does not decline; this result indicates that bounded class weighting can be useful when its value is selected carefully.

Overall, the model does not respond equally to every hyperparameter change. Stable performance depends on a suitable regularization level and a reasonably stable mini-batch optimization process; therefore, small-sample FER can be understood as a problem of matching model capacity with regularization strength.

4.6. Error analysis

The confusion matrix of the main VGG-optimized model contains four incorrect predictions. Two anger samples are recognized as sadness, one contempt sample is recognized as fear, and one surprise sample is recognized as contempt. These errors are related to the visual similarity between several expressions. Anger and sadness both belong to negative expressions, and low-resolution grayscale images may show similar mouth shapes or tense facial structures. The confusion between contempt and fear also indicates that categories with fewer training samples remain difficult to distinguish

More obvious problems appear in the weaker baselines. ResNet-small fails to correctly identify sadness and obtains zero sadness precision. ROI-fusion VGG classifies some happy samples as disgust and some surprise samples as sadness; this result suggests that its multi-branch structure is difficult to train reliably with limited data. MobileNetV2 transfer learning also confuses anger with disgust or sadness, which indicates that the pretrained features do not fully match the characteristics of the target dataset.

DOI: 10.54254/2755-2721/2026.GL36462

These results explain the need for class-wise evaluation. High overall accuracy does not guarantee reliable recognition for every expression category; in practical human-centered applications, confusing different negative expressions may lead to an incorrect understanding of the user state and affect later decisions.

4.7. Case study

The qualitative case study considers three representative sample groups, and Grad-CAM visualization can be used to locate facial regions related to classification decisions [19]. The first group includes high-confidence correct predictions for happy, surprise, and disgust; these categories usually contain relatively clear visual patterns. The second group focuses on ambiguous negative expressions, especially anger samples recognized as sadness, where the mouth or eyebrow information may be unclear at a resolution of 48 × 48. The third group examines minority-class errors, such as contempt being recognized as fear; limited training samples may make the corresponding decision boundaries unstable.

For each case, the analysis should provide the input image, ground-truth label, predicted label, softmax probability vector, and Grad-CAM heatmap [19]. The heatmap can show whether the model focuses on meaningful regions, including the eyebrows, eyes, nasolabial folds, and mouth corners. If the highlighted areas mainly appear in the background or other unrelated regions, the error may result from unsuitable feature learning rather than data scarcity alone.

5. Conclusions

This paper addressed the problem of facial expression recognition under limited labeled data, where class imbalance, subtle inter-class differences, and model-data mismatch can undermine generalization. It proposed and evaluated a stability-oriented convolutional framework based on compact VGG-style learning, global average pooling, calibrated supervision, explicit regularization, and momentum-based optimization. The main finding is that small-sample FER does not reward architectural complexity monotonically: lightweight CNNs outperform or match more complex ROI-fusion, residual, and transfer-learning alternatives in the current setting.

The experiments validate this conclusion through comparative evaluation, ablation studies, sensitivity analysis, and error diagnosis. The VGG-optimized model achieves 97.30% accuracy and 95.64% macro-F1, while LeNet reaches 98.65% accuracy and 97.56% macro-F1. In contrast, ROI-fusion VGG and MobileNetV2 transfer learning drops to 56.08% and 83.78% accuracy, respectively. Ablations further show that label smoothing and L2 regularization are beneficial, while focal loss [18] and naive class weighting may overcorrect the boundary. Sensitivity results indicate that moderate regularization and stable batch-level optimization are critical for reliable performance.

Future work should extend this study in three directions. First, repeated-seed evaluation and fixed split files should be used to quantify statistical variance. Second, external FER datasets should be introduced to test whether the observed capacity-control principle generalizes beyond the current data distribution. Third, Grad-CAM-based case studies should be integrated to verify whether compact models rely on semantically meaningful facial regions rather than dataset artifacts.

References

[1] Gao, W., & Jin, H. (1997). Analysis and recognition of facial expression image. *Chinese Journal of Computers, 20*, 782–789.

DOI: 10.54254/2755-2721/2026.GL36462

[2] Feng, X., & Zhang, J. P. (2017). Facial microexpression recognition: A survey. *Acta Automatica Sinica, 43*(3), 333–348.
[3] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215.
[4] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
[5] Li, Y., Lin, X. Z., & Jiang, M. Y. (2018). Facial expression recognition with cross-connect LeNet-5 network. *Acta Automatica Sinica, 44*(1), 176–182.
[6] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
[8] Sun, X., Pan, D., & Ren, F. (2016). Facial expression recognition using ROI-KNN deep convolutional neural networks. *Automation Journal, 42*(6), 883–891.
[9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19).
[10] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520).
[11] Tan, M., & Le, Q. (2019, May). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105–6114). PMLR.
[12] Liu, F. Y., Wang, S. H., & Zhang, Y. D. (2018). Survey on deep belief network model and its applications. *Computer Engineering and Applications, 54*(1), 11–18.
[13] Akula, A., Budha, G., Bingi, G., Chanda, U., Borra, A. R., Yadav, D. B., & Saravanan, M. (2026). Emotion recognition from facial expressions using CNNs. *International Journal of Engineering & Extended Technologies Research (IJEETR), 8*(1), 120–125.
[14] Jiang, Y. (2025). Static facial expression recognition based on depth perception and visual large model feature guidance. *Sichuan University*. https://doi.org/10.27342/d.cnki.gscdu.2025.000099
[15] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.
[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).
[17] Yu, J. (2025). *Research on facial expression recognition algorithm based on data augmentation and improved DHVT.* Heilongjiang University. https://doi.org/10.27123/d.cnki.ghlju.2025.002063
[18] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980–2988).
[19] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).