A Facial Emotion Recognition Model Based on an Improved MobileViT

Zhang Aobo

***Abstract*— To address the limitations of conventional facial emotion recognition (FER) models — namely inadequate precision in modeling expression-related features, insufficient responsiveness to subtle micro-expression regions, and constrained computational resources for mobile deployment— this paper proposes a lightweight yet high-performance FER model termed LMFER (Lightweight MobileViT for Face Emotion Recognition). Built upon the MobileViT backbone, LMFER performs structural optimization by introducing an improved Coordinate Attention (CA) module and an Emotion Attention Mechanism (EAM) into the first two network stages, respectively. The CA module enhances salient-region extraction from both spatial and semantic perspectives via directional spatial encoding and the incorporation of a channel-attention component (i.e., the Channel Attention in CBAM), enabling joint modeling across spatial and channel dimensions. The EAM strengthens the responses to fine-grained, emotion-critical regions (e.g., eyebrows, eyes, and mouth corners) by generating multi-channel emotion maps and applying a Softmax-based normalization strategy. Moreover, EAM supports the integration of facial landmarks as semantic priors to guide attention toward expression-discriminative areas more precisely. In addition, Dropout regularization is introduced after the fully connected layer to mitigate overfitting, and training stability as well as generalization are further improved by adopting the AdamW optimizer and a dynamic learning-rate scheduling strategy. Experimental results demonstrate that LMFER achieves an accuracy of 95.08% on the RAF-DB dataset, outperforming MobileViT by 4.78%, 4.55%, and 4.67% in precision, recall, and F1-score, respectively, while maintaining only 2.1M parameters. LMFER also exhibits clear advantages over mainstream approaches such as ResNet50, MobileNetV2, ShuffleNetV2, GoogLeNetV1, and VGG-16. Overall, LMFER delivers notable improvements in recognition performance while retaining favorable model compactness and deployment adaptability, indicating strong potential for practical applications in human－computer interaction and affective computing.lease download TEMPLATE HELP FILE from the website.**

Keywords: facial emotion recognition; LMFER; MobileViT; emotion attention mechanism; coordinate attention mechanism

I. INTRODUCTION

Facial emotion recognition (FER) is a key technology in domains such as human－computer interaction, mental health assessment, and security surveillance. Its core objective is to capture subtle variations in human affect by modeling facial expression cues. Emotion recognition involves multiple research areas, including artificial intelligence, biological sciences, and psychology, and has become a prominent research branch within artificial intelligence[1]. In addition, emotion recognition technologies exhibit broad application prospects and can be extended to healthcare and tele-education, where they may assist in diagnosing mental health conditions, improving educational outcomes, and promoting intelligent development in related fields[2]. Traditional approaches typically rely on handcrafted feature extractors (e.g., LBP and HOG) combined with shallow classifiers (e.g., SVM and random forests). However, these methods suffer from limited representational capacity and high sensitivity to variations in illumination and head pose, making them inadequate for emotion-semantic understanding in complex real-world scenarios. With the advancement of deep learning, convolutional neural network (CNN)-based models (e.g., ResNet and MobileNet) have substantially improved recognition accuracy through end-to-end learning. Nevertheless, most such models contain a large number of parameters and are therefore difficult to deploy on mobile devices. Meanwhile, emotional cues are characterized by both local sensitivity and global semantic correlations. Conventional CNNs are limited in long-range dependency modeling and often entail excessive computational and parameter costs[3]. In contrast, pure Transformer architectures can capture global context but tend to incur high computational complexity and may lose fine-grained local details.

In recent years, the lightweight hybrid architecture MobileViT has demonstrated efficient performance in mobile vision tasks by integrating the local inductive bias of CNNs with the global modeling capability of Transformers. However, it still faces challenges when applied to FER: the attention focusing mechanism for emotion-relevant regions (e.g., eyebrows/eyes and mouth corners) has not been specifically optimized[4] and is thus susceptible to background noise; overfitting can easily occur during high-dimensional feature fusion, impairing generalization; and subtle changes in dynamic expressions require more robust nonlinear activation functions.[5] In response to these issues, a series of improvements have been explored in the literature. Howard et al. proposed a lightweight network for mobile and embedded devices, namely MobileNetV1, by replacing standard convolutions with depthwise separable convolutions, thereby reducing computation and improving efficiency while maintaining accuracy[6]. Liu et al. designed a global module to emphasize holistic facial characteristics and a local module to capture fine-grained regional features[7]. Sun proposed a lightweight image generation method based on MobileViT[8]. In addition, Cao proposed a

Manuscript received February 26, 2026

Zhang Aobo , Software Engineering, Dalian Jiaotong University, Dalian, China

multi-scale MobileViT model[9]; Li presented a lightweight deep learning solution emphasizing high accuracy, portable design, and low latency[10]. To better capture global and contextual information, Tao designed a multi-scale branch module to enlarge the receptive field and proposed an improved MobileViT-based approach[11].

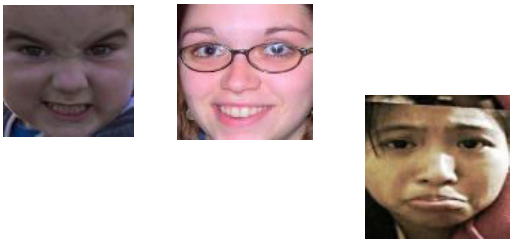
In this paper, we propose a lightweight FER model termed LMFER (Light MobileViT for Face Emotion Recognition). The main contributions are threefold. First, we design an Emotion Attention mechanism and an improved Coordinate Attention mechanism. By weighting input features with emotion maps, the proposed attention modules enhance the feature responses of expression-sensitive regions. Specifically, the Coordinate Attention mechanism performs adaptive pooling along the height and width dimensions of the input features, enabling the model to attend to different spatial dimensions independently and thus capture key information more precisely[12]. Second, we incorporate Dropout regularization in the fully connected layer and adopt a Warm-up + Cosine Annealing dynamic learning-rate schedule to effectively mitigate overfitting. Finally, we replace SiLU with the Swish adaptive activation function to strengthen nonlinear representational capacity. Experiments demonstrate that LMFER achieves an accuracy of 95.08% on the RAF dataset with only 2.1M parameters, providing an efficient solution for real-time affective computing on mobile devices. Compared with existing methods, LMFER significantly outperforms classical network architectures such as ResNet50, MobileNetV2, ShuffleNet v2, and ResNet in terms of precision, recall, and F1-score. In the future, the proposed model can be further optimized and deployed on mobile devices, offering new technical support for practical FER applications.

II. Materials and Methods

A. Dataset Acquisition

To ensure data correctness and reliability, this study employs a facial emotion recognition dataset derived from the publicly available RAF (Real-world Affective Faces Database). The RAF dataset was collected and curated by professional researchers and contains facial expression images captured in diverse real-world scenarios. Owing to its realism and breadth, it has been widely used in facial emotion recognition research. RAF includes emotion annotations for seven basic categories—anger, disgust, fear, happiness, sadness, surprise, and neutral—and comprises a total of 15,339 face images that have been uniformly resized to 100 × 100 pixels. The images exhibit substantial diversity in illumination conditions, head pose, age, and gender, thereby providing a representative characterization of the complexity of emotional expression in real-world settings.

In this study, we construct the experimental dataset by selecting six typical emotion categories (anger, disgust, fear, happiness, sadness, and surprise) together with the neutral expression. Representative samples from each category are shown in Fig. 1. All images in RAF are rigorously annotated: the emotion label of each image is independently assessed by multiple annotators and finalized through consensus, ensuring high label accuracy and reliability. In addition, the dataset provides facial landmark annotations, which facilitate preprocessing procedures such as face alignment and feature extraction.



0.Anger 1.Happiness 2.Sadness

Figure 1. Examples of Facial Expression Categories

B. Indicator Augmentation

To address challenges in facial emotion recognition—such as large intra-class variability, imbalanced sample distributions, and the diversity of illumination and head poses in real-world scenarios—this study adopts a set of data augmentation strategies to improve the model′s generalization capability. Specifically, preprocessing operations including brightness adjustment, Gaussian noise injection, and horizontal flipping are applied, expanding the dataset to 20,699 images. The augmented dataset is then split into training, validation, and test sets with proportions of 70%, 20%, and 10%, respectively. Detailed statistics of the sample distribution across categories are provided in Table 1.

Table 1. Number of Images per Category Before and After Augmentation

Category	Images	Images
	(Original)	(Augmented)
Anger	867	1350
Happiness	5957	8204
Neutral	3204	4950
Sadness	2460	3801
Surprise	1619	2501
Fear	355	1548
Disgust	877	1355
Total	15339	23709

III. Main Modules

A. MobileViT Module

MobileViT is a hybrid architecture that integrates convolutional neural networks (CNNs) with Transformer techniques[13]. It effectively combines the spatial inductive bias of CNNs with the advantage of Vision Transformers in global feature modeling[14], thereby preserving model compactness while substantially improving classification efficiency[15]. The core of MobileViT consists of an MV2 module and a MobileViT module, which are connected in a cascaded manner. The MV2 module is designed based on the MobileNetV2 block. The overall architecture of MobileViT is illustrated in Fig. 2.

The MV2 module integrates depthwise separable convolution, inverted residual structure, and linear bottleneck design, which jointly reduce the parameter scale while enabling efficient extraction of local features. Specifically, the module first employs a 1×1 pointwise convolution to expand the feature dimensionality, enhancing fine-grained representations. It then applies a 3×3 depthwise convolution to extract per-channel features with reduced parameter cost. Finally, another 1×1 pointwise convolution is used to project features back to a lower dimensional space, thereby reducing the output size. To avoid information loss associated with low-dimensional representations, a linear activation function is adopted after the final 1×1 pointwise convolution. In addition, the residual connection in MV2 is enabled only when the stride is 1 and the input and output dimensions are consistent, preventing feature degradation. When the stride is 2, a sequential connection is used to perform downsampling across feature stages.

As the key innovation of the model, the MobileViT module first captures local spatial information via a 3×3 convolution layer, and then employs a 1×1 convolution layer to expand the number of feature channels and learn linear combinations across channels. Next, the features are mapped into a d-dimensional embedding space, and the feature map is partitioned into N non-overlapping image patches to form a sequence. The sequence is processed by a Transformer encoder, which models global relationships among patches and enables attention to global contextual information. To mitigate potential information loss across patches, the encoded patch representations are reshaped back to the original spatial layout, producing the reconstructed feature map. Finally, a fusion module integrates local and global features to generate the final output feature Y.

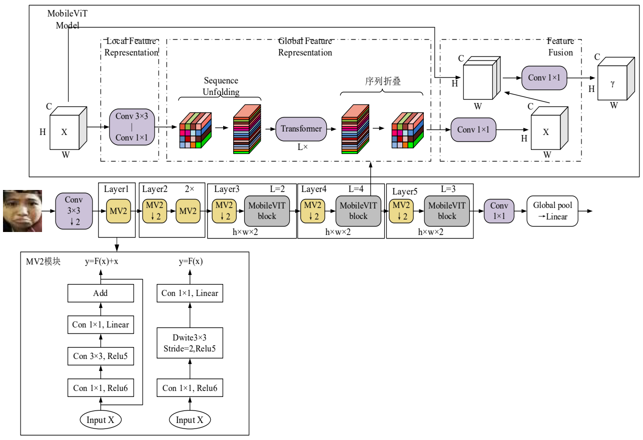


Figure 2. MobileViT Model Architecture with Key Components

B. Construction of the LMFER Network Model

To improve the accuracy of facial emotion recognition while meeting the requirements of mobile deployment, this study proposes LMFER (Light MobileViT for Face Emotion Recognition) by introducing several innovations on top of the MobileViT architecture. First, an Emotion Attention (EmotionAttention) mechanism is embedded into the input feature extraction stage of the original MobileViT network. Specifically, EmotionAttention performs spatial reweighting of the input features using emotion maps, thereby strengthening feature responses in expression-sensitive regions such as the eyebrows/eyes and mouth corners. This design enables more effective capture of local expression details and their global semantic correlations. Moreover, coordinate attention modules and emotion attention modules are appended after both Layer1 and Layer2 of the original MobileViT. By applying adaptive pooling along the height and width dimensions, the model can attend to salient cues along different spatial axes independently, which facilitates more effective extraction of informative patterns and improves the discrimination of visually similar emotional characteristics, ultimately enhancing classification accuracy. Second, to improve global feature modeling while reducing the computational overhead introduced by multi-head attention, LMFER replaces the multi-head self-attention module in the original Transformer with PoolFormer, a pooling-based token-mixing structure. PoolFormer encodes global semantic information across spatial dimensions via parameter-free pooling operations, which not only strengthens the model′ s ability to capture holistic structural variations associated with emotions but also significantly reduces resource consumption during inference, making it well suited for real-time facial expression recognition.

Finally, a Dropout layer is introduced after the fully connected classification layer, and Label Smoothing is adopted to mitigate overfitting and improve tolerance to ambiguous emotion labels. In addition, the AdamW optimizer combined with the Lookahead strategy is employed to enable more robust gradient updates during training. A dynamic learning-rate scheduling scheme (Warm-up + Cosine Annealing) is further applied to accelerate convergence and enhance generalization performance, thereby improving the robustness and stability of the model under cross-dataset evaluation. The detailed architecture of LMFER is illustrated in Fig. 3.

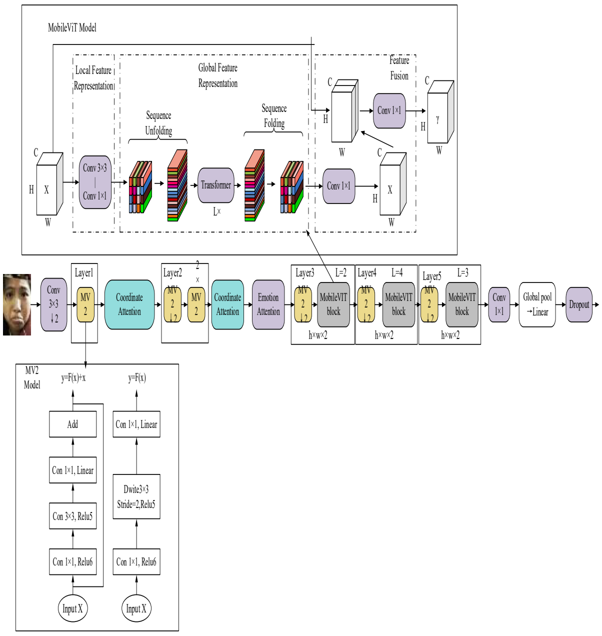


Figure 3. Architecture of the LMFER Model

C. *Incorporating Improved Coordinate Attention and an Emotion Attention Mechanism*

To enhance the sensitivity of facial expression recognition models to micro-expression regions and to strengthen feature representation, this paper introduces and further improves both the Coordinate Attention (CA) mechanism and the Emotion Attention Mechanism (EAM) within the baseline network.

Coordinate Attention (CA) is a lightweight attention design that models spatial dependencies separately along the horizontal and vertical directions, thereby guiding the network to focus on key regions of the input image. However, the original formulation does not explicitly account for semantic interactions across feature channels. To address this limitation, we incorporate a channel-attention module into CA—specifically, the Channel Attention component from CBAM—resulting in a joint modeling scheme over spatial and channel dimensions. This enhancement strengthens the network's responsiveness to semantically informative channels and improves the representation of emotion-relevant cues.

The Emotion Attention Mechanism (EAM) is a spatial attention module tailored for facial emotion recognition. It aims to improve recognition accuracy by generating an emotion map (Emotion Map) to reweight input features, thereby emphasizing expression-critical regions such as the eyebrows/eyes and mouth corners. Conventional approaches typically generate an emotion map using a single-channel Sigmoid function, which provides limited expressive capacity for modeling fine-grained facial affect patterns. In this work, we propose a multi-channel emotion attention mechanism: a lightweight subnetwork produces K spatial attention maps, and a Softmax normalization is applied across the channel dimension to obtain a finer-grained spatial reweighting scheme. Each channel-specific emotion map is then multiplied element-wise with the input feature map. The resulting features are subsequently fused via concatenation followed by a 1 × 1 convolution, yielding an enhanced representation. Moreover, to strengthen the semantic guidance for emotion-map generation, the proposed mechanism can further integrate facial landmark information as a prior, directing attention toward highly discriminative local facial regions. This design suppresses irrelevant background interference and improves both sensitivity and robustness to subtle expression variations.

D. *Activation Function*

In the original MobileViT network, the MV2 module employs the SiLU activation function. In comparison, the Swish activation function[15] offers several advantages, including a self-adjusting threshold, smoother gradient propagation, and improved suitability for capturing subtle variations in facial expressions. By introducing a learnable parameter, Swish can dynamically adjust the effective activation threshold, enabling adaptive modeling of nonlinear relationships in expression features. This property is particularly beneficial for recognizing weak expression changes in regions such as the eyes and mouth corners. Therefore, to further enhance the model′ s representational capacity for complex facial expressions, Swish is adopted to replace SiLU in the original network. The Swish function is defined as:

$$\mathrm{Swish}(x) = x \cdot \sigma(\beta x) \quad (1)$$

Where σ denotes the Sigmoid function and β is a learnable parameter updated via backpropagation, which adaptively modulates the effective activation threshold.

E. *Optimizer*

In conventional deep learning training pipelines, fixed learning rates and simple weight-decay schemes are commonly adopted. Although effective to some extent, these settings often struggle to balance convergence speed and generalization performance. Consequently, an increasing body of work has investigated dynamic learning-rate scheduling to alleviate this trade-off. Among these approaches, the combination of Warm-up + Cosine Annealing represents an advanced learning-rate scheduling strategy that, when paired with the AdamW optimizer, can more effectively improve training stability and final performance.

Specifically, the Warm-up phase refers to the practice of starting training with a relatively small learning rate and gradually increasing it to a predefined maximum. This strategy helps mitigate gradient instability at the early stage of optimization and reduces the risk of gradient explosion that may arise from overly large initial learning rates, thereby enabling a smoother and more reliable training start. As training proceeds, the Cosine Annealing schedule progressively decreases the learning rate following a cosine-shaped trajectory, yielding a smooth decay in the later stages. This quasi-periodic learning-rate evolution can help the optimizer avoid poor local minima, improve exploration in the vicinity of the global optimum, and provide a more stable convergence path during fine-tuning.

AdamW (Adaptive Moment Estimation with decoupled weight decay) is a widely used optimizer in modern deep learning. It combines adaptive step sizes with effective regularization and addresses a known limitation of the original Adam optimizer by decoupling weight decay from the gradient-based parameter update, thereby preventing the unintended weakening of regularization. Through first- and second-order moment estimates, AdamW adaptively adjusts the learning rate for each parameter, offering flexibility when optimizing features at different scales—an advantage that is particularly beneficial for facial emotion recognition tasks with complex parameter distributions. Moreover, weight decay, as a regularization technique, penalizes large parameter magnitudes to control model complexity, suppress overfitting, and improve generalization to unseen samples. By explicitly separating weight decay from the gradient update, AdamW makes the regularization effect more transparent and typically more effective, enhancing training stability and further reducing the risk of overfitting.

IV. Experiments and Analysis

A. Experimental Environment

All experiments were conducted on a workstation running Windows 10, equipped with an Intel® Core™ i7-14700 CPU, an NVIDIA GeForce RTX 3090 (24 GB) GPU, and 32 GB of RAM. Model implementation and training were performed using Python 3.8 with the PyTorch 2.1.0 deep learning framework and CUDA 12.1. The input images were resized to a resolution of 256 × 256. The model was trained using the AdamW optimizer with a learning rate of 1×10^{-4}, a batch size of 64, and 80 epochs.

B. Evaluation Metrics

The evaluation metrics used in this study include Accuracy, Precision, Recall, and the comprehensive metric F1-score.

Model	Accuracy / %	Precision/ %	Recall/%	F1-Score/ %
MobileViT	90.34	90.20	90.50	90.35
MobileViT+Dropout	91.35	91.30	91.40	91.29
MobileViT+CA	92.61	92.55	92.65	92.59
MobileViT+EAM	92.11	92.05	92.15	92.09
LMFER	95.08	94.98	95.05	95.02

Accuracy measures the proportion of correctly predicted samples among all samples. Precision reflects the proportion of truly positive samples among those predicted as positive by the model. Recall denotes the proportion of correctly predicted positive samples among all actual positive samples. The F1-score is the harmonic mean of precision and recall, providing a balanced assessment by jointly considering both metrics. The computation of these metrics is given as follows:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (2)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (3)$$

Model	Accuracy/%	Precision/ %	Recall/%	F1-Score/ %	Model Size /MB
ResNet50	80.34	80.20	80.50	80.40	98.06
VGG-16	81.35	81.30	81.40	81.36	528.14
MobileNetV2	82.61	82.55	82.65	82.60	14.67
GooLeNet v1	82.11	82.05	82.15	82.10	27.14
ShuffleNet v2	84.02	83.98	84.05	84.03	7.98
MobileViT_xs	90.34	90.20	90.50	90.35	2.3
LMFER	95.08	94.98	95.05	95.02	2.1

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (4)$$

$$\mathrm{F1-score} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision+Recall}} = \frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}} \quad (5)$$

TP denotes the number of samples that are actually positive and predicted as positive; FP denotes the number of samples that are actually negative but predicted as positive; FN denotes the number of samples that are actually positive but predicted as negative; and TN denotes the number of samples that are actually negative and predicted as negative.

C. Ablation Study

In the ablation study, to further validate the effectiveness of the proposed network improvements, we evaluate each modification separately and compare the corresponding results. As reported in Table 2, introducing the CA module, the emotion attention mechanism, or the Dropout module individually into the MobileViT baseline leads to consistent improvements across all four evaluation metrics. The best performance is achieved by LMFER, which integrates all three modules simultaneously: the accuracy increases from 90.34% to 95.08%, precision from 90.20% to 94.98%, recall from 90.50% to 95.05%, and F1-score from 90.35% to 95.02%. These results indicate that the CA module, the emotion attention mechanism, and Dropout exhibit a strong synergistic effect in enhancing model performance, thereby corroborating the effectiveness of the proposed improvements.

Table 1. Results of Ablation Experiments

D. Comparative Analysis of LMFER Model Performance

To more comprehensively verify the performance of the improved model, comparative experiments were conducted against ResNet50, VGG-16, MobileNetV2, GoogLeNet v1, ShuffleNet v2, and MobileViT_xxs. All methods were evaluated under the same experimental setting, and comparisons were made in terms of accuracy, precision, recall, F1-score, and model size. The results of the proposed network relative to the competing models are summarized in Table 2.As shown, the improved LMFER achieves an average accuracy of 95.08%, substantially outperforming the other models. Compared with the baseline MobileViT, LMFER improves accuracy by 4.74 percentage points. Moreover, relative to the baseline, LMFER increases precision, recall, and F1-score by 4.78, 4.55, and 4.67

percentage points, respectively, also exceeding the performance of the other compared methods. In addition, LMFER requires markedly fewer parameters than the competing models, demonstrating a favorable trade-off between recognition performance and model compactness.

Table 2. Comparison of Experimental Results Across Different Models

As illustrated in Fig. 4, the accuracy curves of all models across training epochs exhibit an initial upward trend followed by convergence to a relatively stable level. Notably, the LMFER model achieves a comparatively high recognition accuracy of 95.02% and demonstrates the most stable trajectory during training, indicating a clear advantage for facial emotion recognition. The performance of the LMFER model on the facial emotion recognition task is visually demonstrated by the confusion matrix in Fig. 5. Overall, the model achieves high recognition accuracy for most emotion categories. In particular, it performs exceptionally well on “Anger” and “Happiness”, with accuracies reaching 98.5% and 97.8%, respectively. These emotions typically exhibit salient facial manifestations, enabling the model to capture and discriminate their characteristic cues more reliably.In contrast, the recognition accuracies for “Neutral” and “Fear” are relatively lower, at 92.3% and 91.5%, respectively. This may be attributable to the comparatively ambiguous expression patterns associated with these emotions in visual data. Neutral expressions are often subtle and less distinctive than anger or happiness, while fear can partially overlap with surprise in terms of facial appearance, leading to misclassifications. The primary confusion between certain categories is observed between “Neutral” and “Sadness”, suggesting that these two emotions may share similar facial characteristics and remain challenging to distinguish.

In summary, although a small number of misclassifications persist, LMFER maintains strong performance across the majority of emotion categories, demonstrating its effectiveness and stability for facial emotion recognition.

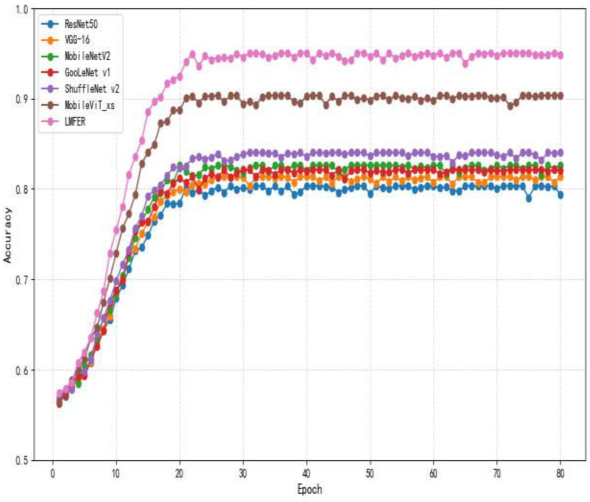


Figure 4. Recognition Accuracy of Different Models

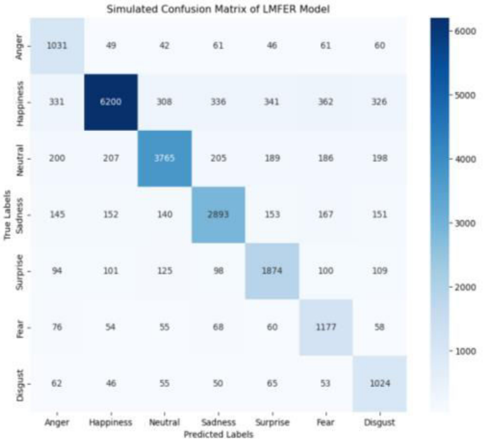


Figure 5. Confusion Matrix

V. CONCLUSION

To address the limitations of conventional facial emotion recognition models—particularly their insufficient sensitivity to micro-expression regions and limited discriminative accuracy — this paper proposes a lightweight yet high-performance facial emotion recognition model, LMFER (Lightweight MobileViT for Face Emotion Recognition). Built upon MobileViT as the backbone, LMFER integrates improved attention mechanisms into the network architecture to enhance both the perception and representation of expression-critical regions.

Specifically, LMFER embeds an improved Coordinate Attention module after the first network stage. By combining directional spatial encoding with a channel-attention branch (i.e., the Channel Attention component in CBAM), the model strengthens its responsiveness to salient regions from both spatial and semantic perspectives. After the second stage, an Emotion Attention Mechanism is further introduced. This module generates multi-channel spatial attention maps and applies Softmax normalization to achieve fine-grained reweighting of expression-sensitive areas (e.g., eyebrows/eyes and mouth corners). Moreover, it supports the incorporation of facial landmark information to improve the semantic validity of attention guidance, thereby further enhancing recognition accuracy and model robustness. To mitigate overfitting and improve generalization, LMFER introduces Dropout after the fully connected layer and adopts the AdamW optimizer with weight decay to effectively control model complexity at the optimization level. Experimental results demonstrate that LMFER consistently outperforms the original MobileViT across multiple metrics, improving accuracy, precision, recall, and F1-score by 4.74%, 4.78%, 4.55%, and 4.67%, respectively, achieving final values of 95.08%, 94.98%, 95.05%, and 95.02%.

In summary, LMFER not only achieves a substantial improvement in emotion recognition accuracy, but also exhibits comprehensive advantages in model compactness, generalization capability, and expression-region modeling, indicating strong potential for practical deployment. Future work will further explore the following directions: (1)

expanding facial datasets under complex real-world conditions to improve robustness; (2) incorporating more diverse attention modules to investigate adaptability across different task scenarios; and (3) studying LMFER′s practical performance in on-device deployment and cross-domain transfer.

ACKNOWLEDGMENT

The author, Aobo Zhang, thanks his supervisor, Professor L. Wang, and the laboratory members for their guidance and assistance throughout this research. He is also grateful to his family for their unwavering support.

REFERENCES

[1] Xue Peiyun, Dai Shutao, Bai Jing, et al. Bimodal Emotion Recognition Based on Speech and Facial Images. Journal of Electronics & Information Technology, 2024, 47: 1-11
[2] Zhang Xuejun, Wang Tianchen, Wang Zetian. A Convolutional Transformer EEG Emotion Recognition Model Based on Multi-domain Information Fusion. Journal of Data Acquisition & Processing/Shu Ju Cai Ji Yu Chu Li, 2024, 39(6).
[3] Zhang Bushi, Fan Hong. Improved MobileViT Algorithm for Few-shot Learning. Journal of Computer Engineering & Applications, 2024, 60(22).
[4] Kang Bo, Qian Yi, Wen Yimin. Image Emotion Recognition Based on Abstract Relational Scene Graphs. CAAI Transactions on Intelligent Systems, 2023, 19(2): 335-343.
[5] Zhang Bo, Wu Yufan. Micro-expression Recognition Algorithm Based on a Dual-branch Lightweight Network. Laser & Optoelectronics Progress, 2024, 61(14): 1437001-1437001-10.
[6] Liu Juan, Wang Ying, Hu Min, et al. Facial Expression Recognition Network Integrating Global Enhancement and Local Attentive Features. Journal of Frontiers of Computer Science & Technology, 2024, 18(9).
[7] Liu Y, Sun W. A lightweight diffusion model for image generation based on improved MobileViT[C]//Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition. 2024: 67-73.
[8] Cao K, Tao H, Wang Z, et al. MSM-ViT: A multi-scale MobileViT for pulmonary nodule classification using CT images[J]. Journal of X-Ray Science and Technology, 2023, 31(4): 731-744.
[9] Li X, Feng J, Cai J, et al. Fv-mvit: Mobile vision transformer for finger vein recognition[J]. Sensors, 2024, 24(4): 1331.
[10] Tao X, Wei L, Tang S, et al. Image Recognition of Alzheimer′s disease Based on Improved MobileVit Network[C]//2024 43rd Chinese Control Conference (CCC). IEEE, 2024: 8002-8007.
[11] Tang, S., & Wei, L. (2024, April). Micro-expression Recognition Based on Improved MobileViT. In 2024 4th International Conference on Computer, Control and Robotics (ICCCR) (pp. 61-65). IEEE.
[12] Sun, J., Zhang, F., Liu, H., & Hou, W. (2024). Research on Improved MobileViT Image Tamper Localization Model. Computers, Materials & Continua, 80(2).
[13] Zhang, Y., & Zhan, Q. (2024). Welding defects recognition based on DCP-MobileViT
[14] Wang B, Yang H, Zhang S, et al. Identification of Multiple Diseases in Apple Leaf Based on Optimized Lightweight Convolutional Neural Network[J]. Plants, 2024, 13(11): 1535.

Aobo Zhang Aobo Zhang was born in Dalian, Liaoning Province, China, in 2001. He received his B.E. degree in Software Engineering from Dalian Jiaotong University in 2023, where he was recognized for his academic performance. He is currently pursuing his M.E. degree in Software Engineering at the same university, with an expected graduation in 2026. His research interests include machine learning, large language models, multimodal AI, and natural language processing.During his master's studies, Zhang has gained extensive research experience through multiple industry internships and projects. As an AI intern at Neusoft Education Technology Group (2025), he participated in the implementation of Retrieval-Augmented Generation (RAG) technology, building knowledge bases with FAISS indexing and fine-tuning Qwen and ChatGLM3 models using LoRA and P-Tuning v2 techniques, achieving significant improvements in accuracy and response speed. He also served as a Multimodal LLM Developer at Shenghe Technology (2024), where he worked on deploying and optimizing the Donut model for medical receipt recognition and information extraction, generating structured JSON data.Zhang has published a first-author paper titled "Prediction of Shared Bicycle Entry and Exit Flow based on LSTM Model" (2024). He has also completed several notable projects, including a web-based intelligent Q&A system for education using RAG technology, and a sentiment analysis system for public opinion monitoring. His technical expertise spans Python, PyTorch, LangChain, and various LLM fine-tuning techniques.

His academic excellence has been recognized with multiple awards, including the Second Prize in the Yongchuang Cup National College Creative Competition (2023), a Third-Class Scholarship (2023), and the National Computer Rank Examination Level 3 Certificate (2025). Beyond academics, Zhang has demonstrated leadership and teamwork as a member of Dalian Jiaotong University's men's basketball team for five years, competing in multiple university league championships and winning third place in the 3×3 Gold League Dalian Station (2019). He also served in the Academic Department of the School of Software, organizing forums on AI development.