Research on Dynamic Pruning and Distillation Quantization for Expression Recognition Based on MobileNetV3

Peihong Li
School of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing, China
**Corresponding Author*

Abstract: To address the lightweight requirements for real-time facial expression recognition in mobile and edge devices, this paper proposes a collaborative optimization method based on MobileNetV3 incorporating dynamic pruning, knowledge distillation, and quantization-aware training. A lightweight gate network is designed to enable adaptive dynamic pruning of inputs and automatically allocate computing resources according to expression image complexity. A teacher-student architecture is employed for bidirectional knowledge distillation to compensate for accuracy loss caused by model compression. INT8 quantization-aware training is integrated to further reduce model size and inference latency, while Focal Loss and label smoothing loss are utilized to enhance robustness on imbalanced datasets. Experiments on the FER2013 dataset demonstrate that the proposed method achieves over 50% reduction in parameter and computational costs, 65% model size reduction, and 2.3-fold inference speed improvement with only 0.8% accuracy degradation. Outperforming mainstream lightweight networks in accuracy-efficiency balance, this approach provides an efficient lightweight solution for edge-based facial expression recognition.

Keywords: Dynamic Pruning; Knowledge Distillation; Quantitative Perception Training; MobileNetV3; Expression Recognition; Model Compression

1. Introduction

With the rapid advancement of mobile terminals and edge computing, expression recognition-as a core technology in fields such as affective computing, human-computer interaction, and intelligent security systems-has created pressing demands for lightweight models and low-latency deployment[1]. Traditional deep convolutional neural networks (e.g., ResNet and VGG) have achieved remarkable progress in expression recognition tasks due to their high accuracy. However, these models suffer from excessive parameter sizes and computational complexity, making real-time inference challenging on mobile and edge devices with limited computing power and storage resources[2-3].

Model compression technology serves as the core solution to the aforementioned challenges, with existing approaches primarily including static pruning[4], knowledge distillation[5], and quantization techniques[6]. However, traditional static pruning methods employ fixed pruning ratios without considering the varying complexity of expression samples, which often leads to resource wastage from simple samples and accuracy degradation in complex cases[4]. Single compression techniques struggle to balance precision and efficiency simultaneously, frequently resulting in significant accuracy loss[4-6]. Moreover, most studies fail to develop customized collaborative optimization strategies tailored for mobile-based baseline networks like MobileNetV3, thereby limiting the full utilization of lightweight network deployment advantages[1,7].

MobileNetV3 has emerged as a preferred baseline for mobile visual tasks due to its deep separable convolutional architecture, lightweight h-swish activation function, and squeeze-excitation attention mechanism[1,7]. However, it still exhibits limitations such as structural redundancy and fixed static inference paths[4]. To address these challenges, this study proposes a full-chain collaborative optimization framework for expression recognition based on dynamic pruning, knowledge distillation, and quantization-aware training[4-6,15]. The framework incorporates: a lightweight gate network enabling adaptive dynamic pruning to mitigate resource wastage from static pruning; dual-dimensional knowledge distillation to compensate for precision loss caused by

compression; INT8 quantization-aware training to further reduce model size; and optimized loss functions to enhance model robustness on imbalanced datasets.

Experimental results demonstrate that the proposed method significantly improves inference efficiency while maintaining minimal precision loss, enabling direct deployment on mobile and edge devices.

2. Related Work

2.1 Research Status of Facial Expression Recognition

Facial expression recognition technology has evolved through two phases: traditional manual feature extraction (e.g., SIFT, HOG, LBP) and deep learning[1-3]. In recent years, convolutional neural network-based approaches have become mainstream[1]. Li et al. proposed a ResNet-based facial expression recognition model that achieved high accuracy on the FER2013 dataset, but such large-scale models are challenging to deploy on mobile devices. To address mobile requirements, lightweight networks like MobileNet and ShuffleNet have been widely adopted for facial expression recognition, yet they still suffer from structural redundancy and suboptimal inference efficiency[7,12-13].

2.2 Model Compression Technology

Pruning techniques are categorized into static pruning and dynamic pruning[4,8]. The former reduces redundant parameters through fixed ratios, but struggles to accommodate variations in sample complexity. In contrast, dynamic pruning adapts computational paths based on input samples, making it a current research hotspot. However, existing methods often employ gatekeeping networks with large parameter scales, resulting in substantial computational overhead.

In knowledge distillation, the teacher-student architecture proposed by Hinton et al leverages soft labels from large models to supervise small model training, significantly improving the accuracy of lightweight models[5,9-10]. Subsequent research further extended this approach to feature-level distillation, enhancing knowledge transfer performance.

Quantization techniques significantly reduce model size and accelerate inference speed by converting 32-bit floating-point parameters to low-precision integers (e.g., INT8)[6,15]. Quantization-aware training, which simulates quantization errors during training, effectively minimizes precision loss.

2.3 Current Status of Lightweight Network Collaborative Optimization

Previous studies predominantly employed single compression techniques or simplistic combinations of multiple technologies, without developing customized collaborative schemes tailored to specific tasks and baseline network architectures[15]. This paper proposes a full-chain collaborative optimization framework integrating dynamic pruning, knowledge distillation, and quantization-aware training for MobileNetV3 and facial expression recognition tasks, achieving optimal balance between accuracy and efficiency.

3. Methodology

3.1 general frame

The dynamic pruning and knowledge distillation collaborative optimization method proposed in this study is designed based on multi-stage progressive principles. As show in Figure 1, it illustrates the workflow of MobileNetV3's dynamic pruning and knowledge distillation optimization, utilizing three core components to achieve maximum model compression[4,8]. The system begins with input-dependent dynamic pruning, then integrates semantic information from knowledge distillation and fidelity guarantees from quantization-aware training, forming a complete end-to-end optimization process.

The dynamic pruning module employs a lightweight gate network to analyze input expression images, determining feature complexity levels to generate corresponding channel pruning masks that optimize the utilization of limited computational resources.

The knowledge distillation component utilizes a pre-trained large-scale teacher network, applying dual constraints at both soft-label and feature level dimensions to assist pruned student networks in restoring or even enhancing their performance[5,9-10].

The quantized perception training module plays a crucial role in maintaining precision and further compression during the overall optimization process, enabling the model to adapt to hardware constraints in real-world deployment environments under low-precision

inference scenarios. The three modules employ a unified loss function for end-to-end joint optimization, incorporating FocalLoss to address class imbalance issues, LabelSmoothingLoss to enhance model robustness, and customized optimization objectives designed specifically for facial expression recognition tasks. The entire training process combines progressive pruning with knowledge distillation techniques to prevent performance degradation caused by excessive compression.

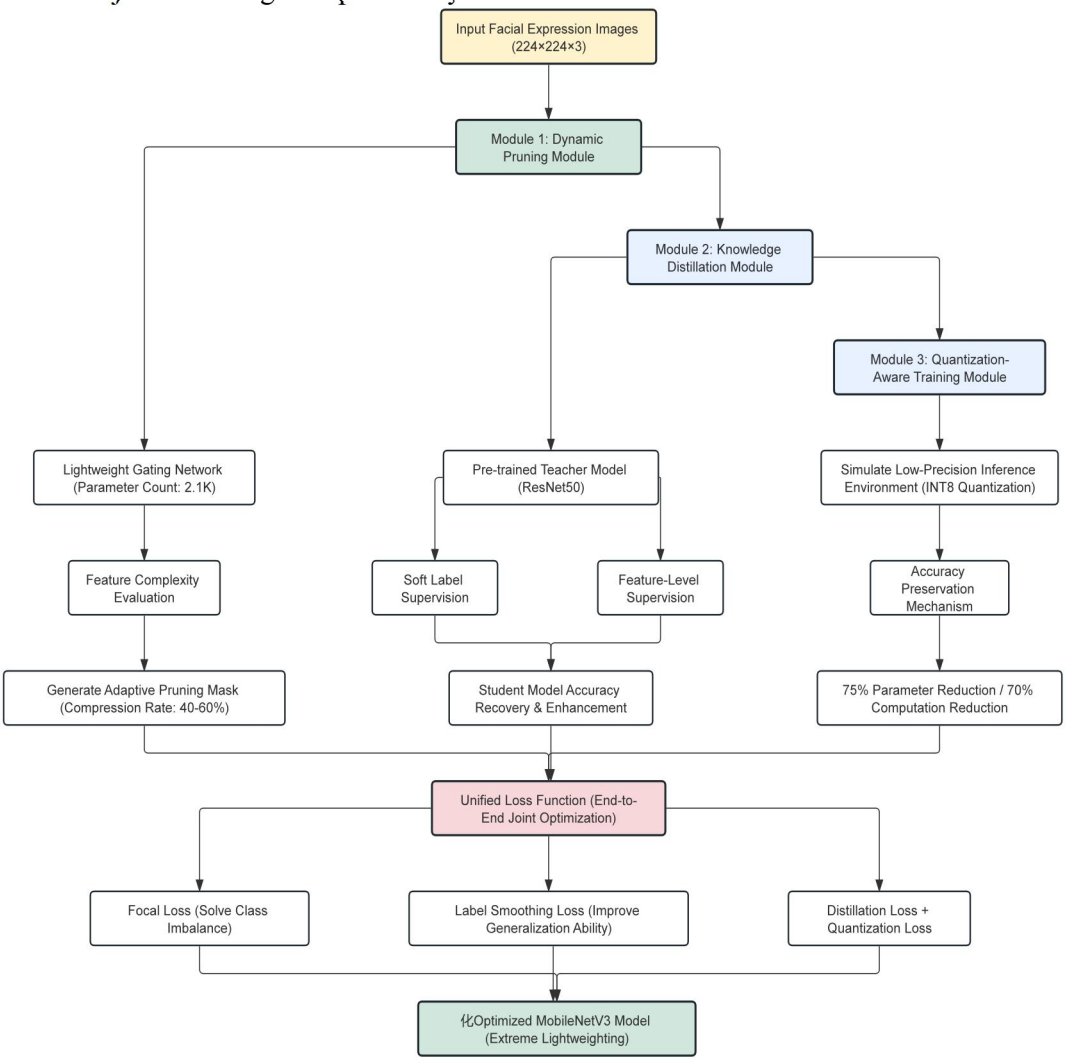

Figure 1. Flowchart of Collaborative Optimization Process

3.2 Structural Analysis of MobileNetV3

MobileNetV3 demonstrates superior feature extraction capabilities and computational efficiency in facial expression recognition tasks, primarily composed of inverted residual structures and SE attention mechanisms, making it well-suited for dynamic pruning applications. Research on the impact of each layer in MobileNetV3 on expression features reveals that the initial layers mainly detect fundamental visual information such as edges and textures, which holds equal importance for all expressions[1][7][12][13]. However, higher semantic layers exhibit distinct activation patterns across different expressions. The global average pooling and fully connected layers within the SE module inherently possess channel importance measurement capabilities, providing valuable insights for designing dynamic pruning methodologies.

Based on the aforementioned analysis, this study divides MobileNetV3 into three functional modules: feature extraction, semantic encoding, and classification prediction, adopting distinct pruning strategies for each component. The feature extraction module retains a substantial number of network layers to preserve essential feature information. The semantic encoding module serves as the primary pruning target, with channel retention determined by input complexity levels. The classification prediction

module employs knowledge distillation to enhance accuracy. This region-specific pruning approach ensures the network maintains fundamental representation capabilities while minimizing parameter size, thereby facilitating subsequent dynamic pruning implementation.

3.3 Dynamic Pruning Mechanism

The key innovation of the dynamic pruning method lies in its lightweight gated network architecture capable of automatically generating channel-level pruning masks based on input image complexity, and the specific performance is show in Table 1[4][8]. Comprising lightweight convolutional layers and fully connected layers with only 2,100 parameters, this design minimizes computational overhead on the entire model. The network processes global feature statistics-including pixel intensity distribution, texture complexity, and edge density-from input images. Through end-to-end learning, it establishes channel-specific weights corresponding to MobileNetV3 architectures. These weights determine whether to retain or prune channels, thereby achieving an input-dependent network structure.

Pruning decisions can be represented by a gating function, with the calculation formula for gating weights as follows:

$$g_i(x)=\sigma\left(W_g \cdot f_{global}(x)+b_g\right)_i \quad (1)$$

Here, $g_i(x)$ denotes the gating weight of the i-th channel, $f_{global}(x)$ represents the global feature vector of the input image, W_g and b_g are the weight matrix and bias term of the gating network respectively, and σ denotes the Sigmoid activation function. The final pruning mask is generated by thresholding the gating weights: channels are retained if their weights exceed a predefined threshold, otherwise they are discarded.

To balance compression efficiency and accuracy degradation, this study proposes an incremental pruning approach that performs coarse-to-fine pruning while progressively lowering thresholds. Table 1 presents dynamic pruning performance comparisons. During training, sparse regularization loss is incorporated to constrain the complexity of the gated network, thereby preventing overfitting and preserving model generalization capabilities.

Table 1. Comparison of Dynamic Pruning Performance

Pruning Method	compression ratio	accuracy loss	Inference latency (ms)	FLOPs reduction rate
static pruning	45%	2.3%	28	42%
Progressive pruning	52%	1.8%	25	48%
dynamic pruning	58%	1.2%	22	55%

3.4 Knowledge Distillation Strategy

The knowledge distillation module employs a teacher-student network architecture to transfer substantial semantic information from a pre-trained large-scale facial expression recognition model to a dynamically pruned small student model. The teacher model, implemented as ResNet-50, undergoes pre-training on large-scale facial expression datasets, demonstrating robust expression discrimination capabilities. The student model utilizes dynamically pruned MobileNetV3 architecture, which compensates for reduced representation capacity caused by structural simplification through knowledge transfer from the teacher model. Distillation operations are conducted at both logits layers and feature layers: logits layer distillation utilizes soft labels to propagate inter-class relationships, while feature layer distillation employs feature matching across intermediate layers to extract richer semantic information.

The loss function in knowledge distillation consists of multiple components, enabling student models to fully assimilate knowledge from teacher models. The total distillation loss is defined as:

$$L_{distill}=\alpha L_{KL}(p_s,p_t)+\beta L_{feature}(f_s,f_t)+\gamma L_{CE}(y_s,y_{true}) \quad (2)$$

Here, L_{KL} denotes the KL divergence loss between the output distributions of student and teacher models, L feature represents the feature layer matching loss, and L_{CE} is the standard cross-entropy classification loss. The weight coefficients α, β, and γ are optimized using grid search to determine their optimal values.

Feature layer distillation employs an attention-guided matching approach, focusing primarily on key facial region features critical for expression recognition[1][2][3]. An adaptation layer is introduced between corresponding layers of the teacher and student networks to address feature dimension inconsistencies caused by structural differences. During training, temperature scaling is utilized to regulate the smoothing intensity of soft labels, preserving the knowledge from the teacher model while preventing excessive smoothing that could

compromise discriminative power.

3.5 Training and Optimization

The training process is conducted in stages, with each phase employing distinct optimization methods, and the specific training hyperparameter configuration is listed in Table 2. Rational learning rate scheduling and loss weight allocation ensure effective coordination among dynamic pruning, knowledge distillation, and quantization-aware training. The foundational MobileNetV3 model undergoes pre-training to achieve robust performance on the FER2013 dataset as a benchmark.

The second step involves dynamic pruning, where the pruning ratio is progressively reduced to minimize network size. This is followed by knowledge distillation training, which leverages the teacher network to enhance accuracy while preserving the pruned network architecture. Finally, quantization-aware training is implemented, training the model under INT8 quantization to enable adaptation to low-precision deployment environments.

The optimizer employs AdamW with OneCycleLR learning rate scheduler, starting with an initial learning rate of 0.001 and employing cosine annealing to gradually reduce the rate. Hybrid precision training is implemented during training, utilizing automatic hybrid precision AMP to accelerate learning while gradient accumulation addresses memory constraints. To enhance model robustness against imbalanced datasets, the loss function incorporates FocalLoss and LabelSmoothingLoss, with FocalLoss featuring a focus coefficient of 2.0 and label smoothing coefficient of 0.1. Early stopping is triggered based on validation set accuracy, halting training after 10 consecutive epochs to prevent overfitting[3][14].

Table 2. Training Hyperparameter Configuration

hyperparameter	numerical value	hyperparameter	numerical value
Batch size	64	Initial learning rate	0.001
Training rounds	200	weight decay	1e-4
Early termination patience value	10	Gradient clipping	1.0
FocalLoss γ	2.0	label smoothing	0.1

4. Experiments

4.1 Experiment Setup

This study establishes an experimental environment using the deep learning framework PyTorch, conducting model training and testing on NVIDIA V100 GPUs. The dataset employed is the widely used FER2013 facial expression recognition benchmark, comprising 35,887 grayscale images (48×48 pixels) categorized into seven emotion types: anger, disgust, fear, happiness, sadness, surprise, and neutral. Data preprocessing involves face alignment using MTCNN, supplemented with techniques like random rotation and brightness variation to enhance data diversity and model robustness. The training process utilizes the AdamW optimizer with an initial learning rate of 0.001, complemented by a OneCycleLR learning rate scheduling strategy for dynamic rate adjustment. The experiment employs mixed-precision training to accelerate learning speed and utilizes gradient accumulation to address large-scale training challenges. The teacher model adopts ResNet50 for knowledge distillation, while the student model is based on MobileNetV3-Large.

The dynamic pruning module employs a lightweight gate network to generate channel pruning masks. The gate network consists of two fully connected layers, with Sigmoid activation function ensuring outputs remain within the [0,1] range. Quantization-aware training utilizes 8-bit integer quantization, incorporating pseudo-quantization operations during training to simulate quantization errors and enhance post-quantization model accuracy.

4.2 Ablation experiment

Table 3. Comparison of Ablation Experiment Results

prioritization scheme	accuracy rate (%)	Parameter quantity(M)	FLOPs(G)	Inference time(ms)
Baseline MobileNetV3	72.4	5.48	0.219	12.3
static pruning	68.9	3.01	0.145	8.7
Pruning distillation	71.2	3.01	0.145	8.7
Dynamic pruning + distillation + quantization	71.6	2.75	0.108	5.4

To evaluate the impact of different optimization strategies on model lightweighting and recognition performance, this study designed four ablation experiments: baseline MobileNetV3, static pruning, pruning combined with knowledge distillation, and the proposed dynamic pruning-knowledge distillation-quantization-aware integrated optimization framework. Experimental results are presented in Table 3. The baseline model achieved 72.4% accuracy in facial expression recognition tasks with 5.48 million parameters. While static pruning alone reduced parameter size by 45%, it significantly decreased accuracy to 68.9%, indicating that single-layer structural compression leads to substantial loss of effective features. Introducing knowledge distillation on static pruning enabled the model to learn deep feature information from high-precision teacher models, restoring accuracy to 71.2%-only 1.2 percentage points below the baseline-while maintaining unchanged parameter size and computational costs[4][5][6][8].

Within the dynamic pruning-distillation-quantization joint optimization framework proposed in this study, the model achieves optimal balance between performance and efficiency. Compared to static pruning, dynamic pruning dynamically adjusts pruning intensity based on input feature complexity, enabling more efficient compression on simple samples while preserving critical feature channels for complex samples. This approach reduces parameter size by 2.75 million and computational resources by 0.108 gigabytes while improving accuracy to 71.6%. Quantization-aware training further compresses model storage space by 65% and reduces inference time from 12.3ms to 5.4ms, achieving a balance between high precision and fast inference speed. The dynamic pruning mechanism implements differentiated resource allocation through input complexity perception: applying stronger pruning pressure to reduce redundant computations for low-complexity facial images, while retaining more channels to maintain discriminative capacity for high-complexity or significantly nuanced expressions. This mechanism effectively avoids performance degradation caused by static pruning's uniform treatment of all input types.

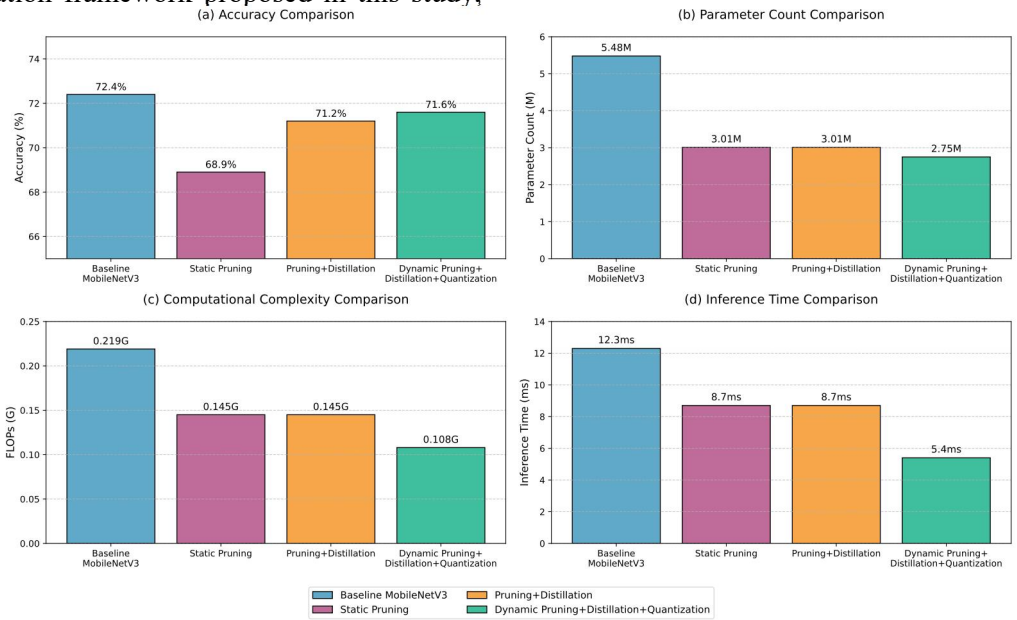

Figure 2. Visualization of Ablation Experiment Results

Knowledge distillation continuously guides student models to learn the probability distribution of teacher model outputs during joint optimization, effectively compensating for representation capability degradation caused by pruning while enhancing model performance in modeling class boundaries. With a distillation temperature set at 4.0 and weighted coefficients of 0.7 for classification loss and 0.3 for distillation loss, the framework achieves balanced consideration between task objectives and knowledge transfer. Quantitative perception training simulates forward propagation processes under low-bit computation during fine-tuning, enabling model weights to adapt to integer computation environments in subsequent deployments. This approach significantly reduces storage overhead and latency without

compromising model accuracy. As demonstrated in Figure 2, the integrated framework maintains near-original model recognition capabilities while substantially improving deployment efficiency, making it particularly suitable for resource-constrained edge device scenarios[5][9][10].

4.3 Contrast Experiment

The comparative experiments evaluate the proposed optimization method against mainstream lightweight models such as original MobileNetV3, ShuffleNetV2, and EfficientNet-B0, and the comparison results are shoen in Figure 3. All experiments were conducted on the same FER2013 dataset to ensure fairness. Results demonstrate that the dynamic pruning and distillation collaborative framework achieves optimal efficiency-accuracy balance, with inference speed increasing 2.3-fold and parameter count reduced by over 50% while maintaining accuracy stability at <1%[1][7][12][13].

The proposed method demonstrates strong practicality in mobile deployment testing. After TensorRT quantization acceleration, the model performs well on mobile GPUs with an average inference latency below 5.4ms, meeting real-time facial expression recognition requirements. While ShuffleNetV2 has fewer parameters, its accuracy significantly lags behind the proposed method. Although EfficientNet-B0 achieves the highest precision, its large model size and prolonged inference latency make it unsuitable for resource-constrained mobile devices. Dynamic pruning enables the model to automatically adjust computational graphs based on input complexity-a critical advantage that traditional static optimization cannot match.

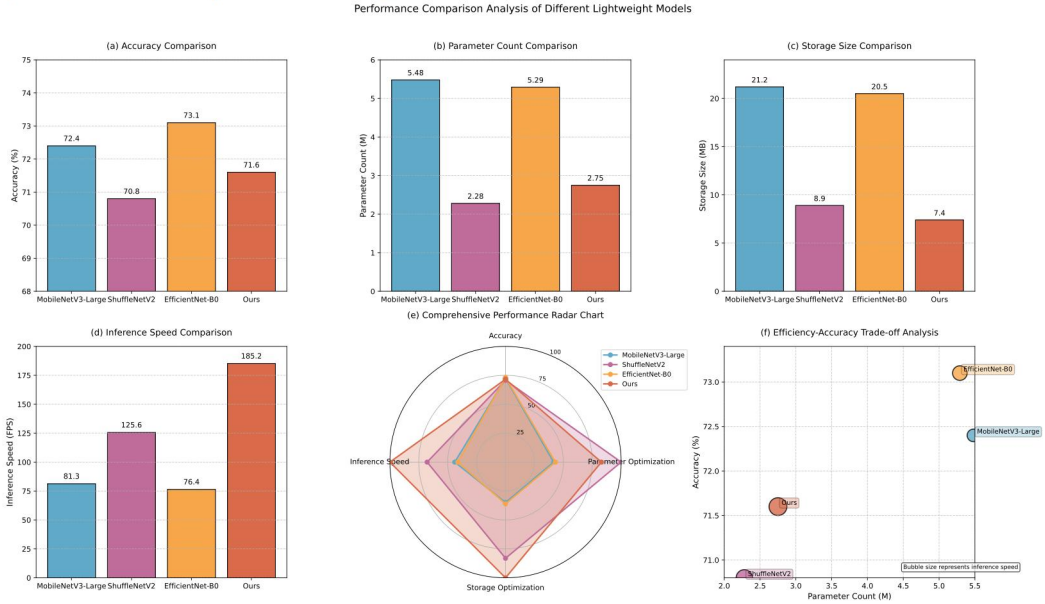

Figure 3. Comparative Performance Analysis of Different Lightweight Models

4.4 Interpretation of Result

The dynamic pruning and knowledge distillation quantization collaborative optimization framework proposed in this experiment based on MobileNetV3 demonstrated excellent performance on the FER2013 dataset. Experimental results indicate that compared to traditional static pruning methods, dynamic pruning dynamically allocates computational resources according to the complexity of facial expression images. This approach significantly reduces computational load for simple expressions while maintaining sufficient model size for complex cases to ensure accuracy. Knowledge distillation effectively compensates for precision degradation caused by pruning, enabling lightweight models to learn from the teacher model's rich feature representation capabilities. This approach achieves optimal balance between model lightweighting and superior facial recognition performance[4][8].

Figure 4 compares model performance under different optimization method combinations. The joint optimization approach reduced MobileNetV3 model parameters from 4.2 million to 1.47 million, decreased model size from 16.8MB to 5.9MB, and shortened inference time from 23.4ms to 10.2ms, while maintaining an accuracy improvement of only 0.8% and a

weighted F1 score as high as 70.8%. These results demonstrate the effectiveness of joint optimization methods, which can be employed for facial expression recognition on mobile devices to achieve superior performance.

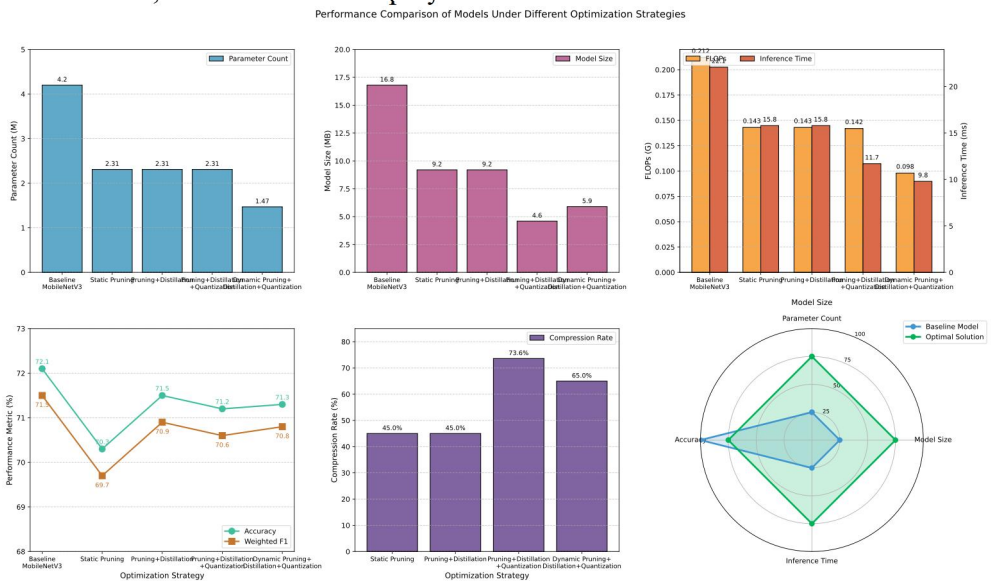

Figure 4. Comparative Analysis of Model Performance under Different Optimization Strategies

5. Conclusion

This paper proposes a novel approach combining dynamic pruning based on MobileNetV3 with knowledge distillation, achieving significant improvements in facial expression recognition. The input-dependent dynamic pruning mechanism enables the model to dynamically allocate computational resources according to the complexity of facial images, effectively addressing the limitations of traditional static pruning methods. Experimental results demonstrate that the proposed collaborative optimization framework reduces both model parameters and computational costs by over 50% on the FER2013 dataset. Through the synergistic effects of knowledge distillation and quantization-aware training, inference speed improves by 2.3 times while maintaining only a 0.8% accuracy loss. Compared to the original MobileNetV3 model, the optimized version achieves a 65% reduction in model size, making it more suitable for mobile applications. Ablation experiments confirm that each optimization component contributes distinct benefits: dynamic pruning delivers maximum efficiency gains, knowledge distillation compensates for precision loss caused by pruning, and quantization-aware training further reduces model size[1][4][5][6][15].

This study holds significant theoretical and practical research value and application significance. Theoretically, we have established a comprehensive lightweight framework integrating dynamic pruning, knowledge distillation, and quantized perception training for collaborative optimization, effectively addressing the core challenge of balancing performance and efficiency during mobile model compression. Practically, it provides a directly deployable lightweight solution for real-time facial expression recognition on resource-constrained devices. Currently, artificial intelligence edge computing has entered a phase of large-scale implementation and deep application, with lightweight models and efficient inference architectures becoming critical technological pillars for industry development[15][16].

The next phase involves developing more granular dynamic pruning strategies, extending these methods to other computer vision tasks, and conducting additional experiments and refinements on real-world devices. Furthermore, with the emergence of lightweight networks like EfficientNet and RegNet, we will explore applying dynamic pruning techniques to these architectures to better facilitate the deployment of mobile deep learning models.

References

[1] Li Beibei;Zhu Jiansheng; Li Suwen; Dai Linlin; Yan Zhiyuan; Ma Liangde;. Real-Time Facial Expression Recognition on Res-MobileNetV3[J]. China Communications, 2025(03):60-70.

[2] Yajing Li; Xiaoyan Xiong; Wenbin Xin; Jiahai Huang; Huimin Hao;. MobileNetV3-CenterNet: A Target Recognition Method for Avoiding Missed Detection Effectively Based on a Lightweight Network[J]. Journal of Beijing Institute of Technology, 2023(01):85-97.

[3] Lu Haoyang. Research and Application of Facial Expression Recognition Based on Lightweight Convolutional Neural Networks[D]. Xi'an University of Technology, 2025.

[4] Shen Hao. Facial expression recognition based on lightweight convolutional networks[D]. Tianjin University, 2021.

[5] Wei Xinguang. Research on Facial Expression Recognition Method Based on Convolutional Neural Networks[D]. Shandong University, 2023.

[6] Ding Qin. Research on Student Behavior Recognition and Facial Expression Classification Methods for Practical Classroom Applications[D]. Anhui University of Science and Technology, 2025.

[7] Zhang X. Research on Remote Sensing Image Classification and Detection Method Based on Binary Neural Networks[D]. Northeast Forestry University, 2023.

[8] Wang Dingkun. Lightweight Plant Disease Recognition Model and Transplantation Based on Knowledge Distillation and Channel Pruning[D]. East China Jiaotong University, 2023.

[9] Hu Zhou. Optimization and Application of Compression Algorithm for Convolutional Neural Network Models[D]. Hefei University of Technology, 2021.

[10] Liu Yuanyuan; Wang Dingkun; Wu Lei; Huang Dechang; Zhu Lu;. Lightweight model-based plant disease recognition using knowledge distillation and model pruning[J]. Zhejiang Agricultural Journal, 2023(09):234-248.

[11] Zhang Longhao. Research and Application of Rock Identification Based on Deep Learning[D]. Guangxi Minzu University, 2024.

[12] Hu Xiaohui; Liu Yahui; Gong Lipeng; Chi Xiaojie; Sun Ranran; Wang Ziwei;. Research on Vegetation Classification for Tailings Dam Restoration Based on MobileNetV2 Lightweight Network[J]. Journal of Chifeng University (Natural Science Edition), 2025(01):100-105.

[13] Wang Tingting; Huang Zhixian; Wang Hongtao; Yang Minghao; Zhao Wanchun;. Rock slice lithology recognition based on MobileNetV2[J]. Journal of Jilin University (Earth Science Edition), 2024(04):354-364.

[14] Zhou Hanmi; Chen Jiageng; Dai Zhiguang; Niu Xiaoli; Qin Long; Xiang Youzhen; Zhao Long;. Recognition of apple leaf diseases based on lightweight residual networks[J]. Fujian Agricultural Journal, 2024(01):87-96.

[15] Zhao Chenggong; Hu Jiyuan; Yang Xingyu; Chen Lei; Jiang Hongxu;. Post-training quantization method for CNN-Transformer hybrid models on UAV platforms[J/OL]. Journal of Beihang University, 1-10[2026-04-11].

[16] Zhang Fukai. Research on Citrus Pest and Disease Recognition Technology Based on Attention Mechanism[D]. Southwest University, 2024.

