# Research on Expression Recognition Algorithm Based on Deep Learning

Ziyue Wang, Hang Yin*, Honghe Xie, Jiaqi Gu

*Abstract*—**Facial expression recognition, as an important research direction in computer vision, has a wide range of applications in human-computer interaction, intelligent monitoring, and mental health assessment. This study proposes a facial expression recognition architecture based on improved YOLOv8, aiming to solve the problem of expression recognition in natural scenes and laboratory environments. In this study, two key improvements are made to the original YOLOv8 model: first, the dynamic convolution module is used instead of the traditional convolution to enhance the model's ability to extract expression features at different scales; second, the global attention mechanism (GAM) is integrated after the C2F module to strengthen the model's ability to model the global semantic information of facial expressions. Comprehensive experimental validation based on three standard datasets, CK+, FER2013, and RAF-DB, shows that the improved model achieves significant performance enhancement in multi-scene and multi-category expression recognition tasks, and the experimental results confirm the effectiveness and practical value of the proposed method in the field of expression recognition.**

*Index Terms*—**facial expression recognition, YOLOv8, dynamic convolution, GAM attention mechanism**

## I. INTRODUCTION

FACE facial expression recognition is a classical problem in the field of computer vision, aiming at predicting basic facial expressions from face images [1]. According to Ekman's research, facial expressions carry 55% of human emotional information [2]. Facial expression recognition, as an important research direction in computer vision, has a wide range of applications in human-computer interaction, distance learning, intelligent monitoring, mental health assessment, virtual reality, and intelligent driving. It is also widely used in more challenging fields[3], such as neuroscience research, social safety and security, and healthcare [4]. Traditional facial expression recognition methods mainly rely on hand-designed feature extraction, such as Gabor Filter, Local Binary Pattern (LBP), etc. [5], and feature extraction usually includes the analysis of color, texture, shape, and other information of the face image. With the development of machine learning technology, facial expression recognition (FER) gradually shifted from manual feature extraction to automatic feature learning [6]. The core of this stage is to use machine learning algorithms to automatically learn feature representations from data to improve the recognition performance, such as Support Vector Machines (SVMs) [7], but their generalization ability in complex scenes is limited. In recent years, advances in artificial intelligence, particularly breakthroughs in deep learning, have driven significant progress in FER, making deep learning-based algorithms a key research focus in academia and industry [8]. Deep learning models such as Convolutional Neural Network (CNN) [9], Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), and attention mechanisms can automatically learn multi-level feature representations [10], which significantly improve the accuracy and robustness of facial expression recognition.

Researchers worldwide have made significant progress in deep learning-based facial expression recognition. Yang Liu et al. proposed a semantic graph-based two-stream network to model semantic relationships between key appearance and geometric changes [11]; Qing Zhu et al. designed a CNN to achieve few-shot recognition by leveraging feature similarity [12]; Hong-Qi Feng et al. combined salient feature filtering with the Vision Transformer (ViT) for feature extraction via light normalization and a CNN [13], these methods performed well in still image expression classification. Although deep learning improves the performance of expression recognition, illumination variations, pose diversity, occlusion, and individual differences still pose challenges. In this context, the You Only Look Once (YOLO) model effectively improves the robustness and real-time performance of expression recognition in real-world scenarios such as complex illumination and pose changes by its multi-scale feature extraction and end-to-end joint optimization, providing a new technical path for expression recognition research. The FER-YOLO model proposed by Hui Ma et al. integrated the SE module to enhance feature extraction and performed well on the RAF-DB dataset [14], while Tejaswi et al. realized an efficient real-time expression classification system based on the YOLO framework, which further validated the potential of this technique in the field of expression recognition [15]. The study improved the YOLOv8 model by using dynamic convolution to enhance the multi-scale feature adaptation ability and added the Global Attention Mechanism (GAM) after the CSP Bottleneck with 2 Convolutions (C2F) module to capture the global contextual information to improve expression recognition performance.

Manuscript received May 12, 2025; revised July 31, 2025.

Ziyue Wang is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning 114051 China (e-mail: 18935265957@163.com).

Hang Yin is an associate professor of University of Science and Technology Liaoning, Anshan, Liaoning 114051 China (corresponding author to provide phone: +86-412-5929815; fax: +86-412-5929805;e-mail: 13842205866@163.com).

Honghe Xie is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning 114051 China (e-mail: xhh20010103@163.com).

Jiaqi Gu is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning 114051 China (e-mail: 13941499086@163.com).

## II. Related Work

### A. Traditional Facial Expression Recognition

Traditional expression recognition methods, which relied heavily on hand-designed feature extraction techniques, were an early paradigm for expression analysis research in computer vision. Their core idea was to achieve expression classification through the extraction of geometric, textural, and motion facial features. This research lineage began with the Facial Action Coding System (FACS) proposed by Ekman and Friesen in 1978, which established a comprehensive framework by decomposing expressions into Action Units (AUs) [2]. Based on this theoretical foundation, researchers developed three major categories of feature extraction methods: in terms of geometric features, the AFA system developed by Ichika Tada was representative of the system, which improved the recognition accuracy by analyzing permanent and transient facial features [16], while the subsequent development of Active Shape Modeling (ASM) and Active Appearance Modeling (AAM) further combined shape and textural characteristics [17]. In texture feature extraction, researchers developed a series of effective algorithms: the Gabor filter proposed by Daugman in 1988 pioneered multi-scale texture analysis [18], followed by the Local Binary Pattern (LBP) proposed by Ahonen et al. in 2006 which further improved the characterization of texture features by encoding local grayscale changes [19]. This temporal evolution demonstrated the developmental trajectory of texture feature extraction techniques, progressing from global analysis to local characterization. Although these traditional methods laid the foundation for FER research, their inherent limitations - including laborious manual feature engineering, sensitivity to illumination variations and pose changes, and poor generalization capability - ultimately led researchers to adopt more robust deep learning approaches.

### B. Machine Learning for Facial Expression Recognition

Expression recognition research underwent a significant paradigm shift from traditional to machine learning approaches, where Support Vector Machine (SVM) and Principal Component Analysis (PCA) served as foundational techniques. Studies have shown that SVM demonstrates superior performance in expression recognition, particularly in small-sample scenarios: Hong-Xu Cai et al. validated SVM's exceptional classification capability using the Facial Expression Recognition System (FERS) integrating multiple features: Angular Radial Transform (ART), Discrete Cosine Transform (DCT), and supplementary descriptors [20]; and Li-Yuan Chen et al. further revealed the SVM was adaptive feature selection law, and found that shape features were outstanding in non-human related scenarios, while radial basis function SVM was more advantageous in human-related scenarios [21]. PCA offered a novel technical approach for facial expression recognition via efficient dimensionality reduction. Arora et al. developed the AutoFER system by integrating PCA with Particle Swarm Optimization (PSO), which achieved remarkable accuracy [22]; and the hybrid PCA-MLP model proposed by Rani et al. optimized the feature extraction through the process to improve the accuracy of children's emotion recognition to a breakthrough level. These machine learning methods not only established new performance benchmarks for expression recognition but more significantly laid a rigorous theoretical foundation and comprehensive methodological framework for subsequent deep learning techniques through approaches like feature optimization and dimensionality reduction.

### C. Deep Learning for Facial Expression Recognition

The rise of deep learning revolutionized the field of expression recognition by achieving automatic expression classification through end-to-end learning. This breakthrough overcame traditional feature engineering limitations, significantly improved recognition accuracy and generalization capability, and demonstrated excellent performance in complex scenarios [23]. Under the deep learning framework, Convolutional Neural Networks (CNN) automatically extract expression features through their convolution-pooling hierarchical structure for efficient classification. Dhvanil Bhagat et al. combined Deep Convolutional Neural Networks (DCNN) with VGG-based pre-trained models to achieve high accuracy on the FER2013 dataset [24], while the CNN system developed by Mangshor et al. achieved real-time recognition of six basic expressions [25]. These studies have shown the advantages of CNNs in expression recognition, though datasets still need optimization to improve generalization capability. Meanwhile, Transformers have demonstrated remarkable potential in expression recognition through their self-attention mechanism. Yan-De Li et al. proposed the FER-former model, innovatively integrating a hybrid CNN-Transformer architecture with multimodal supervision by first designing a heterogeneous domain-guided supervision module to enhance the image features, and then developing a dedicated Transformer encoder to handle multimodal markers, which showed superb capability on multiple benchmark datasets [26]. In addition to advances in model architectures, the application of Data Augmentation [27] and Transfer Learning [28] techniques has further enhanced the generalization capability of deep learning models. Notwithstanding these advances, challenges persist regarding data dependence, interpretability limitations, and substantial computational requirements. Looking ahead, progress in self-supervised learning, multimodal fusion, and lightweight model design promises further breakthroughs in facial expression recognition.

## III. Method

This section details the architecture of enhanced YOLOv8 network [29], systematically presenting its constituent modules and their topological dependencies. We introduce two key architectural innovations to baseline YOLOv8: replacement of conventional convolutions with dynamic convolution modules; and integration of GAM after each C2F module. The subsequent sections will comprehensively detail the implementation methodologies of these enhancements, including their computational advantages and performance implications in facial expression recognition tasks.

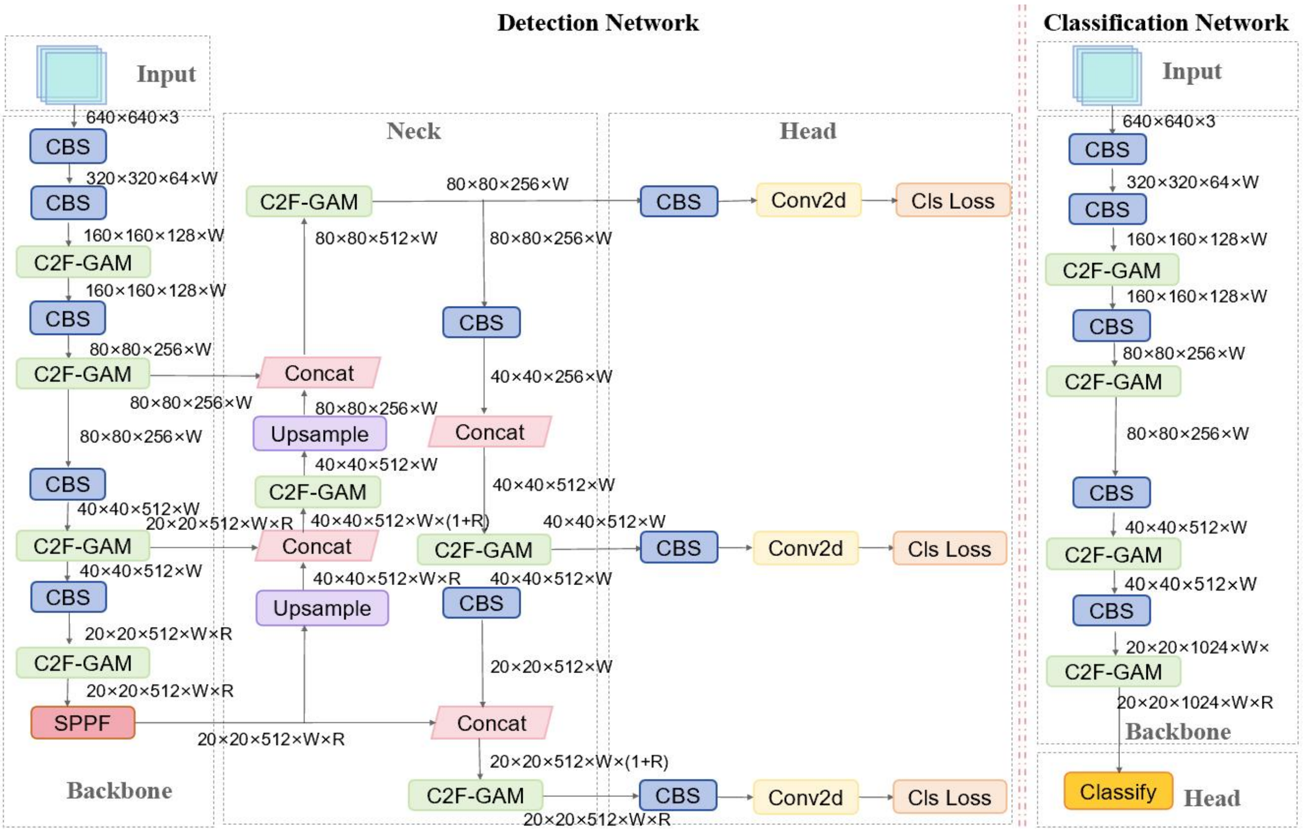


Fig. 1. Overall model diagram. The overall model diagram is divided into two main parts, the detection network and the classification network, and the classification network is further divided into three parts: input, backbone, and head.

### *A. Overall Model Structure*

YOLOv8 as the latest target detection algorithm, has a classification module that achieves target classification through convolution layers (Convolution + Batch Normalization + Sigmoid Linear Unit) and global average pooling [30]. To further improve the performance of the model, this study makes two improvements to the classification module of YOLOv8: the first improvement is to optimize the convolution module of the YOLOv8 classification model, choosing to replace all of the original convolution modules with the Dynamic Convolutional Module [31], to enhance the model's adaptability to diverse expression features; The second improvement is to introduce the GAM [32] after each C2F module further to strengthen the model's focus on key expression regions. The general architecture of the improved YOLOv8 network model is shown in Fig. 1.

### *B. Dynamic Convolution*

In this study, Dynamic Convolution is chosen to replace the ordinary convolution in the original model of YOLOv8, mainly based on its significant advantages in feature extraction capability and adaptability. Dynamic Convolution is adopted to replace conventional convolution in YOLOv8, primarily due to its enhanced feature extraction capability. The fundamental principle involves input-dependent generation of convolution kernel parameters, allowing adaptive adjustment of the feature extraction process [33]. This adaptive mechanism enables the network to capture the local features of the input data more flexibly and improve the expressive ability of the model. Dynamic convolution demonstrates superior generalization capability compared to traditional convolution, exhibiting enhanced adaptability to challenging scenarios including varying illumination conditions, pose variations, and individual differences, while preserving computational efficiency.

The introduction of dynamic convolution significantly improves the classification performance of the expression recognition task and provides a more robust feature extraction mechanism for the model. The dynamic convolution structure is shown in Fig. 2. Where data flow represents the flow path of the input feature map in the model, and model parameter flow represents the generation and update process of the dynamic convolution kernel.

In the traditional convolution operation, the convolution kernel is fixed, and its mathematical expression (1) shows

$$y = W * x + b \tag{1}$$

Where W is a fixed convolution kernel, $x$ is the input feature map, $b$ is the bias term, and $y$ is the output feature map. While in dynamic convolution, the convolution kernel $W$ is no longer fixed, but is dynamically generated based on the input data $x$. Its mathematical expression (2) can be expressed as

$$y = g(W_{(x)} * x + b) \tag{2}$$

Where $W_{(x)}$ is a dynamic convolution kernel generated from the input $x$, $x$ denotes a convolution operation, and $g$ is a nonlinear activation function. According to the above structure diagram, the workflow of dynamic convolution can be described as follows:

1) Input Feature Map Processing: the dynamic

convolution process begins when the input feature map $x$ is fed into the attention mechanism module. This module analyzes the spatial and channel information in the feature map to determine which regions require more focus during subsequent processing.

2) Attention Mechanism Operation: the attention mechanism first computes attention weights based on the input feature map $x$. These weights represent the relative importance of different regions in the feature map. The computed weights are then multiplied with the original feature map, producing a weighted feature map where more significant regions are enhanced while less important ones are suppressed.
3) Dynamic Kernel Generation: the weighted feature map serves as input to the dynamic convolution kernel generation function Π(x). This function is implemented as a multi-layer network containing weight matrices $W_1$ through $W_n$, which progressively transform the input to generate intermediate outputs $\Pi_1$ through $\Pi_n$. These outputs are combined to produce the final dynamic convolution kernel that is specifically adapted to the input's characteristics.
4) Convolution and Output Generation: the generated dynamic kernel performs a convolution operation with the original input feature map $x$. This adaptive convolution process captures the most relevant features from the input. The convolution output then undergoes final processing steps such as activation or normalization to produce the output feature map ready for subsequent network layers.

Dynamic convolution accomplishes input-adaptive feature optimization through the aforementioned four-stage cascaded processing.

### *C. Global Attention Mechanism*

In the YOLOv8 classification class module, the C2F module effectively extracts multi-level features through the cross-stage partial fusion mechanism, but it has some limitations in capturing global contextual information. To solve such problems, this study introduces the GAM attention mechanism after the C2F module to further improve the classification performance of the expression recognition task. GAM attention mechanism is a technique used to improve the feature representation capability of deep learning models. GAM is a technique used to enhance the feature representation capability of deep learning models, the core idea of which is to enable the model to pay more attention to the important regions in the input data by modeling channel attention and spatial attention simultaneously, to improve the efficiency and accuracy of feature extraction. The structure of GAM is shown in Fig. 3.

The GAM attention mechanism consists of two main modules, namely the channel attention module and the spatial attention module. These two modules weigh the feature map from the channel dimension and the spatial dimension, respectively, thus improving the characterization of important features. Here, the main work of channel attention is to assign different weights to each channel to improve the feature representation of important channels. The core idea is to obtain the global information of each channel through Global Average Pooling (GAP), and then generate the channel weights through the fully connected layer, which can reflect the importance of each channel in the task. The structure diagram of channel attention is shown in Fig. 4. The main task of spatial attention is to assign different weights to different spatial locations of the feature map to improve the feature representation of important regions. It captures the spatial context information through convolution operation and generates a spatial weight map, where the spatial weight map reflects the importance of each spatial location in the task. The spatial attention structure diagram is shown in Fig. 5. By combining channel attention and spatial attention, GAM can optimize the feature map from both the channel dimension and the spatial dimension, thus comprehensively improving the feature representation capability of the model.

Assume that the input feature map is $X \in R^{C\times H\times W}$, where $C$ is the number of channels, and $H$, and $W$ are the height and width of the feature map, respectively. The computational procedure of GAM is as follows.

First, we enter the channel attention module and perform global average pooling on the input feature map $X$ to obtain the channel descriptor $Z \in R^{C}$,as in (3)

$$\mathrm{Z_c} = \frac{1}{H\times W}\sum_{\mathrm{i}=1}^{H}\sum_{\mathrm{j}=1}^{W} X_{\mathrm{c}}(\mathrm{i}, j) \tag{3}$$

As in (4), generating channel weights through fully connected layers $\alpha \in R^{C}$

$$\alpha = \sigma\big(W_2\delta(W_1 z)\big) \tag{4}$$

Where $W_1$ and $W_2$ are the weight matrices of the fully connected layer, $\delta$ is the activation function, and $\sigma$ is the sigmoid function. Multiplying the channel weights $\alpha$ with the input feature map $X$ yields the channel attention-weighted feature map $X^{'}$, as in (5)

$$X_C{}^{'} = \alpha C \cdot X_C \tag{5}$$

As in (6), we enter the spatial attention module and perform a convolution operation on the channel-attention weighted feature map $X^{'}$ to generate the spatial weight map $\beta \in R^{H\times W}$:

$$\beta = \sigma\Big(f_{\mathrm{conv}}(X')\Big) \tag{6}$$

Where $\mathrm{f_{conv}}$ is the convolution operation and $\sigma$ is the sigmoid function. Multiply the spatial weight map $\beta$ with the feature map$X^{'}$ to get the final feature map $X^{''}$, as in (7)

$$X^{''}_{C,\mathrm{i,j}} = \beta_{i,j} \cdot X^{'}_{C,i,j} \tag{7}$$

Through the above steps, GAM can optimize the feature map from both the channel dimension and the spatial dimension, thus enhancing the characterization of important features.

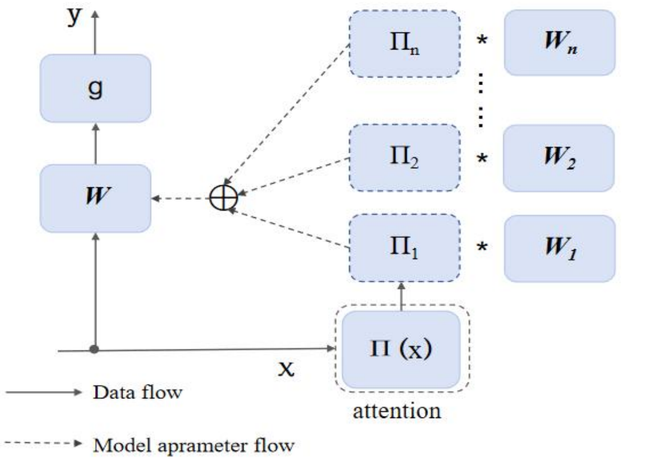


Fig. 2. Dynamic convolution structure diagram. It demonstrates multiple parallel convolution branches $W_1$, $W_2$ ... $W_n$, are dynamically aggregated into a single convolution by means of attention weights $\Pi_1$, Π2 ... $\Pi_n$ dynamically aggregated into a single convolution W for enhanced feature extraction.

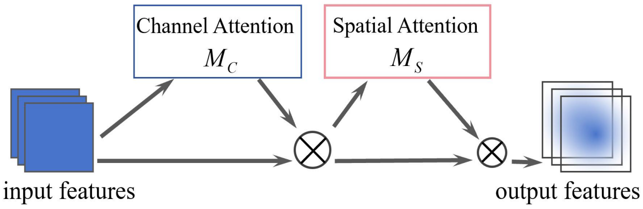


Fig. 3. GAM Attention Structure Chart. The two main components are spatial attention and channel attention.

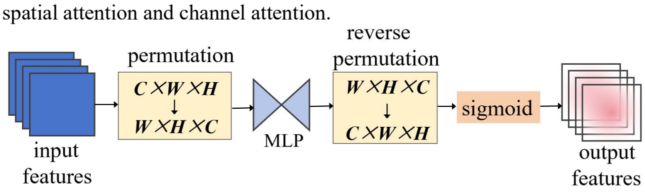


Fig. 4. Channel Attention Structure.

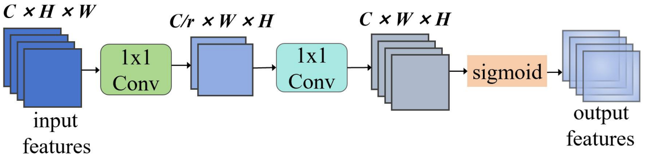


Fig. 5. Spatial Attention Structure.

## IV. Experiments and Results

In this study, we propose an improved YOLOv8 model for facial expression recognition in both natural scenes and laboratory environments, which demonstrates superior performance. To verify its effectiveness and generalization ability, we conducted systematic experimental evaluations on three generic expression datasets, FER2013, CK+, and RAF-DB, and compared them with mainstream methods. The results showed that the improved YOLOv8 model significantly improved the recognition accuracy and robustness, and could effectively cope with the challenges of multi-class expression classification, providing a new general and practical solution for the field of expression recognition.

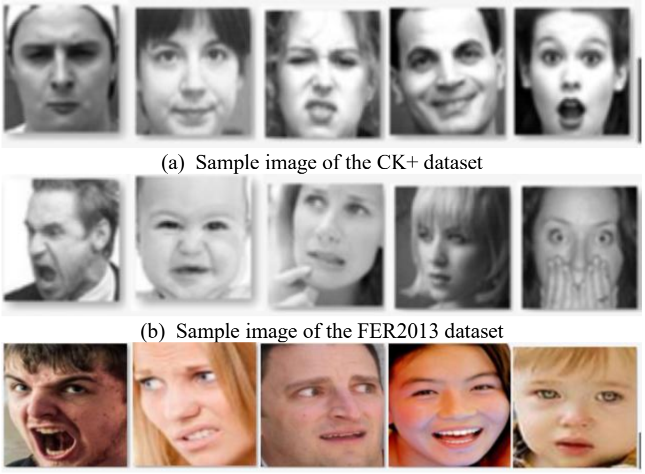
### A. Datasets for Benchmarks

The CK+ dataset [34] is a high-quality facial expression dataset released by Carnegie Mellon University containing 593 video clips from 123 participants covering seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). Each video records dynamic changes from neutral to extreme emotions and contains precise labeling of facial key points. Sample images from the CK+ datasets are shown in Fig. 6a.

The FER2013 dataset [35] is a publicly available facial expression recognition datasets, originally provided by Kaggle, that contains 35,887 48×48 pixel grayscale images labeled with seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutrality). Sample images of the FER2013 dataset are shown in Fig. 6b.

The RAF-DB dataset [36][37] is a publicly available facial expression recognition datasets containing more than 30,000 facial images labeled with seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). These images are captured from real-life scenarios and have a high degree of diversity and naturalness. Sample images from the RAF-DB dataset are shown in Fig. 6c.

In order to further understand the characteristics of the three datasets, CK+, FER2013, and RAF-DB, the label classification of each dataset was statistically analyzed and the corresponding distribution graphs were drawn. The label distribution graphs of the CK+ dataset was shown in Fig. 7a, the label distribution of the FER2013 dataset was shown in Fig. 7b, and the label distribution of the RAF-DB dataset was shown in Fig. 7c.



(a) Sample image of the CK+ dataset

(b) Sample image of the FER2013 dataset

(c) Sample image of the RAF-DB dataset

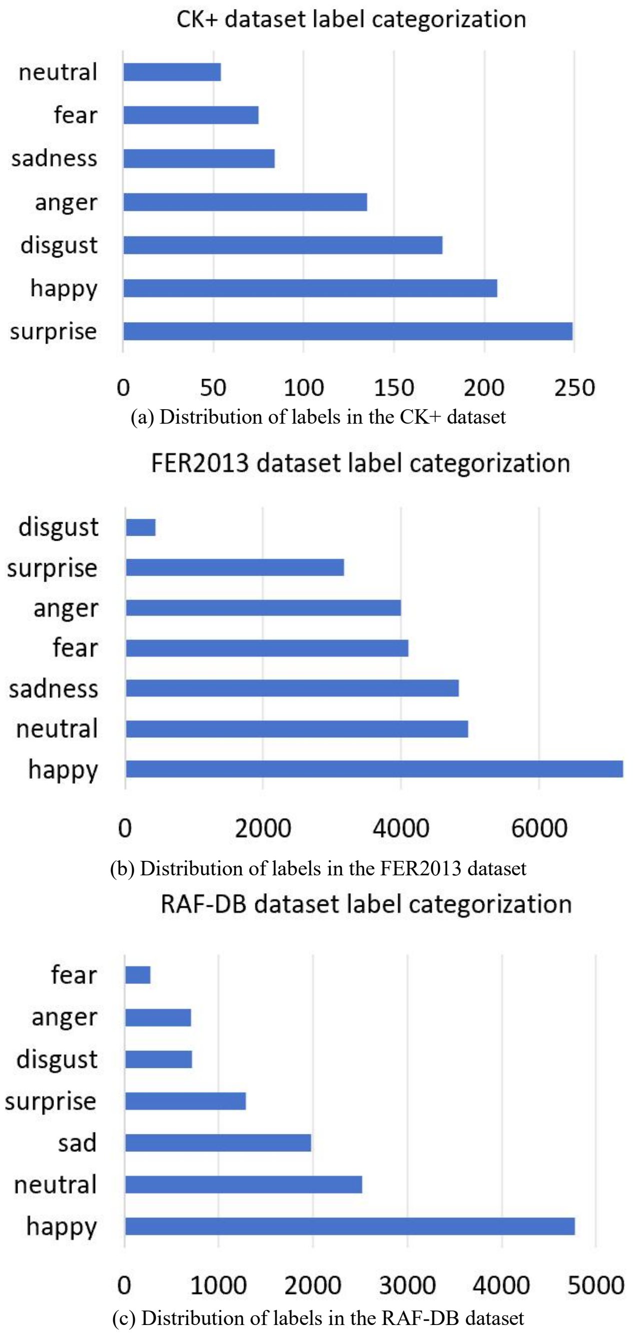
Fig. 6. Examples from the CK+, FER2013, RAF-DB datasets.

### B. Data Augmentation

In this study, data enhancement techniques [38] were applied to the three expression recognition datasets, CK+, FER2013, and RAF-DB, to improve the generalization ability and robustness of the models. Specifically, two main enhancement strategies, random cropping and color jittering were used. Random Cropping improved the model's ability to detect local expression features by extracting different image regions and adjusting their sizes, while Color Jittering improved the model's stability under different lighting conditions by adjusting parameters such as brightness and contrast. The combination of the two enhancement methods significantly enriched the distribution of the training data and alleviated the overfitting problem of the model, while improving the generalization performance of the model in complex scenes, providing a more reliable database for the expression recognition task.

### C. Experiment Settings

In this experiment, the model was implemented based on the Pytorch deep learning framework and utilized an NVIDIA GeForce RTX 4090 GPU as the computing platform. The training parameters were configured as follows: The initial learning rate (lr0) was set to 0.01 to facilitate rapid convergence during the early training stage. The total number of training epochs was 50 to ensure thorough optimization of model parameters. To prevent overfitting, an early stopping mechanism was introduced, which automatically terminated training when the validation performance showed no improvement for several consecutive epochs. The batch size was set to 16 to balance computational resources and training efficiency. All input images were uniformly resized to 64×64 pixels to meet the model's input requirements while reducing computational complexity.



(a) Distribution of labels in the CK+ dataset

(b) Distribution of labels in the FER2013 dataset

(c) Distribution of labels in the RAF-DB dataset

Fig. 7. Distribution of labels from the CK+, FER2013, and RAF-DBSFER datasets. The distribution graph allows us to easily understand the number of expressions in each category.

### D. Analysis

*CK+ Dataset Experiments*

To comprehensively evaluate the performance of the improved YOLOv8 model in this study, comparative experiments with existing mainstream algorithms were conducted on the CK+ dataset. During the experiments, all methods used the same training set, validation set division method, and evaluation metrics to ensure the comparability and fairness of the experimental results. The experimental results were shown in Table I, which showed that the accuracy rate was improved by about 19 percentage points compared with VGG16; the accuracy rate was improved by about 6 percentage points compared with ResNet18; and compared with ShuffleNet, a lightweight network, the method of this study improved the accuracy rate by about 9 percentage points while ensuring real-time performance. Compared with the recently proposed PCARNet algorithm that employed a tapered convolution module and an improved convolution attention mechanism, this study effectively solved the performance degradation problem due to scale changes during expression feature extraction by introducing a dynamic convolution module, thus achieving a 2.5 percentage point improvement in recognition accuracy; compared with the method based on improving the inverted residual structure through the reconfiguration of SandGlass modules Compared with the method based on improving MobileNetV2 by reconstructing the inverted residual structure through the SandGlass module, the GAM attention mechanism adopted in this study could better capture the key region information in the expression features, which significantly improved the model's ability to perceive the subtle expression changes, and the improved YOLOv8 model achieved a recognition accuracy of 97.9% on the CK+ dataset. This result not only verified the effectiveness of the proposed improvement strategy but also showed that the model had strong robustness and generalization ability.

TABLE I
ACCURACY OF DIFFERENT ALGORITHMS ON THE CK+ DATASET

| Method | Accuracy(%) | year |
|---|---|---|
| YOLOv8 | 91.5 | —— |
| VGG16 | 78.5 | —— |
| ResNet18 | 91.5 | —— |
| ShuffleNet | 88.6 | —— |
| PCARNet [39] | 95.4 | 2024 |
| MobileNetV2+SandGlass [40] | 95.9 | 2023 |
| CNN3 [47] | 95.0 | 2024 |
| CNN [48] | 91.4 | 2024 |
| CNN-GCN+CustomLoss | 95.3 | 2025 |
| MAEL-FER [51] | 96.9 | 2025 |
| DCNN-BiLSTM | 97.2 | 2025 |
| EmoNeXt-S | 97.7 | 2025 |
| **Ours** | **97.9** | —— |

*FER2013 Dataset Experiments*

We launched systematic comparison experiments on the FER2013 public datasets to validate the performance of the improved YOLOv8 model proposed in this study for expression recognition in various scenarios. As shown in Table II, compared with the baseline network models such as VGG16, ResNet18, and ShuffleNet, the method in this study improved the recognition accuracy by 14.6, 12.3, and 15.9 percentage points, respectively. Compared to the method using a pre-trained deep learning architecture fusion strategy with a combination of AlexNet, ResNet50, and Inception V3, this study's method achieved a 10 percentage point improvement in accuracy while keeping the model lightweight, which demonstrated that the optimal design of a single network could achieve better performance without increasing the complexity of the model; compared to the CNN + Haar Cascade algorithm, the method in this study achieved a 13.5 percentage point improvement in accuracy to 83.5% accuracy although the latter enhances the feature extraction capability by increasing the number of convolution layers, employing a multi-core ReLU activation function, and incorporating the Haar cascade model. This result fully demonstrated that the improvement strategy proposed in this study achieved a better balance between feature extraction efficiency and recognition accuracy.

TABLE II
ACCURACY OF DIFFERENT ALGORITHMS ON THE FER2013 DATASET

| Method | Accuracy(%) | year |
|---|---|---|
| YOLOv8 | 75.3 | —— |
| VGG16 | 68.9 | —— |
| ResNet18 | 71.2 | —— |
| ShuffleNet | 67.6 | —— |
| CNN + Haar Cascade [42] | 70.0 | 2022 |
| AlexNet + Inception V3 + ResNet50 [41] | 73.5 | 2024 |
| CLCM [45] | 63.0 | 2024 |
| CNN3 [47] | 79.0 | 2024 |
| HybridizedCNN-LSTM | 79.3 | 2024 |
| EfficientFER [50] | 82.4 | 2025 |
| DLFER | 82.1 | 2025 |
| EmoNeXt-S | 74.3 | 2025 |
| **Ours** | **83.5** | —— |

### *RAF-DB Dataset Experiments*

To validate the improved YOLOv8 model for expression recognition in real scenes, we did a series of comparison experiments on the RAF-DB dataset. From the results in Table III, we could see that the improved method improved the accuracy by 35.1, 21.9, and 22.1 percentage points compared to the classical networks, such as VGG16, ResNet18, and ShuffleNet, respectively, which was attributed to the dynamic convolution module that enhanced the extraction of complex expression features and combined with the GAM attention mechanism to make the model focus on the key area features more accurately.

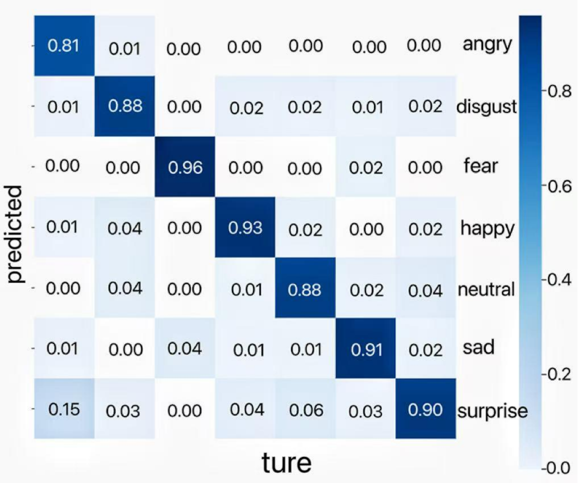


Fig. 8. Confusion matrices of the RAF-DB. The diagonal of the matrix reflects the accuracy of the classification, while the non-diagonal can show common types of misclassifications.

Compared to the method based on VGG19 and augmented with MixCut data, our method improved the accuracy by 2.6 percentage points to 90.4% without complex data augmentation. In addition, compared to PROPOSED+ResNet50 as a method, our model also improved the accuracy by 13.7 percentage points while keeping it lightweight, which further proved the effectiveness of our improvement. This showed that our improvement strategy enhanced the model's ability to recognize expressions in real scenes. To further evaluate the performance of the improved model, we used the tool Confusion Matrix. The confusion matrix visualized how well the model categorized the various expression categories, helping to identify which categories the model performed well on and where it might have been confused. Fig. 8 showed the confusion matrix for the RAF-DB dataset.

The confusion matrix plot showed the performance of the model in various emoji classification tasks, where each row corresponds to the true label, each column corresponds to the predicted label, and the values in the matrix indicated the probability that the model predicted a true label to be in a certain category. Combined with the confusion matrix as a whole, it seemed that the model showed high classification accuracy in most of the emoji categories.

In particular, the model performed well in the categories of Fear and Happy, achieving 0.96 and 0.93 prediction probabilities, respectively. In addition, the model performed well in the categories of Angry, Disgust, Neutral, the predicted probabilities for the categories of "angry", "disgust", "neutral", "sad" and "surprise" were 0.81, 0.88, 0.88, 0.91 and 0.90, respectively, which indicated that the model had good classification ability in these categories as well. However, the model had a 15% probability of misclassifying "anger" as "surprise", which indicated that "surprise" and "anger" might be different in the feature space. This indicated that "surprise" and "anger" might have high similarity in the feature space, which led to some difficulties for the model in distinguishing these two categories. This confounding phenomenon was consistent with the actual situation of human emotion expression, because "surprise" and "anger" might have similar facial expressions or behavioral features in some contexts, which also proved the validity of the model.

TABLE III
ACCURACY OF DIFFERENT ALGORITHMS ON THE RAF-DB DATASET

| Method | Accuracy(%) | year |
|---|---|---|
| YOLOv8 | 82.9 | —— |
| VGG16 | 55.3 | —— |
| ResNet18 | 68.5 | —— |
| ShuffleNet | 68.3 | —— |
| PROPOSED+ResNet50 [43] | 76.7 | 2023 |
| VGG19+MixCut [44] | 87.8 | 2024 |
| CLCM [45] | 84.0 | 2024 |
| ECA [46] | 89.9 | 2024 |
| WCA | 81.7 | 2025 |
| ShuffleNet-V2+ResNet-50 [49] | 89.7 | 2025 |
| FER-MOTION | 90.3 | 2025 |
| **Ours** | **90.4** | —— |

### *Cross Datasets Evaluation*

To comprehensively evaluate the performance of the proposed method, this study conducted systematic experiments on three representative expression recognition datasets, namely, CK+, FER2013, and RAF-DB. The CK+ dataset, as high-quality, small-scale datasets collected in a laboratory environment, could validate the upper bound of the model's performance under ideal conditions; the FER2013 dataset, which contained a large number of expression images under real scenes, could evaluate the performance of the model in real applications; while the RAF-DB dataset provided a rigorous testing environment for the robustness of the model with its rich expression categories and complex background variations. Table IV showed the performance comparison of the cross datasets, and from the experimental results, it could be seen that the improved YOLOv8 model in this study showed significant advantages on all three datasets: on the CK+ dataset, the accuracy rate reached 97.9%; on the FER2013 dataset, the accuracy rate was 83.5%; on the RAF-DB dataset, the accuracy rate was 90.4%. This series of experimental results fully proved that the improvement strategy proposed in this study could effectively improve the expression recognition ability of the model in different scenarios, and had strong generalization and robustness.

TABLE IV
PERFORMANCE COMPARISON OF CROSS DATASETS

| Method | CK+ | FER2013 | RAF-DB |
|---|---|---|---|
| YOLOv8 | 91.5% | 75.3% | 82.9% |
| VGG16 | 78.5% | 68.9% | 55.3% |
| ResNet18 | 91.5% | 71.2% | 68.5% |
| ShuffleNet | 88.6% | 67.6% | 68.3% |
| PROPOSED+ResNet50 [43] | 89.5% | 73.4% | 76.7% |
| PCARNet [39] | 95.4% | 73.7% | 87.5% |
| **Ours** | **97.9%** | **83.5%** | **90.4%** |

TABLE V
ABLATION EXPERIMENTS ON THE CK+, FER2013, RAF-DB DATASET

| Experimental Group | CK+ | FER2013 | RAF-DB |
|---|---|---|---|
| YOLOv8 | 91.5% | 75.3% | 82.9% |
| YOLOv8+DyConv | 92.4% | 78.2% | 86.3% |
| YOLOv8+GAM | 95.3% | 81.9% | 87.2% |
| **Ours** | **97.9%** | **83.5%** | **90.4%** |

### *Ablation Experiment*

To verify the contribution of the innovations proposed in this study to the model performance, we designed an ablation experiment to quantitatively analyze the impact of each innovation on the final results by gradually removing or replacing key modules. Table V showed the data of the ablation experiment for the three datasets of CK+, FER2013, and RAF-DB. It could be seen that when only the convolution module in YOLOv8 was replaced with the dynamic convolution, the classification accuracy of the three datasets of CK+, FER2013, and RAF-DB were improved by 0.9, 2.9, and 3.4, respectively, compared to the original model, which verified that the dynamic convolution could improve classification accuracy by self-adaptively adjusting the convolution kernel weights. This verified that dynamic convolution could significantly improve the expression recognition model's ability to capture complex and subtle expression features and generalization performance by adaptively adjusting the weights of convolution kernels. When the GAM attention mechanism was introduced into the YOLOv8 model, the classification accuracy of the CK+, FER2013, and RAF-DB dataset were improved by 3.8, 6.6, and 4.3, respectively, compared with the original model, which proved that the GAM enhanced the expression recognition model's attention to the key features and classification accuracy by enhancing the capability of capturing the global contextual information. When both the dynamic convolution module and the GAM were invoked, the model achieved optimal performance, which indicated that the dynamic convolution module and the GAM were complementary.

## V. CONCLUSION

In this study, we propose a facial expression recognition architecture based on improved YOLOv8, which significantly improves the model's expression recognition performance in natural scenes and laboratory environments by introducing a dynamic convolution module and GAM (Global Attention Mechanism). The dynamic convolution module enhances the model's ability to adapt to multi-scale expression features, while the GAM effectively captures the global contextual information of facial expressions. Through systematic experimental evaluations on three widely used benchmark datasets for expression recognition, namely CK+, FER2013, and RAF-DB, the improved YOLOv8 model performs well in the expression recognition task, with a significant increase in recognition accuracy. The experimental results show that the proposed method achieves 97.9% accuracy on the CK+ dataset, 83.5% on the FER2013 dataset, and 90.4% on the RAF-DB dataset, which are all better than the existing mainstream methods. These results validate the effectiveness of the proposed improvement strategy and show that the model has strong robustness and generalization ability, and can effectively deal with the challenges of multi-category expression classification in different scenarios. The research in this study provides a new solution in the field of expression recognition with a wide range of applications.

In future research, we plan to further explore the fine-grained feature information of facial expressions and work on optimizing the model structure to reduce its complexity. Specifically, we will try to design a new network architecture that focuses on improving the ability to recognize subtle expressions, thus achieving higher classification accuracy in complex scenes.

## ACKNOWLEDGMENT

Ziyue Wang sincerely appreciates the invaluable guidance and constructive suggestions provided by Supervisor Hang Yin.

## REFERENCES

[1] Xing Guo, Yu-Dong Zhang, Si-Yuan Lu, and Zhi-Hai Lu, "Facial Expression Recognition: A Review," Multimedia Tools and Applications, vol. 83, no.8, pp23689-23735, 2024.

[2] Paul Ekman, and Wallace V. Friesen, "Facial Action Coding System," Environmental Psychology & Nonverbal Behavior, 1978.

[3] Gan Chen, Jun-Jie Peng, Wen-Qiang Zhang, Kanrun Huang, Feng Cheng, Hao-Chen Yuan, and Yan-Song Huang, "A Region Group Adaptive Attention Model for Subtle Expression Recognition," IEEE Transactions on Affective Computing, vol. 14, no.2, pp1613-1626, 2021.

[4] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer, "Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries," IEEE Transactions on Industrial Informatics, vol. 18, no.8, pp5619-5627, 2022.

[5] Andrea Caroppo, Alessandro Leone, and Pietro Siciliano, "Comparison Between Deep Learning Models and Traditional Machine Learning Approaches for Facial Expression Recognition in Ageing Adults," Journal of Computer Science and Technology, vol. 35, pp1127-1146, 2020.

[6] Jian-Hua Liu, and Lei Tang, "A Review of Facial Expression Recognition Technologies," Information and Communications Technology and Policy, vol. 48, no.8, p89, 2022.

[7] Jian-Wei Zheng, Xin-Mei Liu, and Jun-Ling Yin, "Pain Expression Recognition Based on LBP and SVM", Computer Systems & Applications, vol. 30, no.4, pp111-117, 2021.

[8] Amjad Rehman Khan, "Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges," Information, vol. 13, no.6, p268, 2022.

[9] Mustafa Can Gursesli, Sara Lombardi, Mirko Duradoni, Leonardo Bocchi, Andrea Guazzini, and Antonio Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," IEEE Access, 2024.

[10] S. Vinod Kumar, G. Sunil, Ramy Riad Hussein, S. Manju Vidhya, and S. Meenakshi Sundaram, "Attention Based ConVnet-Recurrent Neural Network for Facial Recognition and Emotion Detection," Lecture Notes in Engineering and Computer Science: Proceedings of The 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), pp1-5, 2024.

[11] Yang Liu, Xing-Ming Zhang, Jin-Zhao Zhou, and Lun-Kai Fu," SG-DSN: A Semantic Graph-Based Dual-Stream Network for Facial Expression Recognition," Neurocomputing, vol. 462, pp320-330, 2021.

[12] Qing Zhu, Qi-Rong Mao, Hong-Jie Jia, Elias Nii Ocquaye Noi, and Juan-Juan Tu, "Convolutional Relation Network for Facial Expression Recognition in the Wild with Few-Shot Learning, " Expert Systems with Applications, vol. 189, p116046, 2022.

[13] Hong-Qi Feng, Wei-Kai Huang, and Deng-Hui Zhang, "Facial Expression Recognition Method Combining Salient Feature Selection and Vision Transformer (in Chinese)", Journal of Computer Engineering and Applications, vol. 59, no.22, 2023.

[14] Hui Ma, Turgay Celik, and Heng-Chao Li, "FER-YOLO: Detection and Classification Based on Facial Expressions," in Proceedings of the International Conference on Image and Graphics, pp28-39, 2021.

[15] K. Tejaswi, D. Mokshith, E. Sai Pradeep, and K. Manoj Kumar, "Facial Expression Recognition Using YOLO," in 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp1-8, 2023.

[16] Y-I. Tian, Takeo Kanade, and Jeffrey F. Cohn, "Recognizing Action Units for Facial Expression Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no.2, pp97-115, 2001.

[17] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active Appearance Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no.6, pp681-685, 2001.

[18] John G. Daugman, "Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no.7, pp1169-1179, 2002.

[19] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no.12, pp2037-2041, 2006.

[20] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp886–893, 2005.

[21] Li-Yuan Chen, Chang-Jun Zhou, and Li-Ping Shen, "Facial Expression Recognition Based on SVM in E-Learning," IERI Procedia, vol. 2, pp781–787, 2012.

[22] Malika Arora and Munish Kumar, "AutoFER: PCA and PSO Based Automatic Facial Emotion Recognition," Multimedia Tools and Applications, vol. 80, no.2, pp3039–3049, 2021.

[23] Doha Taha Nor El-Deen, Rania Salah El-Sayed, Ali Mohamed Hussein, and Mervat S. Zaki, "Multi-label Classification for Sentiment Analysis Using CBGA Hybrid Deep Learning Model," Engineering Letters, vol. 32, no.2, pp340-349, 2024.

[24] Dhvanil Bhagat, Abhi Vakil, Rajeev Kumar Gupta, and Abhijit Kumar, "Facial Emotion Recognition (FER) Using Convolutional Neural Network (CNN)," Procedia Computer Science, vol. 235, pp2079–2089, 2024.

[25] Nur Nabilah Abu Mangshor, Norshahidatul Hasana Ishak, Muhammad Haiqal Zainurin, Nor Aimuni Md Rashid, Nur Farahin Mohd Johari, and Nurbaity Sabri, "Implementation of Facial Expression Recognition (FER) Using Convolutional Neural Network (CNN)," in 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC), pp92–96, 2024.

[26] Yan-De Li, Ming-Jie Wang, Ming-Lun Gong, Yong-Gang Lu, and Li Liu, "FER-former: Multimodal Transformer for Facial Expression Recognition," IEEE Transactions on Multimedia, 2024.

[27] Guo-Fang Chen, Zhao-Ying Chen, Yu-Liang Wang, Jin-Xing Wang, and Han-Qing Li, "Apple Flower Detection Method Based on Data-Augmented Deep Learning (in Chinese)", Journal of Chinese Agricultural Mechanization, vol. 43, no.5, p148, 2022.

[28] Ying Li, Peihua Song, "Research Progress of Transfer Learning in Medical Image Classification (in Chinese)", Journal of Image and Graphics, vol. 27, no.3, pp672-686, 2022.

[29] Rejin Varghese and M. Sambath, "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," in Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp1-6, 2024.

[30] Luyl-Da Quach, Khang Nguyen Quoc, Anh Nguyen Quynh, Hoang Tran Ngoc, and Nguyen Thai-Nghe, "Tomato Health Monitoring System: Tomato Classification, Detection, and Counting System Based on YOLOv8 Model with Explainable MobileNet Models Using Grad-CAM++," IEEE Access, vol. 12, pp9719-9737, 2024.

[31] Yin-Peng Chen, et al, "Dynamic Convolution: Attention Over Convolution Kernels," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp11030-11039, 2020.

[32] Yi-Chao Liu, Zong-Ru Shao, and Nico Hoffmann, "Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions," arXiv preprint arXiv:2112.05561, 2021.

[33] Yufei Li, et al, "Omni-Dimensional Dynamic Convolution Feature Coordinate Attention Network for Pneumonia Classification," Visual Computing for Industry, Biomedicine, and Art, vol.7, no.1, p17, 2024.

[34] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp94-101, 2010.

[35] Ian J. Goodfellow, et al, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013, Proceedings, Part III, pp117–124, 2013.

[36] Shan Li, Wei-Hong Deng, and Jun-Ping Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp2852–2861, 2017.

[37] Li Shan and Weihong Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," IEEE Transactions on Image Processing, vol. 28, no.1, pp356-370, 2018.

[38] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar, "A Survey of Synthetic Data Augmentation Methods in Machine Vision," Machine Intelligence Research, vol. 21, no.5, pp831–869, 2024.

[39] Hui Qi, Xi-Peng Zhang, Ying Shi, and Xiao-Bo Qi, "A Novel Attention Residual Network Expression Recognition Method," IEEE Access, vol. 12, pp24609–24620, 2024.

[40] Chun-Man Yan, Xiang Zhang, and Qing-Peng Wang, "Facial Expression Recognition Based on Improved MobileNetV2 (in Chinese)" , Computer Engineering & Science, vol. 45, no.6, p1071, 2023.

[41] Resmi K. Reghunathan, Vineetha K. Ramankutty, Amrutha Kallingal, and Vishnu Vinod, "Facial Expression Recognition Using Pre-trained Architectures," Engineering Proceedings, vol. 62, no.1, p22, 2024.

[42] Ozioma Collins Oguine, Kanyifeechukwu Jane Oguine, Hashim Ibrahim Bisallah, and Daniel Ofuani, "Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction," arXiv preprint arXiv:2206.09509, 2022.

[43] Swadha Gupta, Parteek Kumar, and Raj Kumar Tekchandani, "Facial Emotion Recognition Based Real-Time Learner Engagement Detection System in Online Learning Context Using Deep Learning Models," Multimedia Tools and Applications, vol. 82, no.8, pp11365–11394, 2023.

[44] Jia-Xiang Yu, Yi-Yang Liu, Rui-Yang Fan, and Guo-Bing Sun, "MixCut: A Data Augmentation Method for Facial Expression Recognition," arXiv preprint arXiv:2405.10489, 2024.

[45] Mustafa Can Gursesli, Sara Lombardi, Mirko Duradoni, Leonardo Bocchi, Andrea Guazzini, and Antonio Lanata, "Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets," IEEE Access, 2024.

[46] Cheng-Yan Yu, Dong Zhang, Wei Zou, and Ming Li, "Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition," IEEE Transactions on Affective Computing, 2024.

[47] Gaurav Meena, Krishna Kumar Mohbey, Ajay Indian, Mohammad Zubair Khan, and Sunil Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," Multimedia Tools and Applications, vol.83, no.6, pp15711-15732, 2024.

[48] Jagendra Singh, Akansha Singh, Krishna Kant Singh, Bechoo Lal, Harsh Verma, Niranjan Samudre, and Harsh Raperia, "Real-Time Convolutional Neural Networks for Emotion and Gender Classification," Procedia Computer Science, vol.235, pp1429-1435, 2024.

[49] Yang-Bo Chen, Chun-Yan Peng, Xuan Wang, and Yu-Hui Zheng, "Self-learning weight network based on label distribution training for facial expression recognition," IET Image Processing, vol. 19, no.1, pe13326, 2025.

[50] Mehmet Emin Konuk, and Erdal Kılıç, "EfficientFER: EfficientNetv2 Based Deep Learning Approach for Facial Expression Recognition," 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), pp1-7, 2025.

[51] G.Balachandran, S.Ranjith, G.C.Jagan, and T.R.Chenthil, "MAEL-FER: a multi-aspect enhancement learning framework for robust facial emotion recognition through integrated learning modules," International Journal of Machine Learning and Cybernetics, pp1-32, 2025.

**Ziyue Wang** is a postgraduate student at School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051 China. Her current research interests include deep learning and computer vision.

**Hang Yin** is Associate Professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051 China. His current research interests include computer vision, big data management and information Security.

**Honghe Xie** is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051 China. His current research interests include deep learning, network embedding, and meta human.

**Jiaqi Gu** is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051 China. Her current research interests include deep learning and computer vision.