Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048688

ARTICLE

E2E-MFERC: A Multi-Face Expression Recognition Model for Group Emotion Assessment

Lin Wang[1], Juan Zhao[2], Hu Song[3] and Xiaolong Xu[4,*]

[1]Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing, 210042, China

[2]School of Network Security, Jinling Institute of Technology, Nanjing, 211169, China

[3]State Grid Jiangsu Electric Power Company Limited, Nanjing, 210000, China

[4]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, 210042, China

*Corresponding Author: Xiaolong Xu. Email: xuxl@njupt.edu.cn

Received: 15 December 2023 Accepted: 01 March 2024 Published: 25 April 2024

ABSTRACT

In smart classrooms, conducting multi-face expression recognition based on existing hardware devices to assess students' group emotions can provide educators with more comprehensive and intuitive classroom effect analysis, thereby continuously promoting the improvement of teaching quality. However, most existing multi-face expression recognition methods adopt a multi-stage approach, with an overall complex process, poor real-time performance, and insufficient generalization ability. In addition, the existing facial expression datasets are mostly single face images, which are of low quality and lack specificity, also restricting the development of this research. This paper aims to propose an end-to-end high-performance multi-face expression recognition algorithm model suitable for smart classrooms, construct a high-quality multi-face expression dataset to support algorithm research, and apply the model to group emotion assessment to expand its application value. To this end, we propose an end-to-end multi-face expression recognition algorithm model for smart classrooms (E2E-MFERC). In order to provide high-quality and highly targeted data support for model research, we constructed a multi-face expression dataset in real classrooms (MFED), containing 2,385 images and a total of 18,712 expression labels, collected from smart classrooms. In constructing E2E-MFERC, by introducing Re-parameterization visual geometry group (RepVGG) block and symmetric positive definite convolution (SPD-Conv) modules to enhance representational capability; combined with the cross stage partial network fusion module optimized by attention mechanism (C2f_Attention), it strengthens the ability to extract key information; adopts asymptotic feature pyramid network (AFPN) feature fusion tailored to classroom scenes and optimizes the head prediction output size; achieves high-performance end-to-end multi-face expression detection. Finally, we apply the model to smart classroom group emotion assessment and provide design references for classroom effect analysis evaluation metrics. Experiments based on MFED show that the mAP and F1-score of E2E-MFERC on classroom evaluation data reach 83.6% and 0.77, respectively, improving the mAP of same-scale You Only Look Once version 5 (YOLOv5) and You Only Look Once version 8 (YOLOv8) by 6.8% and 2.5%, respectively, and the F1-score by 0.06 and 0.04, respectively. E2E-MFERC model has obvious advantages in both detection speed and accuracy, which can meet the practical needs of real-time multi-face expression analysis in classrooms, and serve the application of teaching effect assessment very well.

KEYWORDS
Multi-face expression recognition; smart classroom; end-to-end detection; group emotion assessment

1 Introduction

With the continuous development of educational informatization, the pace of smart classroom construction is gradually accelerating, especially the empowerment of artificial intelligence technology, which has also brought new opportunities for the development of smart classrooms. In this context, classroom effect evaluation is becoming increasingly important to ensure the steady development of education outcomes. Deep learning technologies are widely applied in classroom effect evaluation, and the related computer vision technologies are receiving more and more attention. In modern education, the intelligent analysis of students' participation and emotional behaviors is of vital importance to the evaluation of teaching effects. Psychological research shows that in human face-to-face communication, the proportion of language information transmission is about 7%, while non-verbal information such as facial expressions accounts for over 55% [1]. By studying multi-face expression recognition technology, we can enable intelligent devices and human-computer interaction systems to understand and interpret changes in students' facial expressions and emotions more accurately, intelligently and naturally. Emotion analysis and evaluation based on this can gradually become an important parameter for evaluating teaching effectiveness.

In smart classrooms, recognizing students' group facial expressions in a classroom environment can provide important references for evaluating teaching quality and learning outcomes. By analyzing the distribution of students' group facial expressions, we can understand the overall learning status, participation, and emotional attitudes of the students, thereby evaluating the emotional impact of teaching activities on the overall students, examining students' emotional feedback under different teaching methods, and improving teaching design accordingly to effectively promote the development of teaching quality. In addition, multi-face expression recognition can monitor students' emotional changes in real time, helping teachers understand students' learning status in real time, respond to negative emotions, and interact well with students. Meanwhile, distinguishing individual and group expression situations also helps to discover different reactions of different students to teaching activities and conduct personalized teaching guidance.

Earlier facial expression recognition relied on manually designed features and classifiers [2], with poor accuracy and robustness. Since the popularity of deep learning, various convolutional neural networks and recurrent neural networks have shown great advantages in facial expression feature representation and classification, improving recognition performance [3,4]. With the introduction of network architecture design, attention mechanisms, data augmentation and other technologies, facial expression recognition has extended from basic expressions to more refined expression analysis, and the technology has become more intelligent and generalizable. However, facial expression recognition research has been mainly focused on recognizing expressions from single persons under well-controlled conditions [5–7].

In contrast, multi-face expression recognition faces difficulties such as expression variations, occlusions, high synchronization requirements, and high performance demands [8]. Most existing

multi-face expression recognition methods adopt multi-stage processing, which is complex and inefficient. The accuracy and efficiency of multi-face expression recognition are limited, and there lacks high-quality and highly targeted multi-face expression datasets [9–11]. Increasingly more research applies facial expression recognition techniques to smart classrooms for analyzing student status and evaluating teaching quality, demonstrating the necessity of combining facial expression recognition and smart classrooms. However, in this field, methods based on deep learning have been extensively studied, but problems like low dataset specificity, insufficient training samples, and poor model generalization still restrict further development. Especially for multi-face expression recognition, most current solutions require additional face detection models, lacking real-time performance, and end-to-end solutions are still not mature enough. Moreover, in-depth research is lacking regarding group emotion assessment and teaching guidance driven by students' multi-face expressions, with most work staying at the experimental stage [12–15].

In summary, the existing work faces the following main problems:

- The quality and applicability of existing datasets are limited, resulting in insufficient generalization capability of facial recognition algorithms.
- Multi-face expression recognition algorithms mostly adopt multi-stage processing, which is complex in workflow. End-to-end solutions are not mature enough and the recognition efficiency and real-time performance need to be improved.
- The application effects in actual scenarios are poor, lacking combination with practical applications in teaching effect assessment.

Therefore, we propose an end-to-end multi-face expression recognition algorithm model for smart classrooms (E2E-MFERC), achieving end-to-end multi-face expression detection (MFED). We constructed a multi-face expression dataset from real classroom scenarios to support model research, and applied the model to smart classroom group emotion assessment, providing design references for classroom effect analysis evaluation metrics.

More specifically, the main contributions of our work are summarized as follows:

- We constructed a multi-face expression dataset in real classrooms: MFED, containing 2,385 images and a total of 18,712 expression labels, collected from smart classrooms. The dataset was accurately annotated and split using the hold-out method. MFED provided high-quality and highly targeted datasets for algorithm research. We also provided the commonly used VOC and TXT object detection annotation format labels, as well as labels in face rectangle segmentation annotation format, which can be used by related research on multi-face expression recognition, face detection, etc.
- We proposed an end-to-end multi-face expression recognition algorithm model for smart classrooms: E2E-MFERC. We used RepVGGBlock, SPD-Conv and C2f_Attention modules to form the backbone of E2E-MFERC; C2f_Attention is improved from the C2f module; AFPN feature fusion technology is combined with RepVGGBlock in the neck for representation enhancement; the output head size is scaled down. Through module optimization and ingenious combination, the advantages of each part are leveraged, making the algorithm model fast, efficient, high-performing, and easy to deploy. It realizes an efficient end-to-end multi-face expression detection solution, improving recognition efficiency and real-time performance.
- We applied the model to smart classroom group emotion assessment scenarios, used E2E-MFERC for multi-face expression recognition, designed a group emotion assessment scheme

based on multi-face expression recognition results, and provided calculation methods of evaluation metrics for classroom effect analysis, further demonstrating the practical value of the model.

2 Related Work

2.1 Facial Expression Recognition Techniques

Our research is based on facial expression recognition. Facial expression recognition techniques have evolved from traditional machine learning methods to deep learning methods. Earlier methods used manual feature extraction and classifiers for facial expression recognition, such as Lyons et al. [16] proposed a facial expression classification method based on tagged elastic graph matching, 2D Gabor wavelet representation and linear discriminant analysis for expression classification. Shan et al. [2] studied facial expression recognition based on local binary patterns (LBP). These early methods all had problems like low accuracy and poor robustness. With the rise of deep learning, facial expression recognition techniques have made qualitative leaps. Deep convolutional neural networks, recurrent neural networks and others have shown powerful capabilities in facial expression feature learning and classification tasks. For example, Zhang et al. [4] used a dual-layer recurrent neural network (RNN) model to provide an effective way to utilize the spatial and temporal correlations of input signals for emotion recognition. Khorrami et al. [17] demonstrated that convolutional neural networks (CNN) can achieve strong performance in facial expression recognition tasks. However, the effects and performance of facial expression recognition were poor, unable to meet higher requirements. By exploring combinations of neural networks with long short-term memory (LSTM), attention mechanisms, etc., to handle impacts from variations in image angles, lighting, and more, the performance of facial expression recognition has been improved. For instance, Liu et al. [6] proposed the AU-aware deep networks (AUDN) modeled with AU attention modules to guide the network to learn subtle facial changes. Jaiswal et al. [7] used a combination of convolutional and bidirectional long short-term memory neural networks in a deep learning way to jointly learn the shape, appearance and dynamics of facial expression features. Meanwhile, facial expression recognition tasks have also expanded from basic expression recognition to micro-expression and complex expression recognition. For example, Liu et al. [18] designed a multi-stream convolutional neural network (MSCNN) for micro-expression recognition, using eulerian video magnification (EVM) and optical flow to amplify and visualize the subtle motion changes in micro-expressions, and extracting masks from optical flow images. However, these methods have simple network structures and poor performance on complex tasks, and the facial expression datasets used are small in scale, resulting in insufficient model generalization capability. In recent years, researchers have also tried introducing new technologies like self-supervised learning, keypoint labeling, feature fusion, etc., into facial expression recognition, in order to alleviate data dependency and improve overall performance. Taherkhani et al. [19] trained convolutional neural networks in a semi-supervised manner for classification tasks; Haghpanah et al. [20] realized real-time facial expression recognition based on facial keypoints combined with neural networks. Hu et al. [21] extracted features from facial expression images using deep separable convolutional modules, fused the features to expand receptive fields, and obtained richer facial feature information. In summary, driven by the advancement of deep learning, facial expression recognition techniques are developing towards more accurate, comprehensive and intelligent directions, laying an important foundation for building intelligent teaching environments. However, current techniques focus mainly on single-person facial expression recognition in controlled environments. How to achieve accurate multi-face expression analysis for complex application scenarios remains a research emphasis.

2.2 Multi-Face Expression Recognition

Multi-face expression recognition is an important requirement for smart classroom scenarios. Compared with single-face expression recognition, multi-face has difficulties such as expression variations, occlusions, high synchronization requirements, performance demands, etc. Most multi-face expression recognition techniques adopt a multi-stage processing approach, first performing ace detection and localization, then extracting expression features for each face, and finally achieving expression classification. For example, Jung et al. [22] adopted an ensemble method, first conducting face detection and alignment, then classifying expressions; Li et al. [23] first performed face detection and keypoint extraction, followed by expression classification; In a similar face mask detection study, Kareem et al. [24] effectively applied a sequence model of Haar cascade classifiers to the construction of a face detector, which was used to identify the presence of face masks on faces based on face detection, and applied it to reduce the risk of COVID-19 transmission. Such methods have relatively complex processing flows and lower robustness, and accumulated detection errors also affect recognition results. Currently, some single-stage methods achieve facial expression recognition through multi-module fusion, focusing on feature expression, and introducing attention mechanisms or using contrastive learning methods. For example, Zhao et al. [25] proposed the global multi-scale local attention network. This network consists of three important components: A feature extractor, a multi-scale module, and a local attention module for facial expression recognition. Chen et al. [26] proposed a feature fusion residual attention network that integrates global and local expressive features through a feature fusion module and establishes residual links between the input and output. Xia et al. [27] proposed a three-level hierarchical structure based on Transformer, which combines multi-scale spatio-temporal aggregation for dynamic facial expression recognition. These methods improve the overall efficiency of the model. However, overall, multi-face expression recognition still faces problems such as insufficient training data and occlusion interference. Its accuracy and efficiency need to be improved. There are still many challenges to achieve truly adaptive, end-to-end, and efficient multi-face expression recognition in complex environments.

2.3 Multi-Face Expression Recognition for Smart Classrooms

For the complex classroom environment, researchers have proposed applicable multi-face expression recognition methods. For example, Bie et al. [28] improved YOLOv5 based on feature enhancement ideas, effectively extracted and fused features, and applied it to classroom teaching scenarios to recognize students' facial expressions. Trabelsi et al. [29] proposed an improved multi-head attention-based facial expression recognition model to identify students' expressions in the classroom. Many researchers focused on data augmentation and the combination of multi-source heterogeneous information, fusing multi-modal information such as audio, text, and body gestures for emotion understanding. And evaluate the teaching effect of smart classrooms from multiple aspects such as student behavior analysis, teaching quality monitoring, and student emotional state, to provide a basis for educational decision-making. For instance, Palash et al. [30] proposed an interpretable multi-modal emotion recognizer with situational knowledge, using visual information for human emotion recognition and explanation. Gupta et al. [31] evaluated three modalities based on deep learning methods from real-time video streams, including facial expressions, blink count, and head motion, to predict student status. Chen et al. [32] proposed a class expression recognition model based on spatio-temporal residual attention network, and used deep convolutional neural networks to capture student behaviors, combining student expressions and behaviors to intelligently evaluate classroom status. These methods have preliminarily improved the accuracy of facial expression recognition in classroom scenarios, and have demonstrated the possibility and urgency of in-depth integration

of facial expression recognition and analysis technologies with smart classrooms. However, the introduction of more information undoubtedly reduces efficiency. In addition, due to the changing factors such as individual differences and scene changes in the classroom, the effects and real-time performance of multi-face expression recognition for smart classrooms need to be improved. At the same time, the application effects in actual scenes are poor due to the limitations of dataset specificity.

Today, the rapid development of deep learning has enabled algorithm models such as Single Shot MultiBox Detector (SSD) [33], Faster-RCNN (FR-CNN) [34], Region-based Fully Convolutional Networks (R-FCN) [35], You Only Look Once (YOLO) [36] and others to achieve fruitful results in object detection tasks. In particular, the YOLO algorithm, owing to its single-shot regression characteristic, has a series of advantages such as high speed, good performance, and simple network structure. In recent years, it has received increasing attention and research, and more and more practitioners have gradually applied it to fine-grained object detection tasks, which also inspired our algorithm design.

3 Dataset Construction

Most publicly available facial expression datasets are single-face datasets, while attempts to synthesize multi-face datasets lack details and authenticity, and cannot fully support algorithm research for smart classroom scenarios. Therefore, we constructed a high-quality multi-face expression dataset in real classrooms: MFED, to promote research on multi-face expression recognition techniques for smart classroom scenarios. In the process of constructing the dataset, we have fully considered and designed data collection, image preprocessing, and expression label annotation to ensure the quality of the data to the greatest extent possible and meet our needs for features such as multiple faces, real scenes, diversity, high quality, and easy promotion in the dataset.

3.1 Data Collection

Considering the specificity and practicality of the dataset, this study chose university students in real smart classrooms as data collection objects, collecting 2,721 images and screening out 2,385 valid images with 18,172 facial expression labels. The collection involved 81 university students, divided into 10 groups of 6–10 people, and crossover replacement between groups was carried out on this basis, forming 34 group sessions. Student positions were also alternated within groups, corresponding expressions were provided in response to different classroom content and evaluation feedback needs, effectively increasing data diversity and ensuring reliability. In selecting collection objects, we fully considered factors like student gender, nation, age, glasses-wearing, hat-wearing, facial occlusion, etc., to maximize coverage of the dataset. For the collection environment, we chose 5 real smart classrooms as collection scenarios, and considered different time periods, complex lighting conditions, positions, angles, atmospheres, classroom content, etc., to retain the authentic classroom environment and process to the greatest extent. The whole dataset collection lasted 3 weeks, using intelligent classroom cameras, cameras, webcams and other devices for image capture.

3.2 Image Processing

Out of the originally collected 2,721 images, we performed filtering to exclude low-quality images, such as those that were blurry, heavily occluded, or had high content redundancy. As a result, we retained 2,385 high-quality and valid images. To ensure consistency, these images were uniformly resized to a width of 3,024 pixels, serving as input data for multi-face expression analysis. Fig. 1 shows partial images in the dataset.

Figure 1: Sample images of our MFED dataset

3.3 Expression Label Annotation

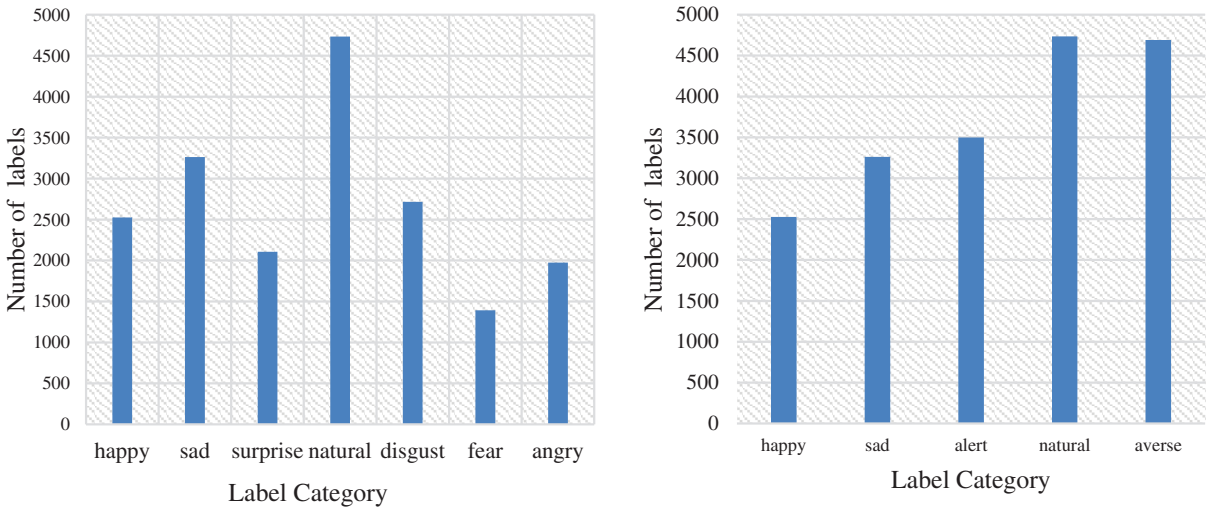
After image data preprocessing, we manually annotated all images. To meet the end-to-end multi-face expression recognition requirement, we simultaneously annotated the location coordinates and expression categories for multiple faces in each image. The annotation format for a single facial expression is: (category_label,x,y,w,h), where category_label is the numeric code corresponding to the expression category. In the basic class data, the 7 expressions of happy, sad, surprise, natural, disgust, fear, angry correspond to numeric codes 0–6, respectively; in the classroom evaluation data, the 5 expressions of happy, sad, alert, natural, averse correspond to numeric codes 0–4, respectively. x and y are the normalized horizontal and vertical coordinate values of the center point of the annotated object. w and h are the normalized width and height values of the bounding box of the annotated object. The calculation methods for each parameter are shown in Eq. (1), (X1,Y1) and (X2,Y2) are the coordinates of the top-left and bottom-right corners of the bounding box of the annotated object. W and H are the width and height of the original image, respectively. The annotation data of multiple facial expressions in each image are saved as multi-face expression labels in TXT files corresponding to the images, with one annotation per line.

$$x = \frac{X_1 + X_2}{2W} \quad y = \frac{Y_1 + Y_2}{2H} \quad w = \frac{X_2 - X_1}{W} \quad h = \frac{Y_2 - Y_1}{H} \tag{1}$$

For expression category annotation, first according to mainstream research, we included 7 basic expressions: Happy, sad, surprise, natural, disgust, fear, angry, as the foundation categories. Meanwhile, to meet research needs, we also used 5 expression categories commonly seen in classroom scenarios that have important impacts on classroom evaluation: Happy, sad, alert, natural, averse, to process the annotations. These 5 basic expression categories, on the one hand, can well cover the main expression categories of students in the classroom, and balance positive and negative expressions, while also considering intermediate states, which effectively reflect students' emotional states in the classroom: "Happy", "Sad", "Alert" are mainly facial expressions in response to special teaching

scenarios, content and forms; "Natural" is an ordinary neutral expression state, while "Averse" is a negative expression state. By analyzing the matching degree between expression states and teaching scenarios, content and forms, classroom effects can be effectively analyzed. On the other hand, these 5 expressions have clear distinctions, which facilitates model recognition, and the annotation difficulty is not high, which benefits application promotion. Analyzing student emotions and evaluating classroom effects based on these 5 expressions, compared to simple expressions like "Attentive" and "Distracted", can more comprehensively and profoundly analyze students' real emotional states, which essentially interprets "Attentive" or "Distracted" states. And these expressions are relatively hard to fake, which facilitates accurate recognition. These indicators can also more accurately and flexibly adapt to classroom effect analysis under different teaching scenarios, content and forms.

To ensure annotation quality, preliminary annotations were first conducted by the collection subjects on their own expressions. On this basis, 3 personnel who were not collection subjects conducted secondary independent labeling on all images. For labels with major disputes, the final labels were determined through cross-checking with the subjects themselves and comprehensive evaluation with other subjects in the same image. Finally, the average of the 3 annotation results was taken as the standard label for the sample, provided in both TXT and VOC label formats. Statistically, the 2,385 images contain a total of 18,712 face samples. The distribution of annotated instance quantities is shown in Fig. 2. The distribution of the center coordinates and bounding box sizes of the annotated objects is shown in Fig. 3. Subfigure (a) shows the distribution of the x-axis and y-axis values of the center coordinate of the annotated object, while subgraph (b) shows the width and height distribution of the bounding box of the annotated object. It can be seen that the distribution of annotated object positions is uniform, the sizes are reasonable, mainly small objects, conforming to reality.



(a) Basic expression classes (b) Classroom evaluation expression classes

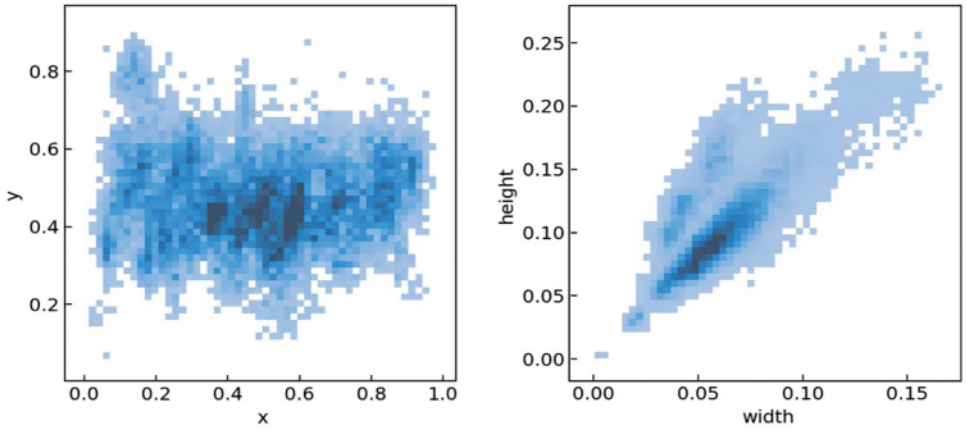
Figure 2: Bar chart of annotated instance quantities in MFED

4 Method

4.1 Framework

Currently, most multi-stage methods first introduce face detection or segmentation models to extract multi-face and convert them into single face, and then use expression recognition models for

classification and recognition. The use of multiple steps and models involves multiple encode-decode processes and relies on multiple loss constraints to obtain high-quality prediction results, which limits the improvement of algorithm efficiency and generalization capability, and cannot meet the real-time needs in actual application scenarios. In order to achieve end-to-end multi-face expression recognition and solve the above problems, we need to find a way to transform the complex multi-step face detection and expression classification problem adopted in a large amount of previous research into one step. For this purpose, inspired by the model framework design of the YOLO algorithm, we propose an end-to-end multi-face expression recognition algorithm model for smart classrooms: E2E-MFERC. It treats different facial expressions on the same image as detection objects, and simultaneously performs location, size detection and classification, converting multiple encode-decode processes into one solution. This allows greater room for improvement in both algorithm efficiency and accuracy.



(a) Center coordinate distribution (b) Bounding box size distribution

Figure 3: Distribution graphs of center coordinates and bounding box sizes of annotated objects in MFED

The overall framework of E2E-MFERC is shown in Fig. 4. Considering the characteristics of facial expression objects like small feature differences, high fine-grained recognition requirements, low obviousness of regional differences, category overlaps, and requirements like high efficiency, lightweight, and multiple demands on recognition effects in actual application of the algorithm. Our backbone structure adopts RepVGGBlock, SPD-ConvBlock and the improved C2f_Attention module to leverage the advantages of each part, expand the receptive field, reduce model parameters, while enhancing feature representation capability, making the model highly performant, with good recognition effects and high deployment efficiency. C2f_Attention is optimized from the Cross Stage Partial Network Fusion (C2f) module by introducing Polarized Self-Attention mechanisms. The application of attention mechanisms enables the model to focus on features and subsets more important for the multi-face expression recognition task. To further enhance the model's representation capability and recognition performance and strengthen detection of small objects, we adopt AFPN feature fusion technology in the model neck structure, introduce RepVGGBlock for representation enhancement, and scale down the output head size.

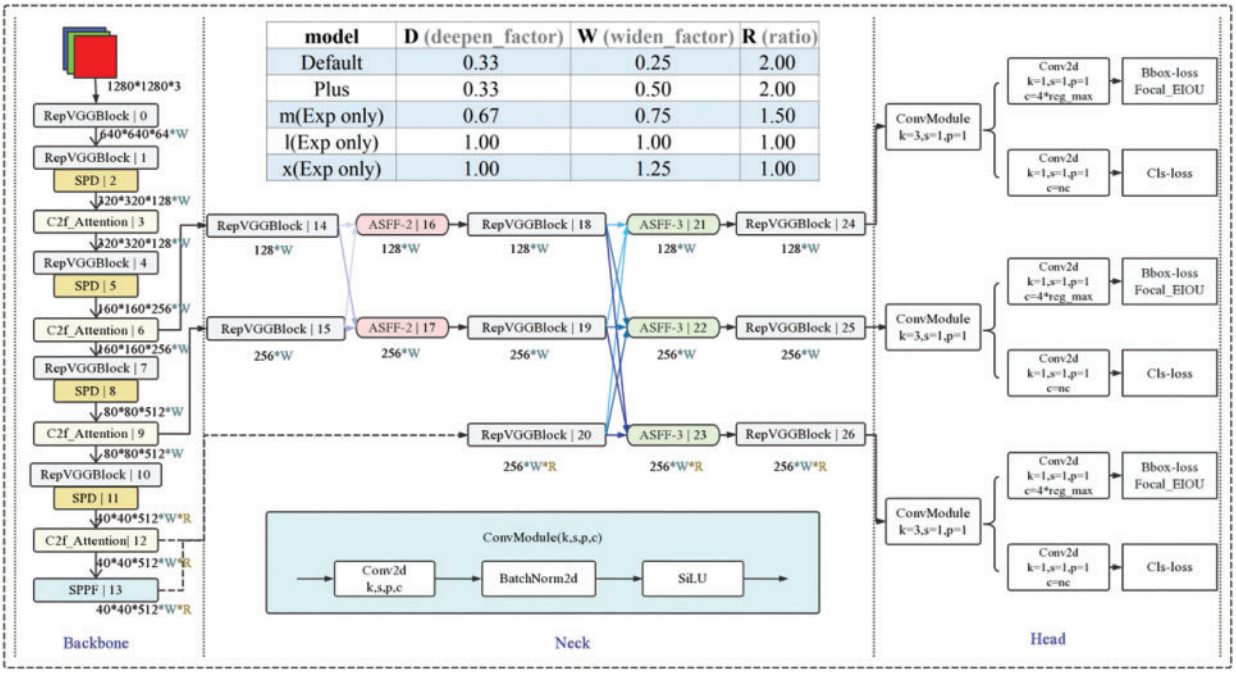


Figure 4: Framework of E2E-MFERC

E2E-MFERC uses the backbone network structure to extract features from preprocessed images or videos and obtain feature maps. Feature maps of three different scales are selected from the feature maps, and feature fusion is performed through the progressive feature pyramid network structure in the neck part to obtain three new feature maps. The new feature maps are input into the decoupled prediction branches in the head part to obtain the predicted boxes and confidences. Non-maximum suppression is performed on the predicted boxes to obtain multi-face expression recognition results. Finally, we calculate the group emotion indices based on the multi-face expression recognition results to obtain the classroom effect evaluation metrics. Next, we will elaborate the construction characteristics and advantages of each structure of E2E-MFERC in detail.

4.2 Representation Enhancement

In order to meet the important requirements of speed and accuracy for multi-face expression recognition oriented to smart classrooms, we introduced RepVGG Block and SPD-Conv modules in the model to enhance representational capability and improve the overall performance of E2E-MFERC. RepVGG is a lightweight network structure, whose innovation lies in drawing on the idea of residual neural network's (ResNet) residual structure design [37]. We use RepVGGBlock in both the backbone and neck of the model. During training, a modular structure with branched paths and final element-wise addition is used. This residual structure can perform high-quality feature extraction. During inference, it is converted into an equivalent convolution to significantly reduce the model's computational cost and memory consumption. This ensures accuracy while effectively improving algorithm efficiency. The use of this module provides more room to expand network width and depth under the same computational cost, allowing greater space to improve network performance. The module structure of RepVGGBlock and the conversion from multi-path models during training to single-path models during inference is illustrated in Fig. 5. 1×1 convolution layers are converted into

3 × 3 convolution layers by setting the parameters to zero; identity layers are first converted into 1 × 1 convolution layers by setting the convolution kernel parameters of the current channel to 1 and other channels to 0, then further converted into 3 × 3 convolution layers; finally, merging several 3 × 3 convolution layers can convert the multi-branched into equivalent convolutions.

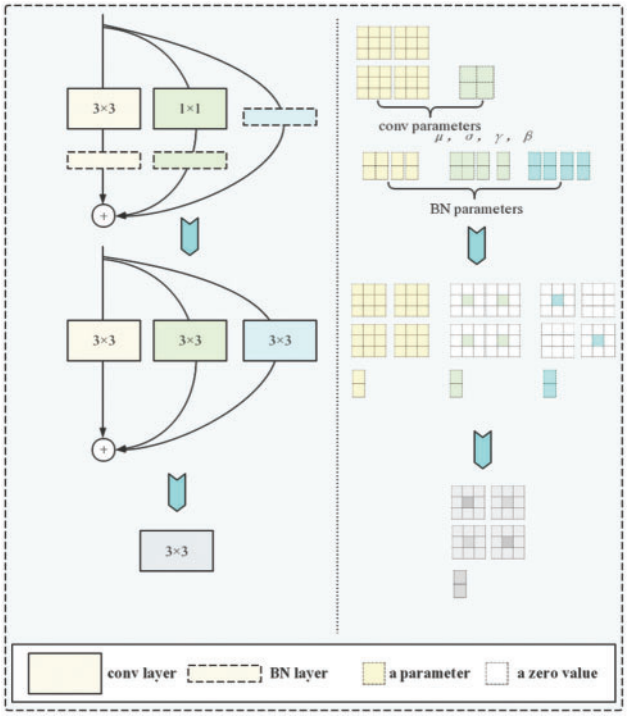


Figure 5: Structure and path conversion of RepVGGBlock

When combining the convolution layer and batch normalization layer, the convolution formula is:

$$Conv(x) = Q * x + b \tag{2}$$

where Q is the weight, and b is the bias.

The batch normalization layer formula is:

$$BN(x) = \gamma \times \frac{x - mean}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{3}$$

where γ and β are learning parameters, *mean* is the mean of the batch sample data, σ is the variance, ε is an extremely small but non-zero number.

Substituting the convolution layer result into the $BN(x)$ result is:

$$BN(Conv(x)) = \gamma \times \frac{Q * x + b - mean}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{4}$$

Let $BN(Conv(x)) = y$, Q_{fused}, b_{fused} be as, respectively:

$$Q_{\text{fused}} = \frac{\gamma \times Q * x}{\sqrt{\sigma^2 + \varepsilon}} \tag{5}$$

$$b_{fused} = \frac{\gamma \times (b - mean)}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{6}$$

The final result can be obtained with:

$$y = Q_{fused} * x + b_{fused} \tag{7}$$

Similar to YOLO, most algorithms perform downsampling operations like strided convolutions or pooling with increased strides, which can cause problems like loss of fine-grained information and inefficient feature representation learning, affecting algorithm performance. We introduce the SPD-Conv module, which consists of a spatial-to-depth (SPD) layer and a non-strided convolution (Conv) layer [38], enhancing the model's perception capability and feature representation capability to better solve these problems. This module introduces a spatial depth operation, whose mapping is:

$$f_{0,0} = X[0{:}\ S{:}\ scale, 0{:}\ S{:}\ scale], f_{1,0} = X[1{:}\ S{:}\ scale, 0{:}\ S{:}\ scale], \ldots,$$

$$f_{scale-1,0} = X[scale - 1{:}\ S{:}\ scale, 0{:}\ S{:}\ scale];$$

$$f_{0,1} = X[0{:}\ S{:}\ scale, 1{:}\ S{:}\ scale], f_{1,1}, \ldots,$$

$$f_{scale-1,1} = X[scale - 1{:}\ S{:}\ scale, 1{:}\ S{:}\ scale];$$

$$\ldots$$

$$f_{0,scale-1} = X[0{:}\ S{:}\ scale, scale - 1{:}\ S{:}\ scale], f_{1,scale-1}, \ldots,$$

$$f_{scale-1,scale-1} = X[scale - 1{:}\ S{:}\ scale, scale - 1{:}\ S{:}\ scale]. \tag{8}$$

In our backbone network, RepVGGBlock is ingeniously combined with SPD-Conv to replace commonly used strided convolutions or pooling, avoiding the negative impacts of losing fine-grained information and inefficient feature representation learning. By introducing more extensive contextual information, the convolution kernels can more comprehensively capture the relationships between features, improving the model's perception capability and feature representation capability, thereby enhancing the extraction and understanding of subtle features. This adapts to the requirements of facial expression objects having small feature differences and high fine-grained recognition needs.

In order to effectively avoid image distortion caused by cropping and scaling operations on image regions, and solve the problem of convolutional neural networks extracting graph-related redundant features, the model adopts the spatial pyramid pooling—fast (SPPF) module. The structure of this module is shown in Fig. 6. After passing through a ConvModule, three MaxPooling operations are performed in parallel. The feature map without MaxPooling and the feature map obtained after each additional MaxPooling are concatenated, and finally output after passing through another ConvModule. It should be noted that our algorithm has actually focused on lightweight design, so we use this module when the algorithm model is relatively large.

4.3 C2f Optimization

The C2f module structure is shown in Fig. 7. First pass through a ConvModule, the structure of ConvModule is shown in Fig. 4, consisting of a convolution layer + Batch Normalization layer + SiLU activation function; then use the chunk function to evenly split the output into two vectors; input the latter half to the Bottleneck Block, which contains n Bottlenecks, concatenate the output of each Bottleneck to a list, so the concatenated output channel number is $0.5 \times (n + 2)$; then output through a ConvModule, so the output is h×w×c_out. Compared with traditional detection algorithms, C2f has

significant advantages in both detection accuracy and speed, which plays a vital role in the construction of our algorithm model.

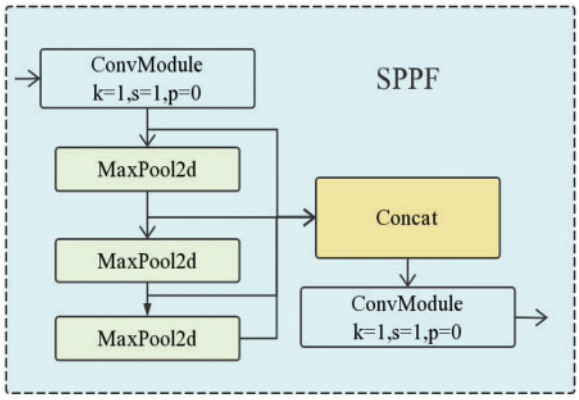


Figure 6: Structure of SPPF

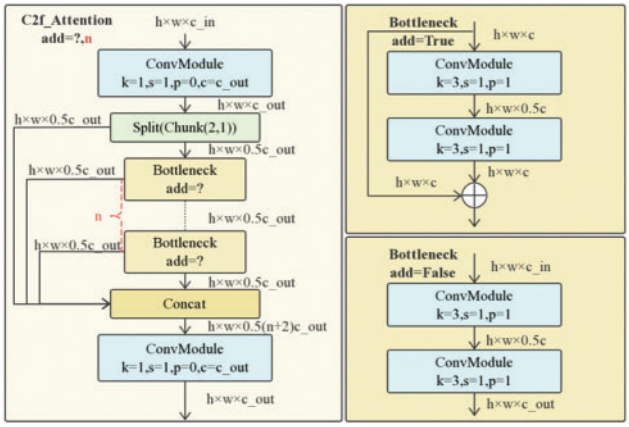


Figure 7: Structure of the C2f module

Attention mechanisms originate from research on human vision. In cognitive science, due to bottlenecks in information processing, humans selectively focus on part of the available information while ignoring other visible information. To reasonably utilize the limited visual information processing resources, humans need to select specific parts in the visual area and concentrate attention on them. For example, when performing facial expression recognition, humans usually only pay close attention to the relevant facial information. Considering the important requirement of key information extraction for multi-face expression recognition in complex actual scenarios, the introduction of attention mechanisms through weight allocation can help algorithm models selectively focus on parameters that have greater impacts on final results like humans, improving the overall performance of the algorithm. After comparing several attention mechanisms suitable for algorithm requirements, we introduced the Polarized Self-Attention (PSA) module in the C2f module for optimization. PSA is a self-attention mechanism, whose main feature is decomposing the attention vector into two subspaces, highlighting pixel classifications from the channel perspective, while detecting pixel positions belonging to the same semantics as much as possible from the spatial perspective. It uses an orthogonal method to make the information more complete without increasing computational complexity, reducing computational

and optimization difficulties, improving model training efficiency, and is highly compatible with our algorithm requirements. The detailed structure of this module is shown in Fig. 8. Its characteristic is completely folding features in one direction while maintaining high resolution in its orthogonal direction.

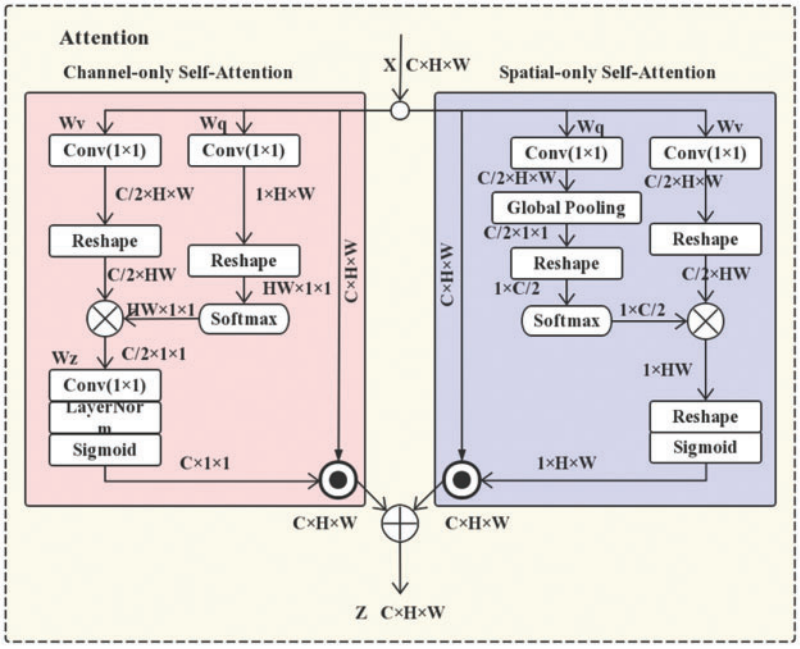


Figure 8: Structure of polarized self-attention block

The calculation method for the channel-only branch is:

$$A^{ch}(X) = F_{SG}\left[W_{z|\vartheta_1}(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X))))\right] \tag{9}$$

The calculation method for the spatial-only branch is:

$$A^{sp}(X) = F_{SG}\left[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X)))\right] \tag{10}$$

W_q, W_k, W_v are 1×1 convolution layers, ϑ_1 is intermediate parameters for channel convolution. σ_1, σ_2, σ_3 are three reconstruction operations. F_{SM} is the softmax operator as:

$$F_{SM}(X) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j \tag{11}$$

F_{GP} is global pooling operation:

$$F_{GP}(X) = \frac{1}{H \times W} \sum_{H}^{i=1} \sum_{W}^{j=1} X(:, i, j) \tag{12}$$

The output of the channel branch Z^{ch} is:

$$Z^{ch} = A^{ch}(X) \odot^{ch} X \in R^{C \times H \times W} \tag{13}$$

The output of the spatial branch Z^{sp} is:

$$Z^{sp} = A^{sp}(X) \odot^{sp} X \in R^{C \times H \times W} \tag{14}$$

We adopt a parallel manner to fuse and output the two branches: $PSA_P(X) = Z^{ch} + Z^{sp}$.

The C2f_Attention module structure optimized with attention mechanism is shown in Fig. 9. Thanks to two features of PSA module, one is Polarized filtering: Keeping high internal resolution in both channel and spatial attention computation while completely collapsing input tensors along their counterpart dimensions, the other is composing non-linearity that directly fits the output distribution of typical fine-grained regression, it effectively enhances the algorithm's ability to extract key information. It has stronger representation capabilities, which significantly facilitates solving several prominent problems in multi-face expression recognition.

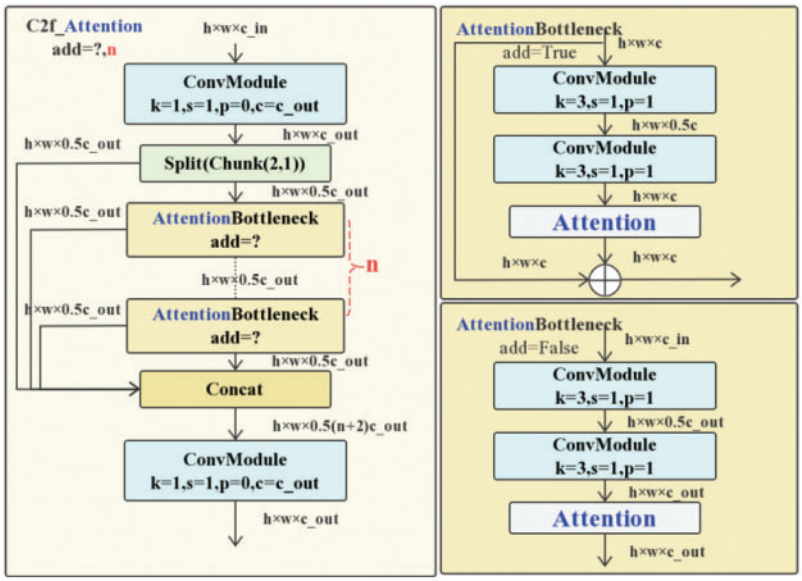


Figure 9: Structure of the C2f_Attention module

Table 1 shows the "add" and "n" parameters corresponding to the 4 C2f_Attention modules in E2E-MFERC.

Table 1: C2f_Attention module parameters

ID	Layers	Out channel	add	n
1	3	128 ∗ W	**TRUE**	**3 ∗ D**
2	6	256 ∗ W	**TRUE**	**6 ∗ D**
3	9	512 ∗ W	**TRUE**	**6 ∗ D**
4	12	512 ∗ W ∗ R	**TRUE**	**3 ∗ D**

4.4 Feature Fusion

In the mainstream YOLO series algorithm models, the common way of feature fusion is upsampling followed by concatenation with adjacent level features. This can cause problems like feature information degradation. For general object detection, the differences between target categories are large, so the negative impacts are not obvious. However, for multi-face expression recognition, the differences between expression objects we want to recognize are small, requiring high-quality feature information passing. Therefore, we adopt the Asymptotic Feature Pyramid Network (AFPN) structure [39] for feature fusion. It first fuses adjacent low-level features, then progressively incorporates high-level features during fusion, rather than crudely directly concatenating and fusing non-adjacent level features with large differences, aiming to maximize high-quality feature information passing. Meanwhile, this method uses the ASFF structure to allocate different spatial weights to non-same-level features during multi-level fusion, increasing the importance of key levels and reducing contradictory information. As shown in Fig. 10, it illustrates 2-level and 3-level feature fusion, respectively.

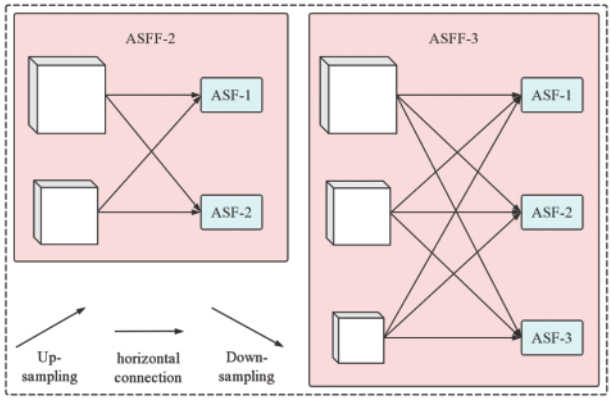


Figure 10: Structure of ASFF

In E2E-MFERC, we first fuse the adjacent low-level features from layers 6 and 9 outputs of the backbone network and after downsampling by RepVGGBlock, using the ASFF-2 module. Then the high-level features from layer 13 output and after downsampling by RepVGGBlock are incorporated through ASFF-3 for fusion. In the multi-face expression recognition scenario for classrooms, facial expression objects are generally small as recognition targets. To improve the detection effects of the algorithm model on small objects, we adjusted the output head size. Referring to the YOLO design, we used RepVGGBlock to control the output head size instead, discarded the large object detection output head sized 20×20, and added the small object detection output head sized 160×160, which effectively improved the overall performance of the algorithm model.

4.5 Loss Function

For the head of E2E-MFERC, we adopted a decoupled head design during prediction, whose structure is shown in the head part of Fig. 4. The classification and detection heads are separated, and the idea of distributional focal loss (DFL) is also used, so the number of channels in the regression head becomes $4 * reg\_max$, to meet the needs of our algorithm to value both classification and detection.

For the loss function, we use varifocal loss (VFL) in the classification head, calculated as:

$$VFL(p,q) = \begin{cases} -q(q\log(p) + (1-q)\log(1-p) & q > 0 \\ -\alpha p^{\gamma}\log(1-p) & q = 0 \end{cases} \tag{15}$$

where q is the label, $q = 0$ for negative samples. It has the characteristics of focal loss, and can effectively focus the network on high-quality samples.

The regression loss function used in the detection head includes DFL and Focal-efficient intersection over union (Focal-EIOU) [40] which performs better for the E2E-MFERC algorithm. Among them, DFL mainly models the box position as a general distribution, allowing the network to quickly focus on the distribution of positions close to the target position. It is calculated as:

$$DFL(S_i, S_{i+1}) = -((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1}) \tag{16}$$

Which means optimizing the probabilities of the closest left and right positions to the label y in the form of cross-entropy, thereby allowing the network to quickly focus on the distribution of areas adjacent to the target position.

For our algorithm improvement, we also compared multiple different mainstream IOU loss functions, and finally chose Focal-EIOU loss with better performance. This loss function considers overlap loss, center distance loss, width and height loss, resulting in faster convergence.The EIOU Loss is calculated as:

$$\begin{aligned} L_{EIOU} &= L_{IOU} + L_{dis} + L_{asp} \\ &= 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \end{aligned} \tag{17}$$

where C_w and C_h are the width and height of the smallest enclosing box covering the two boxes.

Considering the problem of imbalanced training samples also exists in BBox regression, Focal-EIOU loss separates high-quality anchors and low-quality anchors from the perspective of gradients. It is calculated as:

$$L_{\text{Focal-EIOU}} = IOU^{\gamma} L_{EIOU} \tag{18}$$

where $IOU = |A \cap B| / |A \cup B|$, γ is a parameter controlling the degree of outlier suppression. As can be seen from the formula, the higher the IOU, the greater the loss, which is equivalent to a weighting effect and helps improve regression accuracy.

4.6 Model Application

In order to analyze classroom effects, we apply the model to group emotion evaluation scenarios oriented to smart classrooms, and design classroom effect evaluation index calculation methods as parameters to measure teaching outcomes. First, we use E2E-MFERC to recognize students' multi-face expressions. Input sources can be images, videos, camera real-time data, etc. Fig. 11 shows the results of multi-face expression recognition using E2E-MFERC.

When recognizing students' multi-face expressions, the program saves the recognition results in CSV file format for more intuitive analysis of group emotions and to provide source data for calculating classroom effect evaluation indices. Table 2 shows a schematic of multi-face expression detection results per 1 s.

Figure 11: Multi-face expression recognition results

Table 2: Multi-face expression recognition results

Id	Happy	Sad	Alert	Natural	Averse	Sum
1	2	1	2	4	0	9
2	1	1	3	4	0	9
3	2	1	3	2	1	9
...	...	...	...	...	...	...
87	4	0	0	4	1	9
89	5	0	2	2	0	9
90	5	0	1	3	0	9
SUM	260	121	101	314	8	804

We designed index calculation methods and procedures for classroom effect evaluation based on group emotion analysis, using the multi-face expression recognition results by the model as input data. C is the numeric encoding corresponding to the expression category: Happy, sad, alert, neutral, averse correspond to numeric codes 0–4, respectively; the expected proportion parameters P for emotions corresponding to customized classroom content, P_C is the expected proportion for each expression, this parameter reflects whether the feedback to the classroom content meets expectations, rather than simply categorizing expressions as positive or negative; NUM is the total number of expressions, NUM_C is the number for corresponding category; set an allowable deviation factor D, D_C is generally a small number in (0, 0.05], with a default value of 0.05, taking effect when P_C is 0, the product of expression number and it indicates the allowable deviation, also preventing zero denominators.

A is the ratio of actual percentage to set percentage for expression quantities, calculated as:

$$A_C = \frac{NUM_C}{NUM(P_C + D_C)} \tag{19}$$

Set expression penalty parameters M, these parameters mainly evaluate A_C, with default values set as $M_0 = \infty$, $M_1 = 2$, $M_2 = 2$, $M_3 = 2$, $M_4 = \infty$, when A_C exceeds the penalty parameter M_C for that category, the scoring direction or weight intensity will be changed.

G is the weight parameter using a piecewise function to define the scoring intensity and direction more closely to reality, calculated as:

$$G_C = \begin{cases} (A_C)^2 & 0 < \alpha_C \leq 1 \\ (A_C)^{\frac{1}{3}} & 1 < \alpha_C < M_C \\ -(A_C)^{\frac{1}{3}} & \alpha_C \geq M_C \end{cases} \tag{20}$$

S is the final calculated total score, calculated as:

$$S = \frac{\sum_{C=0}^{3} G_C NUM_C - G_4 NUM_4}{NUM} \tag{21}$$

The final score can be used as a parameter for horizontal comparison between different teachers, and for vertical comparison and evaluation of classroom effects over time based on custom scoring levels. Taking the results NUM = [260, 121, 101, 314, 8] shown in Table 2 as input, using default penalty parameters M, and setting different expected proportion parameters P for emotions, the calculation results of index S are demonstrated in Table 3. From the results, it can be seen that the closer the overall classroom emotion proportion is to the set expected proportion parameters, the more positive emotions, the less negative emotions, the better the classroom effect, and the larger the index parameter value. This verifies the applicability of our index calculation design.

Table 3: Calculation results of index S with different emotion expected proportion parameters P

ID	P	D_C	G	S
1	[0.32, 0.15, 0.13, 0.39, 0.01]	0.02	[1.00, 1.00, 0.93, 1.00, 0.99]	0.97
2	[0.50, 0.15, 0.10, 0.25, 0.00]		[0.41, 1.00, 1.07, 1.16, 0.24]	0.87
3	[0.40, 0.05, 0.15, 0.40, 0.00]		[0.65, −1.44, 0.70, 0.95, 0.24]	0.45
4	[0.50, 0.05, 0.05, 0.4, 0.00]		[0.41, −1.44, −1.34, 0.95, 0.25]	0.12

5 Experimental Results and Discussion

5.1 Dataset and Preprocessing

In experiment, we use our self-constructed MFED as the dataset. MFED contains 2,385 images and 18,712 facial expression samples. The expression categories adopt two sets of classification criteria as described previously, and each sample is simultaneously annotated with the precise face rectangle location and size. To facilitate the research of this model and improve the specificity and applicability of the dataset, we adopted the hold-out method to split the dataset into training and validation subsets with a ratio of 8:2.

The allocation of the dataset into basic expressions and classroom scenario expressions categories is shown in Figs. 12 and 13, respectively, with specific data as shown in Table 4. It can be seen that our constructed dataset has sufficient number of images and abundant facial expression samples; the category distribution conforms to the real environment in actual classrooms, and the number of faces contained in each image is also adequate. These ensure sufficient scale and diversity of the dataset, improve its applicability, and provide a solid foundation for algorithm research.

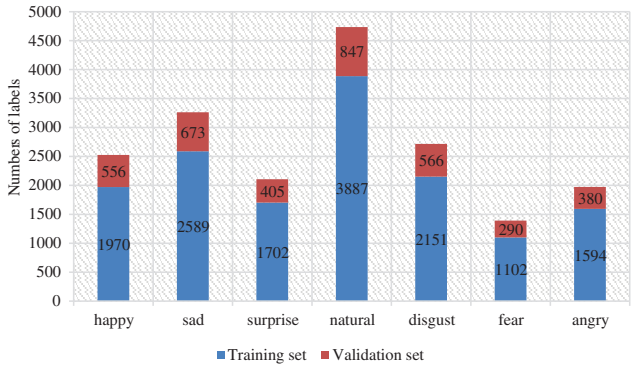


Figure 12: Basic expression category allocation in dataset

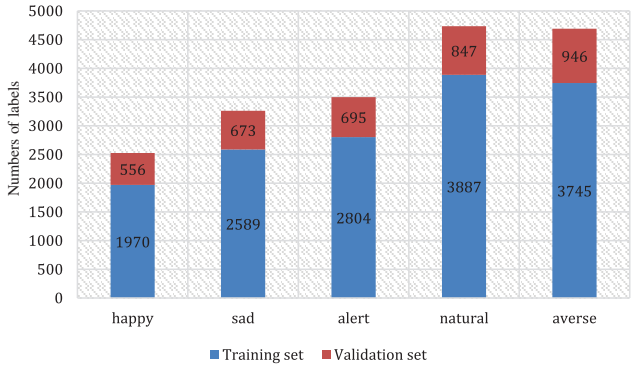


Figure 13: Classroom expression category allocation in dataset

Table 4: Dataset statistical characteristics and allocation details

Parameter		Training set	Validation set	Statistical value
Number of images		1908	477	2385
Number of emoticon samples		14995	3717	18712
Average number of faces		7.9	7.8	7.8
Expression category		7 for base/5 for class		
Category_label-base	Happy	1970	556	2526
	Sad	2589	673	3262
	Surprise	1702	405	2107
	Natural	3887	847	4734
	Disgust	2151	566	2717
	Fear	1102	290	1392
	Angry	1594	380	1974
Category_label-class	Happy	1970	556	2526
	Sad	2589	673	3262
	Alert	2804	695	3499

(Continued)

Table 4 (continued)

Parameter		Training set	Validation set	Statistical value
	Natural	3887	847	4734
	Averse	3745	946	4691

5.2 Implementation Details

The experimental operating system is Windows, the CPU is Intel i7-10700F 2.90 GHz, the memory is 16 G, and the GPU is NVIDIA GeForce RTX 2060 SUPER. The Python version is 3.9.13, and the deep learning framework used is PyTorch 1.10.0. In this experiment, the input image size is $1280 * 1280$. The initial batch size is set to 8, and later adjusted accordingly based on GPU performance to keep the batch size the same under comparable model scales. The AdamW optimizer with an initial learning rate $lr = 1e\text{-}3$ is used. The number of iterations is 200, the number of waits is 40, that is, the training stops when there are no better training metrics in 40 consecutive iterations.

5.3 Evaluation Metrics

To validate the performance of the algorithm, we conducted a series of experiments using the trained E2E-MFERC model on the validation set. In the experiments, we use model size, number of parameters, model computations (FLOPs), single image (per frame) detection time (ms), mean average precision mAP, and multi-class metric F1-score for comprehensive evaluation. The calculation methods for each metric are as follows:

$$precision = \frac{TP}{TP + FP} \tag{22}$$

$$recall = \frac{TP}{TP + FN} \tag{23}$$

$$AP = \int_0^{recall} precision(recall) d(recall) \tag{24}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP \tag{25}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{26}$$

where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. *precision* represents the detection precision for the current class, *recall* represents recall rate, C represents the number of detected expression categories.

5.4 Results Analysis

5.4.1 Comparison Experiment Results

E2E-MFERC architecture design was inspired by the YOLO series of algorithms. We focused on in-depth comparison with v5 and v8 versions of YOLO series which have the highest research popularity and current optimal performance, respectively, including different scales of the corresponding

algorithms. Experiments were conducted on the MFED dataset, for both classification criteria. Note that E2E-MFERC is committed to providing lightweight, high-performance solutions, so the provided model scales are Default and Plus, with number of parameters comparable to YOLO's "n" and "s" scales respectively. E2E-MFERC's "m", "l" and "x" scales are only for experimental comparison purposes, given based on YOLO. The experimental results are shown in Table 5.

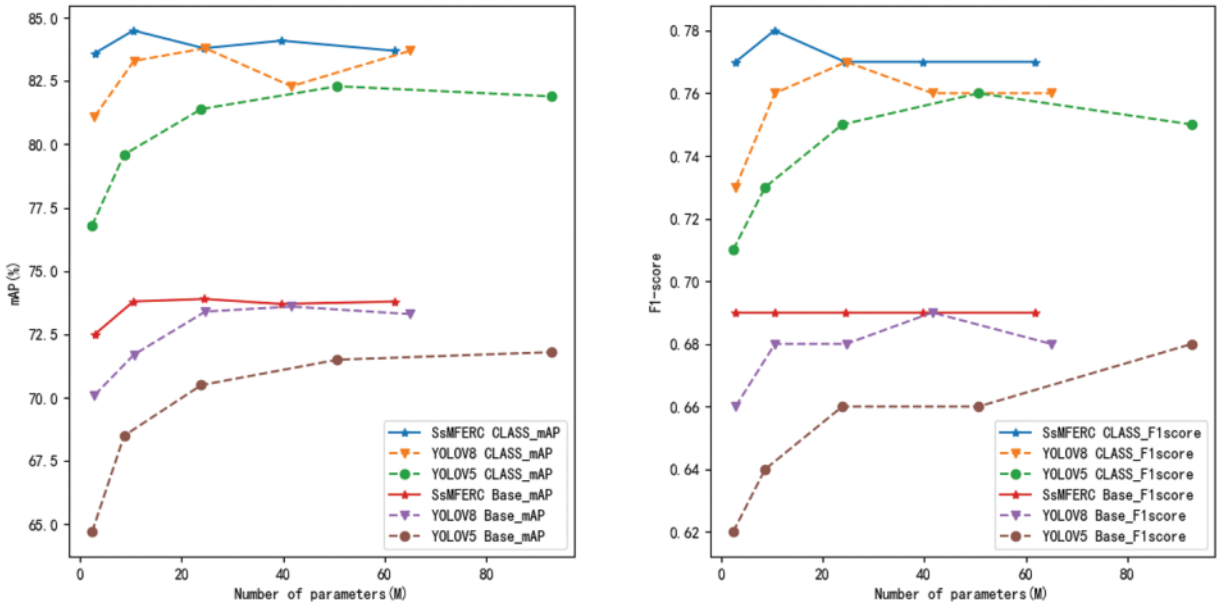
Table 5: Comparison experimental results between E2E-MFERC and YOLO with different scales

ID	Model	Parameters (M)	FLOPs (G)	Model size (M)	Class (5)		Base (7)	
					mAP (%)	F1-score	mAP (%)	F1-score
1	YOLOv5n	2.39	7.8	5.3	76.8	0.71	64.7	0.62
2	YOLOv8n	2.87	8.2	6.3	81.1	0.73	70.1	0.65
3	**E2E-MFERC-Default (ours)**	2.89	7.9	6.3	**83.6**	**0.77**	**72.5**	**0.69**
4	YOLOv5s	8.69	24.2	18.5	79.6	0.73	68.5	0.64
5	YOLOv8s	10.62	28.7	21.5	83.3	0.76	71.7	0.68
6	**E2E-MFERC-Plus (ours)**	10.51	27.9	23.1	**84.5**	**0.77**	**73.8**	**0.69**
7	YOLOv5m	23.89	64.6	50.5	81.4	0.75	70.5	0.66
8	YOLOv8m	24.66	79.1	49.7	83.8	0.77	73.4	0.68
9	*E2E-MFERCm (Exp only)*	*24.41*	*78.7*	*51.6*	*83.8*	*0.78*	*73.9*	*0.69*
10	YOLOv5l	50.68	135.6	106.8	82.3	0.76	71.5	0.66
11	YOLOv8l	41.61	165.4	87.6	82.3	0.76	73.6	0.69
12	*E2E-MFERCl (Exp only)*	*39.74*	*160.0*	*83.9*	*84.1*	*0.77*	*73.7*	*0.69*
13	YOLOv5x	92.72	247.3	195.0	81.9	0.75	71.8	0.68
14	YOLOv8x	65.00	258.1	130.5	83.7	0.76	73.3	0.68
15	*E2E-MFERCx (Exp only)*	*61.85*	*249.4*	*124.3*	*83.7*	*0.77*	*73.8*	*0.69*

The experimental results show that E2E-MFERC achieved the design goals and algorithm requirements of balancing lightweight, high-performance, and superior effects. Overall, under comparable model size and similar number of parameters and computations, E2E-MFERC demonstrated superior recognition performance.

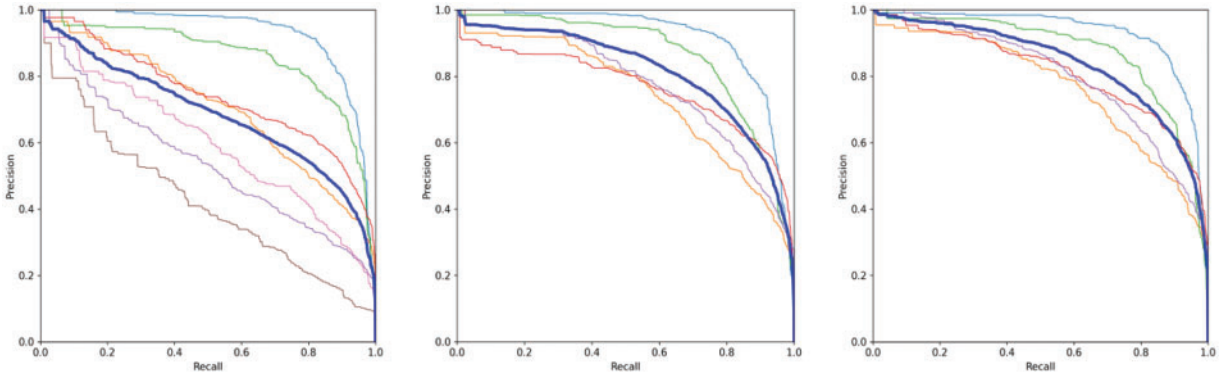
As shown in Fig. 14, the horizontal axis is number of parameters. It can be seen that E2E-MFERC has performance advantages under comparable scales. On classroom evaluation category data, E2E-MFERC-Default improves mAP by 6.8% and 2.5%, respectively over same-scale YOLOv5 and YOLOv8, and improves F1-score by 0.06 and 0.04, respectively; on basic category data, mAP is improved by 7.8% and 2.4%, respectively, and F1-score is improved by 0.07 and 0.04, respectively, showing significant algorithm performance improvement. The mAP and F1-score values of E2E-MFERC-Plus are greater than other algorithms at all scales, and the performance improvement is still significant at the same scale. It achieved the best results among networks of all scales, while having far fewer parameters and computations than larger-scale other algorithms, which also verifies the advantage of lightweight design of E2E-MFERC.

Fig. 15 shows the PR curves of YOLOv5n, YOLOv8n, E2E-MFERC-Default on MFED. The horizontal axis is recall rate R, the vertical axis is precision P, and the curve represents the precision P when recall rate is R. The larger the area in the lower left corner, the better the model performance on the dataset. It can be clearly seen that the model effect of E2E-MFERC is superior to the other two.



(a) mAP (b) F1-score

Figure 14: Comparison charts of experimental results for YOLOv5n, YOLOv8n, E2E-MFERC-Default



(a) YOLOv5n (b) YOLOv8n (c) E2E-MFERC-Default

Figure 15: PR curves of YOLOv5n, YOLOv8n, E2E-MFERC-Default on MFE

Through comparative experiments, we mainly conducted detailed comparisons between E2E-MFERC and mainstream algorithms YOLOv5 and YOLOv8. From the experimental results, it can be seen that under comparable conditions of number of parameters, computations, and model size, the mAP and F1-score metrics of E2E-MFERC have obvious advantages. Under comparable mAP and F1-score metrics, the number of parameters, computations, and model size of E2E-MFERC are significantly less than the other two. The PR curve comparison further verifies the comprehensive performance advantage of E2E-MFERC. Our algorithm model has achieved significant improvements in both speed and accuracy. However, while the lightweight design philosophy brings performance improvements, it also leads to some performance advantage decline when the model scale becomes

larger. Therefore, balancing algorithm efficiency and accuracy, we recommend using E2E-MFERC Default and Plus scales for actual applications.

5.4.2 Ablation Experiments

In the algorithm model design, we constructed the backbone using RepVGGBlock, SPD-Conv and C2f_Attention modules, and used the AFPN structure in the neck for efficient feature fusion, while optimizing the output head size for the actual smart classroom scenarios. To verify the validity and criticality of different module designs in the algorithm, we conducted ablation experiments on both sets of classification criteria in MFED, with results shown in Tables 6 and 7. It should be noted that we referenced the architecture design of YOLOv8 for our algorithm design, so we chose it as the baseline for comparison.

Table 6: Ablation experiment results of E2E-MFERC on MFED classroom evaluation category data

ID	RepVGG Block	SPD-Conv	C2f_Attention	AFPNhead-	Parameters (M)	FLOPs (G)	Time (ms)	mAP (%)	F1-score
1					2.87	8.2	6.6	81.1	0.73
2	√				3.36	6.1	6.5	81.9	0.74
3	√	√			3.92	10.6	7.8	82.1	0.75
4	√	√	√		4.05	10.9	9.6	82.5	0.75
5	√	√	√	√	**2.90**	**7.9**	**7.4**	**83.6**	**0.77**

Table 7: Ablation experiment results of E2E-MFERC on MFED basic expression category data

ID	RepVGG block	SPD-Conv	C2f_Attention	AFPNhead-	Parameters (M)	FLOPs (G)	Time (ms)	mAP (%)	F1-score
1					2.87	8.2	6	70.1	0.65
2	√				3.36	6.1	6.1	71.0	0.67
3	√	√			3.92	10.6	7.8	71.5	0.67
4	√	√	√		4.05	10.9	9.6	72.1	0.68
5	√	√	√	√	**2.90**	**7.9**	**7.6**	**72.5**	**0.69**

The results show that thanks to its unique branched structure and branch fusion during inference, the introduction of RepVGGBlock reduces computations while improving both mean average precision and F1-score. The SPD-Conv module, used to compensate for the negative effects of convolution downsampling operations with stride 2 on feature information, plays an important role in improving recognition accuracy. C2f_Attention is an optimization based on the C2f structure. By introducing attention mechanisms, it enhances the network's ability to extract important features, and promotes improvements in mAP and F1-score in experiments. The neck based on the AFPN structure along with RepVGGBlock for adjusting the output head, not only improves algorithm accuracy, but also significantly reduces parameters, computations and detection time, even slightly lower than the baseline network in parameters and computations. The ingenious combination of modules enables E2E-MFERC to achieve a higher balance of accuracy and speed. This also lays the technical foundation for applying the algorithm to complex smart classroom environments and performing real-time multi-face expression analysis under limited hardware conditions.

5.4.3 Experiment Results of Different Attention Mechanisms in C2f_Attention

In order to optimize the C2f module, we investigated several currently high-performing attention mechanisms, and screened out several that were more compatible with our research needs and data, including CBAM [41], SimAM [42], CoTAttention [43] and Polarized Self-Attention [44], etc. Among them, Polarized Self-Attention has sequential and parallel implementations. We introduced the above attention mechanisms respectively into the C2f_Attention module of E2E-MFERC, and conducted comparative experiments on the MFED dataset.

The results in Table 8 show that using the parallel implementation of Polarized Self-Attention in the C2f_Attention module can achieve the highest mAP and F1-score, with overall optimal performance compared to other attention mechanisms. Fig. 16 compares Epoch-mAP curves of several schemes, with training Epoch as the horizontal axis and mAP as the vertical axis. Considering all metrics comprehensively, the parallel Polarized Self-Attention, with its feature of computing global and local dependencies in parallel, enhanced the ability of the C2f_Attention module to focus on and represent key information in facial areas. Therefore, we chose the parallel Polarized Self-Attention as the final scheme for optimizing the C2f_Attention module.

Table 8: Experiment results of different attention mechanisms (AM) in C2f_Attention

ID	Method	Parameters (M)	GFLOPs (G)	Time (ms)	mAP (%)	F1-score
1	C2f_CBAM	2.9	7.9	13.1	82.3	0.75
2	C2f_SimAM	2.84	7.8	7.6	83.2	***0.77***
3	C2f_COT	3.07	8.5	11.4	82.4	0.76
4	C2f_Polarized-Sequential	2.89	7.9	7.5	82.6	*0.76*
5	C2f_Polarized-Parallel	2.89	7.9	7.5	***83.6***	***0.77***

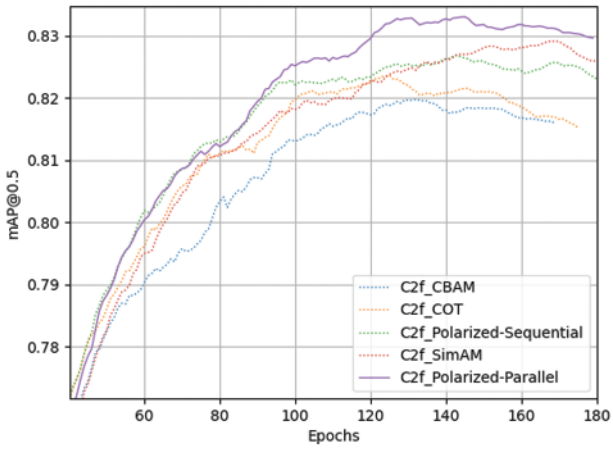


Figure 16: Comparison of Epochs-mAP between different attention mechanisms in C2f_Attention

5.5 Result Demonstration

We designed a prediction program to perform recognition prediction using the trained model on images and videos, display the prediction results in real time, and save the prediction results for subsequent analysis. As shown in Fig. 17, it demonstrates the recognition results of models trained under classroom evaluation and basic expression classification criteria on MFED using E2E-MFERC-Default.

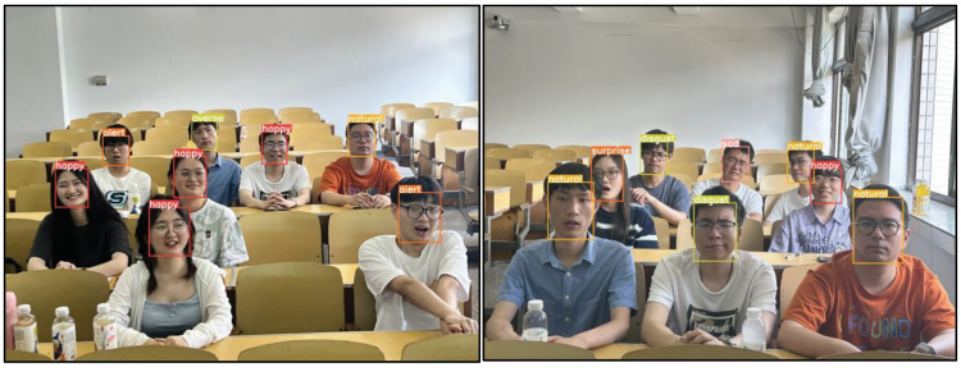


Figure 17: Recognition results of E2E-MFERC-Default model

In addition, we also compared the visualized feature maps of E2E-MFERC-Default and YOLOv8n. Figs. 18 and 19 show a feature map from the corresponding output head of each model when processing the same image on the left side of Fig. 17. Comparing the feature maps shows that our network has stronger representation capabilities through module combination and introduction of attention mechanisms, especially with more concentrated attention on important features related to facial expressions and better extraction results.

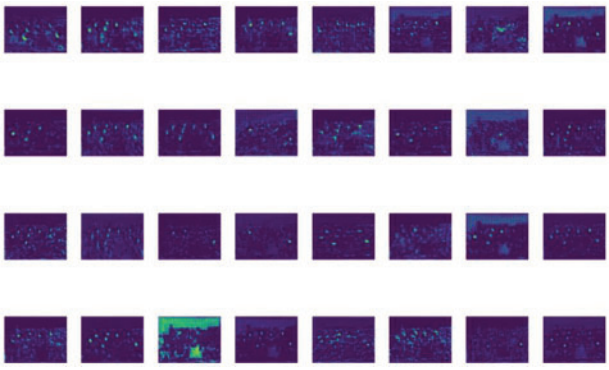


Figure 18: Visualized feature map from output head of E2E-MFERC-Default model

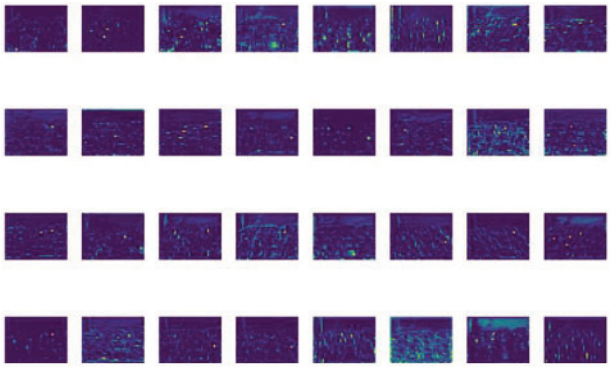


Figure 19: Visualized feature map from output head of YOLOv8n model

6 Conclusion

Analyzing student group emotions by recognizing student facial expression states, evaluating teaching effects, and promoting the improvement of education quality play an important role in the development of smart education. However, the low efficiency, insufficient generalization capability of existing multi-face expression recognition methods, and the lack of high-quality and highly targeted multi-face expression datasets to provide experimental support, as well as the complex multi-stage algorithm process, restrict the improvement of its overall performance. The exploration in actual smart classroom application scenarios is still insufficient, and its practical application value cannot be realized. To this end, we propose the end-to-end algorithm model E2E-MFERC for multi-face expression recognition in smart classrooms, providing a theoretical reference for end-to-end solutions, expanding the technical approach; constructing the large-scale, high-quality and highly targeted dataset MFED to establish data foundations; and applying the model to actual smart classroom scenarios to expand its application value.

In order to study multi-face expression recognition and student group emotion assessment in smart classroom scenarios, we proposed an end-to-end multi-face expression recognition algorithm model for smart classrooms: E2E-MFERC. To simultaneously improve performance while ensuring algorithm speed, we adopted RepVGGBlock and SPD-Conv modules in the backbone, compared different attention mechanisms and introduced Polarized Self-Attention to improve the C2f module, utilized AFPN feature fusion technology combined with RepVGGBlock for representation enhancement and down-scaled the output head size. To provide high-quality and highly targeted data support for algorithm research, we constructed the multi-face expression dataset from real classrooms: MFED, which contains a total of 2,385 images and 18,712 facial expression samples, basic expression and classroom evaluation expression classification criteria data, and provides VOC and TXT label formats. Finally, we applied the model to group emotion evaluation scenarios oriented to smart classrooms, and designed classroom effect evaluation index calculation methods as parameters to measure teaching outcomes.

The experimental results demonstrate that in multi-face recognition oriented to smart classroom scenarios, our proposed algorithm model design is reasonable and shows obvious advantages. Compared with existing mainstream methods, the performance of our algorithm has significantly improved. It has obvious advantages in both algorithm speed and accuracy, providing theoretical and technical support for real-time accurate analysis in smart classroom scenarios. The MFED dataset

we constructed is of high quality, highly targeted, and collected from real scenarios. It not only lays the data foundation for our research, but can also provide experimental data for related research on facial expression recognition, face detection, etc. We also designed a teaching effect evaluation index system based on multi-face expression recognition results, exploring the utility of multi-face expression in smart education effect evaluation, and expanding its application value. Our research is forward-looking. The potential applications and practical value of E2E-MFERC in smart classrooms are not limited to group emotion assessment. It can also be combined with multi-modality for more comprehensive analysis, such as combining speech recognition and text analysis to evaluate students' answer quality, analyzing students' learning mastery; or combining analysis of students' gestures and motions to detect students' interest and participation, evaluating the appeal of teaching activities. Likewise, long-term analysis of students' emotional changes can also be conducted to understand teaching feedback from different periods and different student groups. Combined with learning performance data, automatic diagnosis and targeted improvement of student learning outcomes can be achieved. Thanks to its excellent performance, practical applications can avoid hardware constraints. Even in traditional teaching scenarios, our model can be well promoted by simply adding a computer and image acquisition device.

Of course, our research also has some limitations. In designing E2E-MFERC, we focused on balancing algorithm speed and adopted modules more advantageous for lightweight networks, but the advantages are not obvious when the model size increases. With the development of hardware devices and improved computing capabilities in smart classrooms, research on large-scale algorithms is gradually becoming important. In addition, we only considered a single modality of facial expression data, while multi-modal learning methods are also a trend. For evaluating teaching effects in smart classrooms, our current capabilities are still limited to multi-face expression recognition and related evaluation parameter design. The actual application can be further expanded. In the future, we will focus on the following work:

- Research algorithm models for large-scale multi-face expression recognition with more obvious advantages.
- Try to integrate information including gestures, speech, classroom content, etc., for more accurate and comprehensive analysis using multi-modal data.
- Expand practical application value based on theoretical and technical research, integrating more modal data.

Acknowledgement: We would like to thank the reviewers and the editor for their comments and advices in advance to help us improve the quality of this paper.

Funding Statement: This work was supported by the Science and Technology Project of State Grid Corporation of China under Grant No. 5700-202318292A-1-1-ZN.

Author Contributions: Study conception and design: Lin Wang, Xiaolong Xu; data collection: Lin Wang, Juan Zhao, Hu Song; analysis and interpretation of results: Lin Wang, Juan Zhao, Hu Song; draft manuscript preparation: Lin Wang, Xiaolong Xu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data will be available upon reasonable request to all interested researchers.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Mehrabian, "Communication without words," in *Communication Theory*, Taylor & Francis Group, Sep. 2017, pp. 193–200. doi: 10.4324/9781315080918-15.

[2] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2019. doi: 10.1016/j.imavis.2008.08.005.

[3] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Aug. 2017. doi: 10.1109/TAFFC.2017.2740923.

[4] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019. doi: 10.1109/TCYB.2017.2788081.

[5] S. H. Lee, K. N. Plataniotis, and Y. M. Ro, "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 340–351, Jul. 2014. doi: 10.1109/TAFFC.2014.2346515.

[6] M. Liu, S. Li, S. Shan, and X. Chen, "AU-aware deep networks for facial expression recognition," in *2013 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, 2013, pp. 1–6.

[7] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–8.

[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018. doi: 10.1007/s11263-017-1055-1.

[9] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Vis. Emot. Chall. (AVEC'15)*, New York, NY, USA, 2015, pp. 49–56.

[10] R. Zhao, T. Liu, J. Xiao, D. P. K. Lun, and K. M. Lam, "Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, 2021, pp. 4412–4419.

[11] A. T. Lopes, E. de Aguiar, A. F. de Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, no. 12, pp. 610–628, Jan. 2017. doi: 10.1016/j.patcog.2016.07.026.

[12] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image. Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017. doi: 10.1109/TIP.2017.2689999.

[13] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014. doi: 10.1109/TAFFC.2014.2316163.

[14] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022. doi: 10.1109/TAFFC.2020.2981446.

[15] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimed. Tools. Appl.*, vol. 82, no. 8, pp. 11365–11394, Mar. 2023. doi: 10.1007/s11042-022-13558-9.

[16] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999. doi: 10.1109/34.817413.

[17] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *2015 IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Santiago, Chile, 2015, pp. 19–27.

[18] J. Liu, K. Li, B. Song, and L. Zhao, "A multi-stream convolutional neural network for micro-expression recognition using optical flow and evm," Nov. 2020. doi: 10.48550/arXiv.2011.03756.
[19] F. Taherkhani, A. Dabouei, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Self-supervised wasserstein pseudo-labeling for semi-supervised image classification," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 12262–12272.
[20] M. A. Haghpanah, E. Saeedizade, M. T. Masouleh, and A. Kalhor, "Real-time facial expression recognition using facial landmarks and neural networks," in *2022 Int. Conf. Mach. Vis. Image Process. (MVIP)*, Ahvaz, Iran, Islamic Republic, 2022, pp. 1–7.
[21] Z. Hu and C. Yan, "Lightweight multi-scale network with attention for facial expression recognition," in *2021 4th Int. Conf. Adv. Electron. Mater., Comput. Software Eng. (AEMCSE)*, Changsha, China, 2021, pp. 695–698.
[22] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 2983–2991.
[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019. doi: 10.1109/TIP.2018.2886767.
[24] O. S. Kareem, "Face mask detection using haar cascades classifier to reduce the risk of COVED-19," *IJMSCS*, vol. 2, pp. 19–27, Jun. 2023. doi: 10.59543/ijmscs.v2i.7845.
[25] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Trans. Image Process*, vol. 30, pp. 6544–6556, Jul. 2021. doi: 10.1109/TIP.2021.3093397.
[26] Y. Chen, S. Liu, D. Zhao, and W. Ji, "Occlusion facial expression recognition based on feature fusion residual attention network," *Front. Neurorobot.*, vol. 17, pp. 1250706, Aug. 2023. doi: 10.3389/fnbot.2023.1250706.
[27] X. Xia and D. Jiang, "HiT-MST: Dynamic facial expression recognition with hierarchical transformers and multi-scale spatiotemporal aggregation," *Inf. Sci.*, vol. 644, no. 13, pp. 119301, Oct. 2023. doi: 10.1016/j.ins.2023.119301.
[28] M. Bie, Q. Liu, H. Xu, Y. Gao, and X. Che, "FEMFER: Feature enhancement for multi-faces expression recognition in classroom images," *Multimed. Tools Appl.*, vol. 83, no. 2, pp. 6183–6203, Jan. 2024. doi: 10.1007/s11042-023-15808-w.
[29] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition," *Big Data Cogn. Comput.*, vol. 7, no. 1, pp. 48, Mar. 2023. doi: 10.3390/bdcc7010048.
[30] M. Palash and B. Bhargava, "EMERSK—Explainable multimodal emotion recognition with situational knowledge," Jun. 2023. doi: 10.48550/arXiv.2306.08657.
[31] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e-learning context using deep learning approach," *Multimed. Tools Appl.*, vol. 82, no. 18, pp. 28589–28615, Jul. 2023. doi: 10.1007/s11042-023-14392-3.
[32] Z. Chen, M. Liang, Z. Xue, and W. Yu, "STRAN: Student expression recognition based on spatio-temporal residual attention network in classroom teaching videos," *Appl. Intell.*, vol. 53, no. 21, pp. 25310–25329, Nov. 2023. doi: 10.1007/s10489-023-04858-0.
[33] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Comput. Vis.-ECCV 2016: 14th Eur. Conf.*, Amsterdam, Springer, 2016, pp. 21–37.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.
[35] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," May 2016. doi: 10.48550/arXiv.1605.06409.
[36] A. Nazir and M. A. Wani, "You only look once—Object detection models: A review," in *2023 10th Int. Conf. Comput. Sustain. Global Dev. (INDIACom)*, New Delhi, India, 2023, pp. 1088–1095.

[37] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding and J. Sun, "RepVGG: Making VGG-style convnets great again," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 13728–13737.

[38] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, Berlin, Heidelberg, 2023, vol. 2022, pp. 443–459.

[39] G. Yang, J. Lei, Z. Zhu, S. Cheng, Z. Feng and R. Liang, "AFPN: Asymptotic feature pyramid network for object detection," 2023. doi: 10.1109/SMC53992.2023.10394415.

[40] Y. F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomput.*, vol. 506, no. 9, pp. 146–157, Sep. 2022. doi: 10.1016/j.neucom.2022.07.042.

[41] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Comput. Vis.–ECCV 2018: 15th Eur. Conf.*, 2018, pp. 3–19.

[42] L. Yang, R. Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *38th Int. Conf. Mach. Learn.,* 2021, pp. 11863–11874.

[43] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023. doi: 10.1109/TPAMI.2022.3164083.

[44] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," Jul. 2021. doi: 10.48550/arXiv.2107.00782.