# Classroom Behavior Recognition and Analysis Technology Based on CNN Algorithm

Weihua Qiao
School of Architecture and Planning, Changchun University of Architecture and Civil Engineering, Changchun, 130000, China

***Abstract*—Students' classroom behavior can effectively reflect the learning efficiency and the teaching quality of teachers, but the accuracy of current students' classroom behavior identification methods is not high. Aiming at this research gap, an improved algorithm based on multi-task learning cascaded convolutional neural network architecture is proposed. Through the improved algorithm, a face recognition model is constructed to identify students' classroom behavior more accurately. In the performance comparison experiment of the improved convolutional network algorithm, it was found that the recall rate of the improved algorithm was 88.8%, higher than the three comparison models. The result demonstrated that the improved algorithm performed better than the contrast model. In the empirical analysis of the face recognition model based on the improved algorithm, it was found that the accuracy of the proposed face recognition model was 90.2%, which was higher than the traditional face recognition model. The findings indicate that the model developed in this study is capable of accurately reflecting the students' state in the classroom, thereby facilitating the formulation of targeted teaching strategies to enhance their classroom efficiency.**

***Keywords*—Convolution neural network; multi-task learning; face recognition; classroom; student behavior**

## I. Introduction

In the field of education, students' classroom behavior can directly reflect their learning efficiency and teachers' teaching quality [1]. However, the traditional method has the problem of low accuracy in identifying students' classroom behavior. The advent of sophisticated AI technology, particularly in the domain of computer vision, has led to the emergence of advanced deep learning algorithms, such as Convolutional Neural Networks (CNNs). These algorithms have demonstrated remarkable capabilities in image processing and have yielded novel solutions for classroom behavior recognition [2-3]. In recent years, the education industry has begun to widely adopt intelligent teaching equipment to assist teaching, which not only improves teaching efficiency but also provides the possibility for accurate monitoring and analysis of classroom behavior [4]. Therefore, the development of a deep learning-based classroom behavior recognition technology is of great significance for improving teaching quality and student learning efficiency. Although some studies have achieved some results in using CNN technology for classroom behavior recognition, these methods still have certain limitations. For example, the CNN-based classroom teaching behavior recognition and evaluation method proposed by Li et al., as well as the PSU-CNN model proposed by Sethi and Jaiswal [5]. The traditional CNN algorithm needs to improve its recognition accuracy and efficiency in the face of complex scenes and diverse student behaviors [6]. In addition, most of the existing researches focus on single-task learning and fails to make full use of the advantages of multi-task learning to improve the generalization ability and robustness of the model. Therefore, it is necessary to explore a more efficient and accurate classroom behavior recognition technology.

To solve these problems, an improved algorithm based on Multi-task learning cascade Convolutional neural network (MTCNN) is proposed and applied to students' classroom behavior recognition. By introducing the multi-task learning framework, the MTCNN algorithm realizes the joint optimization of face detection, border regression, and key point detection, and significantly improves the recognition accuracy and efficiency. In addition, the performance of the MTCNN algorithm is further optimized by adjusting the network structure, introducing new activation functions, and using feature selection and dimensionality reduction techniques. Compared with the existing literature, the research method has obvious differences and innovations in algorithm structure and task learning. The primary contribution of this study is the proposal of a technology for recognizing student classroom behavior. This technology is based on an improved MTCNN algorithm, and its effectiveness in improving recognition accuracy and efficiency is verified through experiments. This study not only enriches the application scenarios of Artificial Intelligence (AI) in education but also provides new ideas and methods for future research on classroom behavior monitoring technology.

This paper is divided into six sections. The first section introduces the research background, current research, and the research method. The second section describes classroom behavior recognition and related research on the CNN algorithm. The third section is the construction process of student classroom behavior recognition technology based on an improved MTCNN algorithm. In the fourth section, the performance of the proposed algorithm is verified by experiments and compared with the traditional algorithm. The fifth section is to analyze the experimental results and discuss the related research results. The sixth section summarizes the research results and looks forward to the future research direction.

## II. Related Works

The implementation of AI in educational settings is experiencing a rapid expansion, particularly in the domain of classroom behavior monitoring. This approach can facilitate the generation of precise and time-efficient behavioral insights, thereby assisting educators in enhancing classroom management

and elevating the quality of instruction. CNN has become the mainstream technology of classroom behavior recognition because of its powerful image-processing ability. Li et al. proposed a method based on CNN for the identification and evaluation of classroom teaching behaviors and provided the scientific basis for teaching quality evaluation through accurate analysis of teaching videos [7]. Sethi and Jaiswal used CNN to develop the Prediction of Student Understanding-Convolutional Neural Network (PSU-CNN) model, which predicted students' classroom understanding through facial images and realized real-time feedback on students' learning status [8]. The Ensemble Deep CNN for Assessing (EDFA) model proposed by Gupta et al. used integrated deep CNN to assess the cognitive state of students in an adaptive online learning environment. This model enabled educators to modify their teaching strategies in accordance with the cognitive state of their students, thereby facilitating more effective learning outcomes [9]. Su and Wang also proved the effectiveness of deep learning technology in classroom behavior monitoring [10].

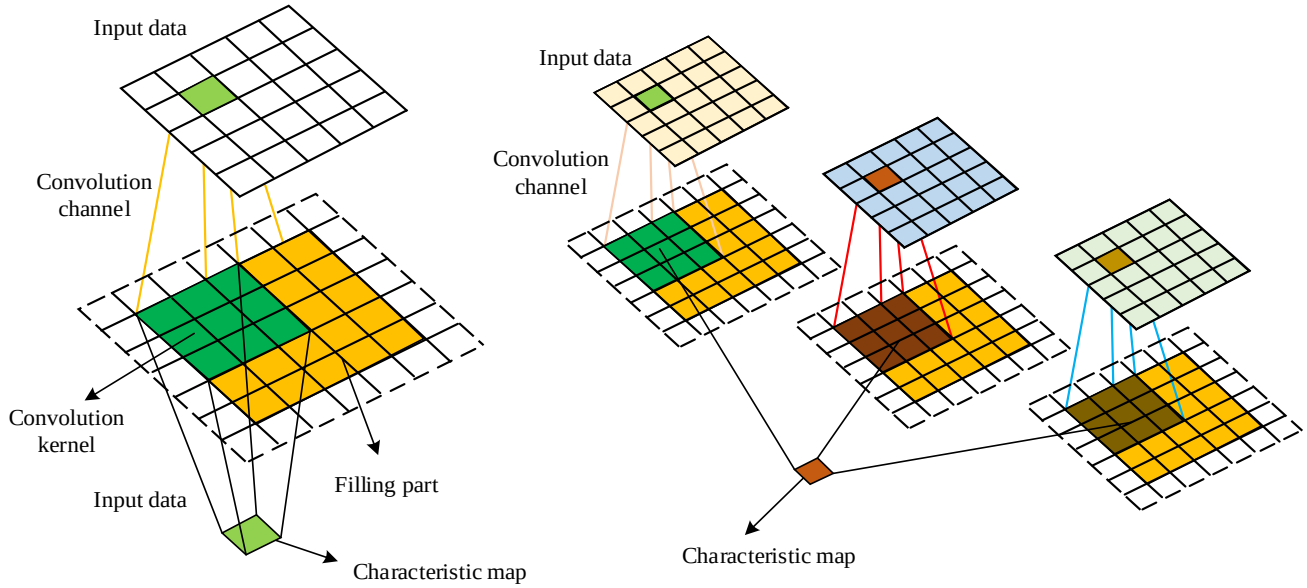
In addition to CNN technology, machine learning and hybrid models also play an important role in classroom behavior monitoring. Lu et al. developed an English online teaching monitoring system based on machine learning, which can analyze students' learning behavior in real-time, provide teachers with teaching feedback, and optimize the online teaching effect [11]. Xu et al. proposed a student online learning behavior monitoring system based on Temporal Shift Module (TSM) behavior recognition and screen recognition, which also provided teachers with feedback on students' learning status [12]. In addition, the CNN and Adaboost fusion model proposed by Hassan et al., and the CNN, Gated Recurrent Unit (GRU), and bidirectional Multi-scale CNN used by Lakshmi et al., were used for human behavior recognition. All these models have further enriched the technical means of classroom behavior monitoring [13-14]. The integration of the Internet of Things and intelligent identification technology provides new possibilities for classroom behavior monitoring. Lin et al. used Internet of Things technology and intelligent image recognition to analyze English classroom behavior and proved the potential of Internet of Things technology in education [15]. This research direction served to enhance the sophistication of classroom behavior monitoring, while simultaneously establishing a robust foundation for the prospective advancement of intelligent education.

To sum up, the application of AI in education and classroom behavior monitoring has shown a diversified trend, covering multiple fields such as CNN-based behavior recognition, machine learning and hybrid models, the Internet of Things, and intelligent recognition. The research work belongs to the category of CNN-based behavior recognition. However, the structure of MTCNN is used to improve the traditional CNN algorithm, thereby optimizing the precision and efficacy of classroom behavior analysis. This work not only enriches the application of AI in education and classroom behavior monitoring but also provides new ideas and methods for future research.

## III. Construction of Student Classroom Behaviour Model Based on CNN Algorithm

### A. CNN Algorithm Combined with Multi-Task Learning

As AI technology develops, target detection based on deep learning has been researched [16]. CNN algorithm is widely used in various image recognition fields because of its excellent performance in image algorithms [17]. For improving the recognition accuracy of student behavior in class, a Face Recognition (FR) model of improved CNN is proposed. The improved algorithm is based on the CNN algorithm and uses multi-task learning to obtain the MTCNN algorithm. CNN algorithm is the most common deep learning algorithm [18]. Convolution usually includes single-channel convolution and multi-channel convolution, as shown in Fig. 1 [19].




(a) Single-channel convolution

(b) Multi-channel convolution

Fig. 1. Two forms of general convolution.

Fig. 1(a) shows a single-channel convolution process. The convolution operation of the input data is performed on a single input channel, and the convolution kernel slides on that channel and applies weights to extract features. Fig. 1(b) shows the multi-channel convolution process, which involves convolution operations of multiple input channels. In this convolution process, each convolution check should have one input channel. There are multiple convolution kernels acting on different channels of the input data at the same time. Each convolution kernel independently extracts the features of a particular channel and then combines these features to form a more complex feature representation. The expression of the convolution operation continuous estimation function $s$ in CNN algorithm is shown in Eq. (1).

$$s(t) = \int x(a)w(t-a)da \tag{1}$$

In Eq. (1), $s$ represents the output signal of convolution operation. $x$ is the input signal, representing the original data or image information. $w$ is the kernel function, also known as the convolution kernel, which is used to weight the input signal. $t$ and $a$ represent the time variables of the output signal and the input signal, respectively. $da$ represents the integral variable and is used to calculate integrals during convolution. The simplified expression is shown in Eq. (2).

$$s(t) = (x * w)(t) \tag{2}$$

The convolution kernel expression is shown in Eq. (3).

$$s(i,j) = (K * I)(i,j)\sum_m \sum_n I(m,n)K(i-m, j-n) \tag{3}$$

$m$ and $n$ respectively represent the effective value range of convolution. $I$ represents the input two-dimensional image. $K$ represents the kernel function of two-dimensional image. To facilitate the application of CNN algorithm in machine learning, Eq. (3) is usually modified, and the expression after the modification is shown in Eq. (4).

$$s(i,j) = (K * I)(i,j)\sum_m \sum_n I(i+m, j+n)K(m,j) \tag{4}$$

Its operation is very similar to the convolution operation, but the change is small within the effective range of $m$ and $n$, which means that when $m$ increases, the input index increases, and the kernel index decreases accordingly, realizing the interchangeability of convolution. The convolution layer of CNN generally refers to two-dimensional convolution operation. Assuming the original image size is set to $D_f \times D_f$ and the convolution core size is set to $D_k \times D_k$. The relationship between the three is shown in Eq. (5).

$$D_f^{'} = (D_f - D_k + 2 \times pad) / stride + 1 \tag{5}$$

In Eq. (5), $pad$ is the filling value, representing the number of pixels added at the edge of the input feature map, which is used to adjust the size of the output feature map. $stride$ is the step length, which represents the stride length when the convolution kernel slides on the input feature map. This parameter affects the size of the output feature map and the granularity of feature extraction. The input layer and convolution layer dimension should be consistent, so it is necessary to select the appropriate step size to influence the extraction of image features. The length calculation of input and output after convolution is shown in Eq. (6).

$$h_0 = \frac{h_i - f + 2p}{s} + 1 \tag{6}$$

In Eq. (6), $h_i$ is the input image width. The width expression of input and output after convolution is shown in Eq. (7).

$$w_0 = \frac{w_i - f + 2p}{s} + 1 \tag{7}$$

In Eq. (7), $f$ is the convolution kernel size. $s$ is the step size, and $p$ is the number of expanded outer layers. By sampling, the pooling layer filters the primary visual features through sampling. Combing the abstract and advanced visual features of the layer, the expression of the whole process is shown in Eq. (8).

$$y_n^l = down(y_n^{l-1}) \tag{8}$$

In Eq. (8), $y_n^{l-1}$ is the $n$ characteristic graph of the output of the $l-1$ th layer network. $y_n^l$ is the $n$ characteristic graph of the pool of the $l$ layer network, and it is the maximum sampling function. The fully connected layer can enhance the nonlinear mapping ability. The neurons used in the previous layer are connected with the neurons in the current network. In the same layer, neurons are not connected. The expression is shown in Eq. (9).

$$o_j^l = f(\sum_{i=1}^{n} X_i^{l-1} \cdot w_{ji}^l + b_j^l) \tag{9}$$

In Eq. (9), $l$ represents the network layer number. $n$ represents the number of network neurons in the $l-1$ layer. $x_i^{l-1}$ is the input value of the $i$ neurons. $w_{ji}^l$ represents the connection weight between the $j$ neurons in the $l$ layer and the $i$ neurons in the $l-1$ layer. $b_j^l$ represents the offset of the $j$ neurons in the $l$ layer. The fully connected layer is

usually composed of linear part and nonlinear part. Among them, the linear part mainly analyzes the input data, and the nonlinear part mainly maps the input data. The overall structure of CNN including the specific fully connected layer structure is shown in Fig. 2.

MTCNN realizes the joint optimization of face detection and key point location by improving the traditional single-task CNN into a multi-task learning framework. It adopts a cascade structure and consists of three networks, P-Net, N-Net, and O-Net, which are respectively responsible for rough detection, candidate region refinement, and final accurate output, which improves detection accuracy and efficiency. In addition, MTCNN also introduces online difficult sample mining to enhance the robustness of the model. In the FR, the MTCNN algorithm initializes the training samples and network weights. The sample set consists of some faces and some non-faces, and the number of samples is N [20]. It inputs the training sample scaling layer image pyramid into the network. Supported by the objective function, the network weight is adjusted by the propagation method [21]. It scales the test image and inputs it into the trained network. Then, the P-Net generates a candidate window and border regression vector. Regression of the bounding box corrects the candidate frame, and NMS overlaps the candidate frame. Finally, it needs to output P-Net and input the improved candidate window and border regression vector into N-Net. It outputs N-Net and inputs the improved results into O-Net to output the final face frame and position. The improved MTCNN network structure is shown in Fig. 3.

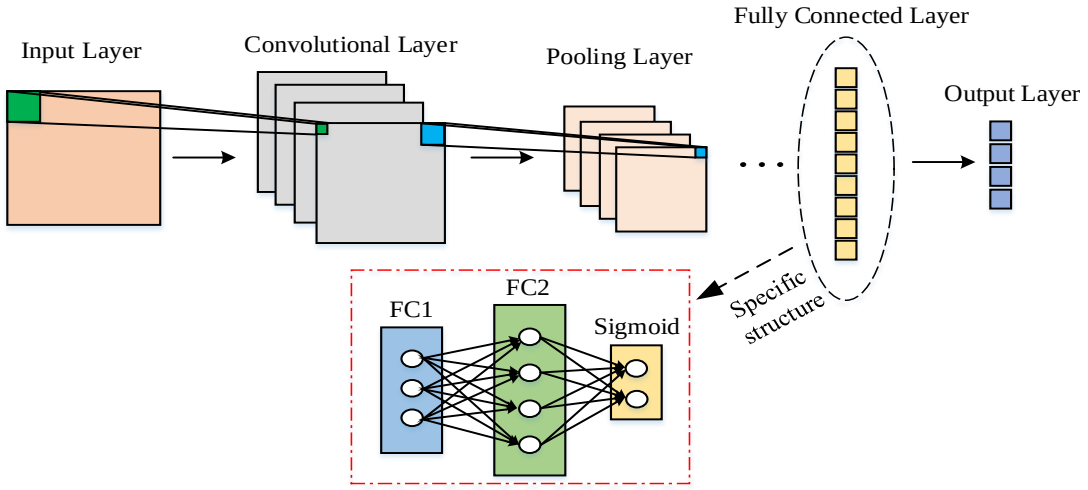



Fig. 2. CNN's overall structure.

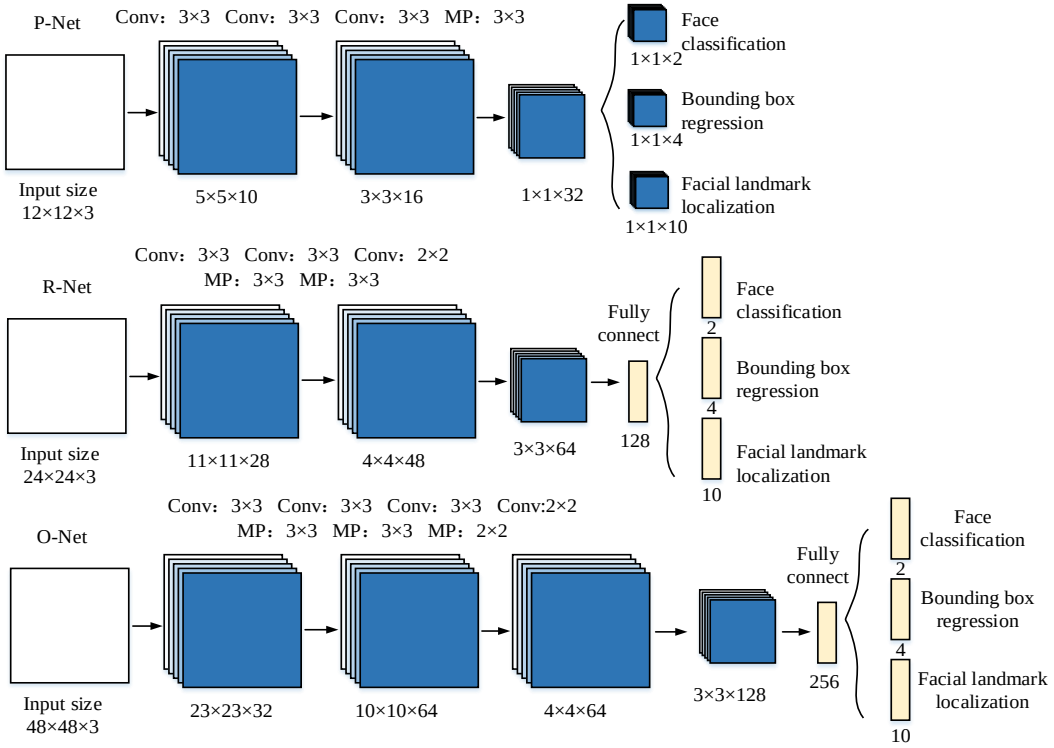



Fig. 3. Improved network structure of MTCNN.

As shown in Fig. 3, the improved MTCNN algorithm first initializes the training samples and network weights. The sample set consists of some faces and some non-faces, and the total number of samples is N. The scaling layer image pyramid of the training samples is input into the network. Combined with the objective function, the propagation method is used to adjust the network weight, and the test image is scaled and input into the trained network. Then, the P-Net is used to generate candidate window and border regression vectors, while bounding box regression is used to correct candidate boxes and Non-Maximum Suppression is used to overlap candidate boxes. Finally, P-Net outputs and inputs the improved candidate window and border regression vector into N-Net, N-Net outputs and inputs the improved result into O-Net, and O-Net outputs the final face frame and position. The ReLU activation function is fast in neural network training, and its function definition is shown in Eq. (10).

$$f(x)\begin{cases}0 \ for \ x<0\\ x \ for \ x\geq 0\end{cases} \tag{10}$$

In neural network training parameters, the ReLU function will not have the problem that the gradient of sigmoid function disappears in error back propagation during model training. Compared with ReLU, the PReLU activation function adds very few parameters. However, the amount of computation does not increase during the whole network training. Especially when the same $a_i$ is used in different ways, the number of parameters will be less. When the error reverse algorithm updates $a_i$, the driving quantity update method is adopted, as shown in Eq. (11).

$$\Delta a_i = \mu \Delta a_i + \varepsilon \frac{\partial \varepsilon}{\partial a_i} \tag{11}$$

Therefore, the activation function of the proposed MTCNN face detection algorithm is the ReLU activation function with parameters.

### B. Construction of FR Model Based on MTCNN

With the increase of students, real-time monitoring of students' classroom status is crucial to the improvement of school classroom quality [22]. To monitor students' discipline in the classroom in real-time and improve the teaching management level of the school, the FR algorithm based on MTCNN is researched and adopted to realize the FR of students in the classroom scene. This method can identify the behavior of students in the classroom. When students behave abnormally, this method can intercept the marker box for the detection target. Then FR is performed on the target in the frame to judge the students' classroom status. This technically supports the improvement of classroom teaching quality. The proposed FR algorithm exists in the whole classroom behavior recognition, and the specific FR algorithm flow is shown in Fig. 4.

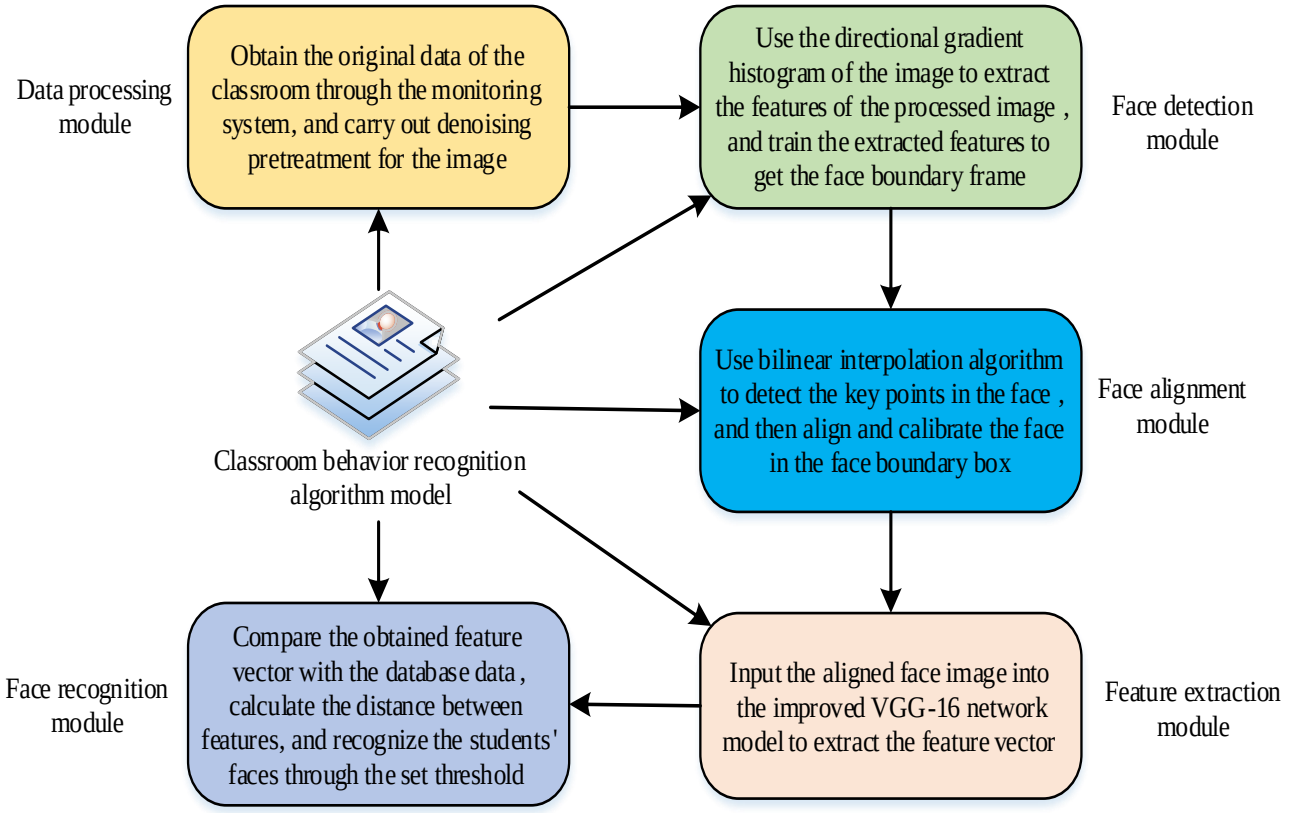



Fig. 4. Structure of classroom behavior recognition algorithm model.

As shown in Fig. 4, the proposed FR algorithm mainly includes five modules. They are the data processing module, detection module, face alignment module, feature extraction module, and FR module. The data processing module needs to obtain the original image of the classroom through the monitoring system and then pre-process the obtained image by framing or noise reduction to ensure that the original image is clear and complete. The face detection module mainly extracts the features of the pre-processed image and inputs the extracted image into the Support Vector Machine (SVM) classifier to train the boundary frame of the face. The face alignment module is mainly used to detect the key points and align the face in the face boundary box. The feature extraction module uses the improved MTCNN model to extract the features of the aligned face image and obtain the feature vector. The final FR module mainly compares the feature vector with the database data, calculates the distance between the features, and recognizes the students' faces through the set threshold. The MTCNN algorithm designs a lightweight structure, which ensures real-time performance. It is a multi-task learning face detection framework, which can simultaneously perform three tasks: face detection, detection frame regression, and face feature point detection. Among them, face detection is solved and described by the cross entropy loss function, whose expression is shown in Eq. (12).

$$L_i = -(y_i^{\det} \log(p_i) + (1 - y_i^{\det})(1 - \log(p_i))) \quad (12)$$

In Eq. (12), $y_i^{\det} \in \{0,1\}$ represents the real label of the $i$ training sample. $y_i^{\det} = 1$ represents the face, otherwise it is non-face. $p_i$ represents the probability that the $i$ training sample is a face. The detection frame regression represents the candidate window loss through Euclidean distance, and its expression is shown in Eq. (13).

$$L_i^{box} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2 \quad (13)$$

In Eq. (13), $y_i^{box} \in R^4$ represents the true border vector of the $i$ training sample. It consists of four elements: the horizontal axis coordinates of the upper left corner, the vertical axis coordinates of the upper left corner, the height, and width. $\hat{y}_j^{box}$ represents the prediction frame vector of the $i$ training sample. Face feature points can be regarded as a group of two-dimensional arrays. The loss of feature points can also be expressed by Euclidean distance, and its expression is shown in Eq. (14).

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2 \quad (14)$$

In Eq. (14), $y_i^{landmark} \in R^{10}$ represents the real face feature point coordinates of the $i$th training sample. There are five points in total and one point for each two coordinates, called 10-tuple. $\hat{y}_i^{landmark}$ represents the predicted face feature point coordinates of the $i$ training sample. This paper applies the MTCNN algorithm to students' classroom FR and puts forward an FR model based on the MTCNN algorithm. It mainly includes five modules: data processing, face detection, face alignment, feature extraction, and FR. In the FR model, the MTCNN algorithm is mainly used to realize the accurate recognition of students' faces in the classroom scene. It can perform face detection, detection frame regression, and face feature point detection at the same time to improve the accuracy and efficiency of FR. The flow of the proposed FR model is shown in Fig. 5.

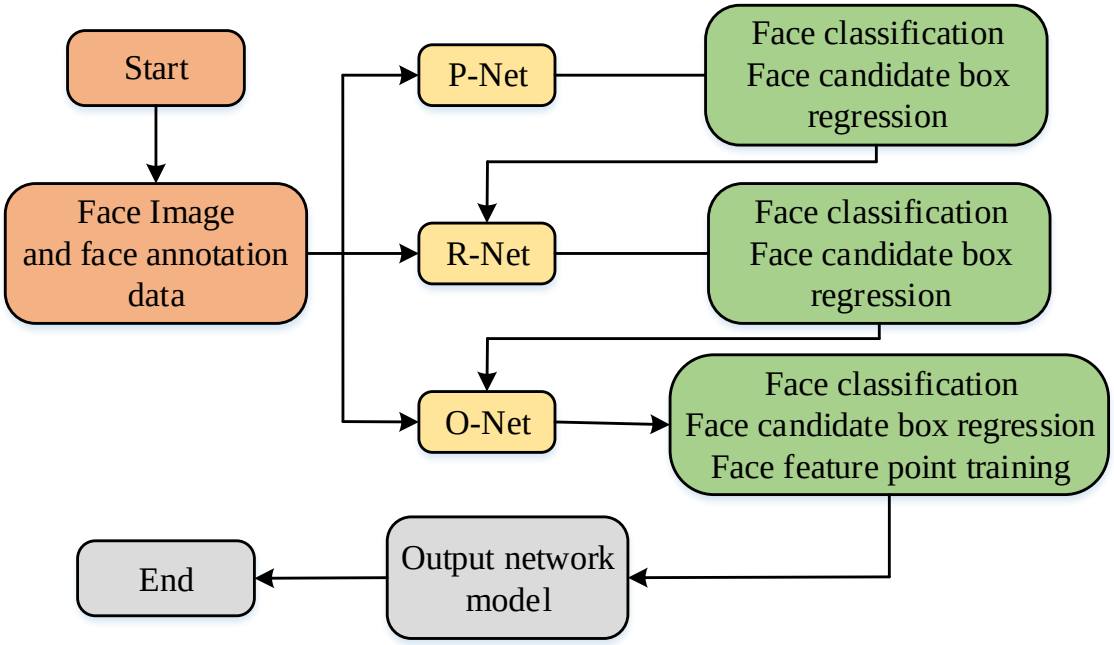



Fig. 5. Classroom FR model flow based on MTCNN algorithm.

In Fig. 5, the workflow of the FR model based on the MTCNN algorithm is as follows: First, an image of any size is input, and after multi-template and multi-scale graph preprocessing, the input image is reduced to 12×12 size and sent to the P-Net. Since the smaller the image, the easier it is to generate candidate regions, the network size for detecting images is set to 12×12. Then the candidate region frame is filtered, and the image is extracted according to the candidate region box and used as the input of R-Net. R-Net makes further adjustments on the border of the candidate regions formed by the previous network to generate more accurate regional recommendations and send the results to the O-Net. It further adjusts the candidate regions to obtain the final face detection structure. At the same time, the coordinates of facial key points will be output to complete the final detection process. The FR model proposed in this study is a multi-task learning framework based on MTCNN. In this model, 5 key points are used in the study to recognize face feature points. These key points constitute 10 coordinate values, that is, 10 attributes. In addition, feature selection and feature dimensionality reduction techniques, such as principal component analysis, are used to optimize feature vectors and improve computational efficiency. Finally, in the MTCNN network structure, this study also designs a multi-layer CNN. The specific number of layers is determined based on task requirements and computing resources to ensure efficient operation and excellent recognition performance of the network. The MTCNN model proposed by the research can recognize students' faces in class. It analyzes students' classroom state and then formulates appropriate strategies to improve students' concentration in class and improve students' classroom learning efficiency.

## IV. Comparative Analysis of FR Algorithm Performance

### A. Experimental Environment Setting

The purpose of this part is to test the performance of the proposed MTCNN and compare its performance with the Visual Geometry Group (VGG) model, CNN model, and Region-based Convolutional Neural Networks (RCNN) model. It takes the loss curve, accuracy, precision, F1 value, and recall rate as the performance comparison indicators for comparative experiments. The experimental environment for the comparison experiment includes a high-performance server equipped with an NVIDIA GeForce GTX 1080 Ti GPU, running the Ubuntu 18.04 operating system, and using the TensorFlow deep learning framework. The hyperparameters of the MTCNN model are set as follows: the learning rates of P-Net, N-Net, and O-Net are 0.01, 0.01, and 0.001 respectively, the training batch size is 128, and the parameters are updated by Adam optimizer. During the training process, data enhancement techniques, including random cropping, rotation, and flipping, are used to increase the generalization ability of the model. The training program is divided into two stages: the pre-training stage and fine-tuning stage. In the pre-training stage, the large-scale FACE dataset WIDER FACE is used for preliminary training, enabling the model to learn the fundamental characteristics of the face. Subsequently, in the fine-tuning phase, the model is further adjusted using a dataset specific to the classroom environment to suit real-world application scenarios. For comparison models, VGG, CNN, and RCNN, similar training strategies and hyperparameter tuning processes are used to ensure that they can achieve full performance within their respective frameworks. After the training is complete, all models are evaluated using the same test set to ensure that the results are fair and comparable. This paper studies the training of four network models in the framework of deep learning. It uses a random gradient descent method to update parameters. The learning rate of model training is set to 0.1, which attenuates exponentially.

### B. Experimental Result

The loss curve is usually used to show the change of the loss value in the training process of the model. It serves as a crucial metric for assessing the efficacy of the model's training. The smaller the loss value, the better the performance of the model.

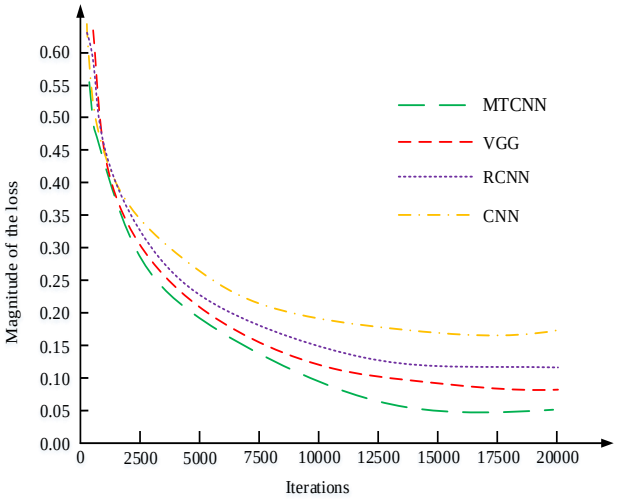



Fig. 6. Comparison of loss curves of four models.

The loss curve of the four models in the deep learning framework is shown in Fig. 6. From Fig. 6, the loss values of the four models went downwards with the increase in the number of iterations. The MTCNN model tended to be stable with the lowest loss value of 0.05, and it tended to be stable when the iterations were 16,100. The loss value of the VGG model that tended to be stable was followed by 0.10, and it tended to be stable when the iterations were 13,900. The loss value of the RCNN model tended to be stable and was only 0.13 higher than that of the CNN model, and it tended to be stable when the iterations were 12,800. The CNN model tended to be stable with the highest loss value of 0.18, and it tended to be stable when the iterations were 12,400. The above results showed that the improved MTCNN model was superior to the other three models in terms of the loss curve dimension. Accuracy is the proportion of the number of correctly classified samples to the total number of samples. The higher the value, the better the classification performance of the model.

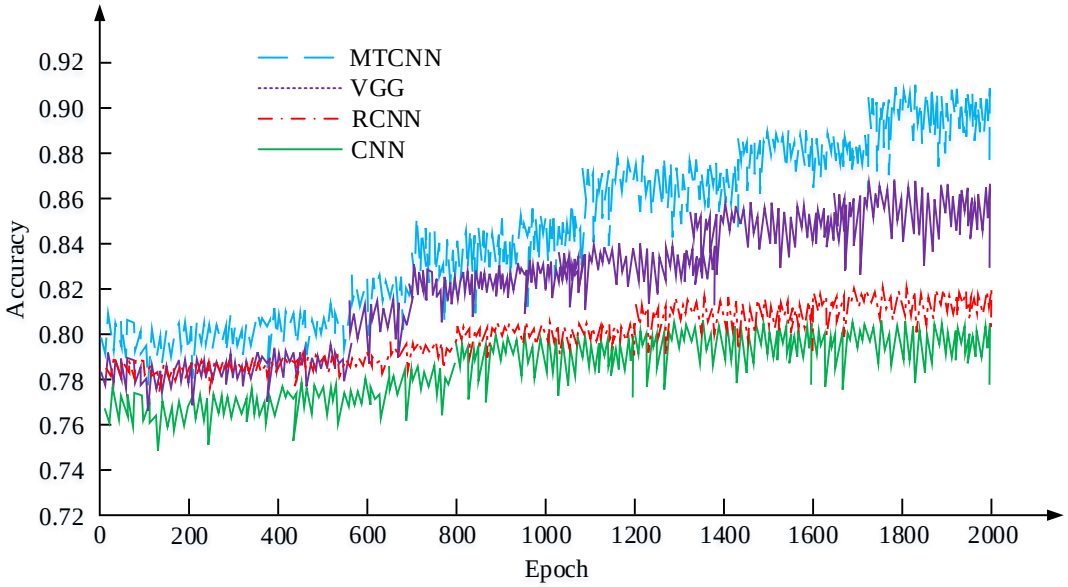
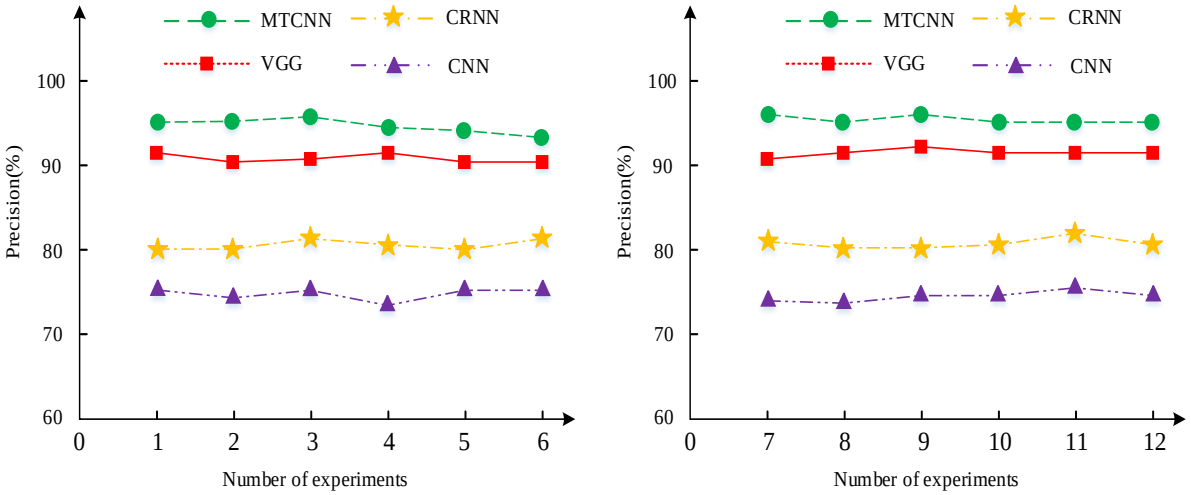



Fig. 7. Accuracy of different models.

The same test set is used for the accuracy of the four, and the results are shown in Fig. 7. The accuracy curve of MTCNN is higher than that of the comparison model. Its accuracy curve shows an upward trend with increasing iterations. In addition, the maximum accuracy of the two-way MTCNN model is 0.91. This is higher than 0.87 for the VGG model, 0.82 for the RCNN model, and 0.79 for the CNN model. The above results indicate that, from the perspective of accuracy, the MTCNN model outperforms the three comparison models. To facilitate a comprehensive comparison of the four models' accuracy, F1 value, recall, and Precision-Recall (PR) curve, a series of tests were conducted on the LFW dataset to train the FR model. Precision refers to the proportion of predicted positive samples that are actually positive samples. The higher the value, the better the classification performance of the model.

The precision results are shown in Fig. 8. Fig. 8(a) is the curve of the previous six comparative experiments. The Precision curve of the MTCNN model in the four network models is higher than that of the other three models. Its average Precision in the first six comparative experiments is 93.5%. This is higher than 90.1% of the VGG network model, 80.1% of the RCNN model, and 76.3% of the CNN. Fig. 8(b) is the Precision curve of the last six comparative experiments. The Precision curve of the MTCNN model in the four network models is higher than that of the other three models. Its average Precision in the first six comparative experiments is 93.6%. This is higher than 90.3% of the VGG model, 80.4% of the RCNN model, and 76.1% of the CNN model. The above results show that the improved VGG-16 network model has the best performance from the perspective of Precision. The recall rate is defined as the proportion of positive samples that are correctly identified as such. The higher the value, the better the classification performance of the model.




（a）Results of the first six experiments

（b）Results of the last six experiments

Fig. 8. Precision comparison results of four models.

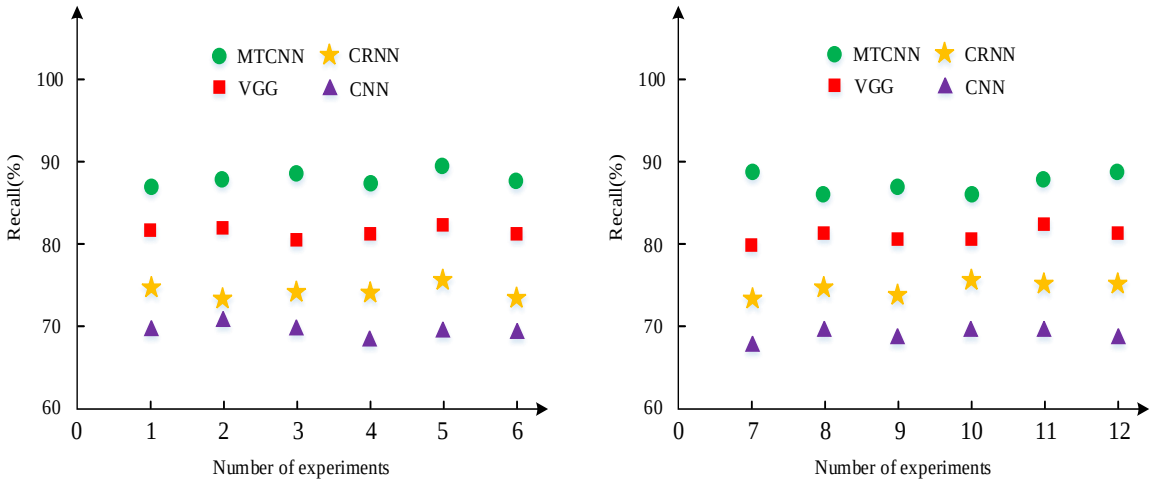



Fig. 9. Comparison results of recall rates of four models.

The recall rates of the four models are shown in Fig. 9. Fig. 9(a) is the results of the previous six comparative experiments. The overall recall rate of the MTCNN model in the four network models is higher than that of the other three models. Its average recall rate in the first six comparative experiments is 88.8%. This is higher than 82.1% of the VGG, 74.3% of the RCNN, and 69.4% of the CNN. Fig. 9(b) is the results of the last six comparative experiments. The overall recall rate of the MTCNN model in the four network models is higher than that of the other three models. Its average recall rate in the first six comparative experiments is 88.6%. This is higher than 81.4% of the VGG, 74.6% of the RCNN, and 68.8% of the CNN. The above results indicate that the improved MTCNN model has the best performance from the perspective of recall rate. The F1 value is the harmonic average of the accuracy rate and recall rate, which is used to comprehensively evaluate the performance of the model. The higher the value, the better the performance of the model.

The results of F1 values are shown in Fig. 10. From Fig. 10, when test samples increase, the F1 values of the four models decrease. When the number of samples to be tested is 50, the four models have good F1 values. However, with the increase of samples, the computational load of the model increases, and the F1 value of some comparison algorithms starts to decrease significantly. Finally, when the number of test samples is 350, the F1 values of the CNN model, RCNN model, VGG model, and MTCNN model are 38.6%, 39.8%, 50.3%, and 61.8%, respectively. The higher the F1 value of an algorithm, the better its performance. Therefore, the above results show that the improved MTCNN is superior to other comparison models from the perspective of F1 value. The PR curve, composed of recall rate and precision, can intuitively demonstrate the average precision value of disparate algorithm models.

The four algorithm's PR curves are shown in Fig. 11. From Fig. 11, the MTCNN model used in this study has the largest area in the PR curve. The MTCNN model has the best effect on student FR detection, with the highest average detection accuracy. Then, the time complexity of the three algorithms is analyzed. This study measures the time required for different models to process the same number of images under the same hardware conditions through experiments.

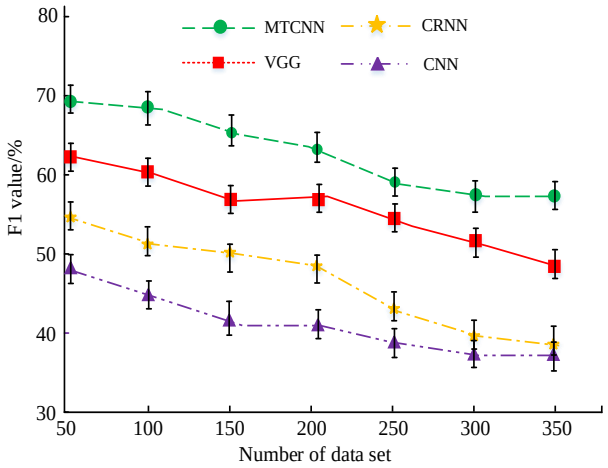



Fig. 10. F1 values of different algorithms.

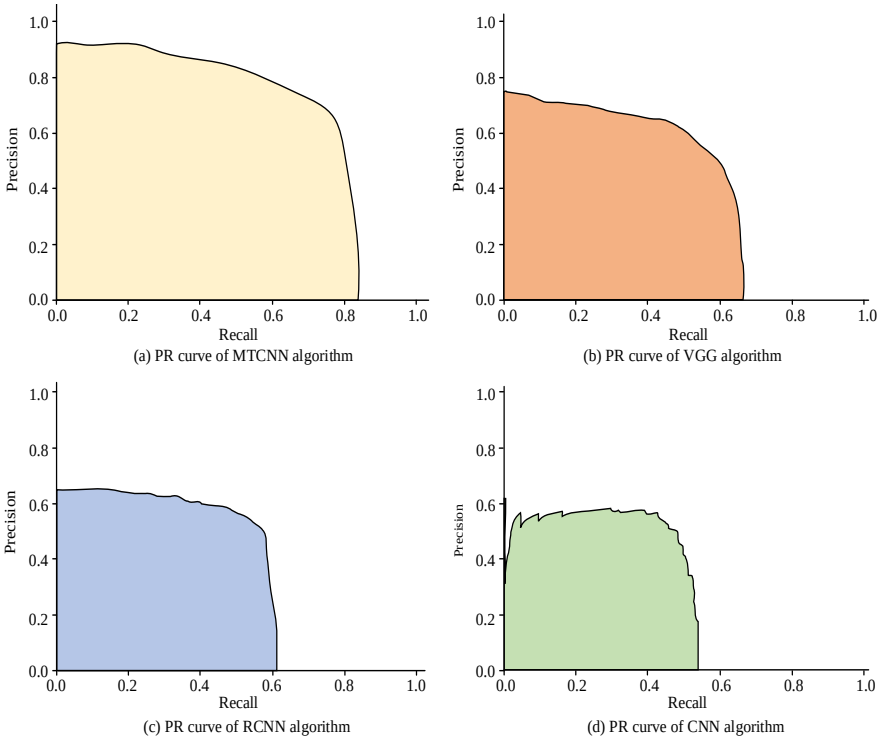



Fig. 11. PR curves of four target detection algorithms.

TABLE I. COMPARISON OF THE TIME COMPLEXITY OF THE THREE ALGORITHMS

| Sample size | Model type | Average processing time (ms) | Standard deviation (ms) | Time complexity evaluation |
|---|---|---|---|---|
| 50 | VGG | 496.3 | 25.3 | Intermediate |
| | RCNN | 748.2 | 30.6 | Higher |
| | MTCNN | 302.5 | 19.8 | Lower |
| 100 | VGG | 991.6 | 48.5 | High |
| | RCNN | 1528.7 | 59.2 | Very high |
| | MTCNN | 589.6 | 31.2 | Intermediate |
| 150 | VGG | 1518.5 | 75.5 | Very high |
| | RCNN | 2244.6 | 78.6 | Extreme height |
| | MTCNN | 887.6 | 39.1 | Intermediate |

The comparison results of the time complexity of the three algorithms are shown in Table I. From Table I, the average processing time of all models shows an upward trend with the increase in the number of samples. Among them, the average processing time of the MTCNN increases from 302.5 ms to 887.6 ms, showing good scalability. In contrast, the average processing time of VGG and RCNN increases more significantly, from 496.3 ms and 748.2 ms to 1528.7 ms and 2244.6 ms, respectively, indicating that they face greater computational challenges when processing large numbers of samples. The average processing time of MTCNN model is lower than that of other models for all sample numbers, and the growth is relatively slow as the sample number increases, showing its potential in practical applications. In summary, the performance of the MTCNN model and the other three models in loss curve, accuracy, precision, F1 value, and complexity are compared. The experimental results show that the MTCNN model has low time complexity while maintaining high accuracy. This is primarily attributable to the lightweight network structure and multi-task learning framework of the MTCNN model, which facilitates expeditious responsiveness in practical applications, thereby addressing the demands of real-time FR. To analyze the practical application effect of the FR model based on the MTCNN model, five classes of students are selected as experimental data sets. The performance of the proposed FR-MTCNN model is compared with traditional models, and the accuracy and precision of the FR model are used as comparison indicators.

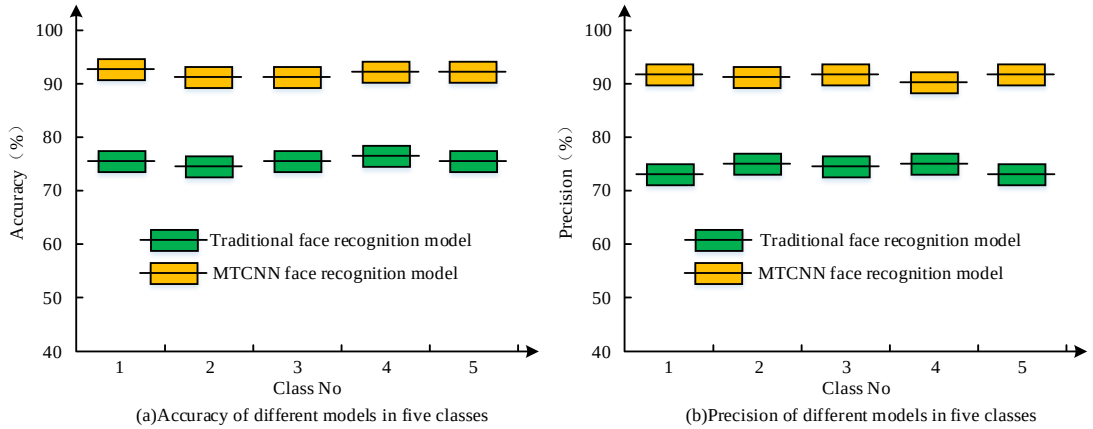



Fig. 12. Comparison results of accuracy and precision of two FR models in five classes.

The specific comparison results are shown in Fig. 12. In Fig. 12, the comparison of accuracy and precision between the two models in 15 categories shows that the proposed FR-MTCNN model generally has higher accuracy and precision than the traditional FR model, with an average accuracy of 90.2% and an average precision of 91.3%. According to the results, the proposed FR MTCNN model is superior to traditional FR model. Applying this model to the classroom can accurately capture students' classroom state, and on this basis, appropriate teaching strategies are formulated to improve students' classroom efficiency. Finally, to more comprehensively verify the accuracy and importance of the proposed method, various indicators are compared with those in [7], [8], and [9]. The comparison results are shown in Table II. The proposed MTCNN method is superior to the methods in references [7], [8], and [9] in all indexes. Among them, the recall rate, accuracy, precision, and F1 value of MTCNN are 6.7%, 4.0%, 3.4%, and 11.5% higher than that of the method proposed in [7]. At the same time, MTCNN has the shortest average time to process 50 images, only 302.5ms, and the lowest loss value is also the lowest, which is 0.05. The above results show that the MTCNN method is more accurate and efficient, and has obvious advantages.

TABLE II. COMPARISON RESULTS OF INDICATORS OF DIFFERENT METHODS

| Validation index | MTCNN | Reference [7] | Reference [8] | Reference [9] |
|---|---|---|---|---|
| Recall Rate | 88.8% | 82.1% | 78.4% | 74.3% |
| Accuracy | 91.0% | 87.0% | 84.2% | 82.1% |
| Precision | 93.5% | 90.1% | 83.3% | 80.1% |
| F1 value | 61.8% | 50.3% | 42.5% | 39.8% |
| Average time to process 50 images | 302.5ms | 496.3ms | 538.5ms | 748.2ms |
| Minimum loss value | 0.05 | 0.10 | 0.12 | 0.15 |

## V. DISCUSSION

The proposed MTCNN model is superior to other models in accuracy and recall rate. This is consistent with the research results of Khan et al. [23]. The reason for this result may be that MTCNN improves the overall detection accuracy and efficiency by simultaneously optimizing the three tasks of face detection, border regression, and key point detection. In addition, the cascade structure gradually screens the candidate regions from coarse to fine, reducing the amount of computation and improving the detection speed. However, the MTCNN model also has some limitations. For example, its adaptability to complex scenes and extreme lighting conditions needs to be improved, and the training time on large-scale datasets is relatively long. While the FR model proposed in the study demonstrates superior accuracy and recall compared to other models, its implementation in an educational setting warrants careful ethical and practical consideration. First, continuous monitoring of students' behavior may have a potential impact on students' psychology. It is not uncommon for students to experience feelings of invasion of privacy, which can lead to elevated stress and anxiety levels. This can have a detrimental impact on their learning efficiency and mental health. Therefore, when implementing such technology, students' feelings must be fully taken into account and appropriate measures must be taken to reduce their psychological burden. Second, the use of FR technology in the classroom involves privacy concerns and possible legal issues. Although strict measures have been taken in data processing and storage to protect students' privacy in this study, legal restrictions and regulatory requirements for FR technology vary in different countries and regions. Therefore, when promoting the application of this technology, it is necessary to strictly comply with relevant laws and regulations to ensure legality and compliance.

In addition, in actual teaching, the interaction between teachers and students is one of the key factors in the quality of teaching. Over-reliance on technological monitoring can weaken this interaction, affecting trust and communication

between teachers and students. Consequently, when integrating such technologies, it is imperative to comprehensively assess their influence on the dynamics of teachers and students and to implement strategies that facilitate constructive engagement between teachers and students. Further research is required to investigate the capacity of diverse models to manage intricate scenarios and mitigate overfitting, as well as to ascertain how these models can be generalized to disparate classroom contexts. In addition, it is necessary to explore the potential of the technology in other applications. For example, in places such as libraries, laboratories, etc. where people's behavior needs to be monitored and managed, the technology may have higher utility and fewer ethical issues. Finally, when the proposed method is applied in sensitive environments such as education, it faces challenges such as privacy protection and educational effectiveness. Therefore, future research should pay more attention to these challenges and explore effective solutions to ensure the stability and reliability of the technology. At the same time, research on the ethical issues of AI technology should be strengthened to promote its healthy development in education.

## VI. Conclusion

To improve the accuracy of the current student behavior recognition model in class, an FR algorithm combining a multi-task learning network and CNN was proposed and applied to the behavior recognition model of classroom students. The proposed MTCNN algorithm was tested in performance. The precision rate, recall rate, and F1 value of the MTCNN algorithm were 93.5%, 88.8%, and 61.8%, respectively, which were better than the three comparison algorithms. In addition, the research also carried out performance comparison experiments on FR models based on the MTCNN algorithm. The accuracy and precision of the proposed FR model were 90.2% and 91.3%, which were far higher than the traditional FR model. In conclusion, the proposed MTCNN algorithm and the FR model were superior to the comparison algorithm and model. Therefore, the FR model based on the MTCNN algorithm can be used to identify and analyze the behavior of students in the classroom to implement corresponding measures to improve classroom quality. The next research direction is to ensure the stability of the classroom student behavior recognition model.

## References

[1] Abed S, Al-Oraifan D, Safar A. Optic disc detection using fish school search algorithm based on FPGA. Journal of Engineering Research, 2019, 7(3):161-177.

[2] Hamrick S A, Richling S M, Brogan K M, Rapp J T, Davis W. Effects of Obtrusive Observation and Rules on Classroom Behavior of Adolescents in a Juvenile Residential Treatment Setting: Behavior Modification, 2021, 45(5):797-821.

[3] Kumar A, Mishra A. Palm Print Recognition: A biometric Identification Technique. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 2021, 10(1):637-640.

[4] Ma B, Fu Y, Wang C, Li J, Wang Y. A high-performance insulators location scheme based on YOLOv4 deep learning network with GDIoU loss function. IET image processing, 2022, 16(4):1124-1134.

[5] Zhao X, Wu B. Algorithm for real-time defect detection of micro pipe inner surface. Applied optics, 2021, 60(29):9167-9179.

[6] Foroughi F, Chen Z, Wang J. A CNN-Based System for Mobile Robot Navigation in Indoor Environments via Visual Localization with a Small Dataset.World Electric Vehicle Journal, 2021, 12(134):1-22.

[7] Li G, Liu F, Wang Y, Guo Y, Xiao L, Zhu L. A convolutional neural network (CNN) based approach for the recognition and evaluation of classroom teaching behavior. Scientific Programming, 2021, 2021(1): 6336773.

[8] Sethi K, Jaiswal V. PSU-CNN: prediction of student understanding in the classroom through student facial images using convolutional neural network. Materials Today: Proceedings, 2022, 62(5): 4957-4964.

[9] Gupta S, Kumar P, Tekchandani R K. EDFA: Ensemble deep CNN for assessing student's cognitive state in adaptive online learning environments. International Journal of Cognitive Computing in Engineering, 2023, 4(2): 373-387.

[10] Su X, Wang W. Recognition and Identification of College Students\Classroom Behaviors through Deep Learning. IEIE Transactions on Smart Processing & Computing, 2023, 12(5): 398-403.

[11] Lu W, Vivekananda G N, Shanthini A. Supervision system of English online teaching based on machine learning. Progress in artificial intelligence, 2023, 12(2): 187-198.

[12] Xu H, Lu M, Qiu L, Xie W, Xu J. Student Online Learning Behavior Supervision Based on TSM Behavior Recognition and Screen Recognition. World Scientific Research Journal, 2023, 9(9): 68-75.

[13] Hassan N M H, Moussa M A, Mahmoud M H M. CNN and Adaboost fusion model for multiface recognition based automated verification system of students attendance. Indonesian Journal of Electrical Engineering and Computer Science, 2024, 35(1): 133-139.

[14] Lakshmi N, Rashmi M, Sathvika M. Using CNN, GRU, and B/idirectional Multiscale Convolutional Neural Networks for Human Behavior Recognition. Turkish Journal of Computer and Mathematics Education, 2024, 15(3): 117-131.

[15] Lin J, Li J, Chen J. An analysis of English classroom behavior by intelligent image recognition in IoT. International Journal of System Assurance Engineering and Management, 2022, 13(3): 1063-1071.

[16] Wu X, Li P, Zhou J, Liu Y. A cascaded CNN-based method for monocular vision robotic grasping. Industrial Robot, 2022, 49(4):645-657.

[17] Dey N, Zhang Y D, Rajinikanth V, Pugalenthi R, Raja N. Customized VGG19 Architecture for Pneumonia Detection in Chest X-Rays. Pattern Recognition Letters, 2021, 143:67-74.

[18] Foroughi F, Chen Z, Wang J. A CNN-Based System for Mobile Robot Navigation in Indoor Environments via Visual Localization with a Small Dataset.World Electric Vehicle Journal, 2021, 12(134):1-22.

[19] Alhussainy A. A New Pooling Layer based on Wavelet Transform for Convolutional Neural Network. Journal of Advanced Research in Dynamical andControl Systems, 2020, 24(4):76-85.

[20] Soffer T, Cohen A. Students' engagement characteristics predict success and completion of online courses. Journal of Computer Assisted Learning, 2019, 35(3):378-389.

[21] Groos L, Kai M, Graulich N. Mimicking Students' Behavior during a Titration Experiment: Designing a Digital Student-Centered Experimental Environment. Journal of Chemical Education, 2021, 98(6):1919-1927.

[22] Sun Y, Xue B, Zhang M, Yen G, Lv J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. IEEE Transactions on Cybernetics, 2020, 50(9):3840-3854.

[23] Khan S S, Sengupta D, Ghosh A, Chaudhuri A. MTCNN++: A CNN-based face detection algorithm inspired by MTCNN. The Visual Computer, 2024, 40(2): 899-917.