Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9027907, 13 pages
https://doi.org/10.1155/2022/9027907

Research Article

An Integrated Model for On-Site Teaching Quality Evaluation Based on Deep Learning

Wei Zhuang,[1] **Fanan Xing**,[1] **Jili Fan**,[1] **Chunming Gao**,[2] **and Yunhong Zhang**[1]

[1]*School of Computer, Nanjing University of Information Science and Technology, Nanjing 210044, China*
[2]*School of Engineering & Technology, University of Washington, Tacoma, WA 98402, USA*

Correspondence should be addressed to Chunming Gao; chunming@uw.edu

Received 16 April 2022; Revised 11 May 2022; Accepted 12 May 2022; Published 24 June 2022

Academic Editor: Kalidoss Rajakani

During on-site teaching for university students, the level of concentration of every student is an important indicator for the evaluation of teaching quality. Traditionally, teachers rely on subjective methods for observing students' learning status. Due to the volume of on-site crowds, teachers are unable to stay on top of the learning status of each student. Meanwhile, because of the subjective evaluation, the results would not be precise. With the fast development of artificial intelligence and machine learning, it is possible to adopt deep learning technology to achieve scientific evaluation of the classroom teaching quality. This paper proposes an integrated evaluation model based on deep learning technology, incorporating YOLOX model, Retinaface model, and SCN model. Among which, YOLOX model is used to detect the area of the students' upper body, Retinaface model is adopted to assess the head-up rate, and SCN model is used to recognize the facial expression. The experimental results have shown that our model can achieve 93.1% object detection accuracy, more than 85% face recognition accuracy, and 87.39% expression recognition accuracy. We further develop a model to use the combination of head-up rate and facial expression scores to jointly evaluate classroom teaching quality. Five teaching professors' evaluations of our classroom video images confirmed that our proposed model is effective in objectively evaluating the on-site teaching quality.

1. Introduction

Education informatization has become a hot topic in the field of educational research. It is an important strategic task for education modernization to accelerate the education reform in the information age and form a modern education management system. With the fast development of artificial intelligence, big data would profoundly change the teaching format and the evaluation methods of teaching quality [1].

In the classroom, teachers are in a one-to-many teaching scenario. It is difficult to stay on top of the learning status of each student, which often leads to incomplete and untimely evaluation of teaching quality. Without an effective and objective mechanism for classroom evaluation, if students show a negative and school-weary learning status in the classroom, teachers will fail to take intervention measures in time. And over time, students' learning efficiency and learning outcomes will be adversely affected. On the contrary, if teachers could find abnormal situations of students' on-site learning and adjust teaching methods in time, it would greatly change students' attitudes in class and therefore provide a strong and important booster for improving the quality of teaching.

The teaching quality of a class is not only closely related to a teacher's teaching ability but also inseparable from students' learning attention. At present, most of the methods used by teachers to stay on top of students' learning status are through traditional evaluation methods such as classroom questions, homework, and tests. There is no doubt that this will cause a lag in teaching quality evaluation. An effective and objective on-site evaluation system will be desirable in improving teaching quality. Fortunately, with the fast development of intelligent surveillance, face recognition technology, and theories related to image processing and pattern recognition [2], the intelligent analysis based on video images has been widely promoted and applied. This has brought a promising opportunity for the innovation of classroom teaching quality evaluation. In this paper, we

present a deep learning model based on joint analysis of students' head-up rate and facial expressions in class, which could provide objective feedback on the quality of classroom teaching.

2. Related Work

Students' head-up rate and emotional expression in classrooms have been considered a direct reflection of how attentive they are for the class. After comparing the individual head-up situation with the standard collective head-up situation, the student's concentration status in a classroom could be evaluated. Therefore, the head-up rate could be used as a quantitative indicator of learning attentiveness. Moreover, the facial expression changes displayed by students during class can be seen as the natural emotional change of students for learning activities. The famous American psychologist Albert Mehrabian has discovered through experiments that in the total expressed information in human communication, verbal information only accounts for 7%, but facial expression information accounts for 55% [3]. Inspired by these findings, scholars have conducted research work on the relationship of classroom head-up rate, students' facial expressions, and classroom teaching quality. Various methods have been explored for automatic analysis of classroom head-up rate, recognition of students' facial expressions, and on-site teaching quality evaluation.

Han et al. [4] proposed a multipose face detection method based on AAM model and CLM model, which realized efficient face information detection and facial expression analysis. This method can track and analyze the overall status of students' attention, and it can also perform targeted statistical analysis of individual students. Wang [5] used MTCNN and FaceNet for face detection and recognition processing, constructed a CNN algorithm expression recognition model, and realized the development of an intelligent education management system based on face recognition. Shi [6] improved the VGG network and the separable convolutional network and established a joint concentration evaluation model based on facial expression and head-up rate. The improved model based on the VGG network achieved better results compared with other algorithms in the FER2013 dataset. Shi verified the model through tests, questions, and teacher-student interviews, indicating that the model has high accuracy and reliability. Zhang [7] proposed a multisize face detection method based on CNN and a facial expression recognition method based on lightweight CNN. Zhang conducted comparative experiments on the WIDER FACE dataset and the FER2013 datasets, respectively, to verify the effectiveness of the method. Zhang built a classroom concentration analysis system based on expression recognition, realizing classroom attendance and real-time concentration analysis tasks. Ge and Liu [8] proposed an emotion recognition evaluation system based on the MASK-RCNN model and the XCeption model to achieve real-time acquisition, processing, and analysis of students' learning status. Zhong et al. [9] proposed a fuzzy comprehensive evaluation algorithm based on face detection. They used OpenCV and Dlib libraries to achieve head posture assessment and used CNN algorithm to train the model on the FER2013 dataset. They evaluated fatigue based on the detection results of eye and mouth closure and then used the fuzzy evaluation method to evaluate students' learning concentration. In the test of the simulated scene, the results were satisfactory. Pan et al. [10] proposed a classroom teaching feedback system based on facial expression recognition, which used the facial expression recognition network combination of CNN and SVM. They conducted comparative experiments with other algorithms to verify the effectiveness of the method and fully demonstrated the usability of the system in the example test. Chen [11] constructed a facial expression recognition model based on a dual RBM-BP neural network. Chen also proposed a text feature analysis model based on Faster R-CNN's sentiment comments, which were experimented on different classes to verify the effectiveness of the system. Wu et al. [12] proposed that the system architecture of the object layer, data layer, technology layer, and application layer should be applied to the four major evaluation and analysis scenarios of classroom language, behavior, emotion, and teaching and achieved good results. Guo [13] proposed a cascading face detection method to detect character information in video sources and proposed a new SCN algorithm for emotion recognition. Guo used the OpenCV framework and CNN to achieve head posture estimation and used fuzzy comprehensive evaluation method to evaluate classroom concentration. Luo et al. [14] classified students' facial expressions based on CNN and compared with SVM-based classification algorithms on their own facial expression datasets, which showed that the method has a high recognition rate and is of great significance for assisting teachers in evaluating teaching quality. Jin et al. [15] proposed an algorithm model based on ResNet50, which can effectively distinguish 7 different expressions in the JEFFE database and obtain the changes in the facial expressions of students in the classroom, thus providing a new method for evaluating the quality of classroom teaching.

In our study, we proposed an integrated evaluation model based on deep learning technology, incorporating YOLOX model, Retinaface model, and SCN model. Among which, YOLOX model is used to detect the area of the students' upper body, Retinaface model is adopted to assess the head-up rate, and SCN model is used to recognize the facial expression. We develop a model to use the combination of head-up rate and facial expression scores to jointly evaluate classroom teaching quality.

3. Evaluation Modelling of On-Site Teaching Quality

Our proposed deep learning-based teaching quality evaluation model examines head-up rate and facial expression score by jointly analyzing students' head-up situation and facial expressions in teaching videos. First, the model of face detection and recognition was constructed by incorporating YOLOX and Retinaface [16, 17]. The VOC2007&VOC2012 datasets and the WIDER FACE dataset were used to train and optimize the parameters

of YOLOX and Retinaface. Then SCN was used as the facial expression recognition model [18], and the model was trained and optimized with the RAF-DB dataset. Finally, a scientific and reasonable teaching quality evaluation model was built by jointly analyzing the head-up rate and facial expression score data. The model building process is shown in Figure 1.

3.1. Face Detection and Recognition. Face detection and recognition is the foundational module of the entire teaching quality evaluation model. This module precisely detects each student in the classroom through an effective detection algorithm, and the accuracy of face detection and recognition will directly affect the performance of the entire evaluation model. Other than knowing exactly where a student's face is, it needs to intercept at the appropriate size. It is also particularly crucial to be able to distinguish between students' heads up and heads down status. Therefore, this paper develops an algorithm that incorporates the face detection and recognition of YOLOX and Retinaface. We use YOLOX to perform face detection on video images and intercept the students' faces in each image. Retinaface is used to further differentiate with head-up students in the images to obtain the head-up rate.

3.1.1. Face Detection. The face detection algorithm used in this paper is YOLOX, which was proposed by the Base Detection Group of Megvii Research Institute [16]. The algorithm is based on YOLO3 and combines the latest results from the target detection academia in recent years. It avoids overfitting the COCO dataset while retaining the easy-to-deploy characteristics of the YOLO series. The main improvements of the algorithm include decoupled head, strong data augmentation, and anchor-free.

In object detection, conflicts between classification and regression tasks are inevitable. Therefore, decoupled heads for classification and localization are widely used in most single-stage detectors. The backbone network and feature pyramid network of the YOLO series are constantly evolving, but the detection heads are still coupled. For each level of FPN features, feature integration is first performed by a 1×1 convolutional normalized activation function to reduce the number of feature channels to 256. And then add two parallel branches for classification and regression tasks, respectively (as shown in Figure 2). Each branch has two 3×3 convolutional normalized activation functions for feature extraction. The Cls branch determines the kind of input feature point by a 1×1 convolution. The other branch is also divided into two parallel branches. Among which, the Reg branch obtains the regression coefficients of feature points by a 1×1 convolution to adjust the prediction frame; and the Obj branch determines whether the feature points have corresponding objects by a 1×1 convolution. Experiments have shown that replacing the coupled head with a decoupled head greatly improves the convergence speed, and the decoupled head is essential for the end-to-end version of YOLO.

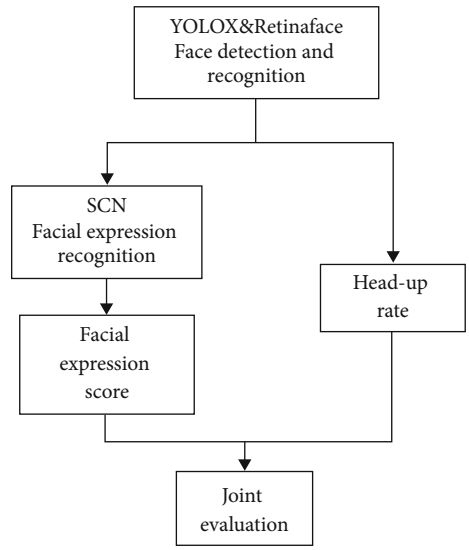

Figure 1: Model building process.

Next, Mosaic and MixUp were used as the enhancement strategy to improve YOLOX performance. Mosaic is an effective enhancement strategy proposed by ultralytics-YOLOv32 and widely used in YOLO series detectors. MixUp was originally designed for image classification tasks and has since been modified to be used mostly for target detection training.

It is well known that the anchor-based mechanism brings the general problem of specific anchor frame facing different datasets, which increases the complexity of the detection header and the number of parameters generated. However, the anchor-free mechanism takes several tricks while significantly reducing the number of parameters. That makes the detector structure simple and achieves a performance comparable to that of the anchor-based mechanism during the training and decoding phases.

3.1.2. Face Recognition. The face recognition algorithm used in this paper essentially identifies the head-up image from those obtained by YOLOX. The Retinaface is a robust single-level face detector proposed by Deng et al. [17]. It takes advantage of joint extrasupervised and self-supervised multitask learning to perform pixel-level face localization on faces of different scales. The Retinaface network structure is shown in Figure 3.

Retinaface uses feature pyramid networks (FPN) to solve multiscale problems, which improves the model's ability to detect small scales while essentially not increasing the computational effort. The outputs of P2 to P5 in Figure 3 are the outputs of C2 to C5 from the residual phase of ResNet50 via the lateral join with the top-down calculation, and P6 is the output of C5 after convolution. ResNet50 is initialized with the network weights pretrained in the WIDER FACE dataset, and the 3×3 convolutional layer of P6 is initialized randomly by the Xavier method. Retinaface uses a contextual module in the feature pyramid to improve the receptive

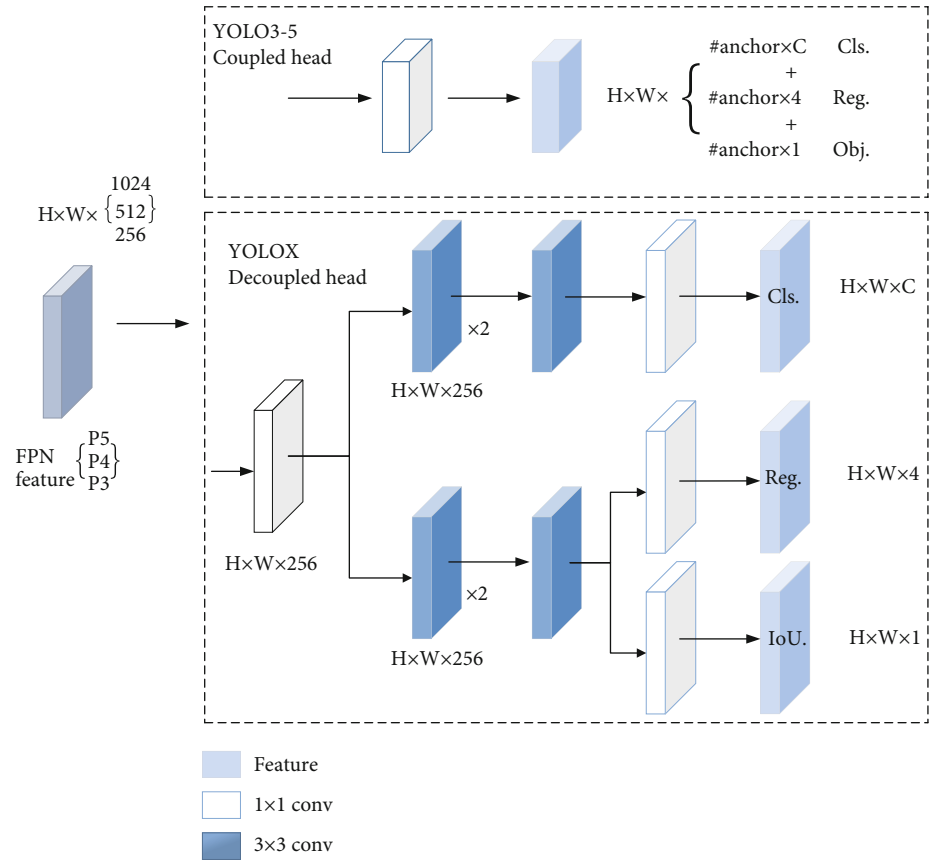

Figure 2: The difference between the YOLO coupled head and decoupled head.

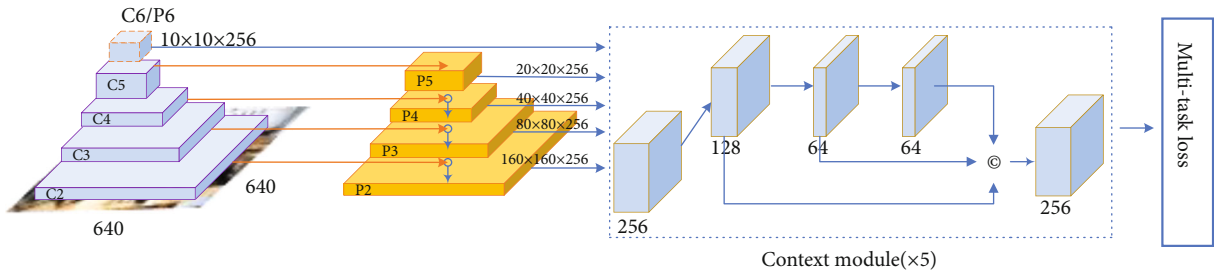

Figure 3: Retinaface network structure.

Table 1: The number of anchors at each level.

Feature pyramid	Stride	Anchor
P2(160 × 160 × 256)	4	16, 20.16, 25.40
P3(80 × 80 × 256)	8	32, 40.32, 50.80
P4(40 × 40 × 256)	16	64, 80.63, 101.59
P5(20 × 20 × 256)	32	128, 161.26, 203.19
P6(10 × 10 × 256)	64	256, 322.54, 406.37

field. For a trained anchor i, the multitask joint loss function is defined as

$$L = L_{\text{cls}}(p_i, p_i^*) + \lambda_1 p_i^* L_{\text{box}}(t_i, t_i^*) + \lambda_2 p_i^* L_{\text{pts}}(l_i, l_i^*) + \lambda_3 p_i^* L_{\text{pixel}}, \tag{1}$$

where $L_{\text{cls}}(p_i, p_i^*)$ is the face classification loss function, p_i is the predicted probability of predicting anchor i for a face, and p_i^* is the true label. Face samples are denoted as 1 and nonface samples are denoted as 0. $L_{\text{box}}(t_i, t_i^*)$ is the face box regression loss function, where t_i and t_i^* represent the coordinates of the prediction box and the real box of the network, respectively. The regression box target is normalized by the smooth-L1 robustness regression function. $L_{\text{pts}}(l_i, l_i^*)$

Table 2: Facial expressions and facial features.

Facial expressions	Facial features
Disgust	Snort, upper lip up, eyebrows drooping, and squinting
Upset	Eyebrows droop, forehead tightly wrinkled, eyelids, and lips strained
Sadness	Squint, eyebrows tighten, corners of the mouth pull down, chin up, or tighten
Fear	The mouth and eyes are open, eyebrows are raised, and nostrils are open
Neutrality	The face is calm and relaxed
Happiness	The corners of the mouth are cocked, the cheeks are raised and wrinkled, the eyelids are contracted, and the tail of the eyes will form "crow's feet"
Surprise	The jaw droops, the lips and mouth relax, the eyes are open, and the eyelids and eyebrows are slightly raised

Table 3: Classroom concentration levels.

Concentration levels	Joint scores
Very unfocused	0-0.25
Unfocused	0.25-0.5
Focused	0.5-0.75
Very focused	0.75-1

is face key point regression loss function, where l_i and l_i^* represent the predicted coordinates and real coordinates of the five face key points, respectively. The normalization of the five face key points also adopts the smooth-L1 function. L_{pixel} is the face dense regression loss function. The loss function adjustment parameters $\lambda_1, \lambda_2, \lambda_3$ are set to 0.25, 0.1, and 0.01, respectively.

The number of anchors at each level is shown in Table 1. Different scale anchor is used from P2 to P6, with smaller scale anchor for smaller targets and larger scale anchor for larger targets. For an input image of size 640×640, the size of the anchor ranges from 16×16 to 406×406. 102,300 anchors are generated in the feature pyramid from P2 to P6. Furthermore, the P2 layer generates 76,800 anchors, accounting for 75% of all anchors.

3.2. Facial Expression Recognition. Facial expression is a natural outpouring of a person's intentions, a way for people to convey their inner changes through various subtle changes in the face. The accuracy of classroom facial expression recognition would directly affect the results of classroom evaluation, so the choice of facial expression recognition algorithm is particularly important. In the past decade or so, significant advances have been made in recognizing facial expressions, tested with datasets generated mainly through field photography or laboratory acquisition. However, for the datasets collected from the classroom, due to unavoidable factors such as crowding and obscuration, in the face recognition process, the front rows are usually clearer, while the back rows are more blurred. So, it is very difficult to achieve high-quality facial expression labeling, and the uncertainty also rises with lower quality video images. These uncertainties will cause the labels to deviate, generating false facial intent feature points. To solve these problems, this paper adopts a self-cure network (SCN) consisting of three main modules: self-attention importance weighting, rank regularization, and relabeling. For a batch of uncertain face images, the facial features are first extracted through the CNN backbone. The self-attention importance weighting module assigns a weight to each image to achieve a lower importance weight for uncertain images. The rank regularization module then sorts these weights in descending order and divides them into high and low importance groups. Finally, the relabeling module uses the relabeling operation to get cleaner and more useful samples from uncertain samples.

3.2.1. Self-Attention Importance Weighting Module. The self-attention importance weighting module is designed to capture the contribution of training samples. Some samples may have higher importance weights, while uncertain samples have lower importance. The facial features of N images are represented by $F = [x_1, x_2, x_3, \cdots, x_N] \in R^{D \times N}$. The self-attention importance weighting module takes F as input and outputs the importance weights for each feature. For the self-attention importance weighting module, it consists of a linear fully connected (FC) layer and an S-shaped activation function that can be expressed as

$$\alpha_i = \sigma\left(W_a^T x_i\right), \tag{2}$$

where α_i is the important value weights of the i-th sample, W_a^T is the parameters of the FC layer for attention, and σ is the sigmoid function.

For attention weights, there are two simple options to calculate the loss weights. The first option is to multiply the weight of each sample by the sample loss. Another option is to solve this problem by alternating minimization. Here adopted is called Logit-Weighted Cross-Entropy loss (WCE-Loss), which can be expressed as

$$L_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\alpha_i W_{y_i}^T x_i}}{\sum_{j=1}^{C} e^{\alpha_i W_j^T x_i}}, \tag{3}$$

where W_j is the j-th classifier and L_{WCE} has a positive correlation with α.

3.2.2. Rank Regularization Module. The self-attention importance weights in the above module can be arbitrary

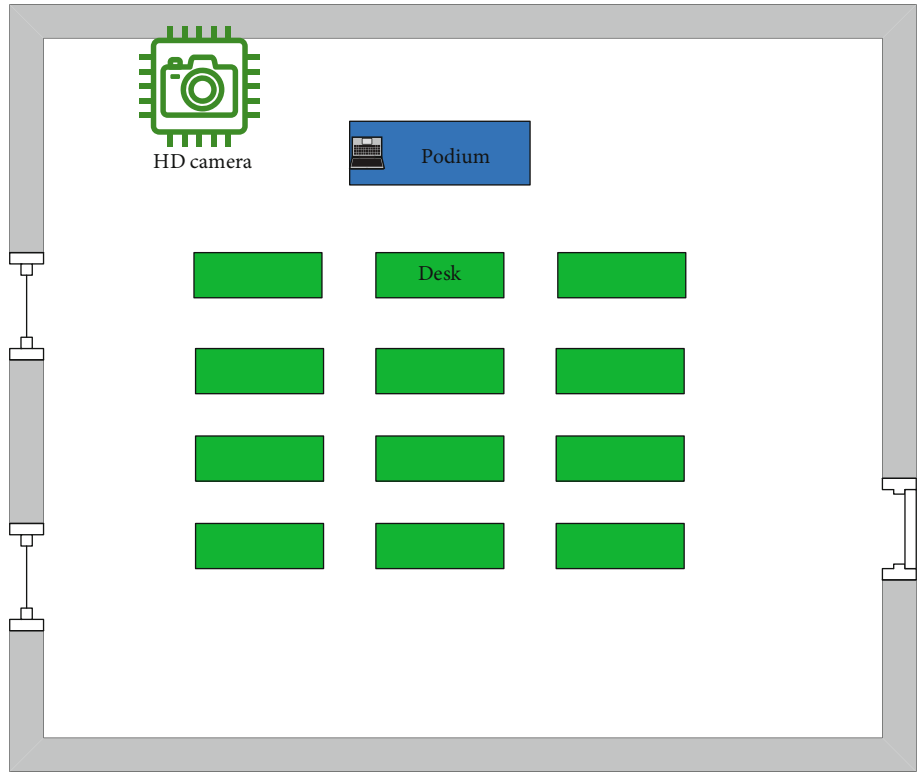

Figure 4: Data collection schematics.

Table 4: Training parameter settings.

Training phase	Epoch	Batch_size	Learning rate	Optimizer
Freeze	50	32	$1e^{-3}$	Adam
Unfreeze	150	16	$1e^{-4}$	Adam

in $(0, 1)$. To explicitly limit the importance of indeterminate samples, a rank regularization loss function is attempted to regularize the attention weights. In the rank regularization module, the attention weights of the learning are sorted first in descending order, and then, they are divided into two groups by ratio β. Rank regularization to ensure that the average attention weight of a high-importance group is higher than the one of low-importance groups. The rank regularization loss (RR-Loss) is defined as follows:

$$L_{\mathrm{RR}} = \max\{0, \delta_1 - (\alpha_H - \alpha_L)\} \tag{4}$$

with

$$\alpha_H = \frac{1}{M}\sum_{i=0}^{M}\alpha_i, \alpha_L = \frac{1}{N-M}\sum_{i=M}^{N}\alpha_i, \tag{5}$$

where δ_1 can be used as a fixed hyperparameter or a learnable parameter, α_H and α_L are the means of the high-importance group of $\beta \times N = M$ samples and the low-importance group of $N - M$ samples, respectively. The total loss function is $L_{\mathrm{all}} = \gamma L_{\mathrm{RR}} + (1-\gamma)L_{\mathrm{WCE}}$ where γ is a trade-off ratio.

3.2.3. Relabeling Module. In the rank regularization module, each small-batch unit is mainly divided into two groups: high importance and low importance. Experiments have found that uncertain samples tend to have lower importance weights, so a strategy is devised to relabel these samples. The difficulty in relabeling these annotations is knowing which annotations are incorrect. The relabeling module considers only samples from low-importance groups and is executed based on probabilities. For each sample, the maximum predicted probability is compared to the probability of a given label. If the maximum predicted probability is higher than one of the sample labels for a given threshold, the sample is assigned to a new pseudolabel. The relabeling module can be defined as

$$y' = \begin{cases} l_{\max}, & \text{if } P_{\max} - P_{\mathrm{gtInd}} > \delta_2, \\ l_{org}, & \text{otherwise}, \end{cases} \tag{6}$$

where y' is the new label, δ_2 is the threshold, $P_{\max}$ is the maximum predicted probability, and P_{gtInd} is the predicted probability of a given label. l_{org} and $l_{\max}$ are the index values of the original given label and the maximum prediction, respectively. In this module, the undetermined sample will receive a lower importance weight, which reduces its negative impact as it is reweighted. Then, it is divided into low-importance groups, which can finally be corrected by relabeling. Those corrected samples may receive higher weights

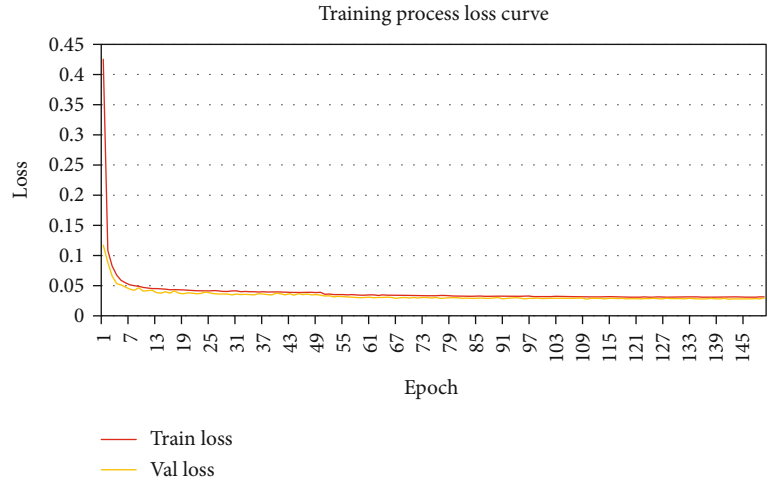

Figure 5: YOLOX training loss function.

Table 5: Performance comparison of each algorithm.

Model	Backbone network	FPS	AP (%)	Parameters	GMACS
YOLO3	Darknet-53	46	91.65	61.52M	27.91
YOLO4	CSPDarknet53	30	92.4	63.94M	25.46
YOLO5-S	Focus+CSP	47	91.9	7.06M	2.95
YOLO5-M	Focus+CSP	38	92.8	21.06M	9.07
YOLO5-L	Focus+CSP	29	93.8	47.06M	20.81
YOLO5-X	Focus+CSP	23	94.2	87.78M	39.43
YOLOX-S	**Focus+CSP**	**45**	**93.1**	**8.94M**	**4.79**
YOLOX-M	Focus+CSP	34	93.6	25.28M	13.23
YOLOX-L	Focus+CSP	27	94.1	54.21M	28.02
YOLOX-X	Focus+CSP	23	94.5	99.07M	50.75

Table 6: AP of WIDER FACE test sets.

Model	Easy	Middle	Hard
ScaleFace	0.867	0.866	0.764
Tiny Faces	0.919	0.908	0.823
S3FD	0.937	0.924	0.852
Faceness-Net-SR-RP	0.717	0.615	0.305
MTCNN	0.851	0.820	0.607
Face R-FCN	0.943	0.932	0.876
PyramidBox	0.956	0.946	0.887
Retinaface-Mobile0.25	0.910	0.880	0.730
Retinaface-Resnet50	**0.950**	**0.940**	**0.850**

in the next period. In this way, the next round of marking is carried out, and the final result is more accurate.

3.3. Datasets

3.3.1. VOC2007&VOC2012. The VOC dataset has twenty object classes: people, birds, cats, cows, dogs, horses, sheep, airplanes, bicycles, buses, cars, motorcycles, trains, bottles, chairs, dining tables, potted plants, sofas, and TVs. It contains a total of 9963 labeled images. This paper uses only images and annotations for the person class in the VOC2007&VOC2012 datasets. In terms of the person class, it provides the corresponding image set, such as classification and detection tasks. These image sets are a subset of those for the main tasks, and all are labeled with a partial layout, such as heads, hands, and feet. They also summarize the number of "person" objects labeled with layout for each image set.

3.3.2. WIDER FACE. The WIDER FACE dataset is a face detection benchmark dataset that is 10 times larger than the largest detected face dataset currently available. It contained 32,203 images and labeled 393,703 faces with height variations in scale, posture, and occlusion. The dataset is organized based on 61 event classes. For each event class, it randomly selects 40%, 10%, and 50% of the data as the training, validation, and test set, respectively. And the dataset uses the same evaluation measures as the PASCAL VOC dataset.

3.3.3. RAF-DB. The Real-World Emotional Face Database (RAF-DB) is a large database of facial expressions that

Table 7: Comparison of results from RAF-DB datasets.

Model	Accuracy (%)
DLP-CNN	84.22
IPA2LT	86.77
gaCNN	85.07
RAN	86.90
SCN	**87.39**

downloads about 30,000 images of various faces from the Internet. Based on crowdsourced annotations, each image is individually tagged by about 40 annotators. The images in the database vary greatly in terms of subject age, gender, ethnicity, head posture, lighting conditions, occlusion (such as glasses, facial hair, and self-occlusion), and postprocessing operations (such as various filters and effects). RAF-DB has a large diversity and rich annotations, including the following: (1) 29672 real-world images; (2) 7-dimensional expression distribution vector for each image; (3) two distinct subsets: single-label subset, including 7 basic emotions, double-label subset, including 12 types of compound emotions; (4) 5 accurate landmark locations, 37 automatic landmark locations, bounding box, ethnicity, age range, and gender attribute annotations; and (5) baseline classifier output for basic mood and compound emotion.

To be able to objectively measure the performance of algorithms, the database is divided into training sets and test sets, where the training set is 5 times the size of the test set, and the expressions in the two sets have almost the same distribution.

3.4. Head-Up Rate. The head-up rate, as the name suggests, refers to the proportion of the amount of heads-up to the total number of a class at a certain time. This paper adopted the analysis method for head-up in [6]. The detection of heads up or not is first done by the YOLOX face detection model, which intercepts the target face from each video image frame. The intercepted images are then fed into the Retinaface face recognition model, which identifies all head-up images at a given moment. There are still inevitable errors throughout the detection and recognition process, due to the effects of light and face occlusion.

The head-up rate analysis module adopted in this paper is in terms of the overall head-up situation, so we only need to calculate the overall head-up rate without needing to consider any individual's head-up rate. Based on YOLOX, the total number of class members can be known, and the number of head-up can be obtained through Retinaface, so the formula for calculating the class head-up rate can be expressed as

$$h_t = \frac{n_t}{N} \times 100\%, \tag{7}$$

where N is the total number of class members and n_t is the number of head-up students of the t-th detection.

3.5. Facial Expression Score. Facial expression score sets different scoring weights to facial expressions that students perform during class. The famous American psychologist Ekman once spent a year watching 200,000 feet of film. He found that regardless of language and culture, the facial muscle changes triggered by these 7 basic emotions (disgust, upset, sadness, fear, neutrality, happiness, and surprise) are roughly the same [19] (as can be seen in Table 2). With comprehensive consideration of the actual teaching environment to simplify the classroom expressions, the seven expressions are categorized and given different weights, so as to facilitate the calculation of subsequent expression scores. Disgust, upset, sadness, and fear are classified as negative, with their weight values set to -1; neutrality is classified as natural, weighted 0; happiness and surprise are classified as positive and given weight 1.

In this paper, we use SCN expression recognition method to record the number of faces in the t-th detected image corresponding to the expression category, then multiply them by different weights, and finally accumulate them as the expression score of this frame image, as shown in the formula:

$$s_t = \begin{cases} n_1 \times (-1) + n_2 \times 0 + n_3 \times 1, & \text{Expression}, \\ 0, & \text{Expressionless}, \end{cases} \tag{8}$$

where s_t is the expression score corresponding to the t-th detected image and n_1, n_2, n_3 represent the number of faces belonging to negative, natural, and positive in the image, respectively. In order to clarify the meaning of the value and facilitate subsequent numerical processing, we normalize the expression score of each frame of the picture, as shown in the formula:

$$s_t^* = \frac{s_t - s_{\text{tmin}}}{s_{\text{tmax}} - s_{\text{tmin}}}, \tag{9}$$

where s_t^* is the expression score of the t-th detected image after normalization, s_{tmax} is the largest score of all images, and s_{tmin} is the smallest.

3.6. Joint Evaluation. To evaluate the classroom teaching quality, teachers should not only pay attention to the content of classroom teaching but also observe the attentive status of students. The traditional method is mostly based on the students' head-up situation to judge the learning status of the classroom, which will undoubtedly cause one-sided judgment, ignoring the emotional changes of most students. In order to assess the teaching situation more objectively and scientifically, this paper proposes a teaching quality evaluation method that combines the analysis of head-up rate and facial expression. After obtaining the head-up rate and expression score through Retinaface model and SCN model,

Figure 6: Face detection.

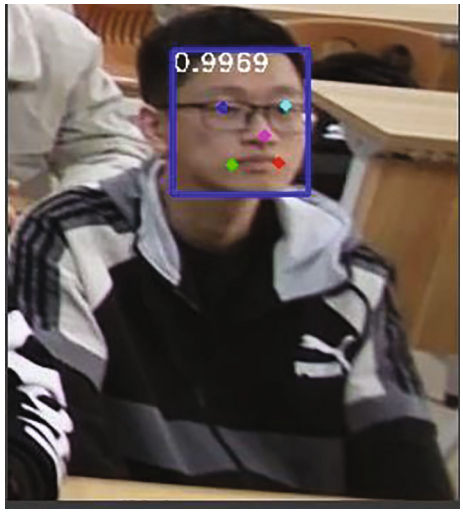

Figure 7: Face recognition.

Figure 8: Facial expression recognition.

respectively, the head-up rate and facial expression score of the t-th detected image are multiplied to obtain the joint score of the image. And then, the combined score of all images is averaged to obtain the overall joint score of the entire video, as shown in the formula:

$$S = \frac{\sum_{t=1}^{M}(h_t \times s_t)}{M}, \tag{10}$$

where M represents the number of detections for the entire video. At the same time, in order to intuitively reflect the status of classroom concentration, this paper categorizes classroom concentration into four levels based on joint scores: very unfocused, unfocused, focused, and very focused (Table 3).

4. Experiment and Performance Analysis

4.1. Experimental Datasets. The datasets used in this study during model training were VOC2007&VOC2012, WIDER FACE, and RAF-DB, respectively. In the testing phase, we used the dataset collected by ourselves. The data collection schematics is shown in Figure 4. The video of students' class activities is captured through a single HD camera, which is set in front of the lectern. We let students simulate a variety of classroom states, so that the subsequent model training is closer to the teaching environment and situation. The resolution of the captured video is 1920×1080, the format is saved as .AVI, and the frame rate is 30 frames per second.

4.2. Experimental Environment. The algorithms were run under Ubuntu 20.04.3 OS environment, based on PyTorch 1.10.0 deep learning framework, and programmed by using Python 3.8. Hardware specifications are as follows: CPU is Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz, GPU is NVIDIA RTX 3090 SUPRIM X 24 GB, and RAM is 32 GB.

4.3. Experimental Processes and Analysis of Results

4.3.1. Face Detection. In this paper, the person class in the VOC2007&VOC2012 datasets was used in YOLOX training with a total of 8566 pictures. The datasets were preprocessed, including random scaling, random cropping, and random brightness change. The pictures were divided into training set and testing set according to the ratio of 9 : 1, and then, 20% of the pictures in the training set were used as the

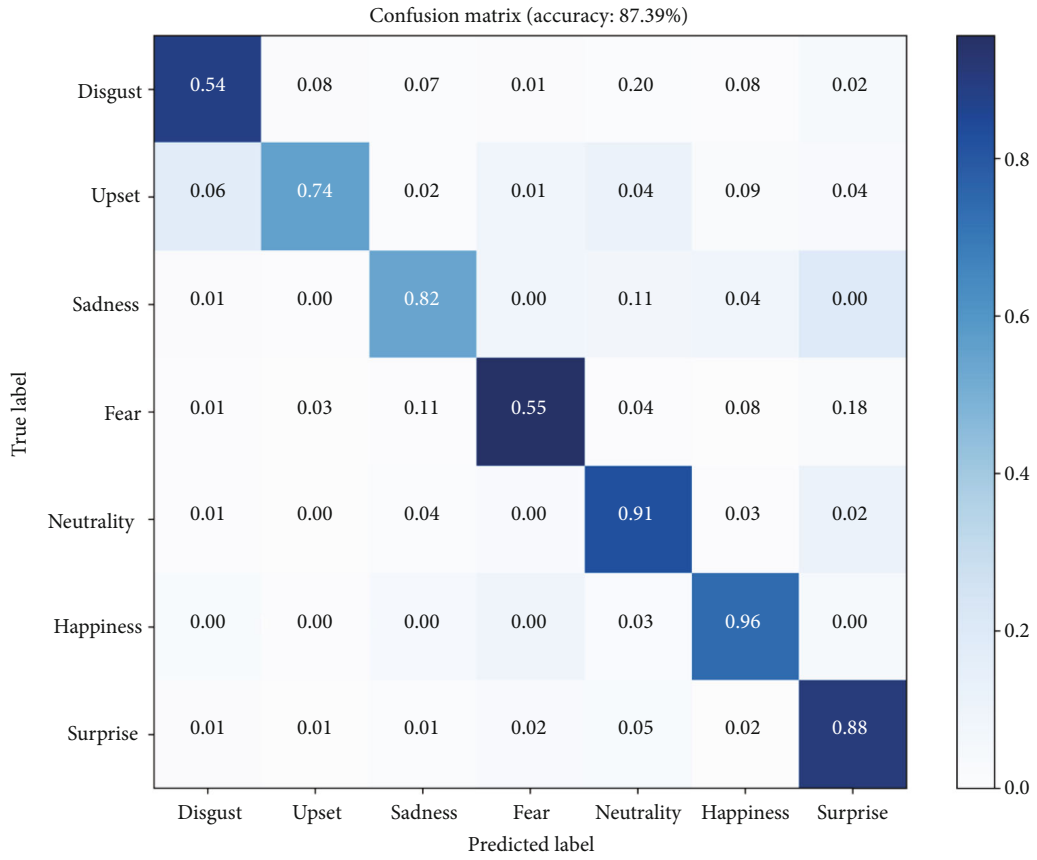

Figure 9: Facial expression recognition confusion matrix.

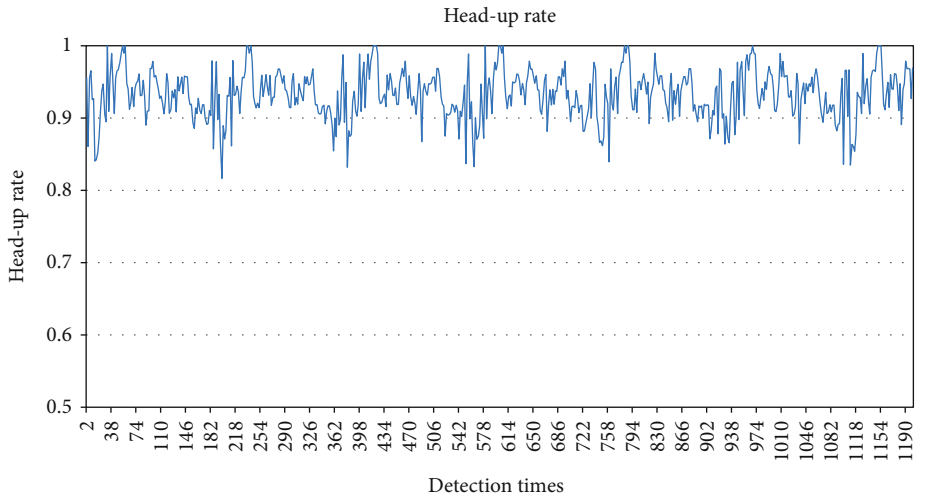

Figure 10: Head-up rate line chart.

verifying set. The training process adopts the frozen training method to improve the training efficiency and to accelerate convergence. The specific relevant parameter settings are shown in Table 4.

The loss curve of the training process is shown in Figure 5. It can be seen that the loss values of the training set and the validation set converge to 0.335 and 0.033, respectively, while the loss values of the training set and the validation set converge to 0.031 and 0.028, respectively.

The detection performance of the model is evaluated by average precision (AP), model size, and FPS. AP refers to the area under the PR (precision-recall) curve. The YOLOX-S

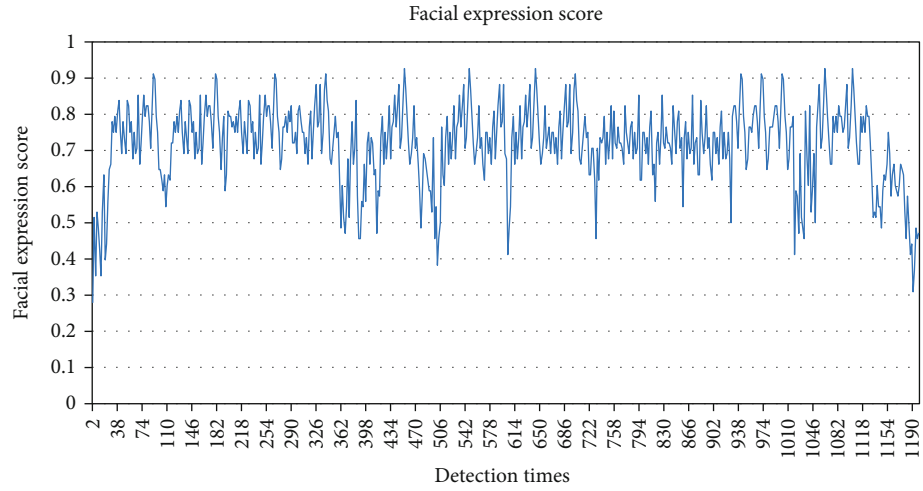

Figure 11: Expression score line chart.

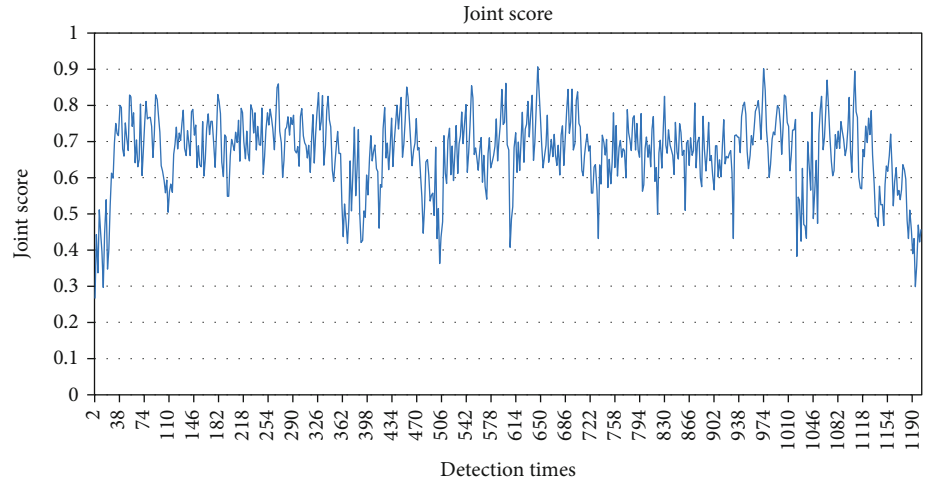

Figure 12: Classroom concentration line chart.

network was compared horizontally and vertically with the same training parameters, and the results are shown in Table 5.

4.3.2. Face Recognition. This paper uses the WIDER FACE dataset for Retinaface training. The training process was SGD, with a momentum size of 0.9, a weight decay of 0.0005, and a batch size of 8. The learning rate starts from $1e^{-3}$ and rises to $1e^{-2}$ after the network updated 5 epochs. Then, at the 34th and 46th epoch, divide it by 10, and the whole training process includes a total of 60 epochs.

To verify the recognition effectiveness of this model, the results of the comparison experiment on the three evaluation subsets of Retinaface-Resnet50 in WIDER FACE under the same training parameters are shown in Table 6 [20–26]. It can be seen that the AP (average accuracy) of Retinaface-Resnet50 on easy, middle, and hard datasets is 0.95, 0.94, and 0.85, respectively, which is a big improvement compared with other models.

4.3.3. Facial Expression Recognition. This paper uses the RAF-DB dataset for SCN training, where the face image is detected and aligned by MTCNN and further resized to 224×224 pixels. SCN is implemented by the Pytorch toolbox, and the backbone network is ResNet18. The training method is the Adam Optimizer optimization model that the learning rate is initialized to 0.1, and at the 15th and 30th epoch being divided by 10. Training stops at the 40th epoch. The relabeling module is included starting at the 10th epoch for optimization, where the relabeling margin δ_2 is set to 0.2 by default.

In order to verify the facial expression recognition performance of this model, the experimental results of the comparison experiments on the RAF-DB dataset of SCN under the same training parameters are shown in Table 7 [27, 28], which shows that the recognition performance of this model is better than using other models.

4.4. Analysis of Performance. To verify the actual performance of the model, we evaluated it with the dataset

collected by ourselves. A teaching video was evaluated repeatedly at 1200 consecutive points, for which the results of face detection, recognition, and expression recognition are demonstrated in Figures 6–8, respectively. It can be seen that this model has excellent detection and recognition performance.

As shown in Figure 9, the recognition accuracy reaches 87.39%. The head-up rate line chart is shown in Figure 10, and the expression score line chart is shown in Figure 11.

From Figures 10 and 11, it can be seen that the overall class head-up rate is more than 0.9, indicating that the class head-up situation is very good. The expression score is above 0.6 most of the time, indicating that the attention status in the classroom is good. However, the overall head-up rate and expression score also fluctuated from time to time, probably due to students taking notes occasionally. Also, because the class size is relatively small, the overall head-up rate and expression score might be greatly affected by some individuals.

Formula (10) was used to calculate the joint head-up rate and expression score for this video. The results are shown in Figure 12.

As can be seen in Figure 12, the joint class attention score rose at the beginning in the time range of 0-50, and the joint score decreased at the end in the time range of 1150-1200. In other words, classroom concentration continued to rise after class started and decreased before class ended. In the time range of 50-1150, the class concentration was above 0.5 most of the time, and the fluctuation showed in this period may be due to students' note-taking activities. The average class concentration of the entire video is 0.6696, which has shown that the class was mainly focused. In our study, we have invited five senior teaching professors to watch the video and conduct manual evaluations. All five teachers' evaluation results have verified and agreed with the results of our integrated evaluation model.

5. Conclusion

This paper is based on the study and optimization of YOLOX, Retinaface, and SCN algorithms. We proposed an integrated evaluation model based on deep learning technology, incorporating YOLOX model, Retinaface model, and SCN model. The experimental results showed that our model can achieve 93.1% object detection accuracy, more than 85% face recognition accuracy, and 87.39% expression recognition accuracy. The model can effectively detect and calculate the head-up rate and expression score of students in a classroom. The model's joint evaluation score correspondents the professors' manual evaluation results. Therefore, our proposed model can assist teachers in evaluating the quality of classroom teaching and objectively analyze the changes in students' learning status in an on-site environment. Future work will be to further refine the algorithms and increase the accuracy of face and expression recognitions.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

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

Acknowledgments

With the consent of all volunteers to disclose the experiment pictures and results, thanks to the volunteers' contribution. This research was funded by the National Natural Science Foundation of China (Grant Number 61802196), the Research Project of Higher Education Reform of Jiangsu Province (Grant Number 2021JSJG250), the Research Project of Higher Education Reform of Nanjing University of Information Science and Technology (Grant Number 2021YBJG12), the Jiangsu Provincial Government Scholarship for Studying Abroad, and the NUIST Students' Platform for Innovation and Entrepreneurship Training Program (Grant Number 202010300080Y).

References

[1] J. Su, R. Xu, S. Yu, B. Wang, and J. Wang, "Idle slots skipped mechanism-based tag identification algorithm with enhanced collision detection," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 5, pp. 2294–2309, 2020.

[2] J. Su, R. Xu, S. Yu, B. Wang, and J. Wang, "Redundant rule detection for software-defined networking," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 6, pp. 2735–2751, 2020.

[3] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*, MIT Press, Cambridge, 1974.

[4] H. Li, L. Yang, Z. Weijia, and S. Peixuan, "Analysis of the teaching effects based on facial expressions in the classroom environment," *Modern Distance Education Research*, vol. 30, no. 4, 2017.

[5] H. Wang, *The Design and Implementation of the Intelligent Education Management Platform Based on Face Recognition*, Beijing Jiaotong University, 2019.

[6] S. Yi, *Research on Evaluation Model of Classroom Attention of Students Based on Face Recognition Technology*, Central China Normal University, 2020.

[7] Z. Jing, *Research on Classroom Focus Analysis Based on Appearance Recognition*, Shanxi University, 2020.

[8] G. Jike and L. Can, "The design and implementation of the classroom environment evaluation system based on emotion recognition," *Office Automation*, vol. 25, no. 17, pp. 43–45, 2020.

[9] Z. M. Chi, Z. Junlang, L. Yangbo, and H. Yuehua, "An online education focus study based on face detection and fuzzy comprehensive evaluation," *Computer Science*, vol. 47, no. S2, pp. 196–203, 2020.

[10] X. Pan, C. Jian, and R. Ma, "Classroom teaching feedback system based on facial expression recognition," *Computer System Application*, vol. 30, no. 10, pp. 102–108, 2021.

[11] C. Xuan, "A study of the teacher classroom evaluation model in the background of artificial intelligence," *Modern Information Technology*, vol. 5, no. 6, pp. 147–149, 2021.

[12] L. Wu, C. Yanan, and C. Yiming, "Framework construction of ai-enabled classroom teaching evaluation reform and technology implementation," *China Audio-visual Education*, vol. 5, pp. 94–101, 2021.

[13] J. Guo, "The design and implementation of the attention detection system based on students' learning process," Beijing University of Posts and Telecommunications, 2021.

[14] W. Luo, G. Kocun, Y. Yang, L. Chang, S. Shoukang, and T. Dongming, "Analysis of students in video stream based on facial expression recognition," *Modern Computers*, vol. 18, pp. 117–121, 2021.

[15] J. Li, J. Zhengxian, H. Lu, Q. Xu, and H. Fangliang, "Evaluation model based on ResNet algorithm," *Journal of Lanzhou University of Arts and Sciences (Natural Science Edition)*, vol. 35, no. 6, pp. 62–66, 2021.

[16] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: exceeding yolo series in 2021," 2021, https://arxiv.org/abs/2107.08430.

[17] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-stage dense face localisation in the wild," 2019, https://arxiv.org/abs/1905.00641.

[18] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

[19] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto:Consulting Psychologists Press, 1978.

[20] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, 2017https://arxiv.org/abs/1706.02863.

[21] P. Y. Hu and D. Ramanan, "Finding tiny faces," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 1522–1530, Honolulu, HI, USA, 2017.

[22] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: single shot scale-invariant face detector," in *2017 IEEE International Conference on Computer Vision, ICCV 2017*, pp. 192–201, Venice, Italy, 2017.

[23] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-Net: face detection through deep facial part responses," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 8, pp. 1845–1859, 2018.

[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[25] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017https://arxiv.org/abs/1709.05256.

[26] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: a context-assisted single shot face detector," in *15th European Conference on Computer Vision*, pp. 812–828, Munich, Germany, 2018.

[27] S. Li, W. Deng, and D. JunPing, "Reliable crowd-sourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861, 2017.

[28] J. Zeng, S. Shan, X. Chen, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 222–237, 2018.