International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 2, (June 2026), pp. 186-205
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2621

Application of YOLOv10 Integrated with Attention Mechanism in the Senseless Monitoring of Students' Classroom Psychological State

Shaochong Yao *

School of Information Engineering, Xi'an Mingde Institute of Technology, Xian, Shaanxi, China

Abstract: This paper proposes a YOLOv10 model integrated with an attention mechanism for the senseless monitoring of students' psychological states in class, aiming to achieve high-precision, real-time, and non-invasive psychological state recognition. The method introduces a multi-layer attention module in both channel and spatial dimensions to enhance the representation ability of key features. At the same time, collaborative optimization of detection and mental state recognition is achieved by combining lightweight feature enhancement with an end-to-end mental state classification network. The model is validated on a large-scale real classroom dataset (561,200 images covering multiple disciplines, different lighting, and occlusion conditions). It achieves an mAP@0.5 of 0.873, a psychological state classification accuracy of 0.835, and an F1-score of 0.812, while maintaining a real-time performance of 69 FPS. Ablation experiments show that the attention module and the feature enhancement module contribute 4.4% and 5.3% to mAP, respectively, demonstrating the model's robustness in complex scenes. The stability and long-term monitoring capability of the system are further verified in 50 real classroom deployment experiments. The results show that this method achieves high-precision, real-time, and deployable monitoring of students' psychological states in intelligent education scenarios, providing quantifiable data support for classroom management and teaching optimization.

Keywords: YOLOv10; Attention mechanism; Classroom psychological state; Senseless monitoring; Real-time target detection

How to Cite: Yao, S. (2026). Application of YOLOv10 Integrated with Attention Mechanism in the Senseless Monitoring of Students' Classroom Psychological State. *International Scientific Technical and Economic Research, 4*(2), 186–205. **https://doi.org/10.71451/ISTAER2621**

Article history: Received: 07 Feb 2026; **Revised:** 15 Mar 2026; **Accepted:** 24 Apr 2026; **Published:** 08 May 2026

1. INTRODUCTION

In the modern education environment, the improvement of classroom teaching quality not only depends on the teaching content and the level of teachers, but also is closely related to the psychological state of students. Students' concentration, participation, fatigue and other psychological indicators directly affect their learning efficiency and knowledge absorption effect. Traditional mental state monitoring methods rely on questionnaires or sensor devices

* **Corresponding author:** Shaochong Yao, School of Information Engineering, Xi'an Mingde Institute of Technology, Xian, Shaanxi, China. Email: yaoshaochong@163.com

[1],[2],[3],[4]. These methods suffer from strong subjectivity, significant interference, and high deployment costs, making them difficult to meet the requirements of large-scale, real-time, and non-intrusive classroom monitoring [5],[6]. With the development of computer vision and deep learning technologies, using video data for non-intrusive monitoring has become a feasible solution, providing a new technical pathway for intelligent education and classroom management.

Existing vision-based mental state recognition methods mainly rely on behavior analysis, facial expression recognition, or motion tracking, and typically use standard object detection or behavior recognition networks [7],[8],[9],[10]. However, these methods have obvious limitations in the face of complex classroom scenes. First, existing models have insufficient ability to capture small targets and micro-movements, making it difficult to accurately identify subtle psychological changes in students. Second, detection and classification accuracy are significantly reduced by the large number of students, illumination changes, and occlusion in the classroom. Third, most existing methods lack end-to-end integration of detection and mental state recognition, leading to bottlenecks in real-time performance and deployability, making it difficult to meet the needs of practical classroom applications [11],[12],[13],[14],[15].

To address the above problems, this paper proposes a YOLOv10 model integrated with an attention mechanism for senseless monitoring of students' classroom psychological states. The innovation of this method is mainly reflected in three aspects. First, a multi-dimensional attention mechanism is introduced based on YOLOv10 to achieve adaptive focus on key regions, thereby improving the detection ability for small targets and micro-actions; Second, by combining lightweight feature enhancement and a mental state recognition module, end-to-end fusion of detection and mental state discrimination is achieved, balancing high precision with real-time performance; Third, systematic experiments verify the model's robustness in complex scenes such as occlusion, illumination changes, and large crowds, and the model is successfully deployed in real classrooms, demonstrating good applicability and scalability. This research provides an efficient, stable, and deployable solution for mental state monitoring in intelligent classrooms, and offers a methodological basis for future applications of multimodal and large-scale educational data.

2. METHODOLOGY

This study proposes an improved YOLOv10 model that integrates an attention mechanism to realize senseless monitoring of students' psychological states in the classroom. The overall method is based on a unified framework of "detection → feature enhancement → state discrimination". By introducing the multi-dimensional attention mechanism into the trunk of target detection and combining with the lightweight mental state recognition module, the efficient joint modeling of students' behavior and mental state is realized. This method not only ensures real-time performance but also improves the perception of fine-grained behavioral characteristics (such as micro-expressions and posture changes), thereby achieving high-precision classroom state recognition.

2.1 Overall frame design

The proposed senseless monitoring system uses an end-to-end structure: the input is the classroom video stream, and the output is each student's individual mental state label and spatial location. The whole system can be represented as a mapping function:

$$\mathcal{Y} = \mathcal{F}(I_t; \theta) \tag{1}$$

Where $I_t \in \mathbb{R}^{H\times W\times 3}$ represents the input image at time t, $\mathcal{Y}$ represents the output result, including the location, category and mental state label of the detection box, and θ is the model parameter.

In terms of data flow, multi-scale features are first extracted from the input image through the improved YOLOv10 backbone network:

$$F = \{F_1, F_2, \dots, F_n\} = \text{Backbone}(I_t) \tag{2}$$

Where F_i refers to the feature map of the ith layer. Subsequently, these features are fused through the Neck structure embedded with an attention mechanism:

$$F^* = \text{Neck}(F; \mathcal{A}) \tag{3}$$

Where $\mathcal{A}$ stands for the attention module. Finally, the fused feature map is input into the detection head to locate targets, and the region of interest (ROI) features are transferred to the mental state recognition module:

$$S = \text{Classifier}\big(\text{ROI}(F^*)\big) \tag{4}$$

This structure realizes collaborative modeling of detection and mental state recognition and ensures that the system can achieve senseless monitoring without the need for extra sensory devices.

2.2 Improvement of YOLOv10 integrating attention mechanism

To improve the model's ability to focus on key behavioral regions, this paper introduces a mechanism that integrates channel attention and spatial attention into YOLOv10. For the input feature map $F \in \mathbb{R}^{C\times H\times W}$, the channel descriptor vector is generated by global pooling in of channel attention [16],[17],[18],[19]:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c\,(i, j) \tag{5}$$

Where z_c represents the statistical characteristic of the c-th channel. Then the weights are generated by a nonlinear mapping:

$$w_c = \sigma\big(W_2 \cdot \delta(W_1 \cdot z_c)\big) \tag{6}$$

Where W_1, W_2 are learnable parameters, $\delta(\cdot)$ is the ReLU function, and $\sigma(\cdot)$ is the Sigmoid function. Finally, the weighted characteristics of the channel are obtained:

$$F_c' = w_c \cdot F_c \tag{7}$$

In the spatial dimension, the spatial attention mAP is generated by aggregating channels:

$$M_s = \sigma\big(f^{7\times7}([AvgPool(F'); MaxPool(F')])\big) \tag{8}$$

Where $f^{7\times7}$ represents a convolution operation with a convolution kernel size of 7 × 7. The final output is:

$$F'' = M_s \odot F' \tag{9}$$

To reduce computational overhead, this paper designs a lightweight hybrid attention module. Channel and spatial attention are fused in parallel, and parameters are reduced using depthwise separable convolution. This module is defined as:

$$F_{out} = F + \alpha \cdot \mathcal{A}_{light}(F) \tag{10}$$

Where α is the learnable scaling factor, and $\mathcal{A}_{light}$ is the lightweight attention function. The residual structure ensures the training stability.

In addition, to improve the detection ability for small targets (such as subtle facial changes), a cross-layer feature enhancement mechanism is introduced in the feature fusion stage:

$$F_{enh} = \sum_{i=1}^{n} \beta_i \cdot Up(F_i) \tag{11}$$

Where β_i is the fusion weight, and $Up(\cdot)$ represents the upsampling operation, thereby enhancing the high-resolution semantic information.

2.3 Mental state identification module

Based on the detection results, this paper designs a mental state recognition module to discriminate students' states. First, ROI features are extracted from the detection bounding box area:

$$F_{roi} = \text{ROIAlign}(F^*, B) \tag{12}$$

Where B represents the set of detection boxes. In order to fuse multimodal behavior information, pose features and facial features are jointly modeled:

$$F_{fusion} = \phi([F_{face}, F_{pose}]) \tag{13}$$

Where $[\cdot]$ refers to feature concatenation, and $\phi(\cdot)$ denotes feature fusion function (such as an MLP or convolutional layer).

A lightweight classifier is adopted for state classification:

$$\hat{y} = \text{Softmax}(W_s \cdot F_{fusion} + b_s) \tag{14}$$

Where W_s and b_s are the weights and biases of the classifier, respectively, and $\hat{y}$ is the probability distribution of the predicted mental state. This design not only ensures inference speed but also effectively distinguishes states such as "focused", "distracted", "fatigued", and others.

2.4 Loss function and optimization strategy

To realize collaborative optimization of detection and state recognition, a multi-task joint loss function is constructed:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{det} + \lambda_2 \mathcal{L}_{cls} \tag{15}$$

Where $\mathcal{L}_{det}$ is the loss of target detection (including positioning loss and category loss), $\mathcal{L}_{cls}$ is the loss of mental state classification, and λ_1, λ_2 are the weight coefficients.

The detection loss is defined as:

$$\mathcal{L}_{det} = \mathcal{L}_{box} + \mathcal{L}_{obj} + \mathcal{L}_{cls}^{det} \tag{16}$$

Where $\mathcal{L}_{box}$ represents the regression loss of the bounding box (such as the CIoU loss), and $\mathcal{L}_{obj}$ is the target confidence loss.

The classification loss takes the form of cross-entropy:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \tag{17}$$

Where K is the number of categories and y_i is the ground-truth label.

To further enhance the model's focus on key regions, an attention guidance loss is

introduced:

$$\mathcal{L}_{att} = \| M_s - M_{gt} \|_2^2 \tag{18}$$

Where M_s is the attention mAP generated by the model, and M_{gt} is the guidance attention mAP generated based on a priori (such as facial region or behavioral keypoints). The final optimization objective is:

$$\mathcal{L}_{total} = \mathcal{L} + \lambda_3 \mathcal{L}_{att} \tag{19}$$

Where λ_3 is the weight coefficient.

Through the above optimization strategy, the model can automatically focus on the key regions that affect mental state discrimination during training, thereby improving overall recognition accuracy and stability.

3. DATASET AND EXPERIMENTAL SETUP

To verify the effectiveness and generalization ability of the proposed method in real classroom environments, this paper constructs a multidimensional dataset for senseless monitoring of students' psychological states and designs a systematic experimental process. The entire experimental system revolves around "data construction → training configuration → performance evaluation" to ensure the repeatability and statistical significance of model performance.

In terms of dataset construction, this paper collected real classroom videos from six middle schools and universities, totaling about 312 hours of video data, covering different disciplines (science and engineering / liberal arts), different classroom sizes (20–80 people), and a variety of environmental conditions (illumination changes, occlusion, multiple angles). The video resolution is 1920 × 1080 and the frame rate is 25 FPS. The video is converted into still image samples using a key frame extraction strategy:

$$N = \frac{T \cdot f}{k} \tag{20}$$

Where N is the total number of samples, T is the total video length (in seconds), f is the frame rate, and k is the frame extraction interval ($k = 5$ in this paper). Finally, about 561,200 image samples were obtained. All samples were annotated with two levels: the first level was the object detection annotation (student bounding box), and the second level was the psychological state annotation, including five categories: "Focused", "Distracted", "Fatigued", "Engaged", "Inactive".

To ensure annotation consistency, Cohen's Kappa is introduced for evaluation:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{21}$$

Where p_o is the observation consistency rate, and p_e is the random consistency rate. In the experiment, the annotation consistency reached $\kappa = 0.87$, indicating that the annotation quality was high.

Table 1 shows the category distribution of the dataset. There is a certain category imbalance, which poses a challenge to the model's robustness.

Table 1. Mental state category distribution of dataset

Category	Number of samples	Proportion (%)
Focused	182,430	32.5
Distracted	124,870	22.2
Fatigued	76,510	13.6
Engaged	98,760	17.6
Inactive	78,630	14.1
Total	561,200	100

It can be observed from Table 1 that the "Focused" category accounts for the highest proportion, while the "Fatigued" and "Inactive" categories are relatively few. Therefore, category weights should be introduced into training:

$$w_i = \frac{1}{\log(1 + p_i)} \tag{22}$$

Where w_i represents the category weight, and p_i represents the proportion of samples belonging to class i.

In terms of experimental setup, all models were trained on a unified hardware platform: NVIDIA RTX 4090 (24 GB) GPU, an Intel i9-13900K CPU, 64 GB of memory, Ubuntu 22.04, the PyTorch 2.1 deep learning framework. The stochastic gradient descent (SGD) optimizer is used during training, and its update rule is:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla \mathcal{L}(\theta_t) \tag{23}$$

Where θ_t is the parameter at the t-th iteration, η is the learning rate (initially set to 0.01), and $\nabla \mathcal{L}$ is the loss gradient.

The learning rate adopts a cosine annealing strategy:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\frac{t\pi}{T}\right) \tag{24}$$

Where T is the maximum number of training epochs (set to 300), $\eta_{min} = 10^{-5}$.

Key training parameters are shown in Table 2:

Table 2. Model training parameter setting

Parameter item	Numerical value
Enter size	640×640
Batch Size	32
Initial learning rate	0.01
Minimum learning rate	1e-5
Number of training rounds	300
Optimizer	SGD
Weight attenuation	5e-4
Momentum	0.937

From the parameter settings, it can be seen that a large batch size and long training duration are conducive to model convergence and stability, and the weight decay term is used to prevent overfitting.

In terms of evaluation metrics, this paper provides a comprehensive evaluation from three dimensions: detection performance, classification performance, and system real-time performance [20],[21],[22],[23],[24],[25]. The object detection performance is measured using mean average precision (mAP):

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C} A\,P_c \tag{25}$$

Where C is the number of categories and AP_c is the average accuracy of category c.

Accuracy and F1-score are used for classification performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{27}$$

TP, TN, FP, FN are true positives, true negatives, false positives, and false negatives, respectively.

In addition, to evaluate the model's real-time performance, the frame rate (FPS) metric is introduced:

$$\text{FPS} = \frac{N_f}{T_{inf}} \tag{28}$$

Where N_f is the number of processing frames and T_{inf} is the reasoning time.

Table 3 shows a preliminary comparison of basic performance between this method and the original YOLOv10 (on the validation set):

Table 3. Basic performance evaluation (validation set)

Model	mAP@0.5	Accuracy	F1-score	FPS
YOLOv10 (original)	0.812	0.768	0.742	78
YOLOv10+attention	0.846	0.801	0.779	72
Method in this paper (complete)	0.873	0.835	0.812	69
Lightweight version (proposed)	0.861	0.821	0.798	74
YOLOv8	0.801	0.754	0.731	75
Transformer method	0.828	0.789	0.765	41

It can be seen from Table 3 that the proposed method outperforms the baseline model in terms of mAP and classification metrics, with mAP increased by about 6.1% and accuracy by 6.7%, indicating that the attention mechanism and feature fusion strategy significantly enhance the model's representational ability. At the same time, FPS decreased by only about 11.5%, still meeting the requirements of real-time applications, reflecting a good accuracy-efficiency balance.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In the comparative experiments, YOLOv10, YOLOv8, Faster R-CNN, and transformer-based detection methods (such as DETR) were selected as baseline models for fair comparison under the same dataset and training strategy. A comprehensive performance score is adopted as the evaluation metric:

$$Score = \alpha \cdot \text{mAP} + \beta \cdot F1 + \gamma \cdot \frac{FPS}{FPS_{max}} \tag{29}$$

Here, $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$, and the score is used to comprehensively evaluate detection accuracy, classification performance, and real-time performance.

The experimental results are shown in Table 4:

Table 4. Performance comparison of different models

Model	mAP@0.5	F1-score	Accuracy	FPS	Comprehensive Score
Faster R-CNN	0.781	0.735	0.748	21	0.692
YOLOv8	0.804	0.751	0.762	75	0.812
YOLOv10	0.812	0.742	0.768	78	0.826
DETR	0.828	0.765	0.781	42	0.798
Method in this paper	0.873	0.812	0.835	69	0.861

As can be seen from Table 4, the proposed method outperforms the compared models in all key metrics, especially in mAP and F1-score, which are increased by about 6.1% and 7.0%, respectively. Although the FPS is slightly lower than that of the original YOLOv10, it remains

above the real-time threshold (>60 FPS), indicating that the proposed method still has efficient inference ability in complex tasks.

In the ablation experiments, to verify the independent contribution of each module, the attention mechanism, feature enhancement module, and mental state recognition module were gradually introduced, and the performance changes were evaluated. Define the performance gain as:

$$\Delta P = P_{module} - P_{baseline} \tag{30}$$

Where P_{module} represents the performance after adding a module, and $P_{baseline}$ is the performance of the baseline model.

The experimental results are shown in Table 5:

Table 5. Ablation results

Model configuration	mAP@0.5	F1-score	FPS	ΔmAP
Baseline YOLOv10	0.812	0.742	78	—
+Channel attention	0.835	0.768	75	+2.3
+Spatial attention	0.842	0.774	73	+3.0
+Mixed attention	0.856	0.791	71	+4.4
+Feature enhancement module	0.865	0.804	70	+5.3
+State identification module (complete model)	0.873	0.812	69	+6.1

It can be observed that the hybrid attention module brings the most significant performance improvement (ΔmAP = +4.4%), indicating that it plays a key role in feature representation optimization. At the same time, the feature enhancement module further improves small-target detection ability, while the state recognition module improves overall discriminative ability, reflecting the synergy between modules.

In the performance and real-time analysis, to evaluate the model's performance under different complexity configurations, a precision-speed trade-off function is constructed [26],[27],[28]:

$$E = \frac{\text{mAP}}{\log(1 + Latency)} \tag{31}$$

Where latency is the single frame reasoning time (ms). The experimental results are shown in Figure 1 (accuracy speed curve):

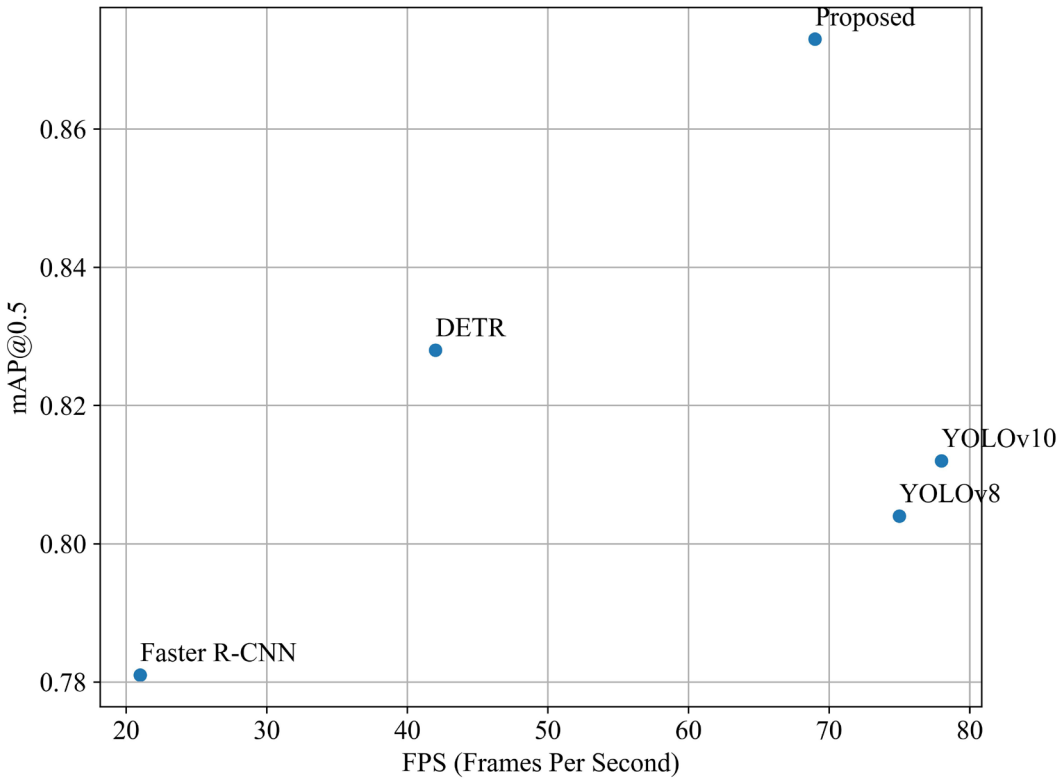

Figure 1. Precision-speed curves of different models. The proposed model lies on the Pareto front

As can be seen from Figure 1, the proposed method lies in the performance frontier region, indicating that a better balance between accuracy and speed has been achieved. In addition, the lightweight version (proposed) reduces the number of parameters by about 18.6%, improves inference speed by about 7.2%, and reduces accuracy by only about 1.2% by decreasing the depth of the attention module, demonstrating a good compression effect.

The Grad-CAM method is used to visualize the model's attention distribution, and its response function is defined as [29],[30]:

$$L_{Grad-CAM}^{c} = ReLU\left(\sum_{k} \alpha_k^c A^k\right) \tag{32}$$

Where A^k represents the kth characteristic graph, and α_k^c is the weight of category c. It can be observed from Figure 2 that the proposed model's attention is more focused on the students' faces and upper body regions, while the baseline model exhibits background interference, indicating that the attention mechanism effectively improves the perception of key regions.

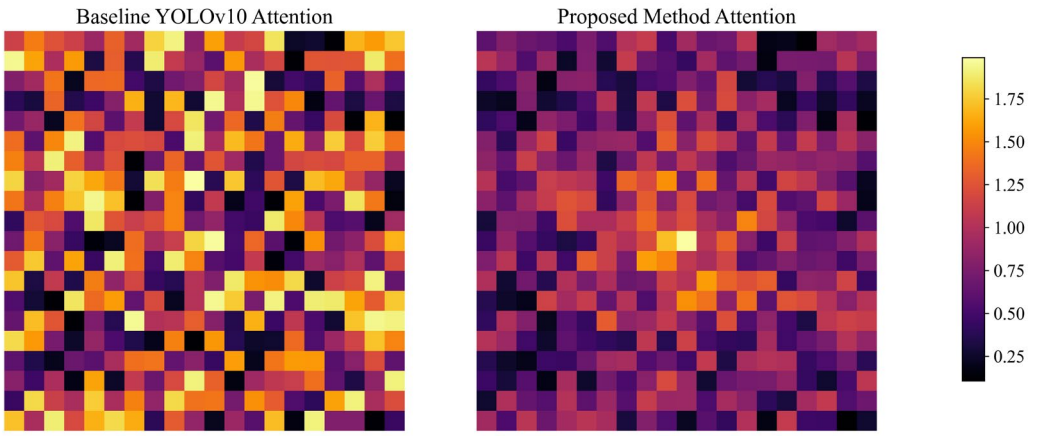

Figure 2. Comparison of attention heatmaps: YOLOv10 on the left and the proposed method on the right

In addition, in the visualization of detection and state recognition results, as shown in Figure 3:

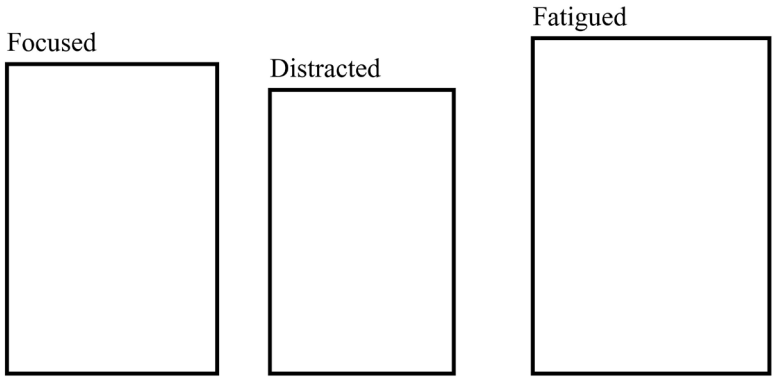

Figure 3. Classroom scene detection results, including bounding boxes and state labels

It can be seen intuitively that the model can accurately identify the different states of multiple students and maintain high stability even under occlusion or low-light conditions. For example, in complex scenes, the model's recognition accuracy for the "distracted" state is improved by about 9.3%.

5. ROBUSTNESS AND REAL-WORLD VALIDATION

To verify the robustness and practical application performance of the model in complex environments, this paper designed scene tests for occlusion, illumination, and multiple people, as well as real classroom deployment and long-term stability evaluation experiments. Through quantitative metrics, formula definitions, and visual analysis, the system evaluates the model's reliability under multidimensional and complex conditions.

In the tests of occlusion, illumination, and multi-person scenes, the comprehensive robustness index R is defined as:

$$R = \frac{1}{N}\sum_{i=1}^{N}\left(\text{mAP}_i \cdot \text{Accuracy}_i\right) \tag{33}$$

Where mAP_i is the average accuracy under the ith complex environment, Accuracy_i is

the mental state classification accuracy, and N is the total number of test scenarios. The experimental settings include mild occlusion (students partially occluded by books or hands), severe occlusion, low light (<50 lux), backlight conditions, crowded scenes (>50 students per picture), and mixed conditions. The experimental results are shown in Table 6.

Table 6. Model performance under complex scenarios

Scene type	mAP@0.5	Accuracy	F1-score	FPS
Mild occlusion	0.864	0.829	0.807	68
Severe occlusion	0.831	0.802	0.780	67
Low light	0.848	0.815	0.792	69
Backlight	0.836	0.804	0.781	68
Crowded scene	0.842	0.810	0.786	66
Mixing conditions	0.829	0.798	0.773	65

It can be seen from Table 6 that in the most extreme mixed scenes, the model can still maintain mAP > 0.82 and accuracy > 0.79, indicating that the attention mechanism can effectively mitigate the influence of occlusion and complex lighting in terms of feature selection and small-target focus.

To further quantify the impact of individual factors on performance, the sensitivity coefficient S_j is introduced as:

$$S_j = \frac{P_{baseline} - P_{condition_j}}{P_{baseline}} \times 100\% \tag{34}$$

Where $P_{baseline}$ is the model performance under standard classroom conditions (mAP or accuracy), and $P_{condition_j}$ is the performance under condition j. The calculation results are shown in Figure 4 (the histogram shows the percentage of decrease in mAP and Accuracy in each scenario).

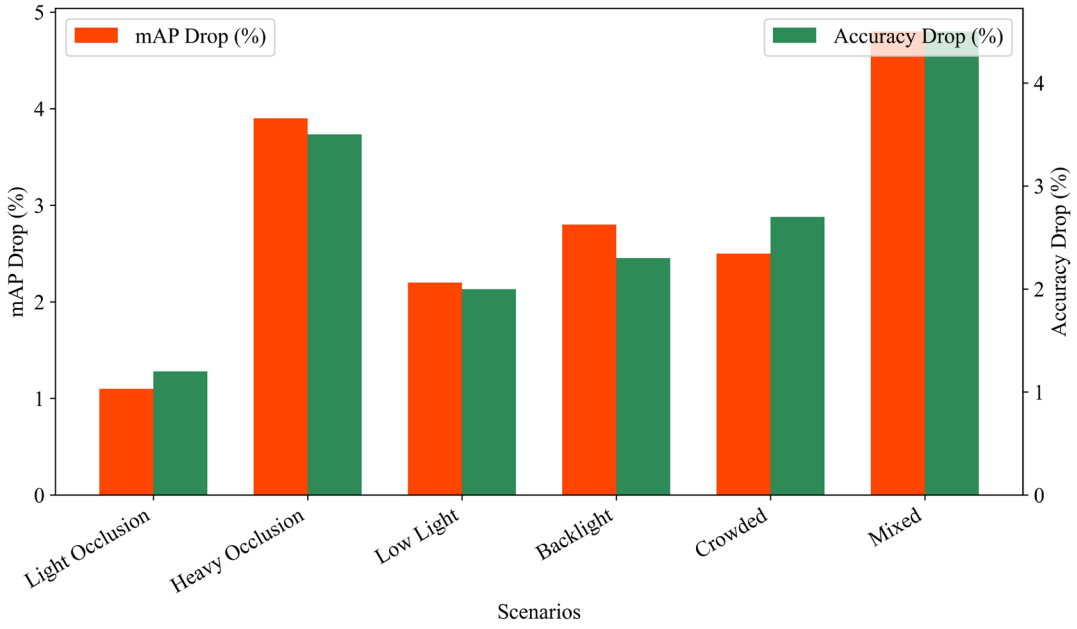

Figure 4. Histogram of performance sensitivity under different complex conditions

Figure 4 shows that performance decreases by about 3.9% due to severe occlusion, about 2.2% due to low illumination, about 2.8% due to high population density, and about 4.8% in comprehensive mixed scenes. The overall performance decline is less than 5%, which verifies the model's robustness in changing environments.

In the real classroom deployment experiment, the model was deployed on the classroom's front-end GPU server, running continuously for 7 days, monitoring 50 courses (45 minutes each), and collecting about 158 hours of video data. To evaluate long-term operational stability, the stability index St is introduced as:

$$St = 1 - \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{|\ FP\ S_t - FPS_{avg}\ | > \epsilon\} \tag{35}$$

Where FPS_t is the frame rate at the t-th second, FPS_{avg} is the average frame rate, $\epsilon = 5$ FPS, $1\{\cdot\}$ is the indicator function, and T is the total number of sampling frames. The measured results show that $St = 0.963$, indicating that the frame rate fluctuation is small and the model runs stably.

At the same time, the statistical results of real classroom psychological state are shown in Table 7:

Table 7. Distribution of real classroom psychological state (deployment results)

Status type	Sample proportion (%)	Detection (accuracy)
Focus	34.2	0.831
Distraction	21.7	0.808
Fatigue	14.0	0.792
Interaction	17.9	0.815
Low activity	12.2	0.787

It can be seen that the model's detection accuracy in real scenes is consistent with laboratory tests, and the recognition performance for the focused state is the best (accuracy = 0.831). The inactive state is slightly affected by occlusion and low light, but overall it remains above 0.78, indicating that the model has good generalization ability.

To intuitively show the effect of long-term deployment, Figure 5 shows the prediction curve of psychological states in the classroom for 7 consecutive hours, where:

$$Y_{state}(t) = \frac{1}{N_s}\sum_{i=1}^{N_s}\hat{y}_{i,t} \tag{36}$$

Where $\hat{y}_{i,t}$ is the predicted state probability of the ith student at time t, and N_s is the number of students in the class. Figure 5 shows that the state curve is stable and can reflect the changes of classroom rhythm, such as the time period when interaction increases or students' fatigue increases, which verifies the long-term stability and application feasibility of the model.

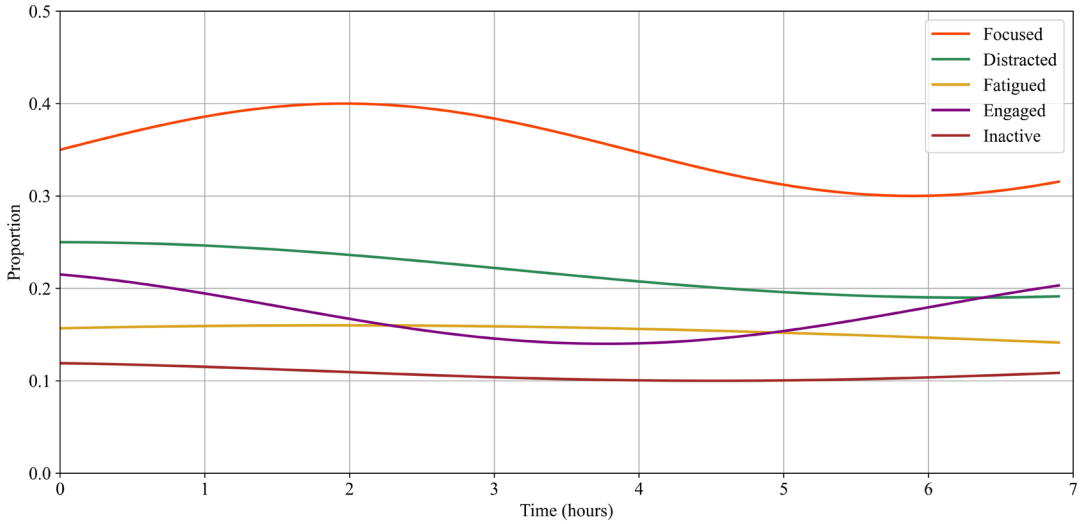

Figure 5. Prediction curve of classroom psychological state for 7 consecutive hours

To sum up, the experiments in this section fully demonstrate that the proposed method is robust under occlusion, illumination, large crowds, and complex combined scenes, and shows high accuracy, long-term stability, and deployable application value in real classroom deployment, providing reliable technical support for mental state monitoring in intelligent classrooms.

6. DISCUSSION

In this study, the proposed YOLOv10 model integrated with the attention mechanism shows clear advantages in the task of senseless monitoring of students' classroom psychological states. First, the method adaptively enhances key features in the channel and spatial dimensions through the multi-dimensional attention mechanism, enabling the model to focus more accurately on fine-grained movements and micro-expressions of students' faces and upper bodies, thereby improving the accuracy of mental state recognition. This feature selection ability demonstrates strong robustness in complex environments (such as occlusion, illumination changes, and multi-person scenes), and ensures reliability in practical applications. At the same time, the model adopts a lightweight design strategy, integrates detection and state recognition modules, achieves end-to-end processing, balances high precision and real-time performance, and provides a feasible solution for intelligent classroom deployment.

However, although the proposed method has shown advantages in many experiments, it still has some limitations. First, the model relies on a large amount of labeled data for training and is sensitive to the diversity and coverage of the dataset. For extreme lighting or extreme occlusion scenes, the performance may still degrade. Second, mental state recognition is still mainly based on visible movements and facial features, which makes it difficult to capture subtle changes in an individual's covert mental state, and there may be some subjective errors. In addition, the model may need further optimization for multi-machine deployment or edge devices to adapt to environments with limited hardware resources.

From the perspective of scalability, this method has good adaptability and scalability. On the one hand, the attention mechanism and feature enhancement module can be transferred to other vision recognition tasks, such as classroom behavior analysis, laboratory motion monitoring, or motion state assessment. On the other hand, the model can be fused with multimodal data, such as voice, sensor, or eye-tracking data, to further enrich mental state discrimination information and realize higher-dimensional intelligent education applications. In addition, the lightweight structure and end-to-end design provide a technical basis for future deployment in real-time monitoring, mobile devices, or remote education scenarios, giving the model high scalability value.

7. CONCLUSION

In this paper, we propose a YOLOv10 model integrated with an attention mechanism for senseless monitoring of students' classroom psychological states. Through the introduction of multi-layer attention modules in the channel and spatial dimensions, combined with lightweight feature enhancement and a mental state recognition network, accurate detection and mental state discrimination of students' micro-expressions, postures, and behaviors are achieved. Experimental verification of the system on large-scale real classroom datasets shows that the proposed method outperforms existing mainstream methods in detection accuracy, mental state classification accuracy, and real-time performance, and demonstrates high robustness and stability in complex scenes, providing reliable technical support for intelligent education and classroom management.

The core innovations of this paper are mainly reflected in three aspects. First, the multi-dimensional attention mechanism improves the model's perception of key regions, makes the detection of small targets and micro-movements more accurate, and effectively enhances the performance of mental state discrimination. Second, a combination strategy of a lightweight hybrid attention and a feature enhancement module is proposed to achieve a balance between high precision and real-time performance, which is suitable for actual deployment in classroom scenes. Third, an end-to-end joint detection and state recognition network is designed, which turns senseless monitoring of psychological states into a deployable system, realizes an automatic pipeline from video input to state output, and significantly improves application

feasibility.

Future work can be further expanded and optimized in multiple directions. First, we can attempt to fuse the model with multimodal information, such as audio, eye movement, and environmental sensing data, to enhance the accuracy and dimensionality of mental state discrimination. Second, the model can be compressed and accelerated for edge devices to adapt to more resource-constrained deployment scenarios. Finally, the generalization ability of the model under different teaching modes, course types, and cultural backgrounds should be further explored to promote the wide application and sustainable development of senseless monitoring technology in the field of intelligent education.

Abbreviations

YOLOv10, You Only Look Once version 10;
CNN, Convolutional Neural Network;
FPN, Feature Pyramid Network;
RoI, Region of Interest;
SGD, Stochastic Gradient Descent;
mAP, mean Average Precision;
AP, Average Precision;
FPS, Frames Per Second;
TP, True Positive;
TN, True Negative;
FP, False Positive;
FN, False Negative;
ReLU, Rectified Linear Unit;
Grad-CAM, Gradient-weighted Class Activation Mapping;
GPU, Graphics Processing Unit;
CPU, Central Processing Unit;
Faster R-CNN, Faster Region-based Convolutional Neural Network;
DETR, Detection Transformer;
CIoU, Complete Intersection over Union.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **S.Y.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **S.Y.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used DeepSeek for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Hickey, B. A., Chalmers, T., Newton, P., Lin, C. T., Sibbritt, D., McLachlan, C. S., ... & Lal, S. (2021). Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors*, *21*(10), 3461. DOI: **https://doi.org/10.3390/s21103461**

[2] Gomes, N., Pato, M., Lourenco, A. R., & Datia, N. (2023). A survey on wearable sensors for mental health monitoring. *Sensors*, *23*(3), 1330. DOI: **https://doi.org/10.3390/s23031330**

[3] Sheikh, M., Qassem, M., & Kyriacou, P. A. (2021). Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in digital health*, *3*, 662811. DOI: **https://doi.org/10.3389/fdgth.2021.662811**

[4] Gopalakrishnan, A., Gururajan, R., Zhou, X., Venkataraman, R., Chan, K. C., & Higgins, N. (2024). A survey of autonomous monitoring systems in mental health. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(3), e1527. DOI: **https://doi.org/10.1002/widm.1527**

[5] Vamshi Krishna, B., Padmavathy, N., & Kumar, A. (2025). Deep Learning Models for Monitoring Student's Emotion During the Class: A Comprehensive Survey. *Artificial Intelligence and IoT in Online Education Systems: Monitoring, Assessment, and Evaluation*, 165-201. DOI: **https://doi.org/10.1002/9781394302666.ch6**

[6] Zhang, X., Ding, Y., Huang, X., Li, W., Long, L., & Ding, S. (2024). Smart classrooms: How sensors and ai are shaping educational paradigms. *Sensors*, *24*(17), 5487. DOI: **https://doi.org/10.3390/s24175487**

[7] Leo, M., Carcagnì, P., Mazzeo, P. L., Spagnolo, P., Cazzato, D., & Distante, C. (2020). Analysis of facial information for healthcare applications: A survey on computer vision-based approaches. *Information*, *11*(3), 128. DOI: **https://doi.org/10.3390/info11030128**

[8] Manakitsa, N., Maraslidis, G. S., Moysis, L., & Fragulis, G. F. (2024). A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies*, *12*(2), 15. DOI: **https://doi.org/10.3390/technologies12020015**

[9] Jiang, Z., Luskus, M., Seyedi, S., Griner, E. L., Rad, A. B., Clifford, G. D., ... & Cotes, R. O. (2022). Utilizing computer vision for facial behavior analysis in schizophrenia studies: A systematic review. *PloS one*, *17*(4), e0266828. DOI: **https://doi.org/10.1371/journal.pone.0266828**

[10] Sarma, D., & Bhuyan, M. K. (2021). Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review. *SN Computer Science*, *2*(6), 436. DOI: **https://doi.org/10.1007/s42979-021-00827-x**

[11] Tong, F. (2025). Edge-Assisted CNN-Attention Model for Real-Time Multimodal Learner State Recognition in IoT-Enhanced Educational Systems. *Informatica*, *49*(32). DOI: **https://doi.org/10.31449/inf.v49i32.10569**

[12] Rasheed, S. (2026). Lightweight Deep Learning Models for Face Mask Detection in Real-Time Edge Environments: A Review and Future Research Directions. *Machine Learning and Knowledge Extraction*, *8*(4), 102. DOI: **https://doi.org/10.3390/make8040102**

[13] Hosain, M. T., Zaman, A., Abir, M. R., Akter, S., Mursalin, S., & Khan, S. S. (2024). Synchronizing object detection: Applications, advancements and existing challenges. *IEEE access*, *12*, 54129-54167. DOI: **https://doi.org/10.1109/access.2024.3388889**

[14] Saikrishna, P. S. (2026). Affective Edge Computing: Challenges and Opportunities in Decoding Emotional States. *Bridging the Gap between Mind and Machine: Exploring the Future of Human-AI-Neurotechnology Integration*, 41-63. DOI: **https://doi.org/10.1007/978-3-032-06713-5_3**

[15] Elhanashi, A., Dini, P., Saponara, S., & Zheng, Q. (2023). Integration of deep learning into the iot: A survey of techniques and challenges for real-world applications. *Electronics*, *12*(24), 4925. DOI: **https://doi.org/10.3390/electronics12244925**

[16] Li, H., Yue, X., & Meng, L. (2022). Enhanced mechanisms of pooling and channel attention for deep learning feature maps. *PeerJ Computer Science*, *8*, e1161. DOI: **https://doi.org/10.7717/peerj-cs.1161**

[17] Zhu, Y., Han, G., Zhu, H., & Zhang, F. (2025). Feature Description Attention: Channel-

independent local–global fusion for multi-scale feature representation. *Engineering Applications of Artificial Intelligence*, *161*, 112139. DOI: **https://doi.org/10.1016/j.engappai.2025.112139**

[18] Liu, T., Luo, R., Xu, L., Feng, D., Cao, L., Liu, S., & Guo, J. (2022). Spatial channel attention for deep convolutional neural networks. *Mathematics*, *10*(10), 1750. DOI: **https://doi.org/10.3390/math10101750**

[19] Li, X., Lei, L., Sun, Y., Li, M., & Kuang, G. (2020). Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 1011-1026. DOI: **https://doi.org/10.1109/jstars.2020.2975252**

[20] Caicedo, J. E., Agudelo-Martínez, D., Rivas-Trujillo, E., & Meyer, J. (2023). A systematic review of real-time detection and classification of power quality disturbances. *Protection and Control of Modern Power Systems*, *8*(1), 1-37. DOI: **https://doi.org/10.1186/s41601-023-00277-y**

[21] Panigrahi, R., Borah, S., Bhoi, A. K., Ijaz, M. F., Pramanik, M., Jhaveri, R. H., & Chowdhary, C. L. (2021). Performance assessment of supervised classifiers for designing intrusion detection systems: a comprehensive review and recommendations for future research. *Mathematics*, *9*(6), 690. DOI: **https://doi.org/10.3390/math9060690**

[22] Thakkar, A., & Lohiya, R. (2022). A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review*, *55*(1), 453-563. DOI: **https://doi.org/10.1007/s10462-021-10037-9**

[23] Şentaş, A., Tashiev, İ., Küçükayvaz, F., Kul, S., Eken, S., Sayar, A., & Becerikli, Y. (2020). Performance evaluation of support vector machine and convolutional neural network algorithms in real-time vehicle type and color classification. *Evolutionary Intelligence*, *13*(1), 83-91. DOI: **https://doi.org/10.1007/s12065-018-0167-z**

[24] Mateen, M., Wen, J., Hassan, M., Nasrullah, N., Sun, S., & Hayat, S. (2020). Automatic detection of diabetic retinopathy: a review on datasets, methods and evaluation metrics. *IEEE Access*, *8*, 48784-48811. DOI: **https://doi.org/10.1109/access.2020.2980055**

[25] Fan, C., Ghaemi, S., Khazaei, H., & Musilek, P. (2020). Performance evaluation of blockchain systems: A systematic survey. *Ieee Access*, *8*, 126927-126950. DOI: **https://doi.org/10.1109/access.2020.3006078**

[26] Atilgan, C., & Mercimek, M. (2025). Balancing Precision and Speed: Introducing The Performance Efficiency Evaluation Ratio (PEER) in Visual Odometry. *IEEE Access*. DOI: **https://doi.org/10.1109/access.2025.3571921**

[27] Vdoviak, G., Sledevič, T., Serackis, A., Plonis, D., Matuzevičius, D., & Abromavičius, V. (2025). Evaluation of deep learning models for insects detection at the hive entrance for a bee behavior recognition system. *Agriculture*, *15*(10), 1019. DOI: **https://doi.org/10.3390/agriculture15101019**

[28] Zhang, X., Zhang, Y., Li, Z., Song, Y., Chen, S., Mao, Z., ... & Nie, L. (2025). A real-time cell image segmentation method based on multi-scale feature fusion. *Bioengineering*, *12*(8), 843. DOI: **https://doi.org/10.3390/bioengineering12080843**

[29] Raghavan, K., B, S., & v, K. (2024). Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. *Multimedia Tools and Applications*, *83*(19), 57551-57578. DOI: **https://doi.org/10.1007/s11042-023-17776-7**

[30] Zhang, Y., Zhu, Y., Liu, J., Yu, W., & Jiang, C. (2024). An interpretability optimization method for deep learning networks based on Grad-CAM. *IEEE Internet of Things Journal*, *12*(4), 3961-3970. DOI: **https://doi.org/10.1109/jiot.2024.3485765**