scientific reports

Check for updates

OPEN

A WAD-YOLOv8-based method for classroom student behavior detection

Lisu Han[1]✉, Xuejian Ma[1], Mengna Dai[2] & Lu Bai[3]

This paper proposes an enhanced YOLOv8 model to address the challenges of complex classroom behavior detection. The model effectively overcomes the limitations of the original YOLO backbone, including the restricted receptive field and insufficient multi-scale feature learning due to fixed convolutional kernels. We introduce a novel CA-C2f module, enabling more comprehensive fusion and adjustment of the receptive field. Additionally, we propose the attention-based 2DPE-MHA module, which enhances the model's ability to capture long-range dependencies, thereby improving detection performance for multi-scale, occluded, and small targets. The model also incorporates a dynamic sampling factor, Dysample, which selectively focuses on regions with rich details, alleviating the potential loss of detail associated with traditional fixed sampling strategies and further boosting model performance. Experimental results demonstrate that the proposed model outperforms existing methods on public datasets such as SCB, SCB2, SCB-S, and SCB-U, with mAP@0.5 improvements of 2.2%, 3.3%, 5.5%, 18.7%, respectively, and mAP@0.5:0.95 improvements of 3.2%, 2.3%, 3.5%, 14.8%. Moreover, the model maintains real-time inference capabilities that surpass those of other object detection models. The application of this model can assist student work managers in effectively monitoring classroom behavior.

Keywords Classroom behavior detection, YOLOv8, Dysample

The advent of the era of artificial intelligence affects all aspects of society and promotes changes in all walks of life. In the field of education, it is manifested in the improvement of teaching efficiency, personalized experience and management process, while big data classroom, a practical direction gradually formed with the combination of big data technology and education, relies on the joint promotion of multiple forces and comes into being[1]. At present, classroom behavior management mainly relies on teacher supervision, which is not only time-consuming and labor-intensive, but also may lead to delayed feedback and reduce teaching efficiency. In contrast, big data classrooms bring significant advantages through automated monitoring and behavioral analysis. With machine learning, administrators can keep track of students' status in real time and fine-tune strategies to make education more efficient, inclusive and equitable. This not only reduces the management burden of teachers, but also improves students' learning autonomy and classroom participation.

As a branch of machine learning, the neural network simulates the human brain to achieve training and consists of a convolutional unit. In 1988, Yann Lecun et al.[2] first proposed LeNet-5, a network model based on CNN, and applied it in the digital recognition task of handwritten checks. Due to the lack of computer computing power and data at that time, the development of deep learning was stagnated. With the development of productivity, in 2012, Alex Krizhevsky[3] proposed AlexNet model and won the champion of ImageNet image classification competition, which promoted the development of various fields based on deep learning, and the development of object detection thus entered the fast lane. The object detection algorithm based on deep learning includes the two-stage detection algorithm proposed earlier and the single-stage detection algorithm proposed later. By training a large number of labeled data for feature learning and extraction, the prediction effect can be obtained.

The two-stage object detection algorithm generates a certain candidate region at first, and then classifies within the candidate region. In 2014, the classical object detection algorithm R-CNN (Region-based Convolutional Network) was proposed by Ross Girshick et al.[4], which uses convolutional neural networks for feature extraction, and then for classification and positioning. The PACAL 2007 dataset achieved the highest detection accuracy at that time. At this time, the object detection algorithm based on deep learning has entered

[1]School of Anesthesiology, Shandong Second Medical University, Weifang 261053, China. [2]School of Stomatology, Shandong Second Medical University, Weifang 261053, China. [3]Shandong Maritime Vocation College, Weifang 261108, China. ✉email: hanlisu783@163.com

the industrial level of application. Subsequently, on the basis of R-CNN, the improved Fast RCNN[5] algorithm and Faster RCNN[6] algorithm were proposed, but the obvious shortcomings of R-CNN were still not changed. On the one hand, because the selective search algorithm is used to obtain candidate boxes, a large number of redundant and overlapping candidate boxes will be generated, and the computing efficiency of the network will be greatly reduced. On the other hand, the input image size is fixed, and the robustness of the whole model is not strong.

In 2015, He et al.[7]. proposed the SPP-NET algorithm to solve the R-CNN candidate frame redundancy problem. This is a spatial pyramid-shaped pooling structure, which can be added to the convolution layer, and the feature maps of different sizes can be pooled at different scales to fix the same size. At the same time, the pyramid structure also reduces the image distortion in the process of pre-processing cropping and scaling, and improves the detection efficiency. In 2017, He Kaiming further proposed the Mask R-CNN[8] algorithm, which uses bilinear interpolation to fill pixels in non-integer regions to solve the problems of large ROI Pooling quantization error and mismatch between feature map and input image in Faster-RCNN, thus improving model performance.

The single-stage object detection algorithm uses convolution and neural network stacking to output the category information and position information of the object directly. In 2015, Redmon J et al.[9]. proposed the YOLO algorithm, pioneering the single-stage detection algorithm, which enabled target detection and classification to be completed in one step and significantly improved the detection speed. In 2016, Liu W et al.[10]. proposed SSD algorithm and combined the advantages of Faster RCNN and YOLO to predict multiple feature maps of different sizes of targets. After that, an improved SSD algorithm FSSD[11] was proposed one after another. In 2017, Redmon J et al.[12] proposed YOLOv2 on the basis of YOLO algorithm, which used joint training to improve detection efficiency. In the same year, Lin TY et al.[13]. proposed the Feature Pyramid structure (FPN), and combined with the idea of feature pyramid, enhanced small target detection YOLOv3[14] was proposed. Subsequently, Retinanet[15] and PANet[16] also learned from FPN's ideas. In 2023, the performance-balanced YOLOv8[17]model was introduced. In 2024, the latest versions of the YOLO series, YOLOv10[18] and YOLOv11[19], were consecutively released.

In recent years, many scholars have proposed their own models tailored to the characteristics of classroom detection tasks. Zejie W[20] proposed a model integrating OpenPose algorithm and YOLO v3 algorithm, which not only improved the accuracy of recognition, but also reduced the consumption of computing resources through model optimization. Lin, F.-C[21] took continuously captured classroom video frames as input, proposed OpenPose skeleton pose data model, combined skeleton pose estimation with character detection for student behavior analysis in smart classroom, and provided real-time feedback for teachers. Chen, H[22] proposed a new Detection model of RT-DETR (Real-Time Detection Transformer) algorithm, and introduced multi-scale attention module (EMA) to enhance the model's ability to perceive targets at different scales. The detection accuracy and robustness of the model in complex scenes are further improved. In 2024, Lin, L[23] introduced MobileNetV3 as a lightweight backbone network to reduce the optimization model parameters to one-tenth of the original, while maintaining high-precision detection performance. Wang Z developed SLBDetection-Net[24], highlighting its significant advantages in improving detection accuracy and efficiency in both closed-set and open-set scenarios. In the same year, he proposed SBD-Net[25], which enhanced the model's ability to handle biased distributions in student behavior and enabled high-precision detection in dynamic and challenging classroom environments.

Related works

In the context of classroom behavior detection, the focus of object detection is mainly on feature extraction and feature fusion to enhance the ability of feature extraction and classification.

Feature extraction

In classroom behavior detection, the core of object detection lies in the effective extraction and fusion of multi-scale feature information, which can more accurately identify and track targets in complex environments, such as students' movements and gestures. Backbone network architectures such as CSPDarkNet[26] and EfficientNet[27] have been shown to improve detection performance, but complex architectures can increase the computational overhead of models. Therefore, the application of technologies such as extended convolutional networks[28] and depth-separable Convolutional networks (DCN)[29] has improved the detection and feature extraction capabilities of objects in complex scenes while maintaining high resolution.

By introducing the spatio-temporal tracking algorithm, the continuity of the target over time can be effectively used to enhance the detection effect of the obscured and deformed target. In addition, Spatio-temporal memory networks (ASTMN)[30] and Efficient Attention networks (EAN)[31] further improve the accuracy and efficiency of detection and tracking through feature fusion and attention mechanisms. Extended convolution is used to avoid the loss of feature details, while DCN enhances the detection performance of irregular and occluded objects through adaptive receptive field adjustment. By introducing global context information, the model can improve the ability of global feature representation for multi-scale targets. The combination of these technologies not only improves the target detection and tracking capability in complex background, but also improves the overall performance of the system.

Feature fusion

The research of feature fusion technology mainly focuses on enhancing the performance of feature extraction and classification by combining multiple features or mechanisms. The introduction of the two-branch structure CNN-Transformer[32], the foreground-aware masking module FAMA[33], and the dual-domain progressive refinement network DPRNet[34] presents a novel fusion approach that effectively combines global information

with local details. This enhances both the representation power and accuracy, especially in tasks such as small object detection. Additionally, CTAFFNet[35] employs an encoder-decoder architecture to facilitate the flow of global information, while SS-BEV[36] uses a cascading method that integrates parallel convolutions and weighted operations, effectively addressing the detection needs for small objects in complex scenarios.It is worth mentioning that CNN-Transformer adopts a two-branch structure, effectively combining global information with local details, which enhances both representation power and accuracy in tasks such as small object detection and target recognition in complex backgrounds. Similarly, the 2DPE-MHA we propose later also follows a two-branch architecture, enabling the model to learn different information and relationships across multiple subspaces, thereby improving its expressive power and flexibility.The attention mechanism dynamically adjusts feature weights according to context and task requirements, enabling the model to focus on important features adaptively. In recent years, attention mechanisms have gradually developed from the early channel and spatial attention to the combination of mixed attention and self-attention mechanisms, such as CBAM[37] and BAM[38] using channel and spatial attention to capture feature representations of different dimensions. However, these methods have high computational complexity and may rely too much on local features, which limits the generalization ability of the model.

To deal with the tradeoff between computational complexity and generalization ability, models such as efficient Pyramid split Attention (EPSA)[39] and Shuffle Attention (SA-Net)[40]have been proposed. EPSA captures multi-scale features through convolution kernels of different sizes to increase the receptor field and improve the feature representation capability, while SA-Net improves the feature interactivity and representation capability at a lower computational cost by reorganizing feature maps and combining channels and spatial attention mechanisms.The spatial representation attention mechanism[41]captures direction-aware feature maps through decomposed pooling operations, and generates attention weights for weighted fusion, effectively preserving spatial positional information and enhancing feature representation capabilities. The CA module we propose later follows a similar approach, with the key difference being that it adjusts the number of channels based on a carefully designed scaling ratio.

Model selection

Compared to previous models, YOLOv8 offers significant improvements in both speed and accuracy, especially excelling in small object detection and complex scenes. Its enhanced network architecture and optimized training strategies make it more efficient for real-time detection tasks, while also increasing robustness. Although YOLOv11 introduces more advanced features for certain tasks, considering YOLOv8's efficiency, stability, and strong performance in classroom behavior detection, this paper selects YOLOv8 as the benchmark network. The improvements in this study include the following aspects: (1) A CA module is proposed and integrated into the backbone network to enhance feature extraction by acquiring multi-scale information and combining different receptive fields. (2) A 2DPE-MHA attention mechanism is adopted to comprehensively improve the model's performance and robustness. (3) Dynamic upsampling technology is employed to reduce model parameters while more accurately restoring image details, ensuring the consistency of depth values in flat regions and effectively handling gradually changing depth values.

We validated the WAD-YOLOv8 framework on four public datasets, achieving average precision mAP scores of 76.3%, 70.6%, 73.6%, and 95.2%, demonstrating the effectiveness of the model. The structure of this paper is as follows: Sect. 2 provides an overview of the relevant technologies used, Sect. 3 presents a detailed description of the implementation methods, Sect. 4 shows the experimental results and their analysis, and finally, Sect. 5 summarizes the main conclusions of the paper and discusses future research directions.

The proposed WAD-YOLOv8 model

WAD-YOLOv8 is a derivative model of YOLOv8, including various scale versions, including n\s\m\l\x, can combine anchor-less architecture with decoupage head, can more effectively deal with different scale targets, and improve the detection ability of small targets and complex scenes through multi-scale feature fusion. WAD-YOLOv8 not only balances target location and classification accuracy, but also reduces computational complexity. It is especially suitable for real-time scenarios such as student behavior detection. Its network structure and detailed modules are shown in Fig. 1.

CA_C2F module

The traditional C2f module is effective in feature extraction. However, for classroom behavior detection tasks, especially when there is a significant difference between near and far targets, the traditional C2f module may struggle to handle the disparity between distant and close student targets. For example, the size difference between a student close to the camera and one farther away can be as much as 20 times, which can result in insufficient capturing of features from distant targets, particularly in the absence of global context awareness. While stacking multiple C2f modules can enhance the network's expressive power, the excessive channel information leads to higher computational costs, and the standard convolutional kernels can only focus on local information, making it difficult to effectively capture global contextual relationships. In contrast, the CA-C2F module dynamically weights features from different channels through an expansion-unequal segmentation-concatenation approach, emphasizing important features while suppressing redundant information. This mechanism enhances the model's ability to perceive global and long-range dependencies, enabling it to better capture complex scenarios involving distant targets in classroom behavior. The basic structure of the CA module is shown in Fig. 2.

CA module uses convolution of different sizes of cavities to obtain different receptive fields, which can better adapt to the shape and size of objects and enhance the robustness of the model. Assuming that the size of the access feature graph is H, W, C, first, the input tensor $X \in \mathbb{R}^{C\times H\times W}$ is convolved by 3×3, and the number of channels is extended to 2 C to obtain the convolutional output $X_{conv} \in \mathbb{R}^{2C\times H\times W}$. Then, the channel

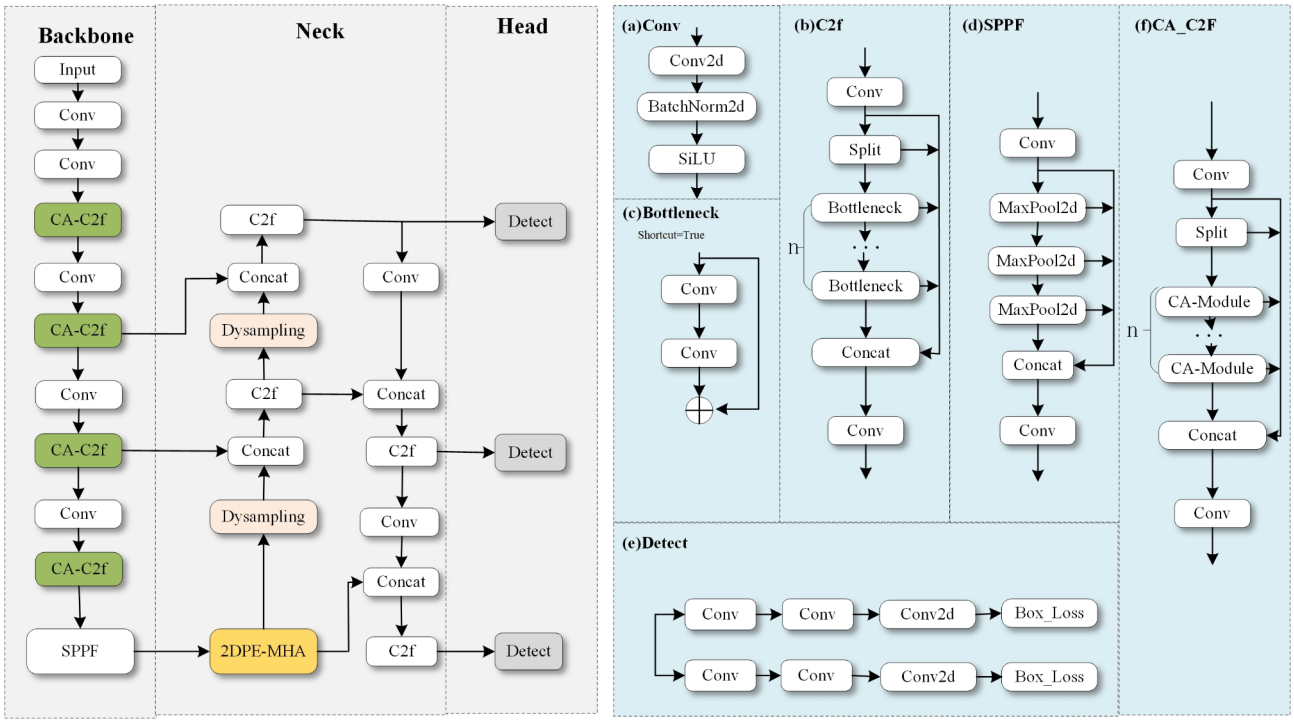

Fig. 1. Basic structure of WAD-YOLOv8.

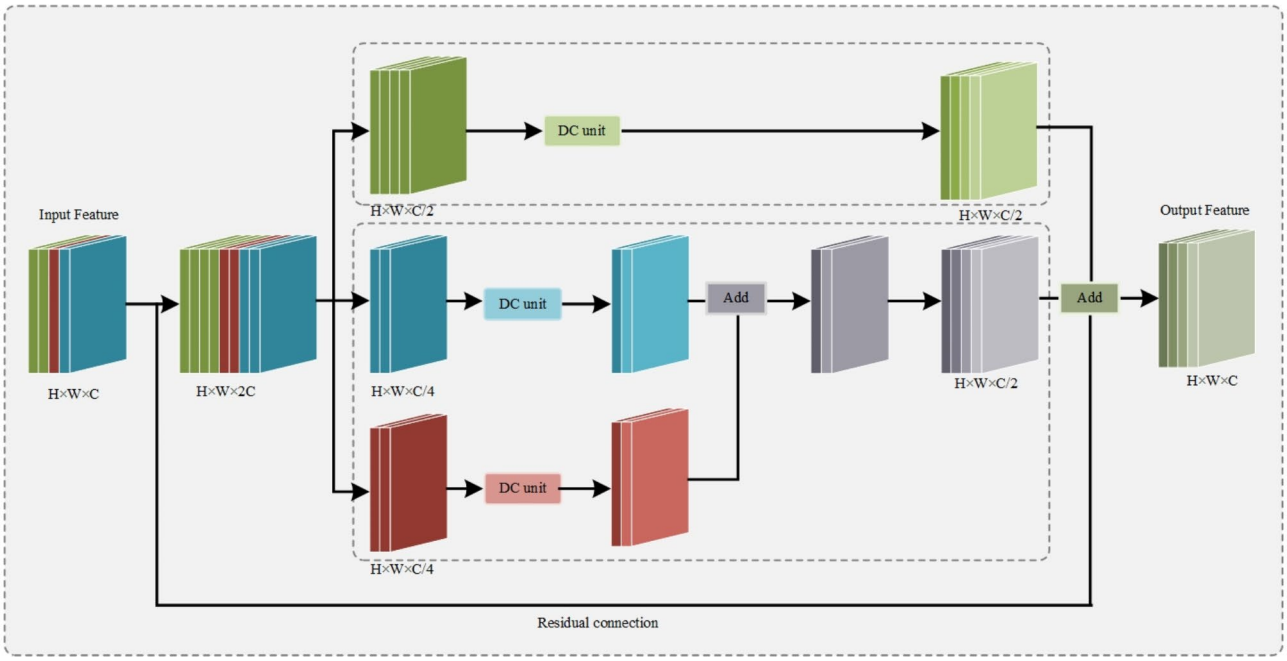

Fig. 2. The basic structure of the CA module.

is divided into three groups: $g1 \in \mathbb{R}^{C\times H\times W}$, $g2 \in \mathbb{R}^{\frac{C}{2}\times H\times W}$ and $g3 \in \mathbb{R}^{\frac{C}{2}\times H\times W}$. Convolution of different void rates is applied to each group, in which group g1 is convolution of 3×3 for $r=1$, group g2 is convolution of 3×3 for $r=2$, group g3 is convolution of 3×3 for $r=4$, and the final output is obtained by channel transformation and weighted fusion to obtain $Y_{final} \in \mathbb{R}^{C\times H\times W}$. First, the global average pooling $z \in \mathbb{R}^{C}$ is calculated:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_{final,c}(i,j) \tag{1}$$

Generate channel weights through the full connection layer:

$$s = \sigma\ (W_2 \delta\ (W_1 z)) \tag{2}$$

Use weights to weight each channel:

$$Y_{attention} = s \cdot Y_{final} \tag{3}$$

The weighted output and input tensors are added, and the residuals are connected.

$$Y_{output} = Y_{attention} + X \tag{4}$$

The output shape is $Y_{final} \in \mathbb{R}^{C \times H \times W}$, which does not change. CA_C2f replaces the Bottleneck module in YOLOv8's C2f module with the CA module.

2DPE-MHA

As is shown in Fig. 3, the 2DPE-MHA attention mechanism we propose is inspired by the multi-head attention mechanism in Transformer[42]. It employs multiple parallel attention heads for self-attention processing, thereby enhancing the model's ability to capture complex dependencies and effectively modeling the relationships between different positions in the input sequence.

However, the traditional multi-head attention mechanism does not inherently include position encoding; it computes attention solely through the mapping relationships between queries, keys, and values. While position encoding can provide some positional information, in the case of two-dimensional spaces, it is typically modeled using a single dimension, which fails to fully capture complex spatial structures. Therefore, the 2DPE-MHA mechanism innovatively introduces two-dimensional position encoding, effectively addressing the shortcomings that may arise when dealing with occlusion and spatial dependencies.

The implementation process of 2DPE-MHA consists of a main branch and a secondary branch. The specific implementation details are as follows:

First, a flattening operation is performed to facilitate the computation of the combination of multi-head attention and position encoding.

$$x = \text{ rearrange } (x, nchw \rightarrow n(hw)c) \tag{5}$$

Next, the input feature map is linearly mapped to obtain the query (Q), key (K), and value (V) vectors.

$$Q = X_r W_q,\ K = X_r W_k,\ V = X_r W_v \tag{6}$$

The main branch implements multi-head attention with num_heads = 8. Each attention head independently applies the self-attention mechanism to compute the input query (Q), key (K), and value (V) vectors. Each head assigns weights to the values by calculating the similarity between the queries and keys, focusing on the most

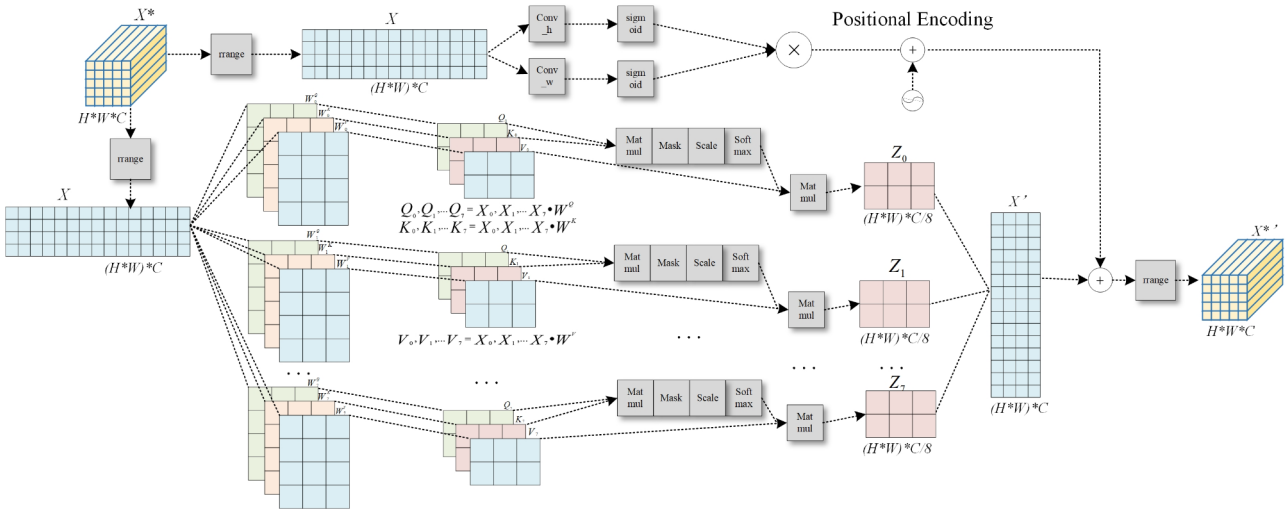

Fig. 3. Basic structure of 2DPE-MHA.

relevant information in the sequence. Different heads can capture various types of dependencies in parallel, and their outputs are then concatenated, enhancing the model's expressive power.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

In the secondary branch, the attention map is determined as follows: a 1×1 convolution is applied to the input feature map to generate attention maps in both the horizontal and vertical directions. These maps are then normalized using the Sigmoid activation function, ensuring that their values are constrained within the range of [0, 1].

$$\begin{array}{ll} \text{attn}_h & = \sigma\ (\text{Conv}_h\ (X)) \\ \text{attn}_w & = \sigma\ (\text{Conv}_w\ (X)) \end{array} \tag{8}$$

Then, the attention maps in the horizontal and vertical directions are element-wise multiplied, and the result is added to a fixed sinusoidal position encoding (PE). In this case, D represents the dimension of the position encoding, which is typically equal to the number of input channels C. The position encoding is generated using sine and cosine functions to encode spatial positions, with different frequencies employed to ensure that each position has a unique encoding.

$$PE_{k,i,j} = \begin{cases} \sin\left(\frac{i}{10000^{\frac{k}{D}}}\right) & \text{for even } k \\ \cos\left(\frac{j}{10000^{\frac{k}{D}}}\right) & \text{for odd } k \end{cases} \tag{9}$$

Here, i represents the row coordinate of the image, j represents the column coordinate, and k is the index of the position encoding dimension. The position encoding is generated using sine and cosine functions to encode spatial positions, with different frequencies used to ensure that each position has a unique encoding. Finally, the main branch and secondary branch are element-wise added, and the two branches are merged.

$$\text{attn}_{\text{final}} = \text{attn} + PE \tag{10}$$

One of the core advantages of the 2DPE-MHA module is the introduction of 2D positional encoding. By embedding positional information in both the horizontal and vertical directions, it overcomes the limitations of traditional multi-head self-attention mechanisms in expressing spatial relationships. This significantly enhances the model's ability to capture long-range dependencies. The module is particularly well-suited for handling occlusion scenarios and detecting distant targets, enabling accurate recognition of occluded and multi-target behaviors in complex scenes.

Compared to standard multi-head attention mechanisms, 2DPE-MHA improves the model's understanding of the spatial layout between targets by leveraging positional encoding, which further optimizes attention distribution. This results in more precise performance, especially in occlusion and small-target scenarios. Additionally, compared to traditional attention mechanisms such as SE and CBAM, 2DPE-MHA not only dynamically focuses on both global and local information but also explicitly models the spatial relationships between targets. This allows it to exhibit stronger robustness and adaptability in complex task scenarios, significantly improving the overall performance of the model.

Dynamic sampling

In the classroom behavior detection project, while traditional upsampling methods (such as nearest-neighbor or bilinear interpolation) can restore image resolution to some extent, they often fail to preserve important details when dealing with complex scenes. These methods may result in generated images or feature maps that lack richness and variety, which is particularly problematic in detecting intricate classroom behaviors. Such deficiencies can reduce the model's sensitivity to behavior details and negatively affect recognition accuracy.

In contrast, the Dynamic Sampling module[43] introduces a more flexible and adaptive sampling strategy. By dynamically adjusting the sampling approach in real-time based on the current generation state, this module can more precisely capture key details and important features. This enables the model to gather more relevant information when processing complex scenes, while also avoiding issues of redundancy and detail loss.

The advantage of the Dynamic Sampling module lies not only in its enhanced flexibility but also in its ability to adaptively adjust the sampling regions and weights based on the current state during the generation process. This improves the quality and expressive power of the images or feature maps used in classroom behavior detection. Unlike traditional fixed sampling strategies, the Dynamic Sampling approach is more efficient, capturing richer contextual information and aiding in the more accurate identification of complex classroom behavior patterns. Therefore, incorporating the Dynamic Sampling module significantly enhances the model's performance in classroom behavior detection, especially in scenarios with varying and subtle behavioral details, where it can better adapt to and extract key features.

The implementation mainly includes the following parts, the first is the formation of offset, through a convolution layer, from the feature map F to generate offset O:

$$\text{O} = \text{Conv}\,(\text{F}) \tag{11}$$

These offsets indicate how the sampling position should be adjusted in the feature plot, and the next step is to calculate the new adopt position (i', j'):

$$i' = i + O[0, i, j],\ j' - j + O[1, i, j] \tag{12}$$

Each original sampling position (i, j) is adjusted according to its corresponding offset, and this dynamic adjustment allows the model to focus more flexibly on different parts of the image during generation. Calculate the weight W for each sampling position, typically using the Softmax function:

$$W = \mathrm{Softmax}\,(F) \tag{13}$$

Softmax normalizes the feature map with an all-inclusive weight sum of 1, making certain areas more important in the generation process, and this weighting mechanism helps the model emphasize key areas.

The resulting image G is obtained by weighted average:

$$G = \sum_{(r', j')} W\left[i', j'\right] \cdot F\left[i', j'\right] \tag{14}$$

By combining the calculated sampling position and weight, a new image is generated, which effectively integrates the information of different areas of the feature map to produce a more delicate and rich image.

Dynamic sampling has the following advantages: It can better capture the details in the image because it dynamically adjusts the sampling point based on the currently generated features; With adaptive weight allocation and sampling strategy, the model can generate more diversified images, instead of being limited by fixed sampling strategy. The model can flexibly adjust its generation strategy according to the change of input characteristics, so as to adapt to different generation tasks.

Experiment results and discussion

Dataset

Dataset benchmarks used for training and validation of the model were selected from public datasets SCB, SCB2, SCB-S and SCB-U[44]. These four datasets are designed to study and analyze student behavior in the classroom.

The SCB dataset focuses on the annotation of various classroom behaviors, including 11,248 annotations, the annotation object is the behavior of raising hands, and a total of 4,001 images are included, covering the interaction and performance of students in different scenarios, aiming to provide a reliable basis for the behavior detection model.

On this basis, the SCB2 dataset was expanded to include 18,499 comments on the behavior of raising hands, reading and writing, including a total of 4,266 images, adding more behavior categories and scenes, improving the diversity and applicability of the dataset, so as to better support the study of behavior recognition in complex scenes.

The SCB-S dataset focuses on specific classroom behaviors, including 25,810 annotations related to hand raising, reading, and writing, across a total of 5,015 images. This targeted data annotation enhances the model's accuracy in identifying key behaviors. SCB-U, on the other hand, is a more complex dataset, closely mirroring real-world classroom scenarios captured by cameras. It includes 19,768 annotations on behaviors such as hand raising, reading, writing, using a phone, bowing the head, and leaning over the table, across 671 images. All four datasets are randomly divided into training, validation, and test sets with an 8:1:1 ratio, providing balanced and rich data support for researchers in model training and evaluation, thereby advancing the development of intelligent monitoring and analysis technologies in the field of education.

Evaluation metrics

Precision

The Precision (P), also known as the precision rate, refers to the proportion of the model that is truly positive among all predicted positive samples in all predicted positive samples. Generally speaking, the higher the precision rate, the better the classifier.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{15}$$

The variable TP represents the number of true positive samples predicted to be positive, and FP represents the number of false positive samples predicted to be positive.

Recall

Recall (R), also known as recall rate, is the proportion of all predicted positive samples that are actually positive in the total sample.

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{16}$$

The variable FN represents the number of truly positive samples that are predicted to be negative.

Hardware and software name	Version
Operating system	Windows 10
GPU	NVIDIA GeForce RTX 3060
CPU	Intel Xeon E5-2680 v3
Python	3.8.18
Pytorch	1.8.2
CUDA	12.4

Table 1. Experimental software and hardware configuration.

Parameter name	Parameter setting
Input image size (H × W)	640 × 640
Batch-Size	4
Epoch	200
Initial learning rate	0.01
Momentum	0.937
Weight_decay	0.0005

Table 2. Hyperparameter settings.

mAP

AP represents the accuracy value of a category prediction, while mAP (mean Average Precision, mAP) is obtained by averaging the average accuracy of all categories to reflect the accuracy of the entire model. The larger the mAP, the higher the detection accuracy of the model. Otherwise, the lower it is. mAP@0.5 represents the average precision mean at the threshold of 0.5. mAP@0.5:0.95, indicating that the threshold range of the IoU parameter is [0.5:0.05:0.95].

$$\mathrm{mAP}=\frac{\sum_{i=1}^{m}(\mathrm{AP})_i}{m} \tag{17}$$

FPS

Frame rate per second (FPS), that is, the number of images that can be processed per second (in frames per second), can reflect the speed of model detection. The higher the number of images processed per second, the faster the detection speed, the shorter the time required to process an image, the higher the FPS, the faster the detection speed.

Experimental environment and hyperparameter setting

This paper uses deep learning algorithm to detect and identify classroom behavior. During the training and testing of the target detection network model, hardware configuration plays a decisive role in learning speed and capability due to the large scale of the open dataset. In order to reduce training time and improve training speed and capability, GPU hardware is chosen in this paper for experiments. The experimental configuration is shown in Table 1, and some training hyperparameters are shown in Table 2.

The learning rate attenuation method was adopted in the training process, and the initial learning rate controlled the updating speed of the model. In order to ensure the stability and convergence of the model, the entire training cycle was 200 epochs.

In order to quantitatively evaluate the performance of the proposed overall framework, we tested the introduced object detection framework on the SCB-S dataset. Figure 4 shows the different performance metrics of the improved YOLOv8 model on the training set and the validation set.

To evaluate the performance of the WAD YOLOv8 model on the SCB-S dataset, a normalized confusion matrix for object recognition was generated, as shown in Fig. 5. The rows and columns of the confusion matrix represent the actual categories and the predicted categories respectively, and the diagonal values represent the percentage of correct predictions for each category. The PR curve is shown in Fig. 6, where x axis represents training time and y axis represents accuracy and recall rate. Through these curves, it can be observed that the evaluation of the detection performance of the target when the confidence threshold changes: the closer the curve value is to 1, the higher the confidence of the model. As can be seen from Fig. 6, the improved YOLOv8 model performs well in all indicators, which proves its effectiveness.

Ablation study

Through a series of ablation experiments, we evaluated the importance of the CA-C2f, 2DPE-MHA, and Dysample modules within the model to understand their impact on overall performance. Additionally, to demonstrate the versatility of our proposed model, we conducted ablation experiments across four datasets, with the results presented in Table 3. We also compared our model with state-of-the-art object detection frameworks,

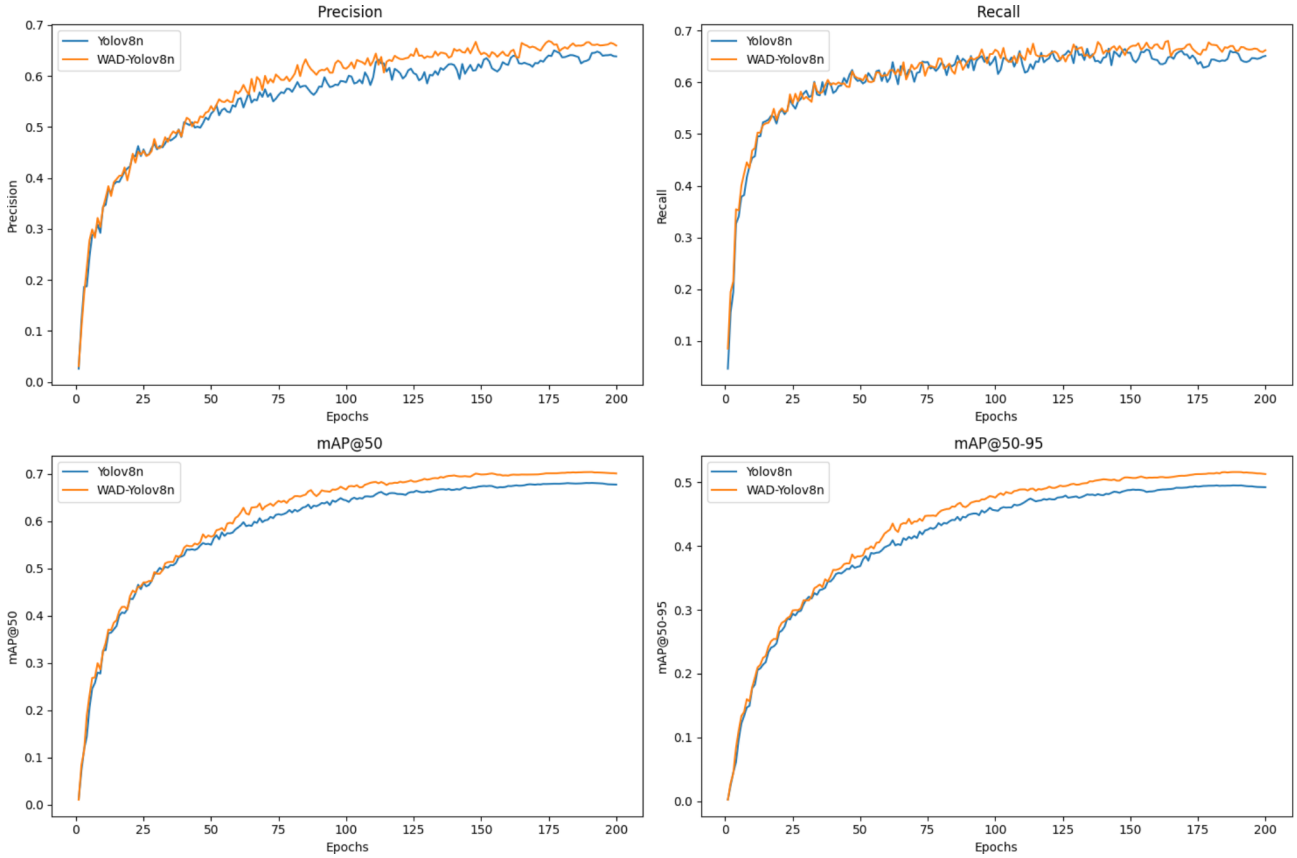

Fig. 4. Improved YOLOv8 performance values.

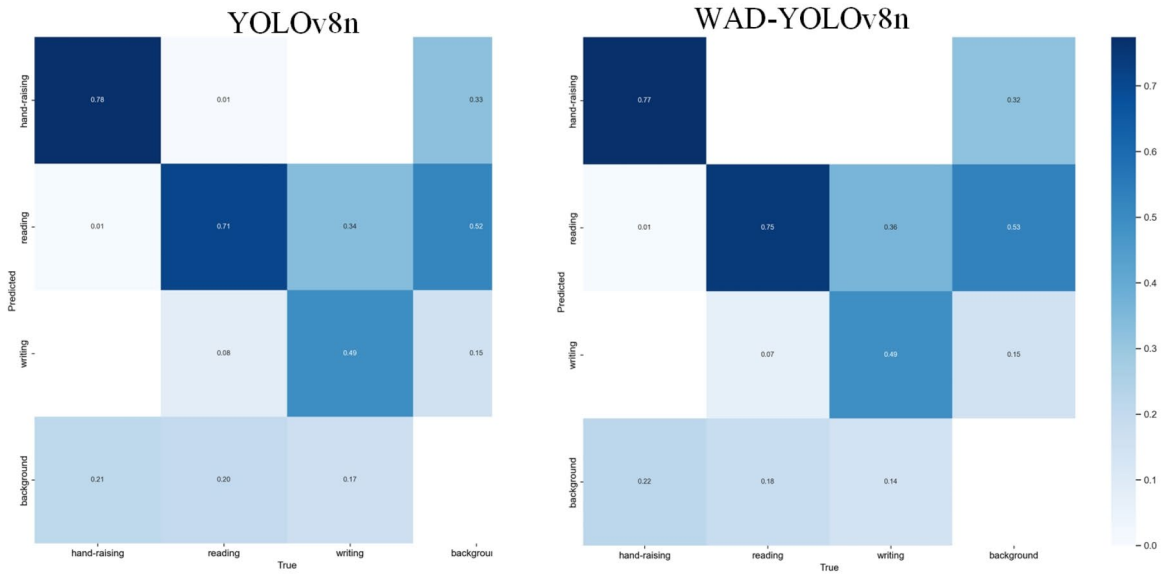

Fig. 5. Normalized confusion matrix.

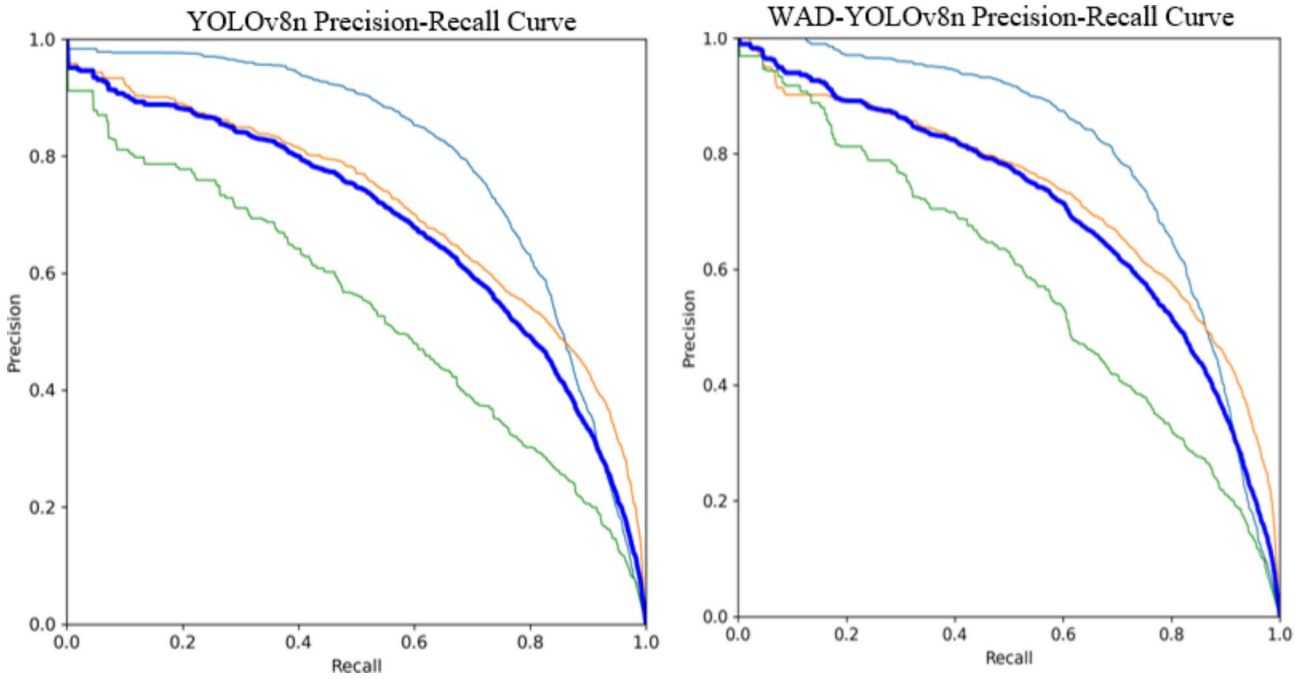

Fig. 6. PR curve.

Model				Parameters		*P* (%)				*R* (%)				mAP@0.5				mAP@0.5:0.95			
B	W	A	D	*N*	G	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4	D1	D2	D3	D4
√				3.0 M	8.1	0.725	0.627	0.646	0.885	0.659	0.663	0.643	0.723	0.741	0.673	0.681	0.765	0.494	0.471	0.495	0.589
√	√			3.3 M	18.6	0.743	0.640	0.664	0.897	0.666	0.672	0.646	0.713	0.749	0.686	0.716	0.926	0.500	0.478	0.516	0.700
√		√		3.6 M	9.9	0.728	0.606	0.63	0.895	0.677	0.671	0.678	0.731	0.745	0.685	0.692	0.932	0.494	0.479	0.505	0.718
√			√	3.0 M	8.1	0.742	0.658	0.648	0.892	0.662	0.626	0.659	0.731	0.743	0.679	0.689	0.930	0.497	0.475	0.500	0.727
√	√	√		3.9 M	20.4	0.758	0.635	0.673	0.887	0.678	0.674	0.684	0.865	0.756	0.696	0.704	0.939	0.514	0.486	0.528	0.733
√	√	√	√	3.9 M	20.4	0.759	0.662	0.694	0.900	0.69	0.682	0.70	0.910	0.763	0.706	0.736	0.952	0.526	0.494	0.530	0.737

Table 3. Ablation test results. 1. B: Base (Yolov8n), W: 2DPE-MHA, A: CA-module, D: Dysample. 2. N: Numble of Parameters, G: Giga Floating-Point Operations per Second (GFLOPS). 3. D1: SCB, D2: SCB2, D3: SCB-S, D4: SCB-U.

and the results show that our framework outperforms other widely used object detection systems in terms of accuracy. These experimental findings provide strong evidence for the effectiveness and adaptability of our framework across various scenarios.

As shown in the Table 3, when the 2DPE-MHA (W) module is added to the network individually, mAP@0.5 increases by 0.8%, 1.3%, 3.5%, and 16.1% on the test sets of each dataset, while mAP@0.5:0.95 increases by 0.6%, 0.7%, 2.1%, and 11.1%, respectively.

When the CA (A) module is added separately, mAP@0.5 increases by 0.4%, 1.2%, 1.1%, and 16.7%, and mAP@0.5:0.95 increases by 0%, 0.8%, 1.0%, and 12.9% on the test sets of each dataset.

Similarly, when the Dysample (D) module is added individually, mAP@0.5 increases by 0.2%, 0.6%, 0.8%, and 16.5%, and mAP@0.5:0.95 increases by 0.3%, 0.4%, 0.5%, and 13.8%, respectively.

Furthermore, when the WAD-YOLOv8 (W + A + D) combination module is added, mAP@0.5 increases by2.2%, 3.3%, 5.5%, and 18.7%, while mAP@0.5:0.95 increases by 3.2%, 2.3%, 3.5%, and 14.8%, respectively. It is clear that the addition of the 2DPE-MHA and CA modules significantly improves mAP@0.5 and mAP@0.5:0.95. Particularly in the case of the W + A + D combination, the individual effect of the Dysample module is not as pronounced, but it achieves the best results when combined with other modules.

On the D4 dataset, which is the most complex and closely resembles real-world scenarios, the performance in terms of mAP@0.5:0.95 generally surpasses that of D1 and D2. This indicates that the improved model has better generalization capability and effectively enhances the network's performance. The test set visualization results of YOLOv8 and the improved WAD-YOLOv8 on D3 and D4 datasets, as shown in Figs. 7 and 8, demonstrate a significant improvement in detecting multiple, occluded, and small targets.

It can be seen that in the first row, WAD-YOLOv8 addresses the missed detection issue of YOLOv8, successfully detecting the learning states of all four students, whereas YOLOv8 only detects two.

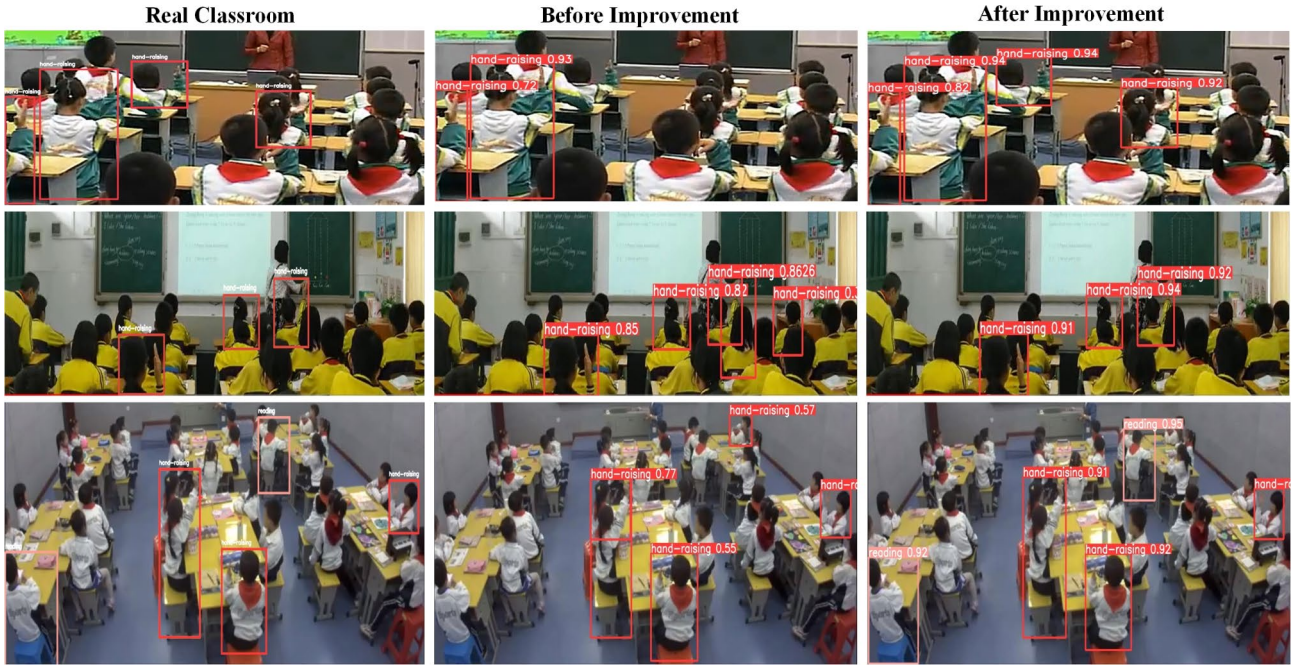

Fig. 7. Real-time monitoring in simple scenarios.

In the second row, WAD-YOLOv8 overcomes YOLOv8's false detection problem, resolving the issue of misdetection for two students.

In the third row, a clear improvement in accuracy is evident, with WAD-YOLOv8 significantly outperforming YOLOv8, achieving higher precision across all detections.

In simulated real-world and slightly more complex scenarios, it can be observed that the YOLOv8 model exhibits false positives and missed detections, and small objects at long distances are not detected. However, WAD-YOLOv8 effectively addresses these issues and achieves higher accuracy.

In order to compare their performance in feature extraction in classroom behavior detection, the two were added to the YOLOv8 and WAD-YOLOv8 networks respectively for thermal map visualization comparison, as shown in Fig. 9.

It can be seen that the highlighted part of the WAD-YOLOv8 network in the right image is more fully displayed. The network not only pays attention to the information of channel dimension, but also pays attention to the information of spatial dimension, which solves the problem of insufficient network feature extraction, and highlights some missing small target defects (marked with red boxes). The problem of misdetection and missing detection of small target is solved effectively.

Performance comparison with state-of-the-art methods

In order to quantify the performance of the WAD-YOLOv8 model in different scenarios, we conducted comparative training and testing with the mainstream target detection model on the open data set SCB-S, as shown in Table 4 below.

The experimental results on the SCB- dataset demonstrate that the WAD-YOLOv8 series outperforms YOLOv8 models of various sizes, including YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Specifically, the WAD-YOLOv8 models show improvements of 2.4%, 1.8%, 1.1%, and 2.0% in mAP@0.5, with an average increase of 1.825%. Similarly, mAP@0.5:0.95 increases by 0.9%, 1.4%, 0.8%, and 0.6%, with an average improvement of 0.925%. Additionally, compared to traditional detection models such as SSD, Fast RCNN, and YOLOv3, WAD-YOLOv8 achieves significant gains in detection accuracy. In terms of overall performance, WAD-YOLOv8 outperforms the latest YOLOv11 network, and also demonstrates notable advantages over recent models like VWEYOLOv8 and CSB-YOLO (pruning + distillation).

Conclusion and future work

We propose the WAD-YOLOv8 model, which focuses on addressing various challenges in complex classroom behavior detection. Compared to existing techniques, our model demonstrates significant advantages in multi-scale feature learning, occluded target detection, and attention to detailed regions. Traditional YOLO models (e.g., YOLOv5, YOLOv7) are limited by their backbone network structures, which constrain their performance in complex scenarios such as overlapping targets, occlusion, and small-object detection. Our model overcomes these limitations through several innovative designs: the integration of the CA-C2f module, which combines channel attention mechanisms with a C2f structure, effectively addresses the receptive field limitations caused by fixed convolution kernels in the original YOLO backbone. This module significantly enhances the model's ability to learn multi-scale features, delivering superior performance in complex classroom scenarios with

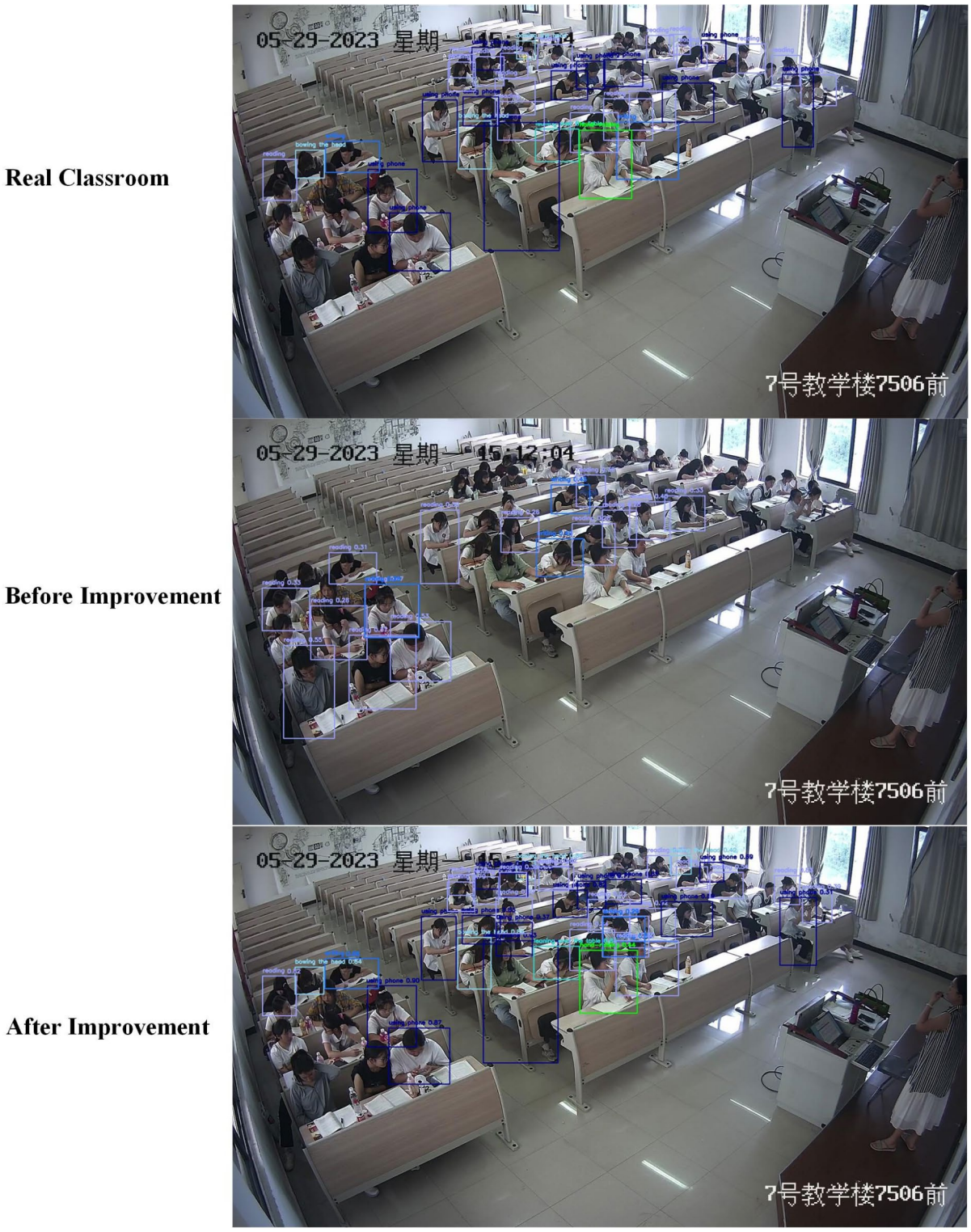

Fig. 8. Real-time monitoring in complex scenes.

targets of varying sizes. The introduction of the 2DPE-MHA module further improves the model's capacity to capture long-range dependencies, making it particularly effective for detecting targets in occluded and distant scenarios while optimizing recognition of occluded and overlapping behaviors in complex scenes. Additionally, our proposed dynamic sampling factor (Dysample) enables the model to adaptively adjust its focus based on the content of the input image. Compared to traditional fixed sampling strategies, Dysample is more flexible, significantly reducing detail loss in regions with rich visual information and substantially improving target recognition accuracy. These innovative designs allow the WAD-YOLOv8 model to demonstrate exceptional

Fig. 9. Visual comparison of heat maps.

Model	mAP@0.5/%	mAP@0.5:0.95/%	Inference time (ms)
SSD[45]	0.678	0.389	28
Fast-RCNN[45]	0.548	0.413	49
Yolov3[45]	0.674	0.489	7.4
Yolov5n[45]	0.671	0.483	8.4
Yolov8n[45]	0.681	0.495	11
Yolov8s[45]	0.728	0.547	7.6
Yolov8m[45]	0.750	0.579	9.5
Yolov8l[45]	0.753	0.591	19.6
Yolov8 × [45]	0.759	0.599	31.9
WAD-Yolov8n	0.736	0.530	13
WAD-Yolov8s	0.752	0.556	15.7
WAD-Yolov8m	0.768	0.593	23.2
WAD-Yolov8l	0.764	0.599	30.4
WAD-Yolov8x	0.779	0.605	48.7
Yolov10[18]	0.684	0.493	30.5
Yolov11[19]	0.687	0.497	29.3
VWEYOLOv8[50]	0.715	0.528	–
CSB-YOLO (prune + distill)[51]	0.711	0.523	–

Table 4. Results of different detection methods on the SCB-S dataset.

performance in complex classroom behavior detection tasks, effectively addressing the limitations of existing methods in critical scenarios.It is worth mentioning that our research can also be extended to the field of mirror detection. Due to factors such as reflection angles and lighting variations, noise and artifacts in mirrored images are often more pronounced than in normal images. This is especially true on reflective surfaces, where noise can interfere with the true edges and structure of objects, affecting detection results. Our proposed fusion strategy enables the module to effectively combine global information with local details, providing enhanced representation power and accuracy for mirror detection tasks in complex backgrounds. The comprehensive application of these innovative measures makes WAD-YOLOv8 perform well on SCB, SCB2, SCB-S and SCB-U, with the improvement of mAP@0.5 and mAP@0.5:0.95 reaching 2.2%, 3.3%, 5.5%, 18.7% and 3.2%, 2.3%, 3.5%, 14.8%, respectively, while maintaining the real-time inference capability. To sum up, the WAD-YOLOv8 model not only exceeds other detection models in performance, but also provides an effective solution for behavior detection in more complex scenarios.

Although the model achieves a good balance between accuracy and speed, it still has certain limitations. On one hand, the improved network structure increases the computational complexity of the model, which may affect real-time performance when deployed on devices with limited computational resources. On the other hand, while the model performs well in specific scenarios, its recognition accuracy may degrade in more complex or extreme environments, such as those with significant lighting changes, heavy occlusion, or non-standard classroom settings.

In the future, we plan to create a specialized dataset that covers a variety of complex scenarios and behaviors. This dataset will include classroom behavior data under varying lighting conditions, angles, and backgrounds to enhance the model's adaptability and generalization in real-world applications. We expect this dataset to improve the model's performance, broaden its range of applications, and provide a more effective solution for complex behavior detection tasks.

Data availability

The datasets analysed during the current study are not publicly available due this research will be submitted for university scientific research achievements in the future but are available from the corresponding author on reasonable request.

Received: 15 November 2024; Accepted: 21 January 2025
Published online: 20 March 2025

References

1. Singh, H. & Miah, S. J. Smart education literature: a theoretical analysis. *Educ. Inform. Technol.* **25** (4), 3299–3328 (2020).
2. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE.* **86** (11), 2278–2324 (1998).
3. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM.* **60** (6), 84–90 (2012).
4. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* 580–587 (2014).
5. Sun, X., Wu, P. & Hoi, S. C. Face detection using deep learning: an improved faster RCNN approach. *Neurocomputing* **299**, 42–50 (2018).
6. Ren, S., He, K., Girshick, R., Sun, J. & Faster, R-C-N-N. Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** (6), 1137–1149 (2016).
7. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** (9), 1904–1916 (2015).
8. He, K. et al. Mask R-CNN. in *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)* 2961–2969 (2017).
9. Redmon, J. et al. You only look once: Unified, real-time object detection. in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 779–788 (2016).
10. Liu, W. et al. SSD: Single Shot MultiBox Detector. in *European Conference on Computer Vision (ECCV)* 21–37 (2016).
11. Qian, H., Wang, H., Feng, S. & Yan, S. FESSD: SSD target detection based on feature fusion and feature enhancement. *J. Real-Time Image Process.* **20** (1), 2 (2023).
12. Redmon, J. & Farhadi, A. YOLO9000: better, faster, stronger. in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7263–7271 (2017).
13. Lin, T. Y. et al. Feature pyramid networks for object detection. in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2117–2125 (2017).
14. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804. 02767* (2018).
15. Lin, T. Y., Priya, G., Ross, G. & Kaiming, H. Dollar Piotr. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42** (2), 318–327 (2020).
16. Liu, S. et al. Path aggregation network for instance segmentation. in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8759–8768 (2018).
17. Ultralytics & YOLOv8. : New - YOLOv8 in PyTorch > ONNX > OpenVINO > CoreML > TFLite. GitHub. (2023). https://github.com/ultralytics/ultralytics
18. Wang, A. et al. Yolov10: real-time end-to-end object detection[J]. (2024). arXiv preprint arXiv:2405.14458.
19. Khanam, R. & Hussain, M. YOLOv11: an overview of the key architectural enhancements[J]. (2024). arXiv preprint arXiv:2410.17725.
20. Zejie, W. A. N. G., Chaomin, S. H. E. N., Chun, Z. H. A. O., **nmei, L. I. U. & Jie, C. H. E. N. Recognition of classroom learning behaviors based on the fusion of human pose estimation and object detection. *Journal of East China Normal University (Natural Science)*, 55 (2022). (2022)(2).
21. Lin, F. C., Ngo, H. H., Dow, C. R., Lam, K. H. & Le, H. L. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors* **21** (16), 5314 (2021).
22. Chen, H., Zhou, G. & Jiang, H. Student behavior detection in the classroom based on improved YOLOv8. *Sensors* **23** (20), 8385 (2023).
23. Lin, L., Yang, H., Xu, Q., Xue, Y. & Li, D. Research on student classroom behavior detection based on the real-time detection transformer algorithm. *Appl. Sci.* **14** (14), 6153 (2024).
24. Wang, Z. et al. Slbdetection-net: towards closed-set and open-set student learning behavior detection in smart classroom of k-12 education[J]. *Expert Syst. Appl.* **260**, 125392 (2025).
25. Wang, Z. et al. SBD-Net: incorporating Multi-level features for an efficient detection network of Student Behavior in Smart Classrooms[J]. *Appl. Sci.* **14** (18), 8357 (2024).
26. Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y. & Hsieh, J. W. and I.-H. Yeh. CSPNet: A new backbone that can enhance learning capability of CNN. *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 390–391 (2020).
27. Tan, M. & Le, Q. V. EfficientNetV2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*. (2021).
28. Yu, F. & Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122*. (2016).
29. Dai, J. et al. and Y. Wei. Deformable Convolutional Networks. *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)* 764–773 (2017).
30. Li, J., Li, J., Zhu, L., Xiang, X. & Huang, T. Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Trans. Image Process.* **31**, 1786–1796 (2022).
31. Yang, H., Wang, J., Xu, W. & Chen, X. Efficient Attention Network: Accelerate Attention by Searching Where to Plug. *arXiv preprint arXiv:14058.* (2020). (2011).
32. Zha, M. et al. Multifeature transformation and fusion-based ship detection with small targets and complex backgrounds[J]. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
33. Zha, M. et al. Weakly-Supervised Mirror Detection via Scribble Annotations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 38(7): 6953–6961. (2024).
34. Zha, M. et al. *Dual Domain Perception and Progressive Refinement for Mirror Detection[J]* (IEEE Transactions on Circuits and Systems for Video Technology, 2024).
35. Dong, X. et al. CTAFFNet: CNN–transformer adaptive feature fusion object detection algorithm for complex traffic scenarios[J]. *Transp. Res. Rec.*, : 03611981241258753. (2024).
36. Shi, P., Pan, Y. & Yang, A. SS-BEV: multi-camera BEV object detection based on multi-scale spatial structure understanding[J]. *Signal. Image Video Process.* **19** (1), 1–13 (2025).
37. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. *in Proceedings of the European Conference on Computer Vision (ECCV)* 3–19 (2018).
38. Wang, F., Hu, H., Shen, C., Feng, T. & Guo, Y. BAM: a balanced attention mechanism to optimize single image super-resolution. *J. Real-Time Image Proc.* **19** (5), 941–955 (2022).

39. Xie, H., Zheng, L. & Li, C. Efficient Pyramid Split Attention Network for Light-weight Visual Recognition. *in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2689–2698 (2021).
40. Huang, X., Zhao, L. & Huang, K. Shuffle Attention: A Lightweight Attention Module for Convolutional Neural Networks. *in Proceedings of the 28th ACM International Conference on Multimedia (MM)* 1867–1875 (2020).
41. Zha, M. et al. A lightweight YOLOv4-Based forestry pest detection method using coordinate attention and feature fusion[J]. *Entropy* **23** (12), 1587 (2021).
42. Vaswani, A. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* (2017).
43. Chen, W. et al. Dynamic Downsampling and Upsampling for Dense Predictions. in *Proceedings of IEEE International Conference on Computer Vision (ICCV)* 12345–12355 (2023).
44. Yang, F. & Wang, T. Scb-dataset3: A benchmark for detecting student classroom behavior. *arxiv preprint arxiv:2310.02522.* (2023).
45. Liu, W., Anguelov, D., Erhan, D., Szegedy, C. & Reed, S. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV), 21–37. (2016).
46. Girshick, R. B. & Fast, R-C-N-N. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448. (2015).
47. Redmon, J., Divvala, S., Girshick, R. B. & Farhadi, A. YOLOv3: An incremental improvement, arXiv preprint arXiv:1804.02767, (2018).
48. Jocher, G. YOLOv5, Ultralytics, 2020. [Online]. Available: https://github.com/ultralytics/yolov5
49. Ultralytics YOLOv8, Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/yolov8
50. Zeng, Y. et al. An improved YOLOv8-based algorithm for student classroom behavior detection under intelligent education. *Comput. Eng.* **50**, 344–355. https://doi.org/10.19678/j.issn.1000-3428.0069597 (2024).
51. Zhu, W. Csb-yolo: a rapid and efficient real-time algorithm for classroom student behavior detection. *J. Real-Time Image Proc.* **21** (4), 140 (2024).

Acknowledgements

This work was supported by "Weifang Medical College Scientific Research Innovation Plan Project" under the background of new medical construction of application-oriented high-quality medical personnel training path exploration (0215500137).

Author contributions

Conceptualization, L.H.; Methodology, X.M.; Software, L.B.; Validation, L.B.; Formal analysis, M.D.; Writing—original draft preparation, L.H.; Writing—review and editing, M.D. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.H.