Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Application of intelligent classroom management system based on deep learning in smart classrooms

Xiaomin Huang

1147648620@qq.com

Jiaxing Vocational & Technical College

Research Article

Keywords: Classroom management, Smart classroom, Deep learning, YOLO11

Posted Date: October 15th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7719324/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Universal Access in the Information Society on October 1st, 2026. See the published version at https://doi.org/10.1007/s10209-026-01349-0.

Application of intelligent classroom management system based on deep learning in smart classrooms

Xiaomin Huang[a*]

[a] School of Marxism, Jiaxing Vocational & Technical College, Jiaxing, 314036, China

* Corresponding Author. E-mail address: 1147648620@qq.com

Abstract: In the prevailing educational paradigm, classroom management predominantly depends on teachers' subjective judgments and static assessments as the primary evaluative approach. This conventional model lacks real-time and dynamic feedback mechanisms, resulting in inherent inefficiencies. To address this problem, this paper develops an intelligent classroom management monitoring system based on deep learning. The system consists of three parts: image acquisition module, image processing module and real-time feedback module. The system detects the behavioral information of students in the classroom through sensor technology and image recognition technology. In the image processing module section, we propose an improved YOLO11 model, the YOLO11-DP model, which effectively achieves real-time and efficient detection of student behavior in the classroom. The YOLO11-DP model significantly enhances detection performance through three key innovations: first, the introduction of a Diverse Branch Block (DBB) module in the C3k2 module effectively overcomes the representation capability bottleneck of the original YOLO11 during the feature fusion stage. Second, the traditional SPPF module is replaced with Poly-Scale Convolution (PSConv), which uses its dynamically varying dilation rate to achieve adaptive multi-scale feature extraction. Finally, the innovative mixed local channel attention mechanism (MLCA) is added, further optimizing the model's detection accuracy and computational efficiency in multi-object scenarios. Results show that the YOLO11-DP-MLCA model demonstrates excellent detection accuracy on the STBD-08 dataset, achieving a mAP@0.5 of 94.2%, and also exhibits good recognition performance in detecting student behavior in real classroom scenarios.

Keywords: Classroom management, Smart classroom, Deep learning, YOLO11

1. Introduction

With the ongoing implementation of the Education Informatization 2.0 initiative (a strategy

advancing the development of “Internet + Education” infrastructure) artificial intelligence is fundamentally transforming modern education ecosystems through its disruptive technological capabilities. This paradigm shift is catalyzing the evolution of teaching management from traditional experience-based approaches to contemporary data-driven, intelligent decision-making systems [1, 2]。This educational transformation has given rise to the “smart classroom” concept, which integrates conventional pedagogical approaches with cutting-edge technological systems. These innovative learning environments are fundamentally characterized by three core elements: digital instructional tools, information and communication technologies, and interactive learning platforms. Their advent signifies a profound paradigm shift in contemporary educational ecosystems. As a new type of teaching space for human-computer collaboration and a key vehicle for the realization of intelligent classroom management, the “smart classroom” is not only concerned with the maintenance of the teaching order, but also promotes the iteration of teaching strategies and a new ecology for the co-development of teachers and students [3, 4].

Current conventional classroom management paradigms remain predominantly reliant on teachers to discipline students' classroom behavior through personal subjective coercion. This approach not only manifests operational inefficiencies but also potentially compromises instructional continuity and pedagogical flow [5, 6]。 With the continuous progress of artificial intelligence, education is gradually shifting to digitalization and intelligence. The digitalization of education promotes the digitalization of classroom management, which in turn promotes the high-quality development of education. The smart classroom paradigm represents a distinct departure from conventional instructional approaches. Educators systematically utilize automated classroom recording capabilities to: document and analyze behavioral patterns, optimize management protocols, and cultivate an enhanced learning atmosphere. This integrated approach ultimately serves to improve overall instructional efficiency and learning outcomes [7, 8]. The intelligent classroom evaluation management system serves as a fundamental component within modern smart learning environments, enabling precise optimization of instructional processes through sophisticated multi-dimensional data fusion and analytical capabilities. As depicted in Fig. 1, the system architecture incorporates two functionally interconnected modules. The learning enhancement module utilizes computer vision technology to continuously monitor and analyze student engagement patterns during instructional activities, while simultaneously integrating multimodal environmental data to construct a

comprehensive real-time representation of classroom dynamics. The student evaluation module provides educators with AI-enhanced analytical outputs that facilitate evidence-based assessment of learner states, thereby empowering instructors to make data-driven adjustments to their pedagogical strategies and classroom management approaches in a timely and responsive manner.

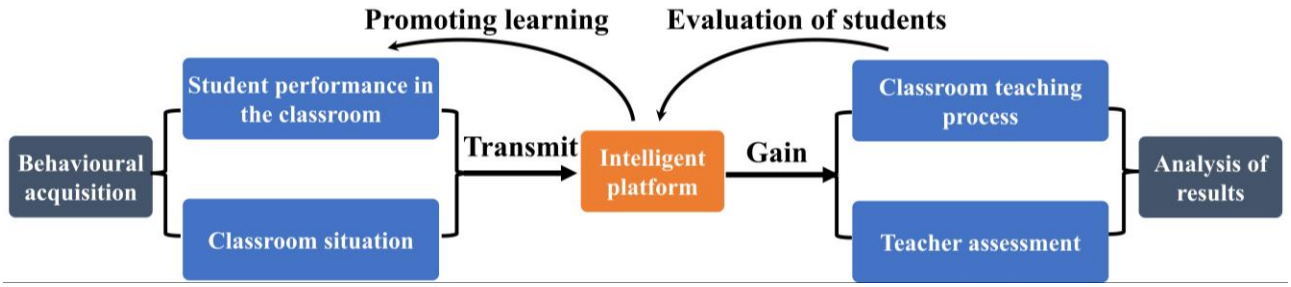


Fig. 1 Intelligent classroom assessment management components

In recent years, there have been many studies on digital classroom management. Classroom monitoring solutions based on wearable devices, such as smart bracelets, can collect students' physiological data in real time to analyze their learning status, but they can infringe on students' privacy rights, thus triggering ethical controversies [9]. Classroom behavior analysis methods based on traditional image processing rely on manually designed features such as HOG and SIFT. There are problems such as low recognition rate and high false detection rate in complex lighting and posture change scenarios, making it difficult to meet the real-time monitoring needs of dynamic classroom environments [10, 11]. Furthermore, classroom management systems utilizing sensor networks incur significant implementation costs owing to the substantial hardware infrastructure required for deployment. These systems are additionally constrained by several technical limitations, including latency in data synchronization and inadequate adaptability to varying environmental conditions [12]. The emergence of deep learning provides an innovative path to break through the technological bottleneck of traditional education management by virtue of its powerful feature representation capability and data-driven learning paradigm [13, 14]。In contrast to conventional image processing techniques, deep learning demonstrates superior capability in autonomously extracting hierarchical feature representations from large-scale educational datasets, exhibiting marked performance advantages in applications such as student behavior recognition and classroom interaction analysis [15, 16].

This paper tackles the issue of low management efficiency in classroom settings by deploying an intelligent detection model based on deep learning to monitor student behavior in real time, thereby

supporting teachers in maintaining classroom order. We propose an enhanced version of the YOLO11 model, named YOLO11-DP, which achieves efficient and real-time detection of student behaviors in classroom environments. To overcome the limitations of YOLO11 in feature fusion, we introduce the Diverse Branch Block (DBB) into the C3k2 module. Furthermore, to enhance the model's feature extraction capacity, the conventional SPPF module is replaced with multi-scale convolutions (PSConv), which dynamically integrate multi-scale receptive fields by combining convolutional layers with varying dilation rates, significantly improving the model's ability to capture discriminative features. Finally, to address the challenges associated with multi-object behavior detection, a Mixed Local Channel Attention mechanism (MLCA) is incorporated, which enhances both detection accuracy and computational efficiency for multiple objects. We have developed a real-time intelligent monitoring system for classroom management and evaluated it in practical scenarios. The experimental results validate the effectiveness of the proposed model in detecting student behaviors within classrooms.

2. Related work

There have been many studies on student behavior detection in the classroom. Gu et al. [17] proposed a fast target detection method based on the FFmpeg codec and MHI-HOG joint features to achieve efficient and intelligent recognition of classroom behaviors. Experiments show that the method has significant improvement in both detection speed and recognition accuracy. Guo et al. [18] proposed an innovative method based on forehead region detection for face detection challenges in complex environments such as classrooms. The method combines a dual classifier cascade framework of extended Haar-Like features and multi-channel color frequency features (MCCFF) to achieve efficient recognition in real scenes with occlusion and pose changes. Hsu et al. [19] proposed a multi-sensor-based intelligent system for monitoring classroom reading concentration, incorporating the artificial bee colony (ABC) optimization algorithm to process physiological signals and evaluate students' real-time engagement levels. While such methods have demonstrated promising results for student behavior detection, conventional image processing approaches remain limited by their reliance on manually engineered feature extraction rules. These methods often exhibit inadequate adaptability and limited generalization capability under challenging conditions such as variable lighting, occlusions, and cluttered backgrounds, resulting in suboptimal performance in complex classroom environments.

In contrast, deep learning algorithms, which automatically extract multi-level features through convolutional neural networks (CNN), Transformer, etc., and can adaptively capture key information,

have been widely used in target detection in the field of computer vision [20]. Liu et al. [21] developed an enhanced YOLOv8n_BT algorithm to address occlusion and small target detection challenges in natural classroom environments. By incorporating a feature enhancement module and adding a dedicated tiny target detection layer, the proposed method achieves significant improvements in behavior recognition accuracy. Wang et al. [22] developed a student classroom behavior detection system based on deformable DETR with Swin Transformer. By integrating a lightweight feature pyramid network (FPN) and CARAFE operator, the proposed method effectively mitigates the loss of location information inherent in traditional CNN approaches under multi-target and multi-scale scenarios. Experimental results demonstrate significant improvements in detection accuracy on real-world classroom datasets. Lu et al. [23] proposed a real-time target detection model named PACR-DETR. The model effectively improves the performance of behavior detection in complex classroom scenarios by integrating a position-aware channel residual network, an efficient residual mixing module, and a multibranch perception cross-level fusion network. Zhu et al. [24] proposed a lightweight CSB-YOLO model for real-time high-precision detection of student behaviors in the classroom by integrating a bi-directional feature pyramid network, an efficiently reparametrized detection head, and self-calibrated convolution. The deep learning algorithm significantly improves the accuracy and robustness of student behavior detection in the classroom through automatic feature extraction and multi-level modelling. Therefore, the further application of deep learning in intelligent classroom management is promoted to provide a reliable solution for breaking the technical bottleneck of complex teaching scenarios and the digital transformation of education.

3. Experimental

3.1 Intelligent Classroom Management System

Fig. 2 shows the module composition of the intelligent classroom management system. The intelligent classroom management system is divided into three modules, namely image acquisition, image processing and real-time feedback. The image acquisition module mainly acquires the behavioral information of students in the classroom through the image sensors installed in the classroom. The image sensor transmits the acquired information to the background image processing system, detects the students' behaviors and outputs the results in real time to the monitor on the teacher's side.

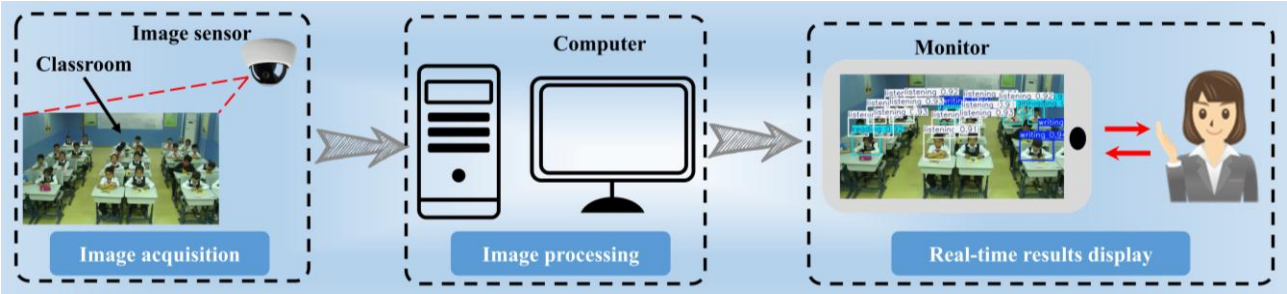


Fig. 2 Module Composition of Intelligent Classroom Management System

3.2 YOLO11 model

In computer vision, real-time object detection has remained a critical and widely studied research domain, with the You Only Look Once (YOLO) algorithm emerging as a prominent solution due to its exceptional performance. Initially introduced by Joseph Redmon et al. in 2015, this groundbreaking approach reformulates object detection as a unified regression problem, enabling end-to-end detection via a single deep neural network. Unlike conventional detection methods, YOLO achieves a significant breakthrough in computational efficiency, effectively fulfilling the stringent requirements of real-time processing [25].

The newest member of the YOLO family line, the YOLO11 is presented by Ultralytics and is designed with an ultra-lightweight architecture that outperforms its predecessor version in terms of speed and efficiency [26]. YOLO11 achieves further optimization in detection accuracy, inference speed, and architectural efficiency. Its core architecture continues the single-stage detection paradigm of the YOLO family, achieving end-to-end efficient detection through a carefully designed backbone network, feature fusion module (Neck), and detection head (Head). The core architectural advancements in YOLO11 primarily involve replacing the original C2f module with a more computationally efficient C3k2 module. This innovative structure builds upon the Cross Stage Partial (CSP) bottleneck design, employing a feature map partitioning strategy where partial features pass through the bottleneck layer before merging with original features, thereby reducing computational complexity while enhancing feature representation. The architecture further substitutes YOLOv8's large convolutional kernels with a dual convolutional structure. While retaining the SPPF module, YOLO11 introduces a novel cross-stage local spatial attention (C2PSA) module that strengthens spatial attention mechanisms, enabling more precise focus on critical image regions. For neck network optimization, YOLO11 similarly replaces C2f modules with C3k2 modules, utilizing parallel

convolutional layers to process input feature maps - an approach that effectively expands network width and improves feature extraction capabilities. The detection head of YOLO11 adopts a dynamic anchor allocation strategy, which uses an adaptive mechanism to match the best prior boxes for targets of different sizes, significantly improving detection stability in dense scenes. In addition, the model introduces a lightweight decoupled head design, which separates classification and regression tasks to avoid mutual interference in feature learning, while combining a task-specific context module to enhance task awareness.

3.2 Diverse Branch Block Fusion Method

Fig. 3 shows the structure of the Diverse Branch Block (DBB) module. In the training phase of the DBB module, a multi-branch design is used. Each branch is equipped with bulk normalization and nonlinear activation. This multi-branch design enriches the feature representation and enables the model to learn a wider range of features during training. In the inference phase, the entire multi-branch structure is reduced to an equivalent single-branch structure through reparameterization, which significantly improves the inference speed while maintaining the model performance [27]。

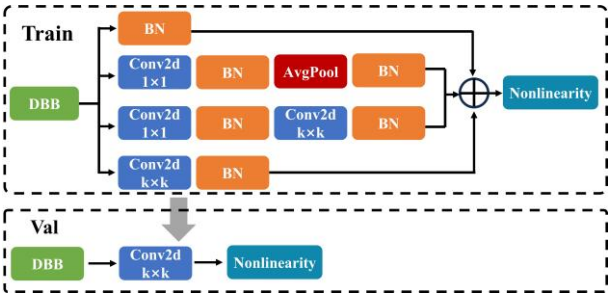


Fig. 3 Diverse Branch Block network structure

The DBB enhances spatial feature representation and improves multi-scale target perception by employing parallel convolutions with varying kernel sizes (including 1×1, 3×3, and 5×5 configurations) alongside dilated convolutions. As demonstrated in the following equation [28] ：

$$O = I * F^{(1)} + I * F^{(2)} + I * F^{(3)} + \cdots$$

where the input features $F^{(1)}$, $F^{(2)}$, and $F^{(3)}$ are 1 × 1, 3 × 3, and 5 × 5, respectively.

In the training phase, the DBB module consists of multiple branches, each of which performs independent convolutional operations. The outputs of these branches are then aggregated through a feature fusion mechanism as shown in the following equation:

$$O = \sum_{i=1}^{n} (I * F^{(i)} + REP(b^{(i)}))$$

$$O_{j,:,:} = \left(\frac{(I * F)_{j,:,:} - \mu_J}{\sigma_j} \right) \gamma_j + \beta_j$$

where n is the number of branches in the DBB module, $F^{(i)}$ is the convolution kernel of the ith branch, and $REP(b^{(i)})$ is the bias of the ith branch, γ_j and β_j are the scaling factor and bias.

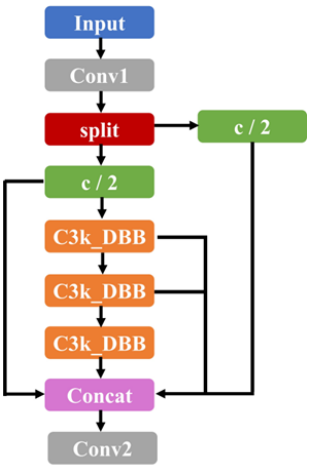


Fig. 4 Improved C3k2 structure

To overcome the limitations of YOLO11 in the feature fusion stage (such as conventional convolution operations being unable to effectively capture features of variable scales, fixed convolution structures limiting adaptive computational complexity during training and inference, and limited non-linear representation capabilities), this paper replaces the standard residual block in the C3k2 module with DBB. DBB achieves the goal of enhancing the expressive power of individual convolutions by enriching the diversity of the feature space through the combination of branches of different scales and complexities. Replacing the Convolution-BatchNorm-Silu (CBS) residual block with DBB effectively enhances feature diversity, resulting in significant performance improvements. The traditional C3k2 module integrates CSPNet with standard residual blocks. However, its reliance on a single 3×3 convolution operation limits the receptive field, hindering its effectiveness in detecting objects across varying scales. In contrast, the DBB module introduces an innovative multi-branch architecture that enables automatic multi-scale feature fusion. This design not only enhances the detection of small objects by capturing finer details but also strengthens global perception for large objects. A distinctive advantage of the DBB module is its improved model robustness. The incorporated pooling branch effectively mitigates image noise, allowing stable performance even with degraded input quality. Meanwhile, the large-kernel convolution branch enhances global context

modeling, which significantly reduces false detections in cluttered backgrounds. Although the DBB structure is more complex than a standard residual block, its carefully optimized parallel computing architecture ensures efficient GPU inference. When combined with techniques such as depth wise separable convolutions, computational efficiency is further improved, enabling significant gains in performance without substantially compromising inference speed. These enhancements allow YOLO11 to achieve superior detection performance across diverse and complex scenarios.

3.3 Poly-Scale Convolution

Poly-Scale Convolution (PSConv) is an innovative convolution operation module designed to enhance the feature extraction capabilities of traditional convolution by dynamically combining multi-scale receptive fields. Figure 5 shows the network structure diagram of the PSConv module. PSConv achieves multi-scale feature fusion by periodically varying the dilation rate within a single convolution kernel. Unlike traditional convolution or batch-wise dilated convolution, the dilation rate in PSConv changes regularly along the dimensions of the input and output channels, enabling efficient capture of richer scale information without increasing the number of parameters or computational complexity. By integrating convolutional layers with different dilation rates, PSConv achieves efficient and powerful feature extraction capabilities. Its core lies in using low-dilation convolutions to capture local detail information while leveraging high-dilation convolutions to extract global contextual features, thereby comprehensively covering target information at different scales within an image. The PSConv module employs a shared convolutional layer (self.share_conv), which significantly reduces the number of parameters and mitigates the redundancy associated with using separate convolutional layers for each expansion rate. This design not only improves computational efficiency but also reduces storage requirements and computational overhead. Through carefully designed 1×1 convolutional layers (self.cv1 and self.cv2), the module achieves efficient channel adjustment and feature fusion. These lightweight convolutions effectively preserve critical features while compressing the parameter size. Replacing the SPPF module with PSConv in the YOLO11 architecture leads to a notable improvement in model performance. In contrast to the potential loss of fine details caused by the pooling operations in SPPF, the multi-scale convolutional approach based on standard convolution operations demonstrates superior capability in capturing fine-grained details and complex patterns within images. This design offers greater flexibility in feature representation and enhances the overall expressive power of the model. In terms of feature preservation, the convolutional operations in

PSConv avoid the loss of small object details often resulting from the multi-level max pooling in SPPF. Instead, by periodically varying expansion rates along the channel dimension, PSConv better retains fine-grained features. Furthermore, the dynamically adjustable expansion rates in PSConv improve geometric adaptability to irregular objects. The use of high expansion rates enables the module to capture long-range dependencies in occluded scenarios, while supporting end-to-end learning of scale parameters, thereby strengthening the model's adaptability to specific tasks. Fig. 6 illustrates the architecture of the final improved model, termed YOLO11-DP.

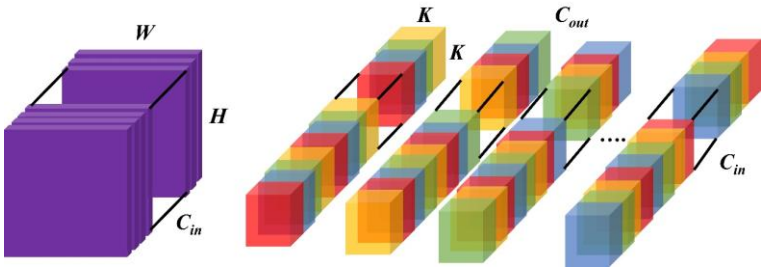


Fig. 5 Network structure diagram of the PSConv module

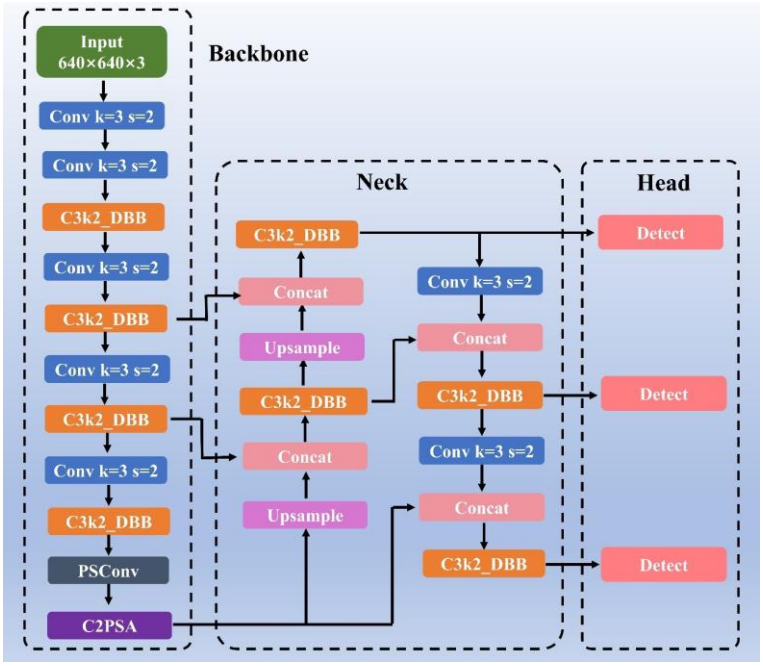


Fig. 6 Architecture of the YOLO11-DP model

4. Experiments and analysis of results

The dataset in this paper is derived from the pre-existing STBD-08 dataset (https://github.com/Whiffe/SCB-dataset.git) [29]. The dataset contains eight target behavioral tests in the classroom are: Writing (Students lower their head and rest it on the desk, taking notes), Reading (Students lower their head and rest it on the desk, taking notes), Listening (Students look up and listen

to teacher in class.), Turning around (Students rotate their bodies to face each other and talk.), Raising hand (Students raise their hands), Standing (Students stand up), Discussing (Several students form a group to discuss questions), Guiding (Teacher guides students). The hardware and software configurations during the experiments are as follows: the CPU model is Intel Core i7-12700K, the GPU model is NAVID RTX 3060, the operating system is Ubuntu 22.04, the version of Python is 3.9, the version of Pytorch is 1.10, and the version of CUDA is 10.2. Table 1 shows the training of the YOLO11 model when the parameter settings.

Table 1 Model training parameter settings

Parameters	Value
Image size	640×640
Train epoch	200
Batch size	32
Initial learning rate	0.01
Optimizer	SGD
Learning Rate	0.937
Momentum	

Model evaluation represents a crucial stage in model development and analysis. Appropriate evaluation metrics enable accurate performance measurement, facilitating model optimization and improvement [30]. Among the most significant metrics are Precision and Recall: Precision measures the proportion of correctly identified positive predictions among all positive predictions made by the model, while Recall quantifies the proportion of actual positive samples that are correctly detected by the model. The mathematical expressions for Precision and Recall are presented as follows [31, 32]:

$$\text{Precision} = \frac{TP + FP}{TP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

where TP denotes a correctly detected real object, FP denotes a misdetection, and FN denotes a missed detection.

This study evaluates the performance of the YOLO11-DP model using the mean Average Precision (mAP) metric. The evaluation process first calculates the Average Precision (AP) for each individual category, then computes the final mAP by averaging these AP values across all categories.

The mathematical formulation for calculating the model's Average Precision is presented as follow. [33]:

$$AP = \int_0^1 p(r)dr$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C}$$

where C represents the total number of detection classes and AP_i represents the average accuracy of class i.

Table 2 Comparison of performance metrics of different detection models

Methods	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
Faster R-CNN	76.9%	74.7%	78.8%	65.8%	21
YOLOv4	78.5%	77.8%	80.2%	69.1%	46
YOLOv5	89.1%	87.1%	91.7%	76.8%	90
YOLOv8	90.3%	86.8%	91.6%	77.1%	98
YOLO11	88.9%	86.5%	**92.3%**	**77.9%**	**128**
YOLO11-DBB	89.8%	87.9%	**93.3%**	**79.2%**	**120**
YOLO11-PSC	89.3%	88.4%	**91.9%**	**77.5%**	**110**
YOLO11-DP	88.8%	86.9%	**93.7%**	**79.6%**	**102**

The performance metrics obtained from different detection models trained on the STBD-08 dataset are presented in Table 2. The results demonstrate that YOLO series models exhibit significant advantages in both detection accuracy and processing speed. Compared to YOLOv4, YOLOv5, and YOLOv8, YOLO11 shows improvements in both mAP@0.5 and mAP@0.5:0.95, while maintaining excellent detection speed at 128 FPS. The enhanced YOLO11-DBB model achieves performance gains, with 1.0% higher mAP@0.5 and 1.3% improved mAP@0.5:0.95 compared to the baseline YOLO11. The enhanced YOLO11-DP model achieves further performance gains, with 1.4% higher mAP@0.5 and 1.7% improved mAP@0.5:0.95 compared to the baseline YOLO11.

Fig. 7 illustrates the evolution of model loss and precision metrics during the training of YOLO11-DP. In the initial training phase, both Box loss and DFL loss remain relatively high, while Precision and Recall values are comparatively low. However, the model exhibits rapid convergence. By approximately Epoch 30, the model stabilizes, achieving Precision exceeding 85.0% and Recall

surpassing 85%, with these metrics maintaining stability throughout subsequent training iterations. Fig. 7(e) presents the training progression of mAP@0.5 and mAP@0.5:0.95 for the YOLO11-DP model, which achieves strong performance on this dataset with mAP@0.5 reaching 92.6% and mAP@0.5:0.95 attaining 79.3%. The P-R curve in Fig. 7(f) further demonstrates that the model maintains high precision across all behavioral action recognitions even as recall increases. These results indicate that the YOLO11-DP model converges rapidly and delivers robust target detection performance. Table 3 shows the evaluation statistics of the training results for different behavioral categories based on the YOLO11-DP model. The number of labels for different behaviors in this dataset are all high, which can meet the training requirements. Among them, the training results for the detection of Writing and Listening behaviors obtained very high average accuracies. mAP@0.5 reached 98.4% and mAP@0.5:0.95 reached 89.8% for the Writing behavior. mAP@0.5 reached 97.9% and mAP@0.5:0.95 reached 90.8% for the Listening behavior, which may be attributed to the high number of labels.

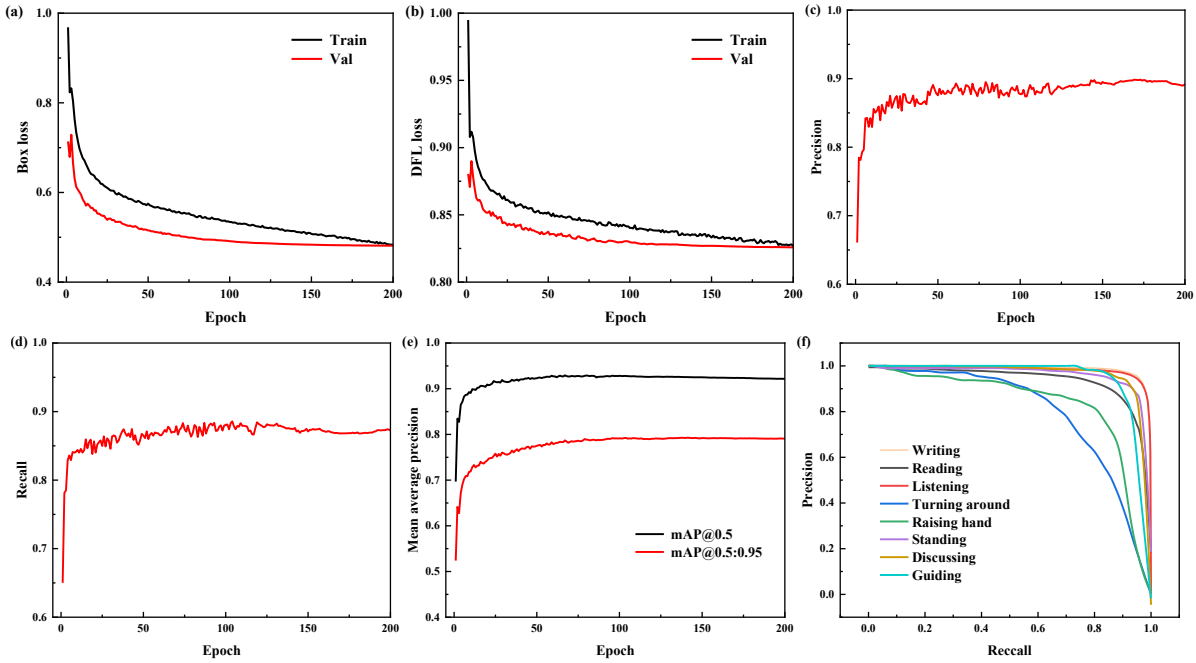


Fig. 7 Training results of student behavior category detection in classroom based on YOLO11-DP model:(a) Box loss during training and validation, (b) DFL during training and validation, (c) Precision, (d) Recall, (e) mean Average Precision, (f) P-R curve

Table 3 Training results for different behavioral categories based on the YOLO11-DP model

Class	Images	Instances	Precision	Recall	mAP@0.5	mAP@0.5:0.95
All	1832	55160	89.0%	87.6%	93.7%	79.6%

Writing	870	15289	95.5%	95.1%	98.4%	89.8%
Reading	1275	12060	89.0%	86.6%	93.7%	82.3%
Listening	1416	24019	93.7%	95.6%	97.9%	90.8%
Turning around	484	1025	78.1%	69.5%	81.0%	70.9%
Raising hand	288	847	81.1%	80.2%	83.4%	66.9%
Standing	499	814	89.9%	93.5%	96.1%	82.3%
Discussing	201	953	93.6%	91.3%	96.3%	78.5%
Guiding	141	144	90.9%	88.9%	94.8%	73.3%

Squeeze-and-Excitation Networks (SE) is an attention mechanism that dynamically adjusts the importance of features through channel dimension. Its core working principle is to compress spatial information through global average pooling, obtain global feature representations for each channel, then use fully connected layers to learn dependencies between channels, and finally generate channel weight vectors to re-calibrate feature maps. This mechanism automatically enhances key channel features relevant to the current detection task while suppressing interference from irrelevant or noisy channels. In the YOLO11 model, the introduction of the SE module significantly improves small object detection performance by adjusting channel-level weights in low-resolution feature maps within shallow networks. This effectively enhances feature channels containing high semantic information but low spatial resolution, thereby improving the ability to retain features of small-scale objects. Coordinate Attention (CA) is an innovative attention mechanism that significantly enhances the performance of object detection models through a unique approach to processing coordinate information. The core innovation of this mechanism lies in decomposing two-dimensional spatial position information into one-dimensional global pooling operations along the X and Y directions, thereby constructing direction-sensitive positional attention maps. This design ingeniously integrates spatial coordinate information with channel attention, effectively addressing the issue of positional information loss caused by translation invariance in traditional convolutional neural networks. In the YOLO11 model, the CA mechanism's unique coordinate attention mechanism can precisely model the geometric structural features of objects. By explicitly encoding spatial positional relationships, it significantly improves the accuracy of bounding box regression. The Global Attention Mechanism (GAM) is a powerful attention architecture that significantly enhances a model’s global perceptual

ability through its unique cross-dimensional interaction design. Employing a cascaded structure that integrates both channel and spatial attention modules, GAM enables collaborative modeling of inter-channel dependencies and long-range spatial relationships. A key aspect of its design is the use of large convolutional kernels (e.g., 7×7) to establish a global receptive field, effectively mitigating the loss of contextual information caused by the limited local receptive fields in deep neural networks. The GAM architecture offers several performance advantages. First, its strong global context modeling capability substantially improves the model's robustness in complex environments, enhancing the discrimination between target objects and background regions. Second, the mechanism facilitates effective integration of multi-scale features, particularly enriching the representation of small-scale objects within high-level features of the backbone network. This leads to a notable increase in the recall rate for small object detection tasks. Fig. 8 shows the structure of the Mixed Local Channel Attention (MLCA) attention mechanism. MLCA is an efficient multimodal attention mechanism that innovatively fuses local spatial perception with global channel interaction to significantly improve target detection performance through hierarchical feature integration [34, 35]. The MLCA mechanism integrates three coordinated components: A local spatial attention branch employing 7×7 depth wise separable convolution to construct dynamically adjustable receptive fields, which adaptively modulates sampling density to reduce computational overhead while resolving boundary ambiguity in dense targets. An innovative lightweight channel attention branch utilizing 1×5 one-dimensional convolutions (replacing traditional fully-connected layers) to enable efficient long-range cross-channel dependency modeling with minimal parameters. A bidirectional feature fusion module that dynamically weights spatial and channel features via gating mechanisms, combining sigmoid activation with residual learning to optimize feature integration while preserving information flow and mitigating gradient vanishing. The integration of the MLCA attention mechanism into the YOLO11-DP model significantly enhances classroom student behavior detection accuracy. Its 7×7 depth wise separable convolutional local spatial attention module precisely captures subtle behavioral features, substantially improving small-target behavior recognition. The channel attention branch leverages cross-channel interactions to effectively discriminate between similar behaviors, reducing misclassification rates in dense scenarios. Furthermore, the mechanism facilitates multi-behavior collaborative analysis, leading to improved composite behavior recognition performance.

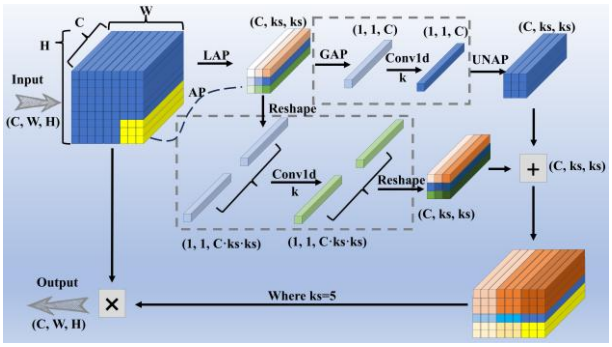


Fig. 8 Structure of MLCA attention mechanism

Table 4 presents the performance metrics of the YOLO11-DP model incorporating different attention mechanisms, including SE, CA, GAM, and MLCA. Comparative analysis reveals that GAM and MLCA achieve the most substantial improvements: GAM enhances mAP@0.5:0.95 by 0.6% over the YOLO11-DP, while MLCA increases mAP@0.5 by 0.5% and mAP@0.5:0.95 by 1.0%. Fig. 9 shows an example of the results of object behavior detection of student behavior in the classroom based on the YOLO11-DP+MLCA model. We selected images of different types of behaviors in the classroom to input to the model for testing. The model obtained good detection results for both all the same behaviors and complex multiple behaviors of students in the classroom. In order to meet the complexity of real classroom situations, we also detected student behaviors in both large and small scenes, depending on the classroom scenarios. The results show that the model also has excellent recognition results for student behaviors in different classroom scenarios.

Table 4 Performance metrics of YOLO11-DP model after adding different attention mechanisms

Methods	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
YOLO11-DP	89.8%	87.9%	93.7%	79.6%	114
+ SE	87.9%	86.4%	91.8%	78.6%	102
+ CA	90.8%	89.2%	92.3%	79.1%	98
+ GAM	90.5%	88.2%	93.3%	**80.2%**	87
+ MLCA	90.7%	88.5%	**94.2%**	**80.6%**	106

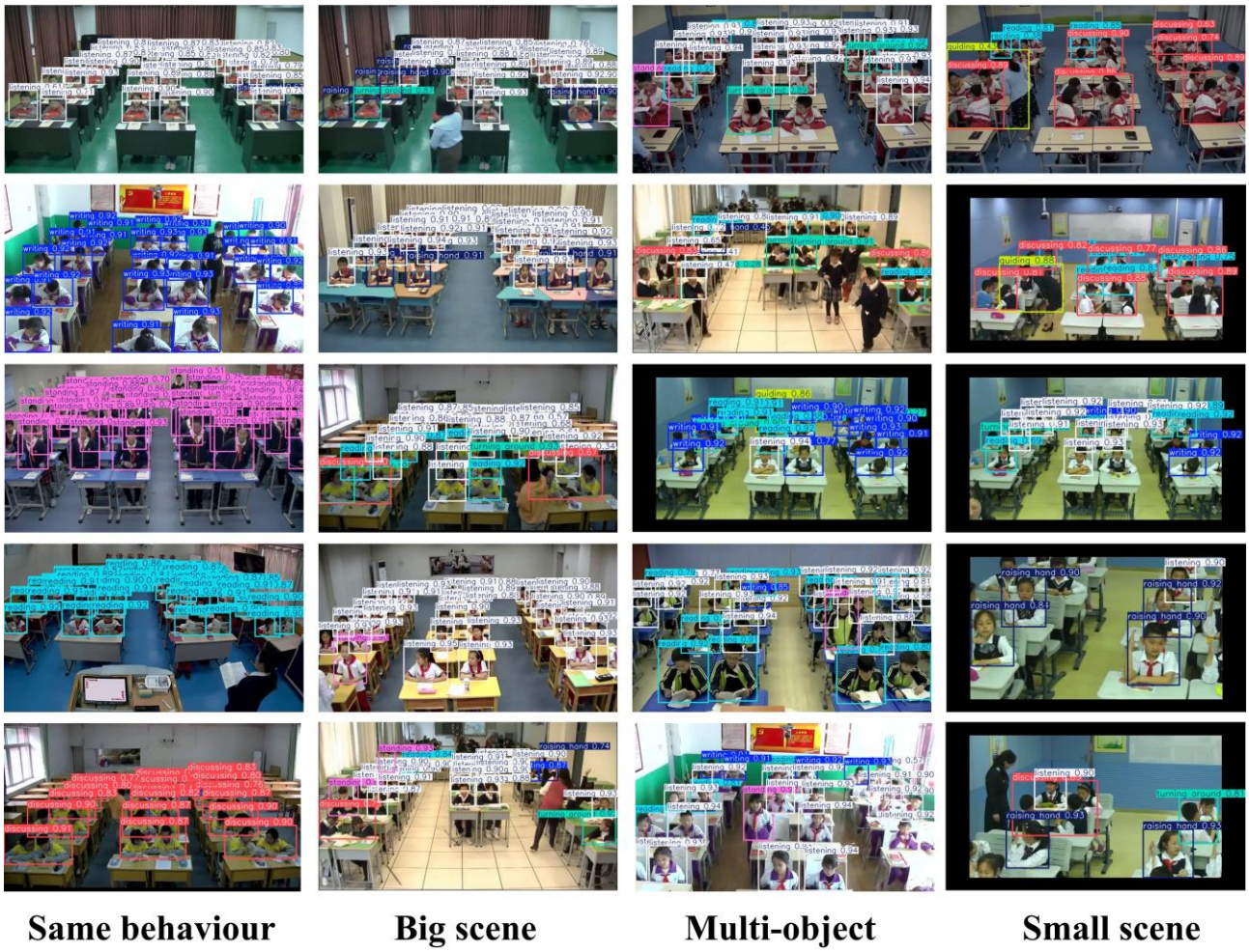


Fig. 9 Example of object behavior detection results based on YOLO11-DP + MLCA model

5. Analysis and discussion

Classroom management is a core part of the educational process, which directly affects the quality of teaching and students' learning effects. Classroom management is a core part of the educational process, which directly affects the quality of teaching and students' learning effects. The development and application of Intelligent Classroom Management System (ICMS) provides an innovative solution to break through the traditional classroom management dilemma. Through data-driven, technology-enabled and process-optimized, ICMS is profoundly changing the mode and effect of classroom management. With the rapid development of artificial intelligence, big data, Internet of Things and other technologies, classroom management is moving towards a new stage of greater intelligence, personalization and collaboration.

To overcome the limitations of conventional classroom management approaches, this research introduces an intelligent classroom management system utilizing the YOLO11-DP+MLCA object detection model. Distinct from prior studies, the proposed system integrates sensor technology with

advanced image recognition capabilities, while employing experimental methodologies to rigorously evaluate the system's detection accuracy and operational effectiveness. Fig. 10 demonstrates the intelligent classroom management system's behavioral recognition performance in real-world scenarios. While exhibiting robust detection capabilities for student behaviors under typical classroom conditions, the system occasionally generates false positives in complex, cluttered environments, as indicated by the orange bounding boxes in Fig. 10(b, d). The current system has deficiencies in recognizing distant and dense student behavioral states in the classroom scene, and miss-detection can occur, as shown by the yellow boxes in Fig. 10(a, c). This can be attributed to the poor recognition and classification ability of the target detection model in targeting dense crowds. The location of the image sensors is also an important cause of this problem. The sensor deployment must achieve maximal coverage of each student's behavioral range within the classroom environment, thereby mitigating occlusion issues and ensuring complete capture of student movement dynamics. From the experimental results, the learning states identified by the intelligent classroom management system are highly correlated with those identified by human observers. In addition, when the system is enabled, teachers can quickly understand students' learning states through the system's real-time test result display interface, which helps teachers to manage their classrooms more effectively. As many researchers have pointed out, the concentration of students' learning is a key factor that affects learning outcomes. And it can be predicted that with the help of the proposed system, students' learning outcomes will be improved.



Fig. 10 Example of recognition of an intelligent classroom management system in a real scenario

6. Conclusions

This paper presents an intelligent classroom management monitoring system based on deep learning and introduces an improved YOLO11 model, termed YOLO11-DP, designed to achieve efficient real-time detection of student behaviors in classroom environments. The proposed model integrates several key enhancements: the Diverse Branch Block (DBB) module enriches feature space diversity and strengthens the representational capacity of convolutional operations through the combination of multi-scale and structurally varied branches. Furthermore, the PSConv module significantly improves feature extraction by dynamically integrating multi-scale receptive fields and incorporating convolutional layers with different dilation rates. On the STBD-08 dataset, the YOLO11-DP model demonstrates outstanding performance, achieving a mAP@0.5 of 94.2% and a mAP@0.5:0.95 of 80.6%. Compared to the original YOLO11 model, mAP@0.5 improved by 1.4% and mAP@0.5:0.95 improved by 1.7%. By further incorporating a mixed local channel attention (MLCA) mechanism to enhance multi-object detection capabilities, the YOLO11-DP+MLCA variant attains additional gains, with mAP@0.5 increasing by 1.9% and mAP@0.5:0.95 by 2.7%. Experimental results confirm the effectiveness of the proposed intelligent monitoring system when

deployed in real classroom scenarios, validating its practical utility in supporting teaching and management tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Funding

The authors received no specific funding for this work.

References

1. Chen, L., Ifenthaler, D., Yau, J. Y.-K., Sun, W. Artificial intelligence in entrepreneurship education: A scoping review. Education + Training. **66**(6), 589-608 (2024). https://doi.org/10.1108/ET-05-2023-0169
2. Marino, M. T., Eleazar, V., Lisa, D., James, B., Jose, B. https://doi.org/-10.1177/01626434231165977
3. Qianyi, Z., Zhiqiang, L. Research on multimodal based learning evaluation method in smart classroom. Learning and Motivation. **84**(101943 (2023). https://doi.org/https://doi.org/10.1016/j.lmot.2023.101943
4. Huang, J., Zhou, D. A scalable real-time computer vision system for student posture detection in smart classrooms. Education and Information Technologies. **29**(1), 917-937 (2024). https://doi.org/10.1007/s10639-023-12365-5
5. Wang, Z., Yao, J., Zeng, C., Li, L., Tan, C. Students' classroom behavior detection system incorporating deformable detr with swin transformer and light-weight feature pyramid network. Systems. **11**(7), 372 (2023).
6. Ngoc Anh, B., Tung Son, N., Truong Lam, P., Phuong Chi, L., Huu Tuan, N., Cong Dat, N., Huu Trung, N., Umar Aftab, M., Van Dinh, T. A computer-vision based application for student behavior monitoring in classroom. Applied Sciences. **9**(22), 4729 (2019).

7. Saini, M. K., Goel, N. How smart are smart classrooms? A review of smart classroom technologies. ACM Comput Surv. **52**(6), Article 130 (2019). https://doi.org/10.1145/3365757

8. Wang, J., Tigelaar, D. E. H., Luo, J., Admiraal, W. Teacher beliefs, classroom process quality, and student engagement in the smart classroom learning environment: A multilevel analysis. Computers & Education. **183**(104501 (2022). https://doi.org/https://doi.org/10.1016/j.compedu.2022.104501

9. López-Jiménez, J. J., Fernández-Alemán, J. L., González, L. L., Sequeros, O. G., Valle, B. M., García-Berná, J. A., Idri, A., Toval, A. Taking the pulse of a classroom with a gamified audience response system. Computer Methods and Programs in Biomedicine. **213**(106459 (2022). https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106459

10. Okazawa, T., Egi, H. Accelchalk: Detecting writing actions with chalk acceleration for collaboration between teachers and students. Collaboration Technologies and Social Computing. 99-106 (2017)

11. Wu, M.-T. Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. Scientific Reports. **12**(1), 3095 (2022). https://doi.org/10.1038/s41598-022-07137-z

12. Sharma, S., Chen, W. Multi-user vr classroom with 3d interaction and real-time motion detection. 2014 International Conference on Computational Science and Computational Intelligence. **2**(187-192 (2014)

13. Pabba, C., Kumar, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems. **39**(1), e12839 (2022). https://doi.org/https://doi.org/10.1111/exsy.12839

14. Zhang, Q., Osman, S. Z. M., Li, H., Lin, X. An empirical study on teaching management and quality feedback based on human behavior detection. Journal of Organizational and End User Computing (JOEUC). **36**(1), 1-16 (2024). https://doi.org/10.4018/JOEUC.350096

15. Zhou, J., Ran, F., Li, G., Peng, J., Li, K., Wang, Z. Classroom learning status assessment based on deep learning. Mathematical Problems in Engineering. **2022**(1), 7049458 (2022). https://doi.org/https://doi.org/10.1155/2022/7049458

16. Wang, R., Chen, S., Tian, G., Wang, P., Ying, S. Post-secondary classroom teaching quality evaluation using small object detection model. Scientific Reports. **14**(1), 5816 (2024).

https://doi.org/10.1038/s41598-024-56505-4

17. Gu, C., Li, Y. Analysis of art classroom teaching behavior based on intelligent image recognition. Mobile Information Systems. **2022**(1), 5736407 (2022). https://doi.org/https://doi.org/10.1155/2022/5736407
18. Meng, G., Qingmei, C., Bo, S., Jun, H. Forehead-based face detection algorithm with multi-feature cascade framework for classroom. ProcSPIE. **10806**(108061N (2018)
19. Hsu, C.-C., Chen, H.-C., Su, Y.-N., Huang, K.-K., Huang, Y.-M. Developing a reading concentration monitoring system by applying an artificial bee colony algorithm to e-books in an intelligent classroom. Sensors. **12**(10), 14158-14178 (2012)
20. Cust, E. E., J., S. A., Kevin, B., and Robertson, S. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. Journal of Sports Sciences. **37**(5), 568-600 (2019). https://doi.org/10.1080/02640414.2018.1521769
21. Liu, Q., Jiang, R., Xu, Q., Wang, D., Sang, Z., Jiang, X., Wu, L. Yolov8n_bt: Research on classroom learning behavior recognition algorithm based on improved yolov8n. IEEE Access. **12**(36391-36403 (2024). https://doi.org/10.1109/ACCESS.2024.3373536
22. Wang, Z., Yao, J., Zeng, C., Li, L., Tan, C. Students' classroom behavior detection system incorporating deformable detr with swin transformer and light-weight feature pyramid network. Systems. **11**(7), (2023)
23. Lu, W., Liu, X., Peng, Y., Kyrarini, M., An, K., Cheng, Y. Pacr-detr: A real-time end-to-end object detector for behavior recognition in various classroom scenarios. IEEE Transactions on Instrumentation and Measurement. **74**(1-20 (2025). https://doi.org/10.1109/TIM.2025.3550238
24. Zhu, W., Yang, Z. Csb-yolo: A rapid and efficient real-time algorithm for classroom student behavior detection. Journal of Real-Time Image Processing. **21**(4), 140 (2024). https://doi.org/10.1007/s11554-024-01515-8
25. Wang, Q., Wang, Q. L. Bt-yolo11: Automatic driving road target detection in complex scenarios. IEEE Access. **13**(72364-72374 (2025). https://doi.org/10.1109/ACCESS.2025.3562747
26. Robledo-Vega, I., Osuna-Tostado, S., Rodríguez-Mata, A. E., García-Mata, C. L., Acosta-Cano, P. R., Baray-Arana, R. E. Embedded vision system for thermal face detection using deep learning. Sensors. **25**(10), 3126 (2025).
27. Yu, X., Yan, P., Zheng, S., Du, Q., Wang, D. Yolov8-wtdd: Multi-scale defect detection algorithm

for wind turbines. The Journal of Supercomputing. **81**(1), 32 (2024). https://doi.org/10.1007/s11227-024-06487-x

28. Yu, F., Zhang, J., Mu, D. Steel defect detection based on yolo-safd. IEEE Access. **13**(77291-77304 (2025). https://doi.org/10.1109/ACCESS.2025.3565587
29. Zhao, J., Zhu, H. Cbph-net: A small object detector for behavior recognition in classroom scenarios. IEEE Transactions on Instrumentation and Measurement. **72**(1-12 (2023). https://doi.org/10.1109/TIM.2023.3296124
30. Zhang, Y., Guo, Z., Wu, J., Tian, Y., Tang, H., Guo, X. Real-time vehicle detection based on improved yolo v5. Sustainability. **14**(19), 12274 (2022).
31. Zhou, G., Liu, M., Wang, H., Zheng, Y. Power equipment image enhancement processing based on yolo-v8 target detection model under msrcr algorithm. International Journal of Low-Carbon Technologies. **19**(1717-1724 (2024). https://doi.org/10.1093/ijlct/ctae122
32. Huo, J., Shi, B., Zhang, Y. An object detection method for the work of an unmanned sweeper in a noisy environment on an improved yolo algorithm. Signal, Image and Video Processing. **17**(8), 4219-4227 (2023). https://doi.org/10.1007/s11760-023-02654-4
33. Xiong, J., Wu, J., Tang, M., Xiong, P., Huang, Y., Guo, H. Combining yolo and background subtraction for small dynamic target detection. The Visual Computer. **41**(1), 481-490 (2025). https://doi.org/10.1007/s00371-024-03342-1
34. Wan, D., Lu, R., Shen, S., Xu, T., Lang, X., Ren, Z. Mixed local channel attention for object detection. Engineering Applications of Artificial Intelligence. **123**(106442 (2023). https://doi.org/https://doi.org/10.1016/j.engappai.2023.106442
35. Lu, F., Yao, S., Sun, G., Zhou, T., Huang, Y. Fms-yolo: A lightweight safety belt detection algorithm for high-altitude workers based on attention mechanism and efficient architecture. Journal of Real-Time Image Processing. **22**(2), 90 (2025). https://doi.org/10.1007/s11554-025-01669-z