Current Science
CS 5 (5), 4828-4840 (2025)
Received 20 July 2025 | **Revised 12 August** 2025 | **Accepted 05 September** 2025 | **Online Available 15 October** 2025
https://doi.org0/10.52845/CS/2025-5-5-36

Current Science

Current Science

OPEN ACCESS

ISSN(O)2795-8639

Research Article

CPID-YOLOv8: Student Behavior Detection Algorithm Based on Improved YOLOv8

Tao Fan

College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201600, China

***Corresponding Author: Tao Fan**

Abstract:
To address challenges such as dense small targets, occlusions, and multi-scale spanning in natural classroom teaching scenarios, an improved YOLOv8 algorithm for student classroom behavior detection, named CPID-YOLOv8, is proposed. Specifically targeting the issue of small target occlusion, the Channel Prior Convolutional Attention (CPCA) was incorporated into the Backbone, enhancing the model's channel response across different scales and capturing more detailed and comprehensive contextual information. Additionally, to tackle the problem of multi-scale spanning, the Parallelized Patch-Aware Attention (PPA) multi-branch feature extraction strategy was utilized, improving the model's ability to select multi-scale features and facilitating the extraction of more effective multi-scale feature information. Furthermore, the Dimension-Aware Integration Diffusion Feature Pyramid Network (DAIDFPN) module was proposed to replace the original Neck layer for feature fusion. This module enables adaptive selection and meticulous fusion of high-dimensional and low-dimensional features. Through its diffusion mechanism, it ensures that each scale feature is provided with detailed context while reducing the model's parameter count. Experimental results demonstrate that the improved CPID-YOLOv8 model exhibits excellent detection performance in student classroom behavior tasks. Compared with the baseline YOLOv8 algorithm, the CPID-YOLOv8 model achieved improvements of 1.7%, 2.1%, 2.2%, and 2.5% in P, R, mAP50, and mAP50-95, respectively, substantiating the effectiveness of the proposed improvements.
Keywords: YOLOv8; student classroom behavior detection; small target; feature fusion; object detection

1. Introduction

In traditional teaching activities, classroom teaching has always been the main form of school education and teaching, and it is also the focus of educational research. Process-oriented assessment of teaching quality in classroom instruction is an important indicator reflecting student engagement and teacher effectiveness, and it is also an important basis for educational evaluation and instructional improvement. In the past, observations and records of student classroom learning behaviors during classroom teaching relied mainly on the subjective judgment of teachers. This method was time-consuming, labor-intensive, and subject to teacher biases, hindering objective, comprehensive, and timely classroom behavior analysis. With the development of artificial intelligence technology, empowering traditional educational formats with computer vision deep learning technology has become a popular trend. Compared to traditional methods of classroom behavior recognition and analysis, this method is highly automated, real-time, and efficient [1]. It can provide accurate and objective data support for subsequent analysis of student classroom teaching behaviors and the construction of a teaching quality evaluation system. This promotes the unification of subjective and objective evaluations, qualitative and quantitative

analyses, ensuring comprehensive and accurate educational and instructional feedback.

In natural classroom teaching scenarios, recognizing students' classroom behaviors based on object detection algorithms is a complex issue. The challenges are highlighted as follows:(1)Real classroom scenes are crowded, and students' behaviors are ambiguous and severely obstructed;

(2)There is a significant disparity in scale between different positions in the classroom, specifically manifested in the large pixel ratio difference between students in the front and back rows [2], as well as variations in camera angles and clarity in different classrooms;(3)In real classroom scenes, there is a certain similarity between different categories of student behaviors [3], such as reading and writing behaviors, or bowing heads and reading;(4)Public datasets for student behaviors in real classroom environments are scarce. Therefore, in order to improve the recall and accuracy of student behavior detection, researching how to detect small targets and address issues of obstruction and non-uniform scale in classroom teaching scenarios has become an urgent issue to be resolved in the field of educational informatization [4].

Among the existing algorithms for object detection, the YOLO algorithm achieves a commendable equilibrium between detection speed and accuracy [5]. YOLO, an end-to-end trained single-stage object detection model, was initially proposed by Joseph Redmon et al. [6] in 2015. YOLOv8, one of the widely used single-stage object detection algorithms, was open-sourced by Ultralytics in 2023. Despite its significant success across multiple domains, we have identified areas for improvement when it comes to the more complex task of detecting students' classroom behaviors. Furthermore, university classrooms, with their inclusion of electronic devices such as mobile phones, tablets, and laptops, create a more complex and variable environment compared to the relatively fixed and sparse student actions observed in primary and secondary school settings. The higher density of people and the greater complexity of movements in these classrooms pose substantial challenges to object detection algorithms. This paper introduces an improved algorithm using YOLOv8 as the baseline model for student classroom behavior detection, named CPID-YOLOv8. The main contributions of the study are as follows:

1. Incorporating the Channel Prior Convolutional Attention (CPCA) module into the Backbone, dynamically allocating attention weights across channel and spatial dimensions to strengthen the channel response to targets of various sizes and enhance detection capabilities at different scales.
2. Using the Parallelized Patch-Aware Attention (PPA) module to replace the C2f module in the Backbone. Utilizing the multi-branch feature extraction strategy and feature fusion attention of the PPA module further enhances the task-related multi-scale feature selection capability, allowing the model to extract more effective multi-scale feature information.
3. We propose the Dimension-Aware Integration Diffusion Feature Pyramid Network (DAIDFPN) module to replace the Neck layer part of the original YOLOv8 model for feature fusion. Experiments have proven that DAIDFPN can effectively enhance the adaptive selection and fine fusion of high-dimensional and low-dimensional features. Through its diffusion mechanism, it ensures that each scale feature is provided with detailed contextual information, which is more conducive to subsequent target detection and classification.

2. Related Works

2.1 Student Classroom Behavior Recognition

At present, many researchers are applying deep learning techniques to the field of student classroom behavior recognition, initiating a series of explorations. The existing algorithms for detecting student classroom behavior are mainly categorized into three types: one is based on video action recognition [7], such as Shou et al. [8] introduced a method for recognizing student classroom behaviors based on time-series classroom images. Their approach enhances the performance of the SlowFast [9] network in detecting student classroom behaviors by integrating an improved Asynchronous Interaction Aggregation network with multi-scale attention modules and incorporating an equalized focal loss function. Another is based on pose estimation algorithms for keypoint detection, such as Zhou et al. [10] utilized OpenPose to extract keypoint information from the human skeleton in images, connecting the keypoints to form a human

skeleton graph that is input into a 10-layer deep convolutional neural network (CNN-10) for identifying student behavior. The third is based on target detection algorithms for recognition, such as Wenchao et al. [11] added an FPN module to construct an improved RF-SSD detection model based on the SSD [12] algorithm, which increases image recognition efficiency and addresses the low efficiency of small target recognition. Wang et al. [13] proposed a system merging deformable DETR with Swin Transformer and light-weight Feature Pyramid Network (FPN) for student classroom behavior detection.

The above three types of algorithms have their respective advantages and disadvantages in detecting student classroom behavior. Algorithms based on video directly extract visual features from classroom videos, thus considering the dynamic characteristics of the spatial-temporal dimension, but require more complex network structures and higher computational complexity. For example, the SlowFast detection algorithm applied to the AVA dataset [14] requires significant manual annotation work. Algorithms based on pose estimation extract human skeleton keypoints and motion features for behavior identification, which perform well for significant behavioral and posture changes but may overlook minor behavioral changes and emotional expressions due to lower sensitivity to details. Additionally, they require substantial computational resources and are not suitable for scenarios with full occlusion or severe obstruction in classrooms. Currently, algorithms based on target detection are considered a promising solution due to their perfect speed-accuracy balance, for example YOLO technology [6], SSD technology [12], faster R-CNN technology [15], RT-DETR technology [16], etc.

2.2 Feature Fusion

When detecting small objects, multiple downsampling operations can lead to the loss of small object information in high-dimensional features, while low-dimensional features may not provide sufficient contextual information. To address this issue, traditional feature fusion modules often use a bottom-up approach to combine high-level semantic information with low-level semantic information, compensating for the lack of low-level detail in high-level features. However, these modules typically use fixed weights, considering only simple weighted averages or other fixed fusion strategies, failing to dynamically adjust channel selection and adaptively fuse high- and low-dimensional features based on the characteristics of the input information. Particularly, using upsampling and simple addition methods can result in the loss of spatial hierarchical information, thereby introducing errors. Additionally, there is a significant semantic gap between feature maps of different depths, which can lead to bottlenecks in information transfer between non-adjacent layers, lacking effective information fusion and thus affecting the network's performance in object detection tasks. To address these challenges, Tan et al. [17] proposed BiFPN, which enhances feature fusion efficiency by adding bidirectional pathways and learnable weights to the traditional FPN. Similarly, Xu et al. [18] introduced RepGFPN, which utilizes reparameterization techniques to improve the efficiency of feature fusion.

3. Methodology

To address the complexities of classroom scenarios and enhance the robustness of the object detection model, we propose a series of targeted improvements to further increase the recall and accuracy of YOLOv8 in detecting student classroom behaviors. CPID-YOLOV8 consists of three parts: backbone, Neck, and Head, as shown in Figure 1. To enrich feature representation, the CPCA attention mechanism is introduced in the backbone feature extraction part, allowing the model to focus more on small object features, as indicated by the green dashed lines in Figure 1, and detailed in Section 3.1. Section 3.2 describes how the PPA multi-branch extraction strategy is used to improve multi-scale feature extraction capabilities, represented by the blue dashed lines in Figure 1. Section 3.3 explains the proposed DAIDFPN method, which replaces the PAN-FPN structure in the Neck part to enhance feature fusion, as shown by the red dashed lines in Figure 1.

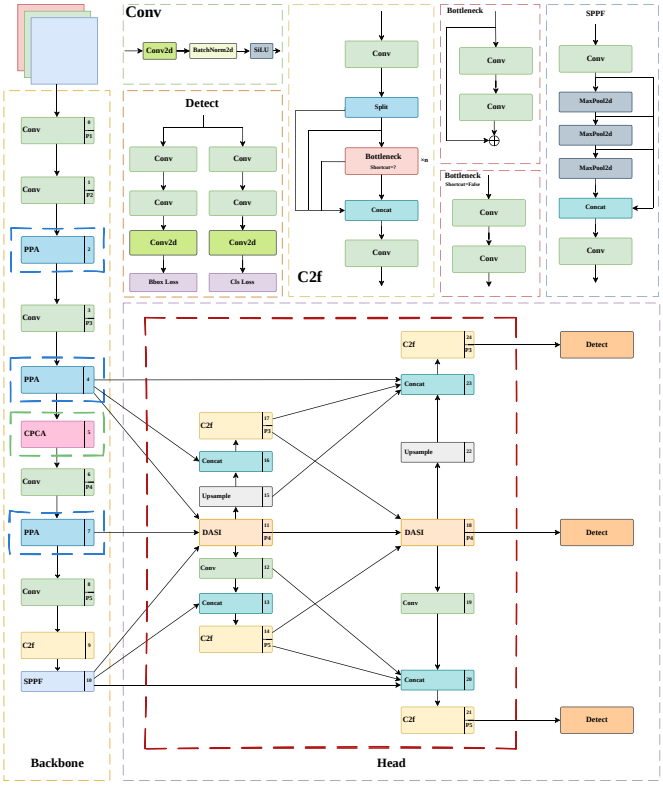


Figure 1 CPID-YOLOv8 structure diagram

3.1 Rich Feature Representation

The CPCA[19] is a lightweight and efficient channel-spatial attention mechanism. It constructs spatial attention using multi-scale depth-wise separable convolution modules [20], dynamically allocating attention weights in both channel and spatial dimensions, effectively extracting spatial relationships while preserving channel priors. This enables the network to more precisely focus on important information channels and regions. Experiments have shown that applying CPCA in computer vision tasks can effectively enhance the feature representation of small and densely occluded targets. Therefore, we plan to introduce CPCA into the YOLOv8 model's backbone component to enhance the model's feature extraction capability for student classroom behavior detection. To better adapt the model to the task of student classroom behavior detection, we conducted structural analysis and comparative experiments on our self-built dataset. By introducing CPCA into different layers of the model to improve performance and comparing the differences, the results of the structural analysis and comparative experiments are shown in

Table 1. Experimental results indicate that adding CPCA to the model's fifth layer yields the best results. The CPCA module sequentially performs channel attention and spatial attention, with its structure shown in Figure 2. The overall process of CPCA is as follows:

$$\begin{aligned} F_c &= CA(F) \otimes F, \\ \hat{F} &= SA(F_c) \otimes F_c. \end{aligned} \tag{1}$$

where $\otimes$ represents element-wise multiplication.

$$CA(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))), \tag{2}$$

where σ denotes the sigmoid function.

$$SA(F) = Conv_{1\times1}(\sum_{i=0}^{3} Branch_i(DWConv(F))), \tag{3}$$

where $DWConv$ represents depth-wise convolution and $Branch_i$, $i \in \{0,1,2,3\}$ represents the i-th branch. $Branch_i$ represents an identity join, which $Conv_{1\times1}$ represents a 1×1 convolutional channel mixing.

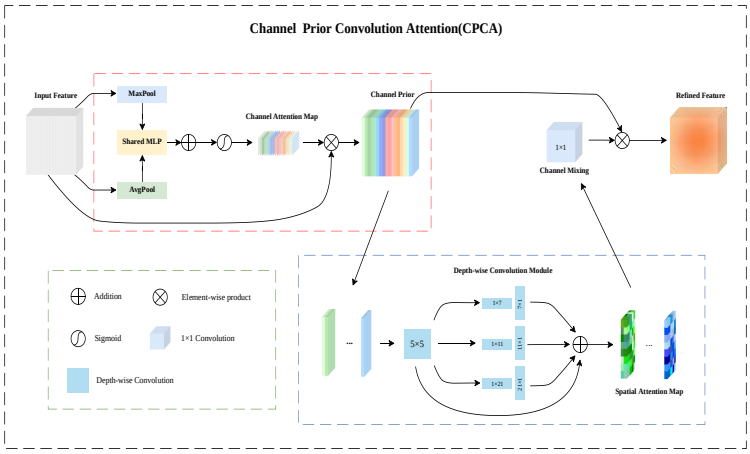


Figure 2 CPCA network structure

Table 1 Experimental results of different integration positions in CPCA
a: Indicates adding CPCA to the model's third layer. b: Indicates adding CPCA to the model's fifth layer. c: Indicates adding CPCA to the model's seventh layer. Experimental results show that adding CPCA to the model's fifth layer yields the best results.

Exp	CPCA location	P(%)	R(%)	mAP50(%)	mAP50-95(%)
1	YOLOv8n(baseline)	79.5	75.3	81.6	61.5
2	Backbone(a)	79.8	**76.6**	83.0	63.0
3	Backbone(b)	**81.2**	75.9	**83.2**	**63.3**
4	Backbone(c)	80.2	76.3	82.8	62.9

3.2 Enhancing Multi-Scale Feature Extraction Capabilities

In traditional small object detection tasks, key information can easily be lost after multiple downsampling operations [8]. The PPA [21] utilizes a multi-branch feature extraction strategy to effectively address this issue. The overall structure of the PPA is shown in Figure 3. The PPA module employs multiple parallel branches to process the input feature map, with each branch responsible for capturing features at different scales. This branching strategy allows the module to focus on both local details and global context simultaneously, thereby improving the identification and localization of small objects.

We replaced the C2f module in the Backbone part of YOLOv8 with the PPA's multi-branch extraction strategy, enabling the model to capture critical information at multiple scales during the feature extraction stage. Considering the model's parameters and computational complexity, structural analysis and comparative experiments have shown that replacing the C2f module in the original YOLOv8 model's 2nd, 4th, and 6th layers with PPA significantly enhances the model's feature extraction capabilities without compromising the balance between accuracy and speed. The results of the PPA structural analysis and comparative experiments are shown in Figure 4.

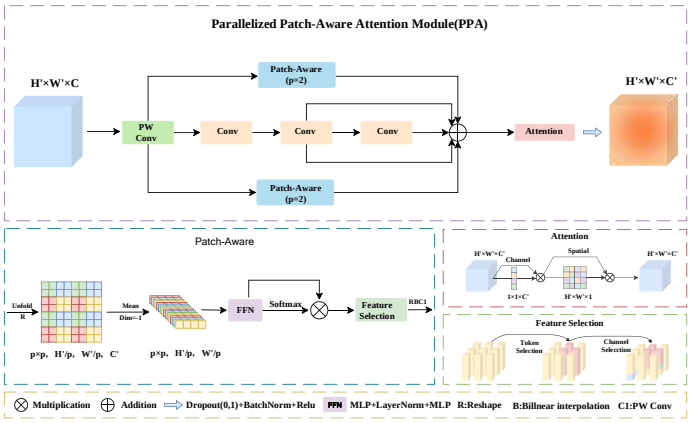
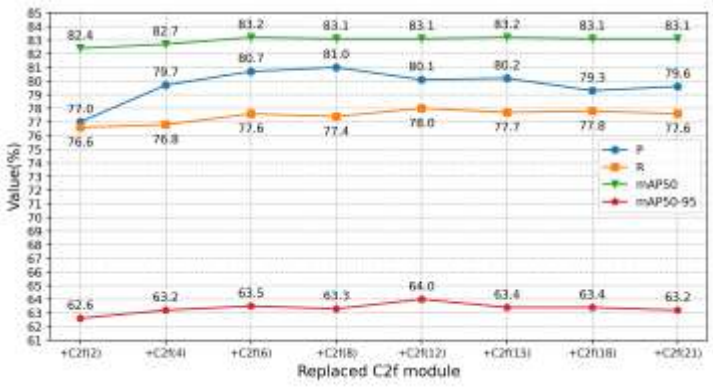


Figure 3 PPA network structure



The x-axis labels in the figure represent the sequential cumulative replacement of C2f modules with PPA modules. For example, '+C2f(a)' indicates that the C2f module in the original YOLOv8 model, corresponding to layer a, has been replaced by a PPA module.

Figure 4 PPA structural analysis experiments

3.3 Strengthening Feature Fusion

Inspired by the DASI [21] module, this paper proposes a feature fusion network called DAIDFPN. Compared to traditional fusion modules, DAIDFPN focuses more on feature quality and adaptability. It can dynamically adjust channel selection according to the characteristics of the input features, making it more flexible to adapt to different scenarios. DAIDFPN utilizes DASI to adaptively fuse features of different scales and then uses a diffusion mechanism to spread features with rich contextual information across various detection scales. The network structure of DAIDFPN is shown in Figure 5.

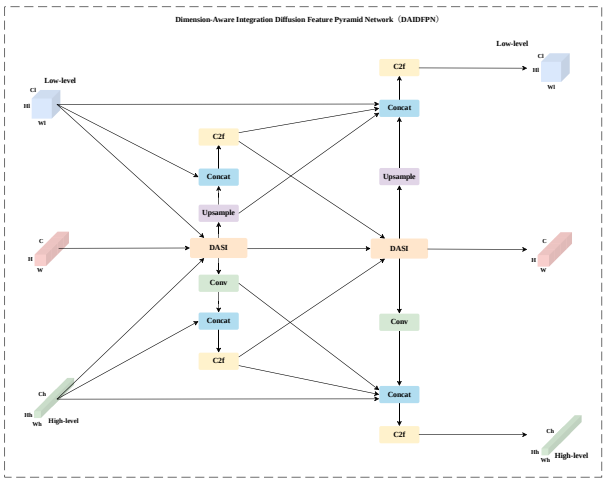


Figure 5 DAIDFPN structure diagram

The structure of the DASI module in DAIDFPN is shown in Figure 6. where h_i, l_i, u_i represent the partitioned features of the high-dimensional, low-dimensional, and current layer features, respectively. The partitions are calculated according to the following formula:

$$\alpha = sigmoid(u_i),\ \ u_i^{'} = \alpha l_i + (1-\alpha)h_i, \tag{4}$$

$$F_u^{'} = \left[u_1^{'}, u_2^{'}, u_3^{'}, u_4^{'}\right], \tag{5}$$
$$\hat{F}_u = \delta(\mathrm{B}(Conv(F_u^{'}))),$$

where $\alpha \in \mathrm{R}^{H\times W\times \frac{C}{4}}$ represents the value of u_i obtained after applying the sigmoid activation function, $u_i^{'}$ denotes the result of selective aggregation on each partition, and after merging along the channel dimension, we get $F_u^{'} \in \mathrm{R}^{H\times W\times C}$. Then, after convolution conv(), batch normalization $\mathrm{B}()$, and Rectified Linear Unit (ReLU) $\delta()$, the final output $\hat{F}_u \in \mathrm{R}^{H\times W\times C}$ is obtained.

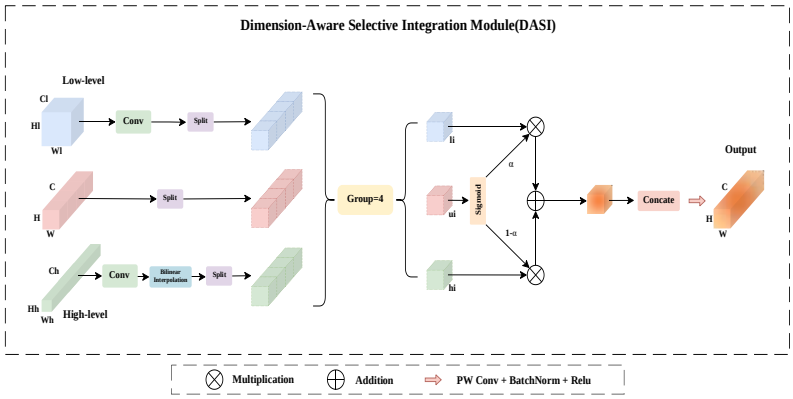


Figure 6 DASI network structure

Dimension-Aware Integrated Diffusion Mechanism in DAIDFPN. DAIDFPN receives three levels of dimensional information as input into the DASI module for adaptive feature fusion, a process we call dimension-aware integration. The fused features are then upsampled and downsampled once, respectively, and concatenated with the original features of the same dimensional level, resulting in a new set of high-level, mid-level, and low-level dimensional information. This process is referred to as dimension diffusion. This process is repeated once more: the newly obtained three levels of dimensional information are fed into the DASI module for further dimension-aware integration and then diffused through upsampling and downsampling. Finally, the original features of the same dimensional level, the features obtained after one round of dimension-aware integrated diffusion, and the features obtained after two rounds of dimension-aware integrated diffusion are concatenated to form the final output of the dimensional features at that level. Compared to traditional feature pyramid networks, which transmit feature information in a bottom-up manner, DAIDFPN reduces the continuous upsampling and downsampling process and fully leverages the DASI module's dimension-aware adaptive fusion characteristics to replace the simple addition of high- and low-dimensional information used in traditional methods. This approach results in less loss of critical information during hierarchical information transfer between non-adjacent levels, thereby effectively improving the performance of subsequent object detection and classification.

Specifically, in the CPID-YOLOv8 model, the DAIDFPN first receives multi-scale features ($F_{P_3} \in \mathrm{R}^{H_l\times W_l\times C_l}$, $F_{P_4} \in \mathrm{R}^{H\times W\times C}$, $F_{P_5} \in \mathrm{R}^{H_h\times W_h\times C_h}$) extracted from the Backbone. These three features of different dimensions are then fed into the DASI module for dimension-aware integration, resulting in the fused feature $\hat{F}_{P_4}^{'}$. Subsequently, $\hat{F}_{P_4}^{'}$ undergoes one up-sampling and one down-sampling, followed by concatenation with F_{P_3} and F_{P_5}, respectively. The concatenated features are then fed into the C2f module, yielding $\hat{F}_{P_3}^{'}$ and $\hat{F}_{P_5}^{'}$. We refer to this process as the dimension diffusion. This way, we obtain a new set of three features with different dimensions: $\hat{F}_{P_3}^{'}$, $\hat{F}_{P_4}^{'}$, and $\hat{F}_{P_5}^{'}$. These new features are then used as inputs to

repeat the process. Finally, the DAIDFPN outputs the results

$\hat{F}_{P_3}^{''} \in \mathrm{R}^{H_l \times W_l \times C_l^{'}}$, $\hat{F}_{P_4}^{''} \in \mathrm{R}^{H \times W \times C}$, and $\hat{F}_{P_5}^{''} \in \mathrm{R}^{H_h \times W_h \times C_h^{'}}$, which are fed into the Detect layer for subsequent object detection and classification tasks.

4. Experiments

4.1 Datasets

The data for this study's college student classroom behavior dataset were sourced from natural classroom videos at Shanghai University of Engineering Science. The dataset was constructed based on real university classroom scenarios, filmed using classroom surveillance cameras. To better adapt to different classroom settings, our dataset involves more than 10 different classrooms, each with varying camera angles, resolutions, and numbers of students. Initially, we obtained videos ranging from 10 to 20 minutes from each of the classrooms. Then, we randomly selected frames from different videos to create the final dataset images (with no repeated images). These extracted images were then input into the annotation tool LabelImg for data labeling. In this study, we categorized common classroom behaviors into eight categories: raising hand, reading, writing, using a mobile phone, looking down, lying on the desk, looking up, and standing. Ultimately, we obtained 1,140 annotated classroom behavior images, containing a total of 20,132 annotation labels. The classroom behavior dataset was divided into training and validation sets in an 8:2 ratio.

4.2 Experimental Environment and Evaluation indicators

4.2.1 Environment Settings

The experimental environment for model training in this study is as follows: the operating system version is Ubuntu 20.04, the CPU is Intel® Core™ i9-10900KF CPU @ 3.70GHz, the GPU is NVIDIA GeForce RTX 4000*2, the memory is 16G, the Python environment is Python 3.8.16, and the Pytorch framework is torch 1.13.1+cu117. The training parameters are uniformly set as follows: 300 epochs of iterative training, batch size set to 16, input image size of 640640, model weights optimized using the stochastic gradient descent (SGD) algorithm, learning rate set to 0.01, momentum parameter set to 0.937, optimizer weight decay set to 0.0005, and other parameters set to default values. For the sake of speed and accuracy, we uniformly chose YOLOv8n as the baseline version for our experiments.

4.2.2 Evaluation Indicators

In this study, precision (P), recall (R), mAP50, and mAP50-95 are used as evaluation indicators, with their respective formulas shown in Equations (6)-(9). Precision (P) refers to the proportion of correct detections among all detections. It indicates the accuracy of object detection. Recall (R) refers to the model's ability to correctly identify the actual objects present, reflecting the model's ability to recognize all object instances in the images. AP is the area under the precision-recall curve, used to measure the average precision of the model at different confidence thresholds. mAP is a key indicator for evaluating object detection performance, integrating the precision of the model across different categories. mAP50 represents the mean average precision at an IOU threshold of 0.5, while mAP50-95 represents the average of the mean average precisions calculated at different IOU thresholds (from 0.50 to 0.95, in increments of 0.05).

$$P = \frac{TP}{TP + FP} \tag{6}$$

where TP represents the number of positive samples predicted as positive samples, and FP represents the number of negative samples predicted as positive samples.

$$R = \frac{TP}{TP + FN} \tag{7}$$

where FN represents the number of positive samples predicted as negative samples.

$$AP = \int_0^1 P(R)d(R) \tag{8}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{9}$$

where n represents the number of classes, and AP_i denotes the AP value for the i-th target class.

4.3 Experimental Results and Analysis

4.3.1 Experimental Results of Improved CPID-Yolov8 Model

In this study, we conducted experiments using an improved CPID-YOLOv8 model on a self-constructed dataset of classroom behavior among

college students. After 300 iterations of training, we obtained the experimental results of the model training. Figure 7 shows the performance metrics of the CPID-YOLOv8 model on the training and validation sets across different numbers of training iterations.

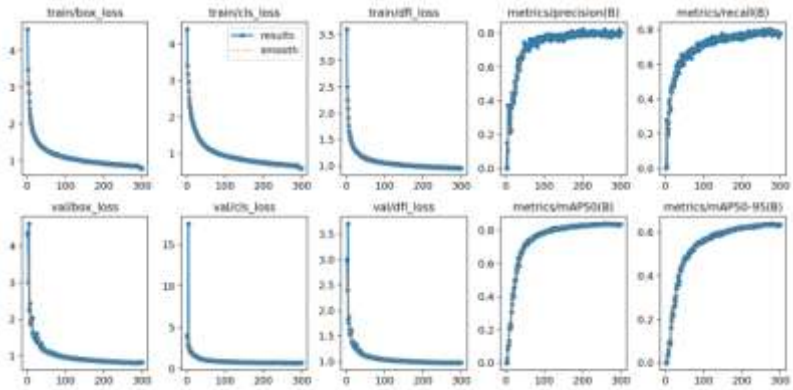


Figure 7 The iterative training results of the CPID-YOLOv8 model

The first three columns in the figure respectively show the CPID-YOLOv8 model's decreasing and stabilizing bounding box loss, classification loss, and objectness loss on the training and validation sets as the number of training iterations increases. The last two columns in the figure respectively illustrate the changes in the P and R curves of the model training iterations. Overall, the CPID-YOLOv8 model demonstrates good fitting performance, high stability, and accuracy in the task of detecting classroom behaviors among college students.

To more intuitively demonstrate the superiority of the CPID-YOLOv8 model, and for privacy considerations, we selected some images from the publicly available SCB-Datasets3 [3]. Figure 8 shows the results of classroom behavior detection performed by YOLOv8 and CPID-YOLOv8 on our selected test dataset. It can be observed that in scenes with dense, overlapping, and multi-scale targets in the classroom, the CPID-YOLOv8 model has higher precision and recall than our baseline YOLOv8 version. This proves that our proposed model can effectively improve the recall and accuracy of target detection, addressing the issues of missed and false detections in classroom behavior monitoring, and thus has practical application value.

Original Image

GroundTruth

YOLOv8-Processed Image

CPID-YOLOv8-Processed Image

Figure 8 Comparison of recognition cases between YOLOv8 and CPID-YOLOv8

4.3.2 Ablation Experiments

To more comprehensively validate the optimization effect of each improved module and further assess the contribution of the improved modules to the model, we conducted ablation experiments. The experimental results are shown in Table 2. The first row of the table represents the experimental results of the baseline YOLOv8 model in this study on the validation set of our self-constructed dataset. According to the ablation experiment results, introducing the CPCA and PPA modules effectively improves the model's multi-scale feature extraction capability. Additionally, replacing the original model's Neck module with The DAIDFPN module effectively reduces model parameters while further improving detection accuracy., indicating that this module effectively enhances the model's feature fusion capability. In summary, ablation experiments validate the effectiveness of our proposed methods.

Table 2 Results of ablation experiment

CPCA	PPA	DAIDFPN	Parameters	P(%)	R(%)	mAP50(%)	mAP50-95(%)
			3007208	79.5	75.3	81.6	61.5
✓			3020536	81.2	75.9	83.2	63.3
	✓		3169942	80.7	77.0	83.1	63.4
		✓	2775608	80.4	77.3	83.0	63.9
✓	✓		3183270	81.5	78.1	83.4	63.0
✓		✓	2788936	81.1	75.9	82.9	62.7
	✓	✓	2938342	81.6	76.4	83.5	63.8
✓	✓	✓	2951670	**81.2**	**77.4**	**83.8**	**64.0**

4.3.3 Comparative Experiments

To further evaluate the effectiveness of the model improvements, we compared CPID-YOLOv8 with several of the most popular current methods, using the same experimental environment for all comparisons. The comparative experiments were conducted using our self-constructed college student behavior detection dataset. The evaluation metrics included the number of model parameters (Parameters), precision (P), recall (R), mean Average Precision (mAP50) at an Intersection over Union (IOU) threshold of 0.5, and the mean Average Precision (mAP50-95) across different IOU thresholds. The experimental results of the models under different algorithms are shown in Table 3. The results indicate that our proposed CPID-YOLOv8 model achieves optimal performance with the same scale of parameters.

Table 3 Comparative experimental results of different models

Model	Parameters	P(%)	R(%)	mAP50(%)	mAP50-95(%)
YOLOv5_n	2504504	80.4	73.3	81.1	60.3
YOLOv6_n	4234536	78.0	74.6	81.1	61.2

YOLOv7-tiny	6033930	78.6	77.6	82.5	63.1
RT-DETR-L	28459700	80.3	**79.2**	79.6	64.0
YOLOv8n+CPCA+PPA+RepGFPN	3468502	80.3	76.4	82.1	62.1
YOLOv8n(baseline)	3007208	79.5	75.3	81.6	61.5
CPID_YOLOv8n	2951670	**81.2**	77.4	**83.8**	**64.0**

4.3.4 Comparison on Public Datasets

To validate the generalization capability of the proposed improved model CPID-YOLOv8, this paper conducts comparative experiments using the public datasets SCB-Datasets3 [3] and VisDrone [22]. The former dataset includes three types of student classroom behaviors: raising hands, reading, and writing. Although the latter dataset pertains to a different domain, its instances share characteristics such as dense scenes, small objects, and large multi-scale variations. Since the VisDrone dataset includes information beyond classroom scenarios, it can further verify the generalization capability of the CPID-YOLOv8 model in other fields. The experimental results are shown in the table 4-5. The results fully demonstrate that the improved model CPID-YOLOv8 proposed in this study has strong generalization capabilities and shows potential for application in other fields characterized by small objects, dense scenes, and multi-scale variations, not just limited to classroom scenarios.

Table 4 Experimental Results on the SCB-Datasets3 Dataset

Model	P(%)	R(%)	mAP50(%)	mAP50-95(%)
YOLOv8n(baseline)	66.4	65.8	70.5	52.0
CPID_YOLOv8n	**70.8**	**67.4**	**73.1**	**55.5**

Table 5 Experimental Results on the VisDrone Dataset

Model	P(%)	R(%)	mAP50(%)	mAP50-95(%)
YOLOv8n(baseline)	79.5	75.3	81.6	61.5
CPID_YOLOv8n	**81.2**	**77.4**	**83.8**	**64.0**

Conclusion

This paper proposes an improved YOLOv8 algorithm for student classroom behavior detection, named CPID-YOLOv8. The algorithm is based on YOLOv8, effectively addressing the issues of missed and false detections caused by small targets, occlusions, and multi-scale problems in classroom scenes. Experimental comparisons show that our proposed CPID-YOLOv8 algorithm improves the P-value, R-value, mAP50, and mAP50-95 by 1.7%, 2.1%, 2.2%, and 2.5%, respectively, on our self-constructed college classroom behavior dataset compared to the baseline YOLOv8 algorithm.

Our improvements focus on three main aspects. To address the occlusion problem of small targets in student behavior detection, we introduce the CPCA module into the Backbone part of the original model. This enhances the model's channel response to targets of different sizes, capturing finer, more global, and richer contextual information. For the multi-scale problem of student behavior in the back row classroom scenes, we replace the C2f module with the PPA module. By using a multi-branch feature extraction strategy and feature fusion attention, the model's multi-scale feature selection capability is strengthened, allowing it to extract more effective multi-scale feature information. Additionally, we propose a feature fusion network, DAIDFPN, to replace the original model's Neck layer. This effectively enhances the model's adaptive selection and fine fusion of high-dimensional and low-dimensional features. Through the diffusion mechanism of the DAIDFPN module, each scale feature possesses detailed contextual information while reducing the model parameters and improving detection accuracy.

In future work, we will continue to reduce network parameters and model computation while ensuring detection performance. We will also further investigate behavior detection in situations

where students in the back row are occluded, continuously improving and optimizing the CPID-YOLOv8 model architecture, and actively promoting the practical application of the CPID-YOLOv8 model in intelligent information-based education.

Reference

1. Wu, B., Wang, C., Huang, W., Huang, D., Peng, H. Recognition of student classroom behaviors based on moving target detection. Traitement du Signal, 2021, 38(1), 215–220.
2. Yang, F., Wang, T., Wang, X. Student Classroom Behavior Detection Based on YOLOv7+ BRA and Multi-model Fusion. International Conference on Image and Graphics, Cham: Springer Nature Switzerland, 2023, pp. 41-52.
3. Yang, F., Wang, T. SCB-Dataset3: A Benchmark for Detecting Student Classroom Behavior. arXiv preprint arXiv:2310.02522, 2023.
4. Liu, Q., et al. YOLOv8n_BT: Research on Classroom Learning Behavior Recognition Algorithm Based on Improved YOLOv8n. IEEE Access, 2024, 12, 36391-36403. doi: 10.1109/ACCESS.2024.3373536.
5. Vijayakumar, A., Vairavasundaram, S. YOLO-based Object Detection Models: A Review and its Applications. Multimed Tools Appl, 2024. https://doi.org/10.1007/s11042-024-18872-y
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788. doi: 10.1109/CVPR.2016.91.
7. Zhu, Y., et al. A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567, 2020.
8. Shou, Z., Yan, M., Wen, H., Liu, J., Mo, J., Zhang, H. Research on Students' Action Behavior Recognition Method Based on Classroom Time-Series Images. Appl. Sci., 2023, 13, 10426. https://doi.org/10.3390/app131810426
9. Feichtenhofer, C., Fan, H., Malik, J., He, K. SlowFast Networks for Video Recognition. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 6201-6210. doi: 10.1109/ICCV.2019.00630.
10. Zhou, J., Ran, F., Li, G., Peng, J., Li, K., Wang, Z. Classroom Learning Status Assessment Based on Deep Learning. Mathematical Problems in Engineering, vol. 2022, Article ID 7049458, 9 pages, 2022. https://doi.org/10.1155/2022/7049458
11. Wenchao, L., Meng, H., Yuping, Z., Shuai, L. Research on Intelligent Recognition Algorithm of College Students' Classroom Behavior Based on Improved SSD. 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI), Beijing, China, 2022, pp. 160-164. doi: 10.1109/CCAI55564.2022.9807756.
12. Liu, W., et al. SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-46448-0_2
13. 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, 2023, 11, 372. https://doi.org/10.3390/systems11070372
14. Gu, C., et al. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6047-6056. doi: 10.1109/CVPR.2018.00633.
15. Ren, S., He, K., Girshick, R., Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv 2015, arXiv: 1506.01497.
16. Lv, W., et al. Detrs beat yolos on real-time object detection. Comput. Vis. Pattern Recognit., 2023. arXiv:abs/2304.08069.
17. Tan, M., Pang, R., Le, Q. V. EfficientDet: Scalable and Efficient Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10781-10790. doi: 10.1109/CVPR42600.2020.01079.
18. Xu, X., Jiang, Y., Chen, W., Huang, Y., Zhang, Y., Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. arXiv preprint arXiv:2211.15444, 2022. doi: 10.48550/arXiv.2211.15444.
19. Huang, H., et al. Channel prior convolutional attention for medical image segmentation.

arXiv preprint arXiv:2306.05196, 2023.
20. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1251-1258. doi: 10.1109/CVPR.2017.195.
21. Xu, S., et al. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. arXiv preprint arXiv:2403.10778, 2024.
22. Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//ICCV 2019: IEEE/CVF International Conference on Computer Vision. 2019: 0-0.