scientific reports

Check for updates

OPEN

Real-time classroom student behavior detection based on improved YOLOv8s

Xiaojing Sheng[1,2]✉, Suqiang Li[3,5] & Sixian Chan[4,5]

The learning capacity of students is significantly influenced by the quality of instruction they receive in the classroom. With the rapid advancement of behavior detection technology, identifying classroom behaviors of students is becoming increasingly common in educational settings. However, the field still faces specific challenges, primarily concerning the accuracy of identifying student behaviors within complex and variable classroom environments, as well as the real-time capabilities of detection algorithms. To address these challenges, we propose an efficient and straightforward algorithm based on the YOLO architecture. A Multi-scale Large Kernel Convolution Module (MLKCM) has been designed to capture feature information across various dimensions through multi-axis pooling, achieving adaptive receptive fields and effectively capturing multi-scale features. This design enhances the network's sensitivity to feature information by incorporating convolution kernels of varying sizes. Subsequently, we introduce a Progressive Feature Optimization Module (PFOM) to segment the channel dimension of the input feature map. This module integrates feature refinement blocks progressively, which not only preserve the refined features but also efficiently aggregate both local and global information. Finally, we conducted comprehensive experiments using the SCB-Dataset3-S and SCB-Dataset3-U datasets. The results demonstrated mean Average Precision (mAP) values of 76.5% and 95.0%, respectively, surpassing other commonly used detection techniques. Additionally, the effectiveness of our approach was validated through ablation studies and visualization of the detection outcomes.

The convergence of advancements in computer vision and artificial intelligence with the enhancement of educational resources has catalyzed the innovative growth of educational systems. A novel way of tracking the actions of students in the classroom is through classroom behavior detection technology, which is essential for enhancing the effectiveness and management of teaching[1–6]. Conventional approaches, largely based on manual surveillance via video cameras, suffer from issues such as visual fatigue, inefficiency, and compromised real-time performance[7,8]. In contrast, the integration of computer vision-based object identification into educational settings has demonstrated the ability to transcend these limitations. By leveraging the automatic extraction of data features, this technology facilitates real-time detection of classroom behaviors, thereby enhancing the quality of instruction and the learning process for students[9–14]. Nonetheless, the technology's broader implementation is hindered by several critical challenges: the homogeneity in student postures and behaviors, the prevalence of occlusions in crowded classrooms, and the dynamic changes in the background environment[15–19].

The widely used object detection algorithms are categorized into one-stage and two-stage methods. Common one-stage algorithms include SSD[20], RetinaNet[21], and the YOLO series[22–30]. In contrast, two-stage methods, such as Fast R-CNN[31] and Faster R-CNN[32], are better suited for applications demanding higher detection accuracy, whereas one-stage models are preferable for real-time, high-precision industrial scenarios. To Find a balance between speed and accuracy is crucial for classroom behavior detection. In order to solve the challenges of small scale and dense distribution faced by student classroom behavior detection. Wang et al.[18] proposed SLBDetection-Net, which focuses on accurately capturing the representation of learning behavior, with special emphasis on multi-scale focusing on key information. This method significantly improves the detection accuracy and performance of the model by designing a learning behavior-aware attention mechanism, extracting key features of learning behavior and capturing complex features of targets at different scales. However, the model has high computational complexity and parameter count, which makes it difficult to deploy and apply the model. Zhang

[1]College of Teacher Education, Quzhou University, Quzhou 324099, China. [2]College of Education, Zhejiang Normal University, Jinhua 321001, China. [3]School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China. [4]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China. [5]Sixian Chan and Suqiang Li: These authors contributed equally to this work. ✉email: sxj8816@126.com

et al.[33] addressed this need by introducing attention techniques to increase feature information inside effective regions and altering the IoU and loss functions to create a lightweight YOLO model with enhanced detection capacity. Although this method achieved good results on the test dataset, it did not test the generalization performance of the model on other datasets. Similarly, Wang et al.[34] optimized the model by combining attention detection techniques with image input optimization, which improved the accuracy of detecting student learning behaviors. This method only performs a simple analysis of the model on a single dataset and does not compare it with mainstream methods, which cannot well reflect the potential of the model in practical applications. Zhao et al. developed CBPH-Net[35] to address problems like occlusion, posture variation, and uneven object scale. It combines a robust feature extraction module to enhance small object detection across scales with PANet and coordinated attention to reduce background noise. This method uses the neck network of the model to detect images of various scales of the model, while our method uses large kernel convolution to enrich the key features of classroom students' learning behavior and capture features of different scales. In response to the problems of complex classroom environment and high similarity between teaching behavior classes, Ma et al.[36] proposed an improved YOLOv7[27] method for large target classroom behavior recognition in smart classroom scenarios. This method optimizes model performance by introducing feature layers and adding attention mechanisms. Jia et al.[37] enhanced the feature extraction capability through the attention mechanism and improved the accuracy of student posture recognition by optimizing the feature map. Wang et al.[38] in order to solve the major challenges faced by existing object detection models, such as occlusion, blur, and scale differences, and the dynamic and complex classroom environment exacerbated these challenges. By introducing a multi-scale deformable transformer for student learning behavior detection, and using large convolution kernels for upstream feature extraction and multi-scale feature fusion, the detection capability of multi-scale and occluded objects was significantly improved, providing a powerful solution for analyzing student behavior. However, the size of the model parameters still needs to be optimized and reduced to achieve lightweight model.

Despite these successes, current methods often struggle to achieve a balance between recognition accuracy and computational efficiency. To address these challenges, our research proposes the Multi-scale Large Kernel Convolution Module (MLKCM) and the Progressive Feature Optimization Module (PFOM) for classroom behavior detection. MLKCM enhances the network's perception of multidimensional feature information, achieves an adaptive receptive field, and effectively captures multi-scale features by combining parallel convolution blocks with multi-axis pooling. To facilitate the efficient capture of both local and global data, PFOM partitions the channel dimension of input feature maps and progressively introduces feature refinement blocks, preserving refined features.

We conducted extensive experiments on SCB-Dataset3-S and SCB-Dataset3-U, which validated the proposed method's effectiveness in classroom behavior detection. The primary contributions of this paper are summarized as follows:

- We propose the MLKCM, which combines parallel convolution blocks with multi-axis pooling to achieve an adaptive receptive field, effectively capturing multi-scale features and enhancing the network's sensitivity to feature information across various dimensions.
- We design a PFOM, which partitions the input feature map along the channel dimension and incrementally integrates feature refinement blocks to retain refined features, thereby effectively aggregating both local and global information.
- Extensive experiments conducted on the SCB-Dataset3-S and SCB-Dataset3-U datasets demonstrate the superior performance of our method, highlighting its potential for future applications in data-driven education.

The overall structure of the article is as follows: The " Related work" section introduces the research on student classroom behavior detection using large kernel convolution blocks and progressive feature optimization modules. The detailed explanation of the large kernel convolution module and progressive feature optimization module designed in this article can be found in the " Methodology" section. The " Experiments and analysis" section includes the experimental results and visual analysis of student classroom behavior detection. The methods proposed in this study are summarized in the " Conclusion" section. The " Discussion" section discussed the limitations of this work and future research directions.

Related work

Large kernel convolution block

With its wide receptive field, the large kernel convolution block is useful for building attention-based feature maps, assessing local contextual information, and using the benefits of both convolutional and attention mechanisms[39]. AlexNet[40] uses multiple convolutional kernels of sizes 5×5 and 11×11 to capture diverse-scale feature information within images. This allows the network to capture a wide range of spatial features and further refine these features in conjunction with smaller convolution kernels. The GoogleNet Inception module[41] introduced a fusion mechanism using multiple concurrent convolutional kernels of different sizes, enhancing feature extraction through multi-scale processing. SPPNet[42] combines large convolution kernels with multi-scale pooling layers, which led to further advancements in feature fusion methods. Building on this, ASPP in DeepLab[43] and the trident structure in Trident Networks[44] apply progressively larger, dilated convolutions to capture detailed features across varying scales effectively. SegNeXt[45] utilizes large kernel convolutions in parallel specifically for visual tasks. For applications that need both small parameter sizes and vast coverage, the deformable convolutional network[46] employs a versatile strategy that strikes a compromise between a large receptive field and lower computing costs and parameters. RepLKNet[47] demonstrated superior performance in object recognition and semantic segmentation by using numerous ultra-large convolutional kernels, outperforming traditional models based on smaller kernels. To address the parameter complexity in

dense convolutional networks while preserving the advantages of large kernel operations, PeLK[48] proposed a new parameter-efficient network architecture for large kernel convolution neural networks. The development of large kernel convolution blocks has significantly advanced a range of tasks, providing a robust foundation for further research in this area.

Progressive refinement module

A well-designed feature extraction architecture is crucial for enhancing the depth and performance of computer vision-based detection tasks because it can learn multi-scale feature representations efficiently and has excellent feature extraction capabilities. In order to develop deeper networks and greatly enhance the feature learning capability. ResNet[49] built residual blocks that can achieve deeper networks and significantly improve the feature learning ability. Wang et al.[28] proposed the reparameterized heterogeneous efficient layer aggregation module. Both the implementation architecture and the convolution design include the utilization of heterogeneous large-scale convolution kernels, which can effectively learn expressive multi-scale feature representations. Ding et al.[50] developed the inception block to improve the multi-branch structure of the convolution block. The performance of the model is greatly enhanced by the average pooling, multi-scale convolution, etc., included in each branch structure. Zhang et al.[51] created the attention mechanism residual block by fusing the advantages of the attention mechanism and the residual block. This allowed the model to concentrate on useful feature information more efficiently, ignore background information, and significantly improve model detection performance. Xue et al.[52] proposed an innovative hierarchical residual framework with an attention mechanism. By adding the attention mechanism to the residual block, the framework's learning and feature representation abilities can be strengthened, enabling it to more precisely recognize and capture defective features. Liu et al.[53] introduced residual blocks with coordinated attention. In summary, feature aggregation through residual blocks has proven highly effective across deep learning models, driving significant improvements in both the precision of detection systems.

Methodology

Overview

Our algorithm, designed as a one-stage detector, comprises three main components: the backbone network, neck structure, and detection head, as illustrated in Fig. 1. The backbone consists of four stages, labeled L1, L2, L3, and L4. At the end of each stage, a PFO module is applied to modulate features and obtain fine-grained information. The neck structure, PAFPN, is tasked with extracting multi-scale features from the feature extraction network, performing preliminary fusion in the shallow layers, and guiding the detection head to obtain diverse output information. Lastly, the detection head analyzes feature representations at various scales to generate prediction results and compute the associated losses.

Multi-scale large kernel convolution module

The design of large kernel convolution kernels has become an excellent method for convolutional neural networks. However, simply increasing the convolution kernel will increase the complexity of the model. As seen in Fig. 2, we proposed a multi-scale large kernel convolution module (MLKCM). The MLKCM serves to enhance the network's ability to capture and process feature information across multiple scales. By integrating both the DCP and MAB components, MLKCM effectively balances the benefits of large kernel convolutions without overly increasing model complexity. The DCP component captures multi-scale dependencies by using large kernels to relate feature information across different scales, while the MAB component recalibrates the extracted features to refine and improve their representational quality.

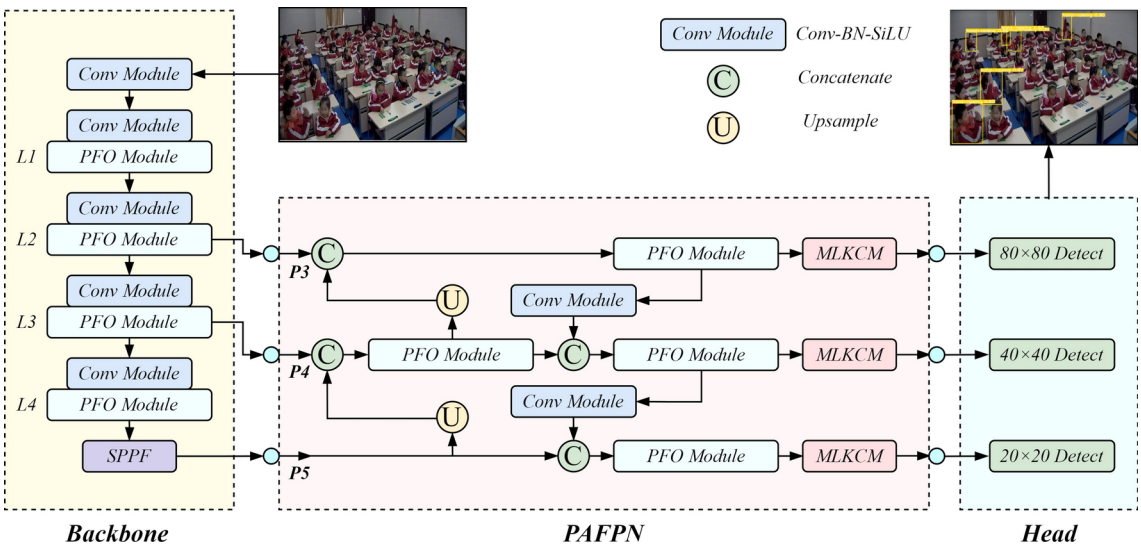

Figure 1. The details of improved framework.

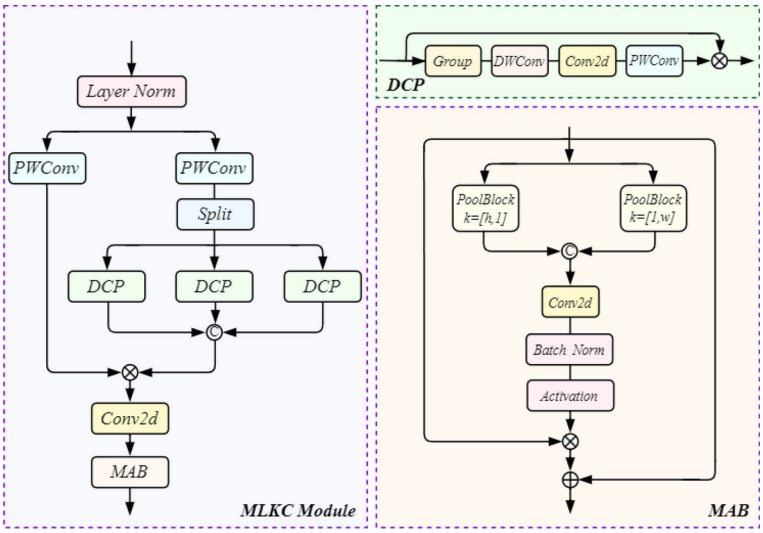

Figure 2. Structure of multi-scale large kernel convolution module.

To maintain feature details and speed up convergence, apply layer normalization in input feature *X*. Then, the feature X_{LN} is divided into *n* groups. For the *i*-th group of features X_i, Parallel large-kernel convolution with the aim to generate an adaptable receptive field and efficiently collect both local and global information. Taking model complexity into account, the convolution kernel sizes of the three parallel DCP blocks we employ are 3, 5, and 7. The formula is given below:

$$X_{LN} = LN(X) \tag{1}$$

$$X_i = Split(F_{PW}(X_{LN})) \tag{2}$$

$$X_{DCP_i} = X_i \otimes F_{PW}(F_{DWD}(F_{DW}(X_i))) \tag{3}$$

$$X_{Cat} = Cat(X_{DCP_i}) \tag{4}$$

$$X_{PC} = F_{Conv2d}(X_{Cat} \otimes F_{PW}(X_{LN})) \tag{5}$$

where $LN(\cdot)$ denotes layer normalization, $Split(\cdot)$ shows that feature *X* is divided in the channel dimension, $Cat(\cdot)$ indicates feature concatenation, $F_{DW}(\cdot)$ represents depth-wise convolution, $F_{DWD}(\cdot)$ stands for depth-wise with d-dilation convolution, $F_{PW}(\cdot)$ represents and a point-wise convolution.

Secondly, the concatenated features X_{PC} are subjected to maxpool and avgpool in the H×W dimension respectively, and the features in the dimension are concatenated to obtain X_{hw}, which is point-multiplied with the input features and then added after convolution layer, *BN* and activation function.

Finally, the layer-normalized feature X_{LN} is multiplied by X_{Cat} after $F_{PW}(\cdot)$ adjusting channels. The expression is:

$$X_{hw} = Cat(Maxpool(F_{Conv2d}(X_{PC})), Avgpool(F_{Conv2d}(X_{PC}))) \tag{6}$$

$$X_{MLKCM} = X_{PC} \oplus (\sigma(F_{Conv}(BN(X_{hw}))) \otimes X_{PC}) \tag{7}$$

where $\sigma(\cdot)$ represents Sigmoid function, $\oplus$ is element-wise summation, $\otimes$ stands for element-wise multiplication.

Progressive feature optimization module

To enhance the effectiveness of feature extraction, we designed the progressive feature optimization module (PFOM), as shown in Fig. 3. PFOM consists of the ghost module and the repeat block. The process begins with channel segmentation, where part of the initial feature information is preserved and treated as refined features. The remaining feature subset undergoes further optimization through a repeatable lightweight module across multiple refinement steps. Finally, the refined features from each step are concatenated. The complete process for the input feature *Y* is described by the following formula:

$$Y_1, Y_2 = Split(Y) \tag{8}$$

$$Y_{GC} = F_{Conv}(F_{Conv}(Y_i) \otimes F_{GConv}(F_{DW}(F_{GConv}(F_{Conv}(Y_i))))) \tag{9}$$

$$Y_{RB} = Cat(F_{Conv}, N \times Y_{GC}(F_{Conv}(Y_i))) \tag{10}$$

$$Y_3 = F_{Conv}(Y_{RB}(Y_2)) \tag{11}$$

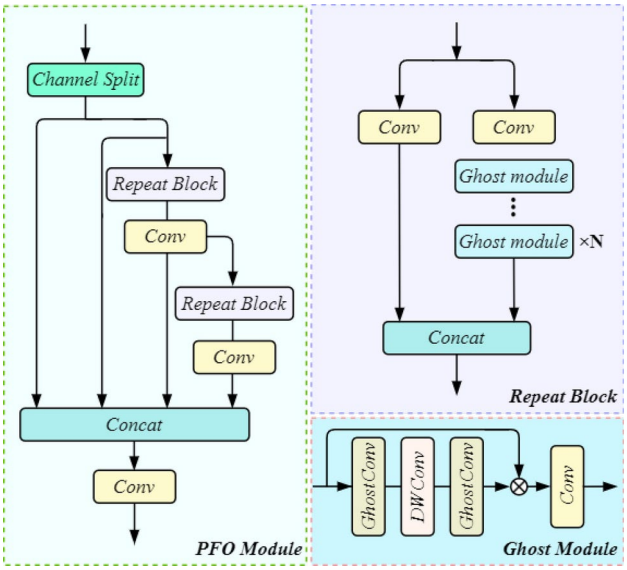

Figure 3. Structure of progressive feature optimization module.

$$Y_4 = F_{Conv}\left(Y_{RB}\left(Y_3\right)\right) \tag{12}$$

$$Y_{PFOM} = Cat\left(Y_1, Y_2, Y_3, Y_4\right) \tag{13}$$

where Y_{RB} is repeat block, Y_{GC} stands for ghost module and Y_{PFOM} represents the refined feature.

Detection head and loss function

The detection part of YOLOv8 adopts an anchor-free decoupled head structure, using two independent branches for object classification and prediction box regression, using binary cross entropy loss (BCE Loss) and distribution focus loss (DFL [54]) and CIoU [55] respectively. This detection structure can improve detection accuracy and accelerate model convergence. YOLOv8 also adopts the Task-Aligned Assigner strategy[56], which selects positive samples according to the weighted scores of classification and regression to achieve dynamic sample allocation.

Experiments and analysis

Experimental environment

To validate the performance of the method, experiments were performed on a Windows 10 system with an Intel XeonGold 6130 CPU and a NVIDIA GeForce RTX3090 24GB GPU. The deep learning framework is PyTorch 2.0.1, CUDA 11.8 and cuDNN 8.6.0. The model is optimized using the SGD optimizer, and the momentum size is set to 0.937. The batch size is set to 8, the initial learning rate is set to 0.01, the training epochs are 300, and the input image size is 640 × 640. To prevent overfitting, we used an early stopping value, which was to stop the model after 50 epochs. In order to ensure the fairness and reliability of the comparative experiments, all models are trained without using pre-training weights.

Datasets

The research methodology presented in this work is assessed utilizing the SCB-Dataset3-S and SCB-Dataset3-U datasets from classroom scenes on a public student behavior detection dataset. SCB-Dataset3-S includes three categories: hand-raising, reading, and writing, with 5015 images and 25810 annotations. Many examples of student conduct in the classroom are depicted in Fig. 4. We assess model performance by splitting the dataset into training and validation sets with a 4:1 ratio. Additionally, we validate our approach on SCB-Dataset3-U to ensure the framework's robustness and consistency across different datasets.

Evaluation metric

The evaluation of the experimental results is based primarily on two metrics: precision and recall, which are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

Figure 4. Example images of the classroom behavior dataset.

where *TP* (True Positives) represents correctly identified positive instances, and *FP* (False Positives) denotes incorrectly identified instances as positive. Similarly, *FN* (False Negatives) corresponds to positive instances that were incorrectly missed. To further assess the method, we incorporate average precision (*AP*) and mean average precision (*mAP*) metrics. Here, *AP* represents the average precision score within a single category, while *mAP* is the mean *AP* across all categories:

$$AP_i = \int_0^1 P(r)\, dr \tag{16}$$

$$mAP = \frac{1}{n} \sum_{i}^{n} AP_i \tag{17}$$

where *i* represents the class, *AP* is the average precision value of all recall, and *mAP* is the average *AP* value of different categories. These metrics provide a comprehensive assessment of both per-category and overall model performance.

Performance evaluation comparison

The proposed algorithm's model performance is compared to mainstream detection methods over the past several years. The SCB-Dataset3-S and SCB-Dataset3-U comparison results are displayed in Tables 1 and 2. The benchmark models encompass ATSS[57], TOOD[56], FoveaBox[58], YOLOX[59], Faster rcnn[32], RetinaNet[21] and YOLO series models. Obviously, the proposed algorithm performs well, achieving the best mAP on the dataset SCB-Dataset3-S, which is 1.9% higher than YOLOv9-s[28], and has exceptional performance in AP of hand raising, reading and writing, reaching 84.9%, 77.8% and 67.0% respectively. To assess the generalization capabilities of our method, additional experiments were conducted on SCB-Dataset3-U. Our method achieved the best performance in mAP, which surpasses the second-ranked model by 0.4%, and also achieved excellent performance in model size and FPS on SCB-Dataset3-U.

However, the approach suggested in this study is comparatively substandard in terms of parameters, computation, and FPS on the SCB-datasets-U dataset when compared to YOLOv8-s, YOLOv9-s and YOLOv11-s. We will concentrate on making the model lighter and enhancing its real-time performance in the future. Table 3 indicates that our model works second best on SCB-Dataset3-S dataset, and our model has fewer parameters than FA-YOLOv9[60]. Our model obtains the best performance on SCB-Dataset3-U dataset. Nevertheless, compared with the compared models, the number of parameters of the model does not reach the optimal effect. Although the number of parameters is slightly higher, the model has stronger generalization ability and adaptability in various scenarios. These results underscore the potential of our method for detecting student behavior in complex real-world applications. In order to explore the impact of the model input size on the model performance, as shown in Table 5. As the input size increases, the mAP increases. Smaller input sizes are limited in spatial information. Due to the low resolution, the detection accuracy of the model will be limited, and the subtle behaviors of students in complex scenes cannot be fully captured. Larger input sizes allow the model to capture more detailed information, thereby improving the accuracy of detecting objects and identifying specific actions. Although the model has the ability to detect at multiple scales and can handle small, medium, and large objects, when the input image is small, the resolution of the image is reduced, making it impossible for the model to obtain enough details when extracting features. This will have a negative impact on the model's ability to detect small or inconspicuous actions, such as sudden movements or other abnormal behaviors of students in the classroom.

Method	mAP (%)	AP (%)		
		Hand-raising	Reading	Writing
ATSS[57]	47.4	56.3	50.2	35.6
TOOD[56]	56.2	71.6	60.4	36.5
YOLOX[59]	73.1	81.4	72.8	65.3
FoveaBox[58]	57.9	72.1	56.3	45.2
Faster rcnn[32]	41.1	49.1	39.2	35.0
RetinaNet[21]	35.1	42.8	37.1	25.4
YOLOv3-tiny[24]	65.9	73.0	68.8	55.9
YOLOV5-s	73.0	82.9	75.0	61.1
YOLOV6-s[26]	73.5	83.0	74.9	62.7
YOLOv7-tiny[27]	71.9	78.5	69.4	**67.8**
YOLOX-s[59]	69.0	78.1	69.4	59.3
YOLOV9-s[28]	74.6	84.1	76.3	63.5
YOLOV10-s[29]	73.5	81.8	75.2	63.5
YOLOv11-s[30]	75.9	**85.2**	76.5	65.8
Ours	**76.5**	84.9	**77.8**	67.0

Table 1. Algorithm performance assessment on SCB-Dataset3-S. Best results are marked in **bold**. Significant values are given in bold.

Method	mAP (%)	Params (M)	GFLOPs	FPS
ATSS[57]	48.5	32.1	48.3	22
Faster rcnn[32]	45.7	41.4	60.1	23
RetinaNet[21]	58.9	36.4	49.5	25
TOOD[56]	79.5	32.0	47.4	15
YOLOv3[24]	94.8	103.7	282.2	79
YOLOv3-tiny[24]	93.2	12.1	18.9	169
YOLOV3-spp[42]	94.6	104.7	283.1	73
YOLOV5-s	94.4	9.1	23.8	170
YOLOV6-s[26]	94.5	16.3	44.0	109
YOLOv7-tiny[27]	78.0	6.0	13.1	74
YOLOX-s[59]	61.5	8.9	26.8	56
YOLOv8-s	94.3	11.1	28.4	170
YOLOv9-s[28]	94.4	7.2	26.7	154
YOLOv10-s[29]	93.5	8.0	24.5	323
YOLOv11-s[30]	94.7	9.4	21.3	217
Ours	**95.0**	22.3	64.5	70

Table 2. Algorithm performance assessment on SCB-Dataset3-U. Significant values are given in bold.

Dataset	Mothod	mAP@0.5	mAP@0.5-0.95	Params (M)
SCB-Dataset3-S	TLB-YOLO[61]	68.1	44.2	3.8
	CSB-YOLO[62]	71.1	52.3	0.7
	YOLOv8-BSAM[63]	69.0	–	–
	FA-YOLOv9[60]	77.8	60.8	25.4
	Ours	76.5	59.7	22.3
SCB-Dataset3-U	CSB-YOLO[62]	74.7	57.6	1.86
	SBD-Net[19]	74.5	57.7	36.5
	Ours	95.0	79.1	22.3

Table 3. Comparative analysis with other methods.

MLKCM	PFOM	mAP (%)	AP (%)		
			Hand-raising	Reading	Writing
		74.0	83.6	75.6	62.7
	✓	75.4	84.5	76.5	65.3
✓		76.1	84.5	**77.9**	65.9
✓	✓	**76.5**	**84.9**	77.8	**67.0**

Table 4. Ablation experiment results. Significant values are given in bold.

Input size	mAP (%)	AP (%)		
		Hand-raising	Reading	Writing
128 × 128	49.5	55.1	52.2	41.2
256 × 256	63.6	74.3	64.4	52.1
512 × 512	73.0	84.1	72.7	62.1
640 × 640	76.5	84.9	77.8	67.0

Table 5. The impact of model input size on model performance.

Ablation experiment analysis

This study employs YOLOv8-s as the benchmark model to rigorously evaluate the efficacy and performance enhancement introduced by our proposed methodologies. To thoroughly assess the impact and performance of our proposed methods.

This paper conducts an ablation experiment on the SCB-Dataset3-S dataset. The results as detailed in Table 4, demonstrate that the integration of MLKCM yields a significant improvement in the mAP by 2.1% over the baseline. Furthermore, the AP of hand-raising, reading and writing are increased by 0.9%, 2.3% and 3.2% respectively. The MLKCM's capability to efficiently preserve multi-scale feature information, maintain the receptive field, and enhance the multi-axis pooling's sensitivity to perceive features of varying dimensions is evident. Upon the incorporation of PFOM, as shown in Table 5, the detection performance across all categories on the SCB-Dataset3-S dataset is further enhanced, with the mAP increasing by 1.4%. PFOM achieves this by partitioning the dimensions, and preserving the original feature information in one segment while optimizing the features in another through a repeatable, lightweight module. This approach effectively captures both local and global information, contributing to the overall enhancement of detection accuracy.

Visualization analysis

This study presents the comparative detection results of the proposed framework and the baseline on SCB-Dataset3-S dataset, as illustrated in Fig. 5, which further demonstrates the performance of our methodology. The predicted results clearly indicate the superior performance of our framework. It exhibits enhanced object placement capabilities and achieves higher prediction accuracy in comparison to the baseline. In a complex classroom scenario, we visualized each single category. For the hand-raising category, we can intuitively see that YOLOv8-s has missed targets. For the writing and reading categories with high posture similarity, our improved method has higher detection accuracy. This demonstrates the robustness and precision of our approach to detecting student behaviors within classroom settings.

Conclusion

Our detection method is tailored to accurately detect students' behaviors within complex and crowded classroom environments, thereby enhancing teaching quality and boosting student learning efficiency. The methodology is structured in a three-tiered approach. First, we introduce the multi-scale large kernel convolution module (MLKCM), which utilizes convolution kernels of various sizes to establish adaptive receptive fields. This configuration effectively captures multi-scale feature information while significantly reducing background noise, resulting in enhanced detection precision. Second, the progressive feature optimization module (PFOM) is incorporated to complement the MLKCM. This module adeptly captures both local and global information, ensuring a comprehensive understanding of the scene and further refining the detection accuracy. Our framework's lightweight design enables real-time monitoring of student behavior in educational settings, facilitating seamless deployment in practical applications without sacrificing accuracy. This approach positions our solution as an ideal tool for enhancing classroom interaction and promoting improved learning outcomes.

Discussion

While our recommended method has demonstrated exceptional performance, there are areas that require further refinement. The datasets utilized for training and evaluation may not encompass all possible scenarios, and factors such as varying ambient illumination, object occlusion, and similar poses can potentially affect the model's real-world application performance. To address these limitations, we plan to compile a more extensive dataset that reflects real-world conditions, incorporating a broader range of environmental factors. This will

Figure 5. Visual comparison on SCB-Dataset3-S dataset.

enable us to tackle the identified shortcomings and enhance the model's effectiveness. For practical applications, real-time detection capabilities must be integrated with on-site model deployment. Future research should focus on rigorously assessing the model's performance in actual settings to ensure its robustness and reliability in real-world scenarios. This will pave the way for the development of more sophisticated and context-aware detection systems that can significantly contribute to improving educational outcomes.

Data availability

The data that support the findings of this study are available from the corresponding author upon request. This study did not report any data. The proposed method was evaluated on two publicly available datasets that are widely used in the field of student classroom behavior detection. The download link is https://github.com/Whiffe/SCB-dataset.

Received: 29 November 2024; Accepted: 17 April 2025
Published online: 25 April 2025

References

1. Wang, R., Chen, S., Tian, G., Wang, P. & Ying, S. Post-secondary classroom teaching quality evaluation using small object detection model. *Sci. Rep.* **14**, 5816 (2024).
2. Yang, B., Yao, Z., Lu, H., Zhou, Y. & Xu, J. In-classroom learning analytics based on student behavior, topic and teaching characteristic mining. *Pattern Recogn. Lett.* **129**, 224–231 (2020).
3. Alnasyan, B., Basheri, M. & Alassafi, M. The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. *Comput. Educ. Artif. Intell.* 100231 (2024).
4. Yu, Z., Xie, M., Gao, J., Liu, T. & Fu, Y. From raw video to pedagogical insights: A unified framework for student behavior analysis. *In Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 23241–23249 (2024).
5. Alruwais, N. & Zakariah, M. Student recognition and activity monitoring in e-classes using deep learning in higher education. *IEEE Access* (2024).
6. Arias, E. M., Parraga-Alava, J. & Montenegro, D. Z. Stress detection among higher education students: A comprehensive systematic review of machine learning approaches. In *2024 Tenth International Conference on eDemocracy & eGovernment (ICEDEG)*, 1–8 (IEEE, 2024).
7. D'Mello, S. K., Lehman, B. & Person, N. Monitoring affect states during effortful problem solving activities. *Int. J. Artif. Intell. Educ.* **20**, 361–389 (2010).
8. Su, X. & Wang, W. Recognition and identification of college students\classroom behaviors through deep learning. *IEIE Trans. Smart Process. Comput.* **12**, 398–403 (2023).
9. Chen, H., Zhou, G. & Jiang, H. Student behavior detection in the classroom based on improved yolov8. *Sensors* **23**, 8385 (2023).
10. Kaul, G., McDevitt, J., Johnson, J. & Eban-Rothschild, A. Damm for the detection and tracking of multiple animals within complex social and environmental settings. *Sci. Rep.* **14**, 21366 (2024).
11. Ren, L., Li, S. & Chen, C. Student classroom behavior detection method based on deep learning. In *2024 4th International Symposium on Computer Technology and Information Science (ISCTIS)*, 104–109 (IEEE, 2024).
12. Yao, F., Chen, X., Jiang, Y. & Jia, W. Dss-yolov7: An enhanced yolov7-based algorithm for classroom behavior detection. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, 281–289 (2024).
13. Zhao, J., Zhu, H. & Niu, L. Bitnet: a lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 101670 (2023).

14. Dang, M. et al. Multi-object behaviour recognition based on object detection cascaded image classification in classroom scenes. *Appl. Intell.* **54**, 4935–4951 (2024).
15. Zhang, S. et al. Msta-slowfast: A student behavior detector for classroom environments. *Sensors* **23**, 5205 (2023).
16. Yang, F. & Wang, T. Scb-dataset3: A benchmark for detecting student classroom behavior. *arXiv preprint* arXiv: 2310.02522 (2023).
17. Li, Y. et al. Student behavior recognition for interaction detection in the classroom environment. *Image Vis. Comput.* **136**, 104726 (2023).
18. Wang, Z., Li, L., Zeng, C., Dong, S. & Sun, J. Slbdetection-net: Towards closed-set and open-set student learning behavior detection in smart classroom of k-12 education. *Expert Syst. Appl.* **260**, 125392 (2025).
19. Wang, Z., Wang, M., Zeng, C. & Li, L. Sbd-net: Incorporating multi-level features for an efficient detection network of student behavior in smart classrooms. *Appl. Sci.* **14**, 8357 (2024).
20. Liu, W. et al. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37 (Springer, 2016).
21. Lin, T. Focal loss for dense object detection. *arXiv preprint* arXiv: 1708.02002 (2017).
22. Redmon, J. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
23. Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).
24. Redmon, J. Yolov3: An incremental improvement. *arXiv preprint* arXiv: 1804.02767 (2018).
25. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint* arXiv: 2004.10934 (2020).
26. Li, C. et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint* arXiv: 2209.02976 (2022).
27. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475 (2023).
28. Wang, C.-Y., Yeh, I.-H. & Liao, H.-Y. M. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint* arXiv: 2402.13616 (2024).
29. Wang, A. et al. Yolov10: Real-time end-to-end object detection. *arXiv preprint* arXiv: 2405.14458 (2024).
30. Khanam, R. & Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint* arXiv: 2410.17725 (2024).
31. Girshick, R. *Fast r-cnn. arXiv preprint arXiv* **1504**, 08083 (2015).
32. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2016).
33. Zhang, Y. et al. Classroom behavior recognition based on improved yolov3. In *2020 International Conference on Artificial Intelligence and Education (ICAIE)*, 93–97 (IEEE, 2020).
34. Wang, Z. et al. Learning behavior recognition in smart classroom with multiple students based on yolov5. *arXiv preprint* arXiv: 2303.10916 (2023).
35. Zhao, J. & Zhu, H. Cbph-net: A small object detector for behavior recognition in classroom scenarios. *IEEE Trans. Instrum. Meas.* (2023).
36. Ma, L. et al. Improving yolov7 for large target classroom behavior recognition of teachers in smart classroom scenarios. *Electronics* **13**, 3726 (2024).
37. Jia, Q. & He, J. Student behavior recognition in classroom based on deep learning. *Appl. Sci.* **14**, 7981 (2024).
38. Wang, Z., Wang, M., Zeng, C. & Li, L. Multi-scale deformable transformers for student learning behavior detection in smart classroom. *arXiv preprint* arXiv: 2410.07834 (2024).
39. Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M. & Hu, S.-M. Visual attention network. *Comput. Vis. Media* **9**, 733–752 (2023).
40. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
41. Szegedy, C. et al. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
42. Purkait, P., Zhao, C. & Zach, C. Spp-net: Deep absolute pose regression with synthetic views. *arXiv preprint* arXiv: 1712.03452 (2017).
43. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
44. Li, Y., Chen, Y., Wang, N. & Zhang, Z. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6054–6063 (2019).
45. Guo, M.-H. et al. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural. Inf. Process. Syst.* **35**, 1140–1156 (2022).
46. Dai, J. et al. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773 (2017).
47. Ding, X., Zhang, X., Han, J. & Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11963–11975 (2022).
48. Chen, H., Chu, X., Ren, Y., Zhao, X. & Huang, K. Pelk: Parameter-efficient large kernel convnets with peripheral convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5557–5567 (2024).
49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
50. Ding, X., Zhang, X., Han, J. & Ding, G. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10886–10895 (2021).
51. Zhang, H. et al. Surface defect detection of hot rolled steel based on multi-scale feature fusion and attention mechanism residual block. *Sci. Rep.* **14**, 7671 (2024).
52. Xue, Z., Yu, X., Liu, B., Tan, X. & Wei, X. Hresnetam: Hierarchical residual network with attention mechanism for hyperspectral image classification. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* **14**, 3566–3580 (2021).
53. Liu, W. et al. Research on fault diagnosis of steel surface based on improved yolov5. *Processes* **10**, 2274 (2022).
54. Li, X. et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020).
55. Zheng, Z. et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **52**, 8574–8586 (2022).
56. Feng, C., Zhong, Y., Gao, Y., Scott, M. R. & Huang, W. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3490–3499 (IEEE Computer Society, 2021).
57. Zhang, S., Chi, C., Yao, Y., Lei, Z. & Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768 (2020).
58. Kong, T. et al. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **29**, 7389–7398 (2020).
59. Ge, Z. Yolox: Exceeding yolo series in 2021. *arXiv preprint* arXiv: 2107.08430 (2021).

60. Nguyen, T.-Q., Tran, H.-L., Tran, T.-K., Phan-Nguyen, H.-P. & Nguyen, T.-H. Fa-yolov9: Improved yolov9 based on feature attention block. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 1–6 (IEEE, 2024).
61. Shen, L., Li, X., Yang, W. & Wang, Q. Tlb-yolo: a rapid and efcient real-time algorithm for box-type classification and barcode recognition on the moving conveying and sorting systems. *Res. Square* (2024).
62. Zhu, W. & Yang, Z. Csb-yolo: a rapid and efficient real-time algorithm for classroom student behavior detection. *J. Real-Time Image Proc.* **21**, 140 (2024).
63. Gu, Y. & Niu, K. Research on classroom behavior analysis based on yolo target detection model. In *2024 4th International Symposium on Computer Technology and Information Science (ISCTIS)*, 730–733 (IEEE, 2024).

Acknowledgements
This research was funded by Science and Technology Plan Project of Quzhou (2023K237).

Author contributions
All authors reviewed the manuscript. Conceptualization: X.S., S.L. and S.C.; Investigation: X.S., S.L. and S.C.; Software and validation: S.L. and S.C.; Writing & original draft preparation: S.L. and S.C.; Formal analysis: X.S. and S.C.; Prepare figures: X.S., S.L. and S.C.; Funding acquisition: X.S.; Interpretation of data: X.S., S.L. and S.C.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to X.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.