IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 6, December, 2024, pp. 3173-3181
Journal homepage: http://iieta.org/journals/ts

A Classroom Interaction Behavior Analysis Method Based on Image Processing and Artificial Intelligence

Die Hu[1], Juan Liu[2], Weili Hu[3*]

[1] School of Foreign Languages, Southwest Medical University, Luzhou 646000, China
[2] Chongqing General Hospital, Chongqing 401147, China
[3] School of Humanities and Management Science, Southwest Medical University, Luzhou 646000, China

Corresponding Author Email: 102139@swmu.edu.cn

https://doi.org/10.18280/ts.410633

Received: 3 July 2024
Revised: 21 November 2024
Accepted: 1 December 2024
Available online: 31 December 2024

Keywords:
Artificial Intelligence, classroom interaction behavior, object detection, YOLOv5, spatiotemporal information fusion, behavior recognition

ABSTRACT

With the rapid development of artificial intelligence technology, methods for classroom interaction behavior analysis based on computer vision and deep learning have gradually become an important direction in educational research. Classroom interaction behavior is a crucial indicator of teaching effectiveness and student learning progress. Traditional classroom observation methods fail to meet the demands of real-time monitoring, accuracy, and comprehensiveness. Image-based interaction behavior analysis can improve the precision and efficiency of classroom assessments through automation. Existing research mainly focuses on the recognition and analysis of interaction behaviors, but challenges such as insufficient detection accuracy, a lack of dynamic spatiotemporal information integration in behavior recognition, and poor algorithm real-time performance still exist. Therefore, improving target detection accuracy and behavior recognition, especially in complex classroom environments, remains a critical research challenge. This paper proposes an improved YOLOv5-based classroom interaction object detection algorithm to address accuracy and real-time issues in recognizing interaction objects in complex classroom settings. Additionally, the paper presents an interaction behavior recognition method based on dynamic spatiotemporal information fusion, which enhances behavior recognition accuracy and robustness by integrating spatiotemporal features. The improved algorithm framework effectively enhances the precision and efficiency of interaction behavior analysis, providing technical support for intelligent evaluation of teaching processes and personalized education.

1. INTRODUCTION

With the rapid development of information technology, the application of artificial intelligence in the field of education has been making continuous progress, especially in the analysis of classroom interaction behavior [1-4]. Classroom interaction, as an important means of information transfer and knowledge sharing between students and teachers during the teaching process, directly affects teaching effectiveness and students' interest in learning [5, 6]. Traditional classroom observation and assessment methods often rely on manual recording and subjective judgment, making it difficult to achieve real-time, comprehensive, and accurate interaction behavior analysis [7-10]. Therefore, classroom interaction behavior analysis based on image processing and artificial intelligence technology has become one of the current hot topics in educational research. By applying computer vision, deep learning, and other technologies, various classroom interaction behaviors can be automatically recognized, providing educators with more scientific and effective teaching evaluation tools.

Research on classroom interaction behavior not only helps improve teaching quality, but also has important significance for personalized education and teaching optimization. Firstly, real-time interaction behavior analysis can help teachers quickly adjust teaching strategies, improve teaching content and methods, thereby enhancing classroom efficiency [11-14]. Secondly, systematic behavior recognition technology helps provide more accurate learning feedback to students, thereby promoting students' self-directed learning and active participation [15, 16]. Finally, with the help of big data and artificial intelligence analysis, scientific decision-making can be provided for education managers, promoting educational reform and innovation [17, 18]. However, despite some existing artificial intelligence-based classroom behavior analysis methods, challenges still remain in terms of algorithm accuracy, real-time performance, scalability, and so on. Therefore, conducting research on classroom interaction behavior analysis based on image processing and artificial intelligence has broad application prospects and profound social significance.

Current research mainly focuses on various aspects of classroom behavior recognition and student behavior analysis, but these studies still have some shortcomings in practical applications [19-22]. Firstly, many existing methods have limitations in detecting interaction objects, especially in

complex classroom environments, where traditional object detection algorithms may struggle to achieve efficient and accurate object recognition [21, 22]. Secondly, existing behavior recognition methods are often limited to static images or single spatiotemporal features, ignoring the integration of dynamic spatiotemporal information in the classroom, leading to lower recognition accuracy in complex scenes [23-27]. In addition, the real-time performance and processing speed of existing algorithms are also insufficient to meet the demands of classroom applications. Therefore, how to improve the robustness of detection algorithms and combine dynamic spatiotemporal features for multidimensional behavior recognition remains an important challenge in current research.

This paper is mainly dedicated to addressing these issues. It proposes an improved YOLOv5-based classroom interaction object detection algorithm aimed at improving the accuracy and real-time performance of interaction object recognition in the classroom. At the same time, this paper also proposes a classroom interaction behavior recognition method based on dynamic spatiotemporal information fusion, aiming to enhance behavior recognition accuracy and robustness by integrating multidimensional spatiotemporal features. The core contribution of this research lies in solving the deficiencies of existing methods in interaction object detection and behavior recognition through algorithm innovation and optimization, thereby promoting the application and development of artificial intelligence technology in the field of education. Through this research, not only can the automation level of classroom behavior analysis be improved, but also more accurate and scientific technical support can be provided for educational evaluation, with significant theoretical and practical value.

2. CLASSROOM INTERACTION OBJECT DETECTION ALGORITHM BASED ON IMPROVED YOLOV5

The traditional YOLOv5 algorithm often faces a trade-off between accuracy and lightweight performance in classroom interaction object detection. To achieve efficient computational performance while ensuring high detection accuracy, this paper proposes an improved YOLOv5-based classroom interaction object detection algorithm, which includes network structure optimization and loss function design. In terms of network structure, the idea of GhostNet is incorporated, and the C3Ghost module is introduced to optimize the redundancy of feature maps. The C3Ghost module effectively reduces the model's parameter count by eliminating redundant feature map information, thereby easing the computational load and improving algorithm speed. To address the potential issue of insufficient correlation between feature map channels in the C3Ghost module, this study introduces the Convolutional Block Attention Module (CBAM) attention mechanism. CBAM can adaptively learn spatial and channel-wise attention information, focusing automatically on feature areas that are crucial for target detection, thereby enhancing the model's ability to extract key features. In classroom interaction object detection, the movements, positions, and interaction behaviors of teachers and students often have strong dynamic and local characteristics.

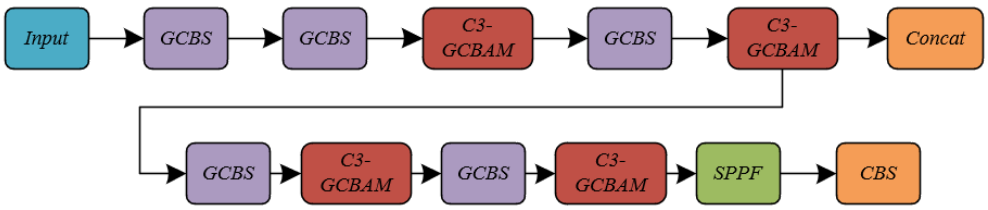


Figure 1. Backbone module of the improved YOLOv5 structure

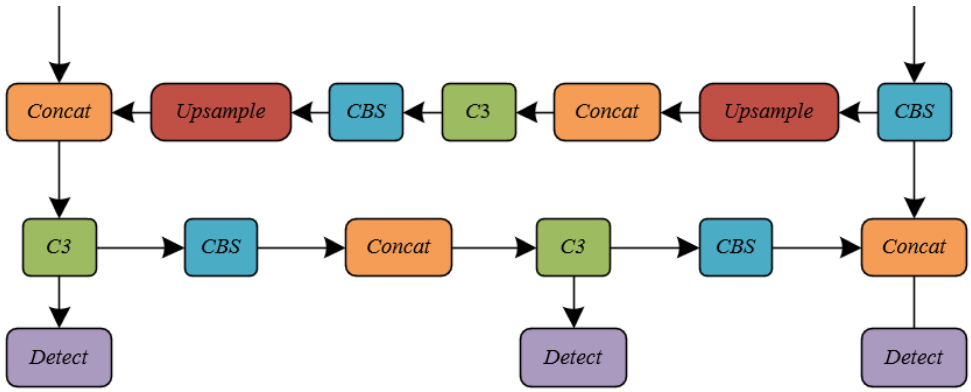


Figure 2. Neck and head modules of the improved YOLOv5 structure

Therefore, CBAM can effectively improve the model's detection accuracy and robustness in complex scenarios. Regarding loss function design, this paper replaces the original YOLOv5 Complete Intersection over Union (CIoU) loss function with the Enhanced Intersection over Union (EIoU) loss function. EIoU, when calculating the target box regression loss, not only considers the overlap between the predicted and ground-truth boxes but also incorporates the shape and positional distribution of the target boxes, enabling more accurate evaluation of the difference between predicted and real boxes in object detection. The EIoU loss function helps accelerate network training convergence and improves the

model's ability to localize different types of interaction objects in dynamic classroom scenes. Figures 1 and 2 show the schematic diagram of the improved YOLOv5 structure.

2.1 Lightweight backbone network

In traditional GhostNet, the core idea of the Ghost Module is to generate low-dimensional feature maps through a small number of standard convolutions, and then use linear transformations to generate more feature maps, thus reducing the computational load while maintaining a high feature representation capability. This design of GhostNet can effectively reduce the model's parameter count and improve computational efficiency. Assuming that the input and output feature map channels are denoted by z and v, the input and output feature map heights by g and g', the input and output feature map widths by q and q', the convolution kernel size by j, the linear operation kernel size by f, and the number of linear operations by t, the theoretical parameter compression ratio of the Ghost module can be calculated as:

$$e_z = \frac{v \cdot z \cdot j \cdot j}{\frac{v}{t} \cdot z \cdot j \cdot j + (t-1) \cdot \frac{v}{t} \cdot f \cdot f} \approx \frac{t \cdot z}{t + z - 1} \approx t \tag{1}$$

The theoretical acceleration ratio of the Ghost module can be calculated as:

$$\begin{aligned} e_t &= \frac{v \cdot g' \cdot q' \cdot z \cdot j \cdot j}{\frac{v}{t} \cdot g' \cdot q' \cdot z \cdot j \cdot j + (t-1) \cdot \frac{v}{t} \cdot g' \cdot q' \cdot f \cdot f} \\ &= \frac{z \cdot j \cdot j}{\frac{1}{t} \cdot z \cdot j \cdot j + \frac{(t-1)}{t} \cdot f \cdot f} \approx \frac{t \cdot z}{t + z - 1} \approx t \end{aligned} \tag{2}$$

Traditional GhostNet performs excellently in general image recognition and classification tasks, but for classroom interaction object detection in dynamic and complex scenes, its ability to eliminate feature map redundancy and capture features still has room for improvement. This is especially important when handling interactions between different roles, where fine-grained feature extraction and higher robustness are critical.

To address this issue, this paper introduces and optimizes the C3Ghost module based on traditional GhostNet to better meet the specific requirements of classroom interaction object detection. The C3Ghost module combines the GhostConv structure of GhostNet with the design advantages of the C3 module to further enhance feature extraction capabilities and computational efficiency. Specifically, the C3 module generates half of the feature information through the Ghost bottleneck and the other half through traditional convolution, then concatenates these two feature maps to maximize the gradient combination difference. This design not only optimizes the feature map generation process but also reduces the propagation of redundant features, thus effectively improving the model's performance in dynamic and complex classroom interaction scenes. Compared to traditional GhostNet, the C3Ghost module reduces computational load while optimizing the feature map generation and information transmission mechanisms, which further improves the model's precise recognition ability for classroom interaction behaviors, especially in complex backgrounds.

Moreover, when applied to classroom interaction object detection, the C3Ghost module can improve the model's performance in handling teacher-student interactions, character positions, and posture changes in complex scenes through more effective feature fusion and information transmission. Traditional GhostNet focuses more on feature extraction in image classification, while the C3Ghost module specifically optimizes adaptability in dynamic scenes, effectively avoiding the loss of detection accuracy caused by feature redundancy or insufficient channel correlation, particularly in multi-person interaction scenarios. This improvement makes the YOLOv5 based on the C3Ghost module more suitable for precise detection of classroom interaction objects, achieving a better balance between real-time performance and accuracy, and providing stronger technical support for classroom interaction behavior analysis.

2.2 Attention mechanism

In classroom interaction object detection, classroom scenes often involve multiple dynamic targets, such as the interaction between teachers and students. The states, postures, and positions of these targets may change over time, leading to complex scene variations. Traditional convolutional neural networks may struggle to effectively capture these changing and local detail information, especially when multiple objects appear simultaneously. Background interference and the similarity between targets may result in decreased detection accuracy.

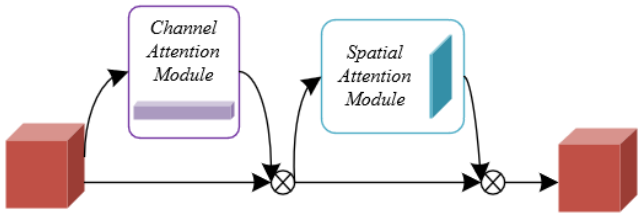


Figure 3. Structure of the CBAM attention mechanism module

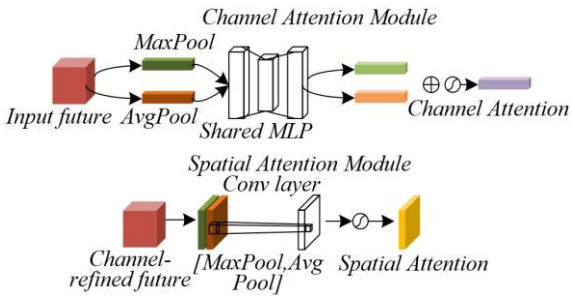


Figure 4. Structure of the channel attention mechanism and spatial attention mechanism

To enhance the model's ability to adapt to these complex scenes, this paper introduces CBAM, as shown in Figures 3 and 4. By adding the CBAM attention mechanism to the C3-Ghost module (i.e., C3-GCBAM), the network can more accurately focus on key information related to classroom interactions. In the channel dimension, CBAM extracts global statistical information of each channel through global average pooling, allowing important feature channels to be weighted and improving their activation. In the spatial dimension, CBAM uses pooling operations at different scales to capture

spatial information of varying granularities, helping the model to focus on feature areas related to the target's position, which is particularly important for dynamic classroom interaction objects.

2.3 Improved loss function

YOLOv5 uses CIoU by default to calculate bounding box loss. Assuming that the diagonal length of the smallest enclosing rectangle of two candidate boxes is denoted by Z, Euclidean distance by ϑ, the center, width, and height of the predicted box by y, μ, g, and the center, width, and height of the ground-truth box by y_{hs}, μ_{hs}, and g_{hs}, and the balance parameter by β, the calculation formula for this loss function is:

$$LOSS_{CIOU} = 1 - IOU + \frac{\vartheta^2(y, y_{hs})}{Z^2} + \beta\left(4/\tau^2\right)\left(\arctan\left(\frac{\mu_{hs}}{g_{hs}}\right) - \arctan\left(\frac{\mu}{g}\right)\right)^2 \quad (3)$$

In classroom interaction object detection tasks, the dynamic nature and posture changes of targets often result in significant changes in the shape and position of the target boxes. Especially when different individuals interact, complex spatial relationships may be involved. For example, when a teacher raises a hand, a student answers a question, or participates in other behaviors, the shape and position of the target may change drastically. This can cause traditional IOU measurement methods, especially the CIoU loss function, to be inadequate in handling such dynamic behaviors. Although the CIoU loss function optimizes bounding box fitting by considering factors like the center point distance and aspect ratio, when the aspect ratio difference between the predicted and ground-truth boxes is large, the CIoU loss function's convergence speed slows down, and it may even lead to instability during the network training process. In such cases, introducing the EIoU loss function can effectively address this problem. By providing more accurate aspect ratio consistency loss, the model can better understand and optimize the relationship between predicted and ground-truth boxes, particularly when dealing with classroom interaction objects that exhibit complex shape and position changes, thereby improving detection accuracy and stability.

Compared to CIoU, the EIoU loss function improves aspect ratio handling. EIoU not only considers the overlap and center distance of target boxes, but also introduces separate consistency loss for width and height, allowing for more precise optimization of the aspect ratio between boxes. Assuming the width and length of the smallest enclosing rectangle of two boxes are denoted by Z^{μ} and Z^{g}, the calculation formula for this loss function is:

$$LOSS_{EIOU} = LOSS_{IOU} + M_{DIS} + M_{ASP} = 1 - IOU + \left(\frac{\vartheta^2\left(y, y^{hs}\right)}{Z^2}\right) + \frac{\vartheta^2\left(\mu, \mu^{hs}\right)}{Z_{\mu}^2} + \frac{\vartheta^2\left(g, g^{hs}\right)}{Z_g^2} \quad (4)$$

3. DYNAMIC SPATIOTEMPORAL INFORMATION FUSION FOR CLASSROOM INTERACTION BEHAVIOR RECOGNITION

In the task of classroom interaction behavior recognition, video data often exhibits strong spatiotemporal dependencies and dynamics. Interaction behaviors between teachers and students, such as speaking, raising hands, and answering questions, change over time. To address this, this paper draws upon the concept of the Time Segmented Network (TSN) and uses sparse sampling techniques to process classroom video data. Specifically, the input video is divided into multiple segments of equal length, and one frame is randomly selected from each segment to form an input sequence containing J frames. This sparse sampling strategy ensures information coverage while reducing the computational cost of redundant frames, thereby improving the computational efficiency of the model.

To further enhance the model's performance in spatiotemporal modeling, this paper proposes a classroom interaction behavior recognition algorithm based on dynamic spatiotemporal information fusion. This is achieved by utilizing a pre-trained 2D ResNet-50 network as the foundation for feature extraction, and embedding motion modules and residual modules to fuse spatiotemporal information. The introduction of the motion module is primarily to capture the dynamic changes of objects and actions in the video. By dynamically modeling spatiotemporal information, the model can analyze the continuity and changing patterns of these behaviors from the temporal dimension, thereby improving recognition accuracy. In addition, the residual module helps to address the issue of gradient vanishing in deep network training, making the model more stable during training and allowing it to capture deeper feature information effectively.

3.1 Motion module

In the proposed algorithm, the core goal of the motion module is to analyze the differences between consecutive frames in the video and extract dynamic changes in human actions, thereby improving the accuracy of recognizing interactive behaviors in the classroom. The structure is illustrated in Figure 5.

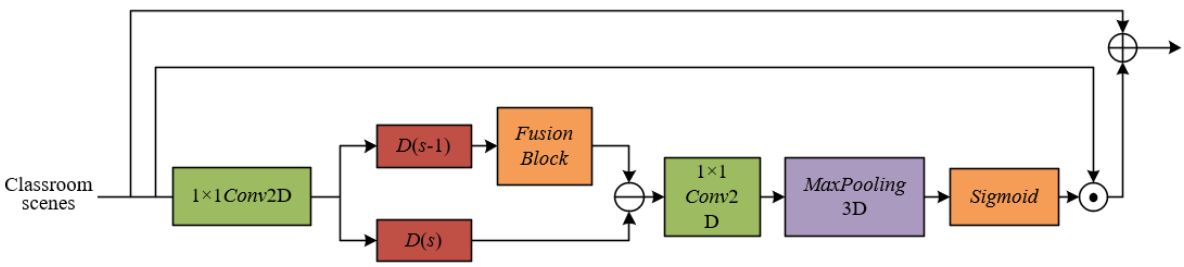


Figure 5. Structure of the motion module

Classroom scenes typically involve diverse actions, such as students raising their hands or interacting with the teacher. These behaviors often exhibit strong spatiotemporal variation. To effectively capture these changes in actions, assume that the shape of the input feature D is [V, S, ZG, Q], where V represents the number of input images, S and Z represent the temporal dimension and feature channels, and G and Q represent the spatial dimensions of the video. Features of size Z/e are represented by D_1, and a 1×1 2D channel convolution is represented by $z_{conv2Dz/e}$. The convolution operation is denoted by *. The motion module first reduces the number of feature channels through the following convolution operation:

$$D_1 = z_{conv2D_{z/e}} * D \qquad D_1 \in E^{(V\times S\times Z/e\times G\times Q)} \tag{5}$$

And calculate the difference between adjacent frames to obtain subtle dynamic change features between frames, that is, by performing a differential calculation between the continuous frames $D_1(s\text{-}1)$ and the processed frame $D'_1(s)$. Assuming that the original action features processed at time s are represented by $D'_1(s)$, and the action features at the previous moment s-1 are represented by $D_1(s\text{-}1)$, the inter-frame difference features are represented by $D_2 \in E^{(V*S*Z*G*Q)}$, where the time dimension is represented by S. The calculation formula is as follows:

$$D_2 = D_1^{'}(s) - D_1(s-1) 2 \le s \le S \tag{6}$$

Next, the motion module uses a 1x1 2D channel convolution operation $z_{conv2Dz}$ to expand the channel number of the inter-frame difference feature, ensuring that the number of channels matches the original input feature. This process enhances the representation of the inter-frame difference features while preventing information loss.

$$D_2^{'} = z_{conv2D_z} * D_2 \qquad D_2^{'} \in E^{(V\times S\times Z\times G\times Q)} \tag{7}$$

The expanded features are then summarized through a 3D global max pooling layer, which helps to reduce redundancy in the spatial dimensions while extracting the most representative contextual information from a broader spatiotemporal receptive field. The global max pooling layer, by performing global maximization in both the temporal and spatial dimensions, can capture the most significant dynamic features. This is crucial for classroom interaction behavior recognition, as key behavioral changes in such scenes are often caused by quick actions, such as a student rapidly raising a hand or a teacher turning around.

$$D_2^{MAX} = \text{maxpool } 3D\left(D_2^{'}\right) \qquad D_2^{MAX} \in E^{(V\times S\times Z\times G\times Q)} \tag{8}$$

Finally, the motion module uses the Sigmoid activation function to calculate dynamic weights for the inter-frame difference features, and performs element-wise multiplication between the attention weights and the original input features D. The core aim of this process is to assign higher weights to the parts of the feature map that exhibit significant dynamic changes, thereby focusing more strongly on regions with noticeable dynamic changes during the recognition process and suppressing features that show little change or are irrelevant to the current task. In this way, the motion module effectively extracts and enhances the key action features in classroom interaction behaviors while filtering out irrelevant background information, improving both recognition accuracy and robustness. The Sigmoid activation function is denoted by σ, and the output feature is denoted by H, with the calculation formula as:

$$H = D + D * \left(\sigma * D_2^{MAX} - \frac{1}{2}\right) H \in E^{(V\times S\times Z\times G\times Q)} \tag{9}$$

3.2 Feature fusion

In the algorithm, feature fusion is a key step in improving the model's performance, especially when dealing with complex classroom scenarios. The goal is to efficiently integrate information from different branches to enhance the model's ability to recognize interaction behaviors. The core idea of feature fusion is to combine the feature information extracted from the correction branch and the context branch, thereby improving the model's ability to perceive and express dynamic spatiotemporal information. Specifically, the original feature D1D1D1 is divided into two branches for processing: the correction branch and the context branch, to optimize and supplement different spatiotemporal features. Figure 6 shows the structure of the feature fusion module.

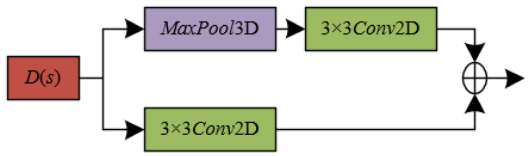


Figure 6. Structure of the feature fusion module

First, the correction branch uses a 3×3 convolution operation to smooth the features. This is because in classroom interactions, the movement of individuals often leads to changes in spatial positions, and the correction branch effectively mitigates these displacements, ensuring spatial consistency between adjacent frames, thereby reducing interference for subsequent recognition tasks. Suppose a 3×3 2D convolution with kernel size 3*3 is denoted by $z_{conv2D3*3}$, the calculation formula is:

$$D_1^{3\times3} = z_{conv2D_{3\times3}} * D_1 \qquad D_1^{3\times3} \in E^{(V\times S\times Z/e\times G\times Q)} \tag{10}$$

Since behaviors in classroom scenarios usually occur within a certain spatiotemporal context, relying solely on local information might not accurately capture the overall characteristics of the behavior. For example, a student raising a hand is often related to the teacher's expression or explanation, and these relationships need to be captured by the context branch through global pooling. By using a 3D global max pooling layer, the context branch can aggregate information across both the temporal and spatial dimensions of the entire video sequence, thus providing more distinctive contextual features for the correction branch.

$$D_1^{MAX} = z_{conv2D_{3\times3}} * \left(\text{maxpool } 3D\left(D_1\right)\right) \tag{11}$$

To ensure the correlation of features extracted from the two branches, and to allow them to remain consistent and

complement each other during feature processing, the model adopts a shared convolution parameter strategy between branches. This approach enables the model to strike a balance between accuracy and computational efficiency, thus enhancing overall performance.

In the final step of feature fusion, the features $D^{3*3}{}_1$ processed by the correction branch and smoothing, along with the context features $D^{3*3}{}_1$ obtained through global pooling, are fused to obtain the final feature $D'_1(s)$.

$$D_1^{'} = D_1^{3\times3} + D_1^{MAX} \qquad D_1^{'}(s) \in E^{(V\times S\times Z/e\times G\times Q)} \tag{12}$$

This fusion method combines multidimensional feature information from different branches, retaining both the smoothed local spatiotemporal features and the globally rich contextual features, thereby enhancing the model's overall ability in classroom interaction behavior recognition. Through this feature fusion process, the algorithm can more accurately recognize complex behavior patterns in the classroom and better handle the challenges posed by spatiotemporal changes in classroom scenarios, ultimately achieving precise recognition of dynamic classroom interaction behaviors.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 7 shows the results of classroom interaction object detection. Table 1 provides the results of the ablation experiments for classroom interaction object detection. Based on the data in Table 1, the performance of the four different models in the classroom interaction object detection task varies. Model 1 (the traditional YOLOv5 algorithm) has 7.12M parameters, 15.9 GFLOPS, and achieves 73.5% in mAP@0.5%. Model 2 introduces the C3Ghost module, which slightly reduces the parameter count to 4.79M and GFLOPS to 11.2. However, the mAP@0.5% only slightly decreases to 72.4%. Model 3 further combines the C3-GCBAM module, resulting in an increase in both parameters (5.23M) and GFLOPS (12.1), with an improvement in mAP@0.5% to 73.6%. Model 4 builds upon Model 3 by adding the EIoU loss function, leading to a slight increase in parameters (5.24M) and GFLOPS (11.8), and achieving the highest mAP@0.5% of 74.8%.

From the experimental results, it is evident that the performance of the model significantly improves as optimization modules are gradually introduced into the traditional YOLOv5 model. Although the C3Ghost module (Model 2) causes a slight performance drop due to the reduction in parameters and computation, the introduction of C3-GCBAM in Model 3 and the EIoU loss function in Model 4 significantly enhance the detection accuracy, particularly in the mAP@0.5% metric. Model 4, in particular, performs the best, indicating the effectiveness of dynamic spatiotemporal information fusion and loss function optimization in improving detection accuracy. Additionally, the moderate increase in parameters and GFLOPS suggests that an appropriate model complexity can effectively improve the accuracy and robustness of classroom interaction object detection.

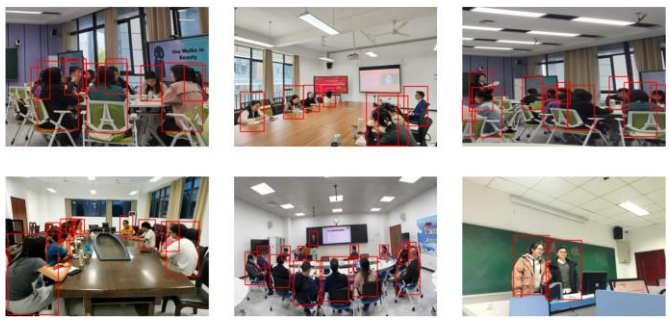
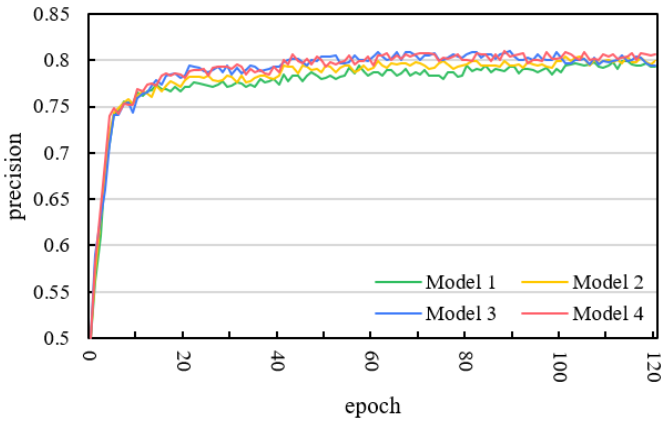


Figure 7. Classroom interaction object detection results

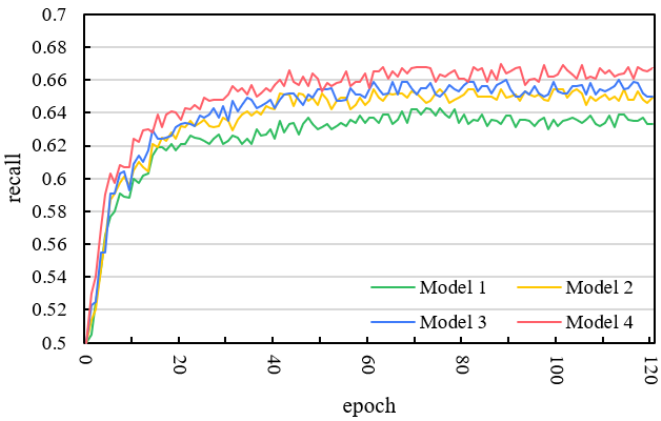
Table 1. Ablation experimental results for classroom interaction object detection

Model Name	*Parameters*	*GFLOPS*	*mAP@0.5%*
Model 1	7.12	15.9	73.5%
Model 2	4.79	11.2	72.4%
Model 3	5.23	12.1	73.6%
Model 4	5.24	11.8	74.8%

Based on the precision and recall data shown in Figure 8, we can observe the gradual optimization of the four models in the classroom interaction object detection task. Model 1 (the traditional YOLOv5 algorithm) initially exhibits relatively low precision and recall, but as training progresses, both precision and recall gradually improve. Specifically, precision increases from 0.50 to 0.78, and recall increases from 0.50 to 0.63, showing some improvement. Model 2 (YOLOv5 + C3Ghost module) achieves a noticeable improvement in precision, with precision increasing from 0.50 to 0.784, and recall increasing from 0.50 to 0.642. Model 2, in particular, shows better improvement in recall compared to Model 1. Model 3 (YOLOv5 + C3-GCBAM module) further optimizes the model's performance, with precision reaching 0.793 after 40 epochs, and recall increasing to 0.649.



(a)



(b)

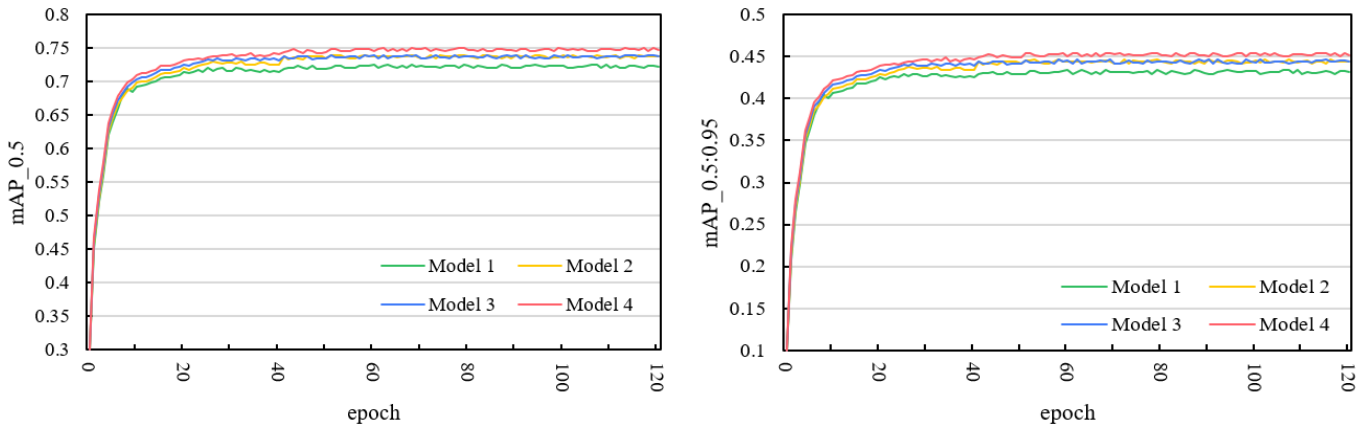


Figure 8. Comparison of classroom interaction object detection model running results

This indicates that the introduction of the C3-GCBAM module enhances the model's feature extraction and fusion abilities, significantly improving overall detection performance. Model 4 (YOLOv5 + C3-GCBAM + EIoU module) performs the best, with precision stabilizing around 0.793, and recall reaching 0.657 in the later stages of training. This confirms the effectiveness of the EIoU loss function in improving both detection precision and recall.

From the mAP_0.5 and mAP_0.5:0.95 data in Figure 8, we can see that with the introduction of different optimization modules, the performance of all four models in the classroom interaction object detection task has shown varying degrees of improvement. Specifically, Model 1 (the traditional YOLOv5) shows an increase in mAP_0.5 from an initial value of 0.30 to 0.72, indicating that the model's precision has gradually improved as training progresses. However, in terms of mAP_0.5:0.95, Model 1's improvement is slower, ultimately reaching only 0.43, suggesting that while the model's precision improves, there is still considerable room for improvement under stricter metrics. Model 2 (YOLOv5 + C3Ghost) demonstrates a more significant improvement in mAP_0.5, increasing from 0.30 to 0.73, and also shows a relatively faster improvement in mAP_0.5:0.95, increasing from 0.10 to 0.44. This indicates that the C3Ghost module has a positive effect on improving the model's fine-grained detection and robustness. Model 3 (YOLOv5 + C3-GCBAM) shows steady improvement in both metrics, with mAP_0.5 reaching 0.73 and mAP_0.5:0.95 stabilizing at 0.44, demonstrating that the C3-GCBAM module effectively enhances the model's feature fusion and discriminative ability. Finally, Model 4 (YOLOv5 + C3-GCBAM + EIoU) performs the best, with mAP_0.5 improving from 0.30 to 0.74, and mAP_0.5:0.95 also showing a significant increase, reaching 0.45. This highlights the important role of the EIoU module in enhancing the model's spatial consistency and optimizing detection accuracy.

From the comparative data in Table 2, the proposed classroom interaction object detection algorithm (Ours), based on the improved YOLOv5 model, outperforms other lightweight algorithms in several performance metrics. Specifically, YOLOX-s achieves an mAP@0.5 of 74.8%, the best performance among multiple lightweight detection models, but its larger parameters (9.25M) and GFLOPS (25.8) lead to higher computational and memory overhead. In comparison, YOLOv5n (parameters: 1.87M, GFLOPS: 4.3) and YOLOv5s (parameters: 7.05M, GFLOPS: 15.9) have smaller model sizes but their mAP@0.5 values are 72.3% and 73.8%, respectively, slightly lower than YOLOX-s. Our model (Ours) has 5.23M parameters and 11.2 GFLOPS, with an mAP@0.5 of 76.2%, significantly higher than YOLOv5n and YOLOv5s, and close to YOLOX-s. This indicates that our algorithm provides higher accuracy while maintaining computational efficiency and real-time performance, demonstrating its advantage in classroom interaction object detection.

Table 2. Comparison of the proposed classroom interaction object detection algorithm with other lightweight algorithms on the dataset

Model Name	Parameters	*GFLOPS*	*mAP@0.5%*
YOLOX-s	9.25	25.8	74.8%
YOLOv5n	1.87	4.3	72.3%
YOLOv5s	7.05	15.9	73.8%
Ours	5.23	11.2	76.2%

The comparative analysis shows that the proposed YOLOv5-based detection algorithm strikes a good balance between detection accuracy and computational efficiency. Although it has fewer parameters and GFLOPS compared to YOLOX-s, it outperforms YOLOX-s in mAP@0.5, indicating that the method is better suited for real-time detection and accuracy in complex educational environments. In comparison to YOLOv5n and YOLOv5s, while these models have smaller computational overhead, they fall short in detection accuracy. Overall, the proposed algorithm not only improves detection accuracy but also optimizes model size and computational complexity, showing great potential for application in the education field and providing crucial technical support for enhancing AI applications in education.

From the comparison data in Tables 3, 4, and 5, it can be observed that the dynamic spatio-temporal information fusion-based classroom interaction behavior recognition model proposed in this paper outperforms existing mainstream models across multiple datasets. On the AffectNet dataset, the proposed model achieves a Top-1 accuracy of 94.5%, significantly higher than other models, such as Spatio-Temporal TSN (92.5%) and TSN (TCNs) (93.7%). On the NTU RGB+D dataset, the proposed model also leads with an accuracy of 85.7%, surpassing TSN (ResNet) (81.2%), TSN(RGB) (82.5%), and Spatio-Temporal TSN (83.7%). On the TEACHER dataset, the proposed model again performs exceptionally, reaching an accuracy of 56.9%, exceeding TSN (ResNet) (47.5%), TSN(RGB) (48.9%), and Spatio-Temporal TSN (53.2%). Overall, the proposed model achieves the best

recognition accuracy across all the comparison datasets, demonstrating that this method can effectively enhance the accuracy of classroom interaction behavior recognition.

Table 3. Comparison of different classroom interaction behavior recognition models on accuracy (AffectNet dataset) (%)

Network Model	Average Accuracy (Top-1)
TSN (RGB)	83.2
Spatio-Temporal TSN	92.5
TSN (TCNs)	93.7
Proposed model	94.5

Table 4. Comparison of different classroom interaction behavior recognition models on accuracy (NTU RGB+D dataset) (%)

Network Model	Average Accuracy (Top-1)
TSN (ResNet)	81.2
TSN(RGB)	82.5
Spatio-Temporal TSN	83.7
TSN (TCNs)	84.9
Proposed Model	85.7

Table 5. Comparison of different classroom interaction behavior recognition models on accuracy (TEACHER dataset) (%)

Network Model	Average Accuracy (Top-1)
TSN (ResNet)	47.5
TSN(RGB)	48.9
Spatio-Temporal TSN	53.2
TSN (TCNs)	54.8
Proposed Model	56.9

Through comparative analysis, it can be concluded that the proposed dynamic spatio-temporal information fusion-based behavior recognition model has a significant advantage in accuracy, particularly in complex classroom interaction scenarios, where it achieves higher accuracy than existing methods. This highlights the importance of dynamic spatio-temporal feature fusion in classroom behavior recognition, effectively addressing the shortcomings of current methods in terms of accuracy and robustness. The proposed model outperforms traditional TSN series models and their improved versions on multiple datasets, validating the effectiveness of spatio-temporal information fusion technology in classroom behavior recognition, and providing a more precise and reliable tool for applying artificial intelligence technology in the field of education. In summary, the recognition model proposed in this paper not only improves recognition accuracy but also provides new ideas and solutions for analyzing classroom interaction behavior in complex scenarios.

5. CONCLUSION

The core research of this paper focuses on two aspects: classroom interaction object detection and classroom interaction behavior recognition. For classroom interaction object detection, this paper proposes an improved YOLOv5-based detection algorithm, incorporating various optimization modules to enhance detection accuracy and real-time performance. Through comparative experiments, the advantages of the proposed method in terms of accuracy and computational efficiency have been verified. Particularly in comparison with lightweight algorithms, this method demonstrates a high mAP value while maintaining relatively low computational overhead, indicating its potential for application in educational scenarios. Meanwhile, for classroom interaction behavior recognition, this paper proposes a recognition method based on dynamic spatiotemporal information fusion. By integrating multidimensional spatiotemporal features, the accuracy and robustness of recognition are significantly improved. Comparative experiments on multiple datasets (AffectNet, NTU RGB+D, TEACHER) show that the proposed behavior recognition model outperforms existing methods in terms of accuracy, validating the effectiveness of spatiotemporal information fusion in complex educational environments.

This research also has certain limitations. First, although the proposed detection and recognition algorithms have achieved good results on multiple datasets, how to further optimize the model's computational efficiency and inference speed, especially on resource-constrained mobile or embedded devices, remains a challenge. Future research directions can be explored in the following aspects: (1) further optimizing the model's computational efficiency, particularly for applications on embedded platforms; (2) expanding the diversity of training datasets to cover classroom interaction behaviors in different countries and educational systems; (3) exploring more advanced spatiotemporal fusion methods, such as deep learning-based adaptive spatiotemporal modeling methods, to enhance the recognition of more complex behaviors; (4) studying how to more closely integrate classroom interaction object detection and behavior recognition to form a stronger real-time feedback system, providing more comprehensive technical support for the intelligentization of education.

FUNDING

This paper was supported by 2023 Medical Education Research Project of the Medical Education Branch of the Chinese Medical Association and the National Medical Education Development Center (Grant No.: 2023B286).

REFERENCES

[1] Wang, S.Y., Cheng, L.M., Liu, D.Y., Qin, J.Q., Hu, G.H. (2022). Classroom video image emotion analysis method for online teaching quality evaluation. Traitement du Signal, 39(5): 1767-1774. https://doi.org/10.18280/ts.390535

[2] Laupichler, M.C., Hadizadeh, D.R., Wintergerst, M.W., Von Der Emde, L., Paech, D., Dick, E.A., Raupach, T. (2022). Effect of a flipped classroom course to foster medical students' AI literacy with a focus on medical imaging: a single group pre-and post-test study. BMC Medical Education, 22(1): 803. https://doi.org/10.1186/s12909-022-03866-x

[3] Jiang, L., Lu, X. (2023). Analyzing and optimizing Virtual Reality classroom scenarios: A deep learning approach. Traitement du Signal, 40(6): 2553-2563. https://doi.org/10.18280/ts.400618

[4] Muthmainnah, M., Siripipatthanakul, S., Apriani, E., Al Yakin, A. (2023). Effectiveness of online informal

language learning applications in English language teaching: A behavioral perspective. Education Science and Management, 1(2): 73-85. https://doi.org/10.56578/esm010202

[5] Ge, Y., Liu, L., Qiu, X., Song, H., Wang, Y., Huang, K. (2013). A framework of multilayer social networks for communication behavior with agent-based modeling. Simulation, 89(7): 810-828. https://doi.org/10.1177/0037549713477682

[6] Perkins, C.J. (2024). Evidence-based classroom observation technique: An interdisciplinary, structured approach to classroom observation. Nursing Education Perspectives, 45(2): 120-121. https://doi.org/10.1097/01.NEP.0000000000001086

[7] Crawford, L.H. (1998). Evaluation of nursing faculty through observation. Journal of Nursing Education, 37(7): 289-294. https://doi.org/10.3928/0148-4834-19981001-04

[8] De Lima, J.Á., Silva, M.J.T. (2018). Resistance to classroom observation in the context of teacher evaluation: Teachers' and department heads' experiences and perspectives. Educational Assessment, Evaluation and Accountability, 30: 7-26. https://doi.org/10.1007/s11092-017-9261-5

[9] Bergin, C., Wind, S.A., Grajeda, S., Tsai, C.L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training? Studies in Educational Evaluation, 55: 19-26. https://doi.org/10.1016/j.stueduc.2017.05.002

[10] Lu, Z., Nishimura, Y. (2024). Telepresence observation for kindergarten classroom rating: A pilot study. IEEE Access, 12: 32181-32191. https://doi.org/10.1109/ACCESS.2024.3368855

[11] Wind, S.A., Jones, E., Bergin, C., Jensen, K. (2019). Exploring patterns of principal judgments in teacher evaluation related to reported gender and years of experience. Studies in Educational Evaluation, 61: 150-158. https://doi.org/10.1016/j.stueduc.2019.03.011

[12] Hou, J., Xu, Y., He, W., Zhong, Y., et al. (2024). A systematic review for the fatigue driving behavior recognition method. Journal of Intelligent & Fuzzy Systems, 46(1): 1407-1427. https://doi.org/10.3233/JIFS-235075

[13] Saqlain, M. (2023). Revolutionizing political education in Pakistan: An AI-integrated approach. Education Science and Management, 1(3): 122-131. https://doi.org/10.56578/esm010301

[14] Lohaus, T., Rogalla, S., Thoma, P. (2023). Use of technologies in the therapy of social cognition deficits in neurological and mental diseases: A systematic review. Telemedicine and e-Health, 29(3): 331-351. https://doi.org/10.1089/tmj.2022.0037

[15] Tang, L., Xie, T., Yang, Y., Wang, H. (2022). Classroom behavior detection based on improved YOLOv5 algorithm combining multi-scale feature fusion and attention mechanism. Applied Sciences, 12(13): 6790. https://doi.org/10.3390/app12136790

[16] Mo, J.W., Zhu, R., Shou, Z.Y. (2023). Detection of students' classroom concentration based on component attention. International Journal of Innovative Computing Information and Control, 19(3): 877-891. https://doi.org/10.24507/ijicic.19.03.877

[17] Mo, J., Zhu, R., Yuan, H., Shou, Z., Chen, L. (2023). Student behavior recognition based on multitask learning. Multimedia Tools and Applications, 82(12): 19091-19108. https://doi.org/10.1007/s11042-022-14100-7

[18] Zong, L., Fang, J. (2024). Deep visual computing of behavioral characteristics in complex scenarios and embedded object recognition applications. Sensors, 24(14): 4582. https://doi.org/10.3390/s24144582

[19] Liu, Q., Jiang, R., Xu, Q., Wang, D., Sang, Z., Jiang, X., Wu, L. (2024). YOLOv8n_BT: Research on classroom learning behavior recognition algorithm based on improved YOLOv8n. IEEE Access, 12: 36391-36403. https://doi.org/10.1109/ACCESS.2024.3373536

[20] Zhang, Y., Zhu, T., Ning, H., Liu, Z. (2021). Classroom student posture recognition based on an improved high-resolution network. EURASIP Journal on Wireless Communications and Networking, 2021(1): 140. https://doi.org/10.1186/s13638-021-02015-0

[21] Zhang, X., Nie, J., Wei, S., Zhu, G., Dai, W., Yang, C. (2024). A study of classroom behavior recognition incorporating super-resolution and target detection. Sensors, 24(17): 5640. https://doi.org/10.3390/s24175640

[22] Chen, H., Guan, J. (2022). Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology. Electronics, 11(23): 3998. https://doi.org/10.3390/electronics11233998

[23] Lin, L., Yang, H., Xu, Q., Xue, Y., Li, D. (2024). Research on student classroom behavior detection based on the real-time detection transformer algorithm. Applied Sciences, 14(14): 6153. https://doi.org/10.3390/app14146153

[24] Li, L., Liu, M., Sun, L., Li, Y., Li, N. (2022). ET-YOLOv5s: Toward deep identification of students' in-class behaviors. IEEE Access, 10: 44200-44211. https://doi.org/10.1109/ACCESS.2022.3169586

[25] Gu, M., Liu, X., Feng, J. (2022). Classroom face detection algorithm based on improved MTCNN. Signal, Image and Video Processing, 16(5): 1355-1362. https://doi.org/10.1007/s11760-021-02087-x

[26] Xiao, G., Xu, Q., Wei, Y., Yao, H., Liu, Q. (2024). Occlusion robust cognitive engagement detection in real-world classroom. Sensors, 24(11): 3609. https://doi.org/10.3390/s24113609

[27] Tang, L., Gao, C., Chen, X., Zhao, Y. (2019). Pose detection in complex classroom environment based on improved Faster R-CNN. IET Image Processing, 13(3): 451-457. https://doi.org/10.1049/iet-ipr.2018.5905