Classroom Anomalous Behavior Detection Based On Improved YOLOv5

Fu Yao

School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China

Corresponding author. E-mail: fuyao2662@163.com

Received:Aug. 07, 2024; Accepted: Sep. 01, 2024

In recent years, there has been an increasing interest in using machine learning methods to improve deep learning-based classroom abnormal behaviour detection tasks, and the theory of fractional order calculus is beginning to be used to enhance the model's ability to describe the features of the data. In this paper, we propose a classroom abnormal behaviour detection method based on fractional order calculus for YOLOv5 to monitor and analyse students' classroom behaviour immediately. A fractional order coordinate attention mechanism is designed in the YOLOv5 feature learning stage to capture the long-range relationship of features in combination with position information, while a hybrid convolutional layer is introduced to achieve computational lightness, and finally an improved loss function is used for training to improve the detection accuracy robustness. In this paper, we use the improved YOLOv5 model based on fractional order differentiation for classroom abnormal behaviour detection. It is experimentally shown that the method proposed in this paper achieves significant performance improvement in classroom behaviour detection, with 5.4 Spf improvement in detection speed and 4.1% improvement in classroom abnormal behaviour accuracy, which is highly observable and provides more targeted intervention and management tools for the education industry.

Keywords: YOLOv5; Fractional order differentiation; Deep learning; Abnormal behaviour detection

http://dx.doi.org/10.6180/jase.202507_28(7).0008

1. Introduction

In the current education field, the use of computer vision and deep learning techniques to detect abnormal behaviours in the classroom is gradually becoming a popular research direction [1–3]. In this study, an efficient and accurate classroom abnormal behaviour detection model is designed and implemented, which can monitor and analyse various behaviours of students in the classroom immediately, including but not limited to inattention, mobile phone use, dozing, and so on [4, 5]. The accurate detection and identification of these abnormal behaviours is of great importance for teachers to carry out effective classroom management and promote students' learning effectiveness [6, 7].

Classroom abnormal behaviour detection belongs to the target detection sub-task, in past work literature [8] is based on face recognition and assesses the teaching and learning situation through expression information, this method has low accuracy in complex scenarios. Recent deep learning based behaviour detection models use the feature learning advantage of convolutional networks to learn target semantic information to help improve the observability of the model [9].Secondly, a hybrid convolutional layer is introduced to achieve computational lightness and reduce arithmetic complexity, the latest research selects the kernel dwell or feature dwell process by considering the amount of data, reduces the memory access required for storing and loading the partial summation reduces arithmetic complexity, improves the model detection speed, followed by model training using an improved loss function, and finally completes the detection frame optimisation to enhance the observability of the results [10, 11].

2. Methods

YOLOv5 Classroom Abnormal Behavior Detection [12–14]. YOLOv5 is a single-stage target detection algorithm that achieves real-time detection by constructing the detection problem as a regression problem [15]. The model uses feature maps of different sizes to predict objects of different sizes, thus enabling multi-scale detection [16]. Its network structure mainly consists of a series of convolutional layers, a batch normalisation layer and a Leaky ReLU activation function [17]. The classroom abnormal behaviour detection model based on YOLOv5 contains three main parts: feature learning module, behaviour recognition and visual display. Among them, the core components in the feature learning module contain the backbone network, the neck network layer and the head output layer.

The schematic structure is shown in Fig. 1. The dotted line on the left is the backbone network, whose role is to extract features from the input image and generate feature maps. Cnov(Convolution) in Backbone is the standard convolution, C3(Cross Stage Partial Network) is the convolution module, and SPPF is the spatial pyramid pooling operation [18, 19]. The dotted part in the middle is the Neck network, which serves to fuse and reorganise the features of the image using the combined bottom-up and top-down feature fusion to improve the detection. UpSample in the Neck network is the upsampling and Concat is the splicing operation. The function of the final Head as the output layer is to predict and classify the bounding box of the input image, and deal with the redundant bounding box by Non-Maximum Suppression (NMS) [20, 21], and finally output the predicted category with the highest confidence level and return the corresponding bounding box coordinates. This process ensures the accuracy of the model in localising and classifying the target objects.

2.1. Coordinate Attention Mechanisms Based on Fractional Order Differentials

Fractional Order Differential Based Detection of Abnormal Behaviour in YOLOv5 Classroom. This subsection improves the model network structure and training by designing a fractional order based coordinate attention mechanism, optimising the loss function, and introducing a lightweight convolution, as well as optimising the detection frame to improve the detection performance of the model.

2.1.1. *Coordinate Attention Mechanism*

In order to make the model more focused on learning to effective feature regions, ensure the utilisation of computing resources, and improve the detection accuracy of the model in complex scenes, an efficient Coordinate Attention (CA) mechanism [22] has been proposed and has been applied to the backbone network of some models in recent work, to enhance the target by improving the effectiveness of the feature learning detection model performance.The CA attention mechanism embeds location information on top of the channel attention and is able to capture long-range dependencies to further enhance the detection effectiveness. By introducing location information, the CA attention mechanism can help the model better understand the relationship between features at different locations and focus attention on important regions. This mechanism is particularly important for dealing with the task of target detection in complex scenes, where a large amount of interfering information may exist and the model needs to be able to accurately locate and identify the target.

2.1.2. *Fractional Order Coordinate Attention Mechanism*

The structure of the FCA attention mechanism is shown in Fig. 2. The input feature maps of size $H \times W \times C$ are globally average pooled in spatial dimension by coordinate information embedding operation and coordinate generation operation.

The pooling kernel size used in the horizontal direction is $H \times 1$, and the pooling kernel size in the vertical direction is $1 \times W$. The feature maps X and Y in the horizontal and vertical directions are obtained as $H \times 1 \times C$, respectively, where C is the number of channels, H is the width, and W is the height, and $C/r \times 1 \times (W + H)$ denotes the weighting formulae Output features are computed in the manner shown in Eqs. (1) and (2).

$$Z_c^h(h) = \frac{1}{W} \sum_{i=0}^{W} x_c(h, i) \tag{1}$$

$$Z_c^w(w) = \frac{1}{H} \sum_{j=0}^{H} x_c(j, w) \tag{2}$$

where $Z_c^h(h)$ and $Z_c^w(w)$ are the results of global average pooling in horizontal and vertical directions, respectively, and x_c is the input feature vector.

After generating the two feature maps X and Y a splicing operation is performed, which is downscaled by a 1×1 convolution kernel, immediately followed by a batch normalisation layer to obtain the feature map f, computed as shown in Eq. (3).

$$f = \sigma\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{3}$$

where the feature map f is of size $C/r \times 1 \times (W + H)$. where F 1 is the batch normalisation result. g^h and g^w are the attention weights in the spatial dimension, respectively.

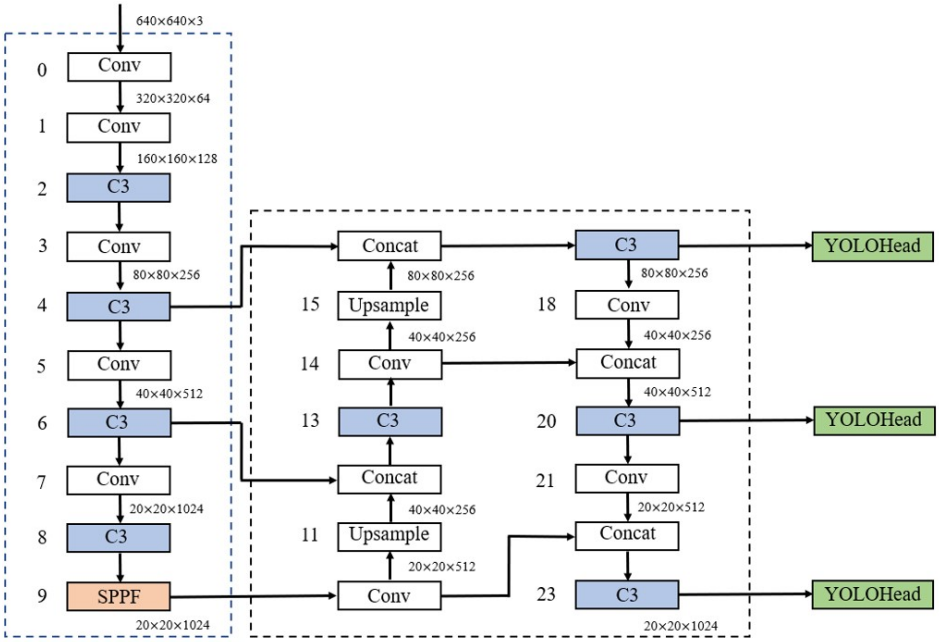

Fig. 1. YOLOv5 Structure.

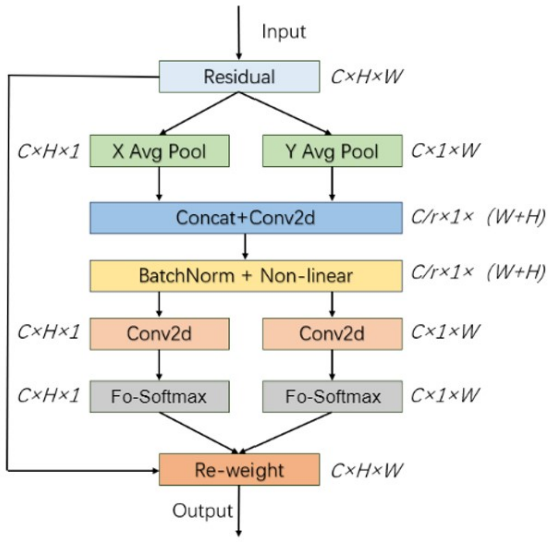

Fig. 2. FCA attention.

The feature graph decomposition of f in the spatial dimension is performed to decompose two feature graphs, f_h and f_w. Next, the similarity is converted into attention weights using the node function that plays a role between similarity computation and weighted combination. softmax function helps to ensure that the attention weights are normalised to correctly represent the importance of the different input elements so that the associations between the input elements are captured efficiently and appropriate attention weights are generated. Where the horizontal attention weight g^h and vertical attention weight g^w are the attention weights generated via the coordinate attention module.

The node function Softmax is processed within the improved CA attention network using fractional order differentiation, where softmax $= p(f)$, which is obtained by first order derivation of $p(f)$:

$$p'(f) = p(f)(1 - p(f)) \tag{4}$$

A 0.5 order derivative of $p(f)$ is obtained:

$$D^{0.3}p\left(Y_S\right) = \frac{1}{\Gamma_{(0.3)}} \int_0^{Y_S} \frac{p^{(1)}(t)dt}{\left(Y_S - t\right)^{0.3}} \tag{5}$$

$$D^{0.3}p\left(Y_S\right) = \frac{1}{\Gamma_{(0.3)}} \int_0^{Y_S} \frac{p^{(1)}(t)dt}{\left(Y_S - t\right)^{0.3}} \tag{6}$$

$$\begin{aligned} D^{0.5}p\left(Y_S\right) &= \frac{1}{\Gamma_{(0.5)}} \int_0^{Y_S} \frac{p^{(1)}(t)dt}{\left(Y_S - t\right)^{0.5}} \\ &= \frac{1}{\Gamma_{(0.5)}} \int_0^{Y_S} p\left(Y_S\right)\left(1 - p\left(Y_S\right)\right)\left(Y_S - t\right)^{-0.5} dt \end{aligned} \tag{7}$$

Similarly, the 1.5 th order derivatives of $p(f)$ are obtained, respectively:

$$D^{0.3}p\left(Y_S\right) = \frac{1}{\Gamma_{(0.3)}} \int_0^{Y_S} \frac{p^{(1)}(t)dt}{\left(Y_S - t\right)^{0.3}} \tag{8}$$

$$D^{0.8}p\left(Y_S\right)=\frac{1}{\Gamma_{(0.8)}}\int_0^{Y_S}\frac{p^{(1)}(t)dt}{\left(Y_S-t\right)^{0.8}} \tag{9}$$

The Softmax first order differential image of the nodal function and its 1.5 order differential image are shown in Fig. 3.

After replacing the integer-order differentiation with the fractional-order differentiation, the convergence of the curve tends to become flat, which is caused by the characteristic of the fractional-order differentiation, which no longer considers the current value singly in the process of convergence, but builds on the historical information of the signal computation, describing some nonlinear and complex features that the function does not have under the integer-order differentiation, and thus can consider the complex pixel from multiple perspectives and in a more stable way. feature probabilities, thus realising the requirement of flexible extraction.

Thus the attentional weights for the spatial dimension can be obtained:

$$g^h=\sigma\left(F_h\left(f^h\right)\right) \tag{10}$$

$$g^w=\sigma\left(F_w\left(f^w\right)\right) \tag{11}$$

where F_h and F_w are transformation functions that transform the feature map f_h and feature map f_w into the same dimension.

$$Z_c^w(w)=\frac{1}{H}\sum_{j=0}^{H}x_c(j,w) \tag{12}$$

2.2. Improvement of the loss function

YOLOv5 adopts YOLOv5 adopts CIoU [23] as the localisation loss function. CIoU is based on the distance intersection and merger ratio based on the addition of the aspect ratio, the specific formula is shown in Eqs. (13) and (14), where b and b^{gt} are the centroid of the prediction frame and the real frame, ρ is the Euclidean distance of the centroid, α is the weight value, V is the aspect ratio of the detection frame, w_{gt} and h_{gt} are the width and height, h_{gt} and h_{gt} are the width and height of the prediction frame. With the introduction of the weight value and aspect ratio, it is possible to take into account the differences in the shape of the target and the importance of different sizes of targets, thus improving the performance and accuracy of the target detection algorithm.

$$L_{\text{CIoU}}=L_{DIoU}+\alpha V=1-IoU+\frac{\rho^2\left(b,b^{gt}\right)}{c^2}+\alpha V \tag{13}$$

$$V=\frac{4}{\pi^2}\cdot\left(\arctan\frac{w_{gt}}{h_{gt}}-\arctan\frac{w}{h}\right)^2 \tag{14}$$

Since CIoU only reflects differences in aspect ratios, it does not accurately reflect the true difference between the respective widths of the bounding box lengths and their confidence levels. This means that in some cases, CIoU may lead to a certain degree of ambiguity. In this paper, we use EIoU (Efficient IoU loss) [24], which is based on CIoU, and measures the positional matching of the target detection frames by introducing centroid offset and aspect ratio offset to accurately measure the degree of positional matching between the target detection frames and the real bounding box, to speed up the regression of prediction frames and to make the regression process pay more attention to high-quality anchor frames, so as to improve the regression accuracy of the prediction frame. The EIoU loss function is calculated as shown in Eq. (15).

$$L_{EIoU}=1-IoU+\frac{\rho^2\left(b,b^{gt}\right)}{c^2}+\frac{\rho^2\left(w,w^{gt}\right)}{c_w^2}+\frac{\rho^2\left(h,h^{gt}\right)}{c_h^2} \tag{15}$$

where c_w and c_h are the width and height of the smallest outer rectangle of the prediction and real boxes.

2.3. Hybrid Convolutional Lightweight Improvement

Due to the network training characteristics, this paper introduces hybrid convolution to achieve model lightweighting and reduce the parameters of the network structure while maintaining the relatively excellent performance of the network. The traditional convolution operation introduces a large number of parameters and computational overheads when dealing with high-dimensional input data, and each convolution kernel needs to learn a set of weight parameters for the convolution operation with a channel, so this channel-by-channel convolution leads to an increase in the number of parameters, and when the number of channels of the input data is large, the number of parameters grows linearly, which increases the complexity of the model and the computational overheads. For the traditional convolution operation, it also introduces a large number of parameters and computational overheads when dealing with high-dimensional input data. To solve this problem, this paper introduces Depthwise Separable Convolution(DeepthSepConv), which decomposes the convolution operation into Depthwise Convolution and Pointwise Convolution to achieve hybrid Convolution [25, 26].

In order to balance accuracy and speed, GSCnov is introduced [21]. GSConv consists of a hybrid convolution of a standard convolution, a depth-separated convolution, and a shuffle operation.

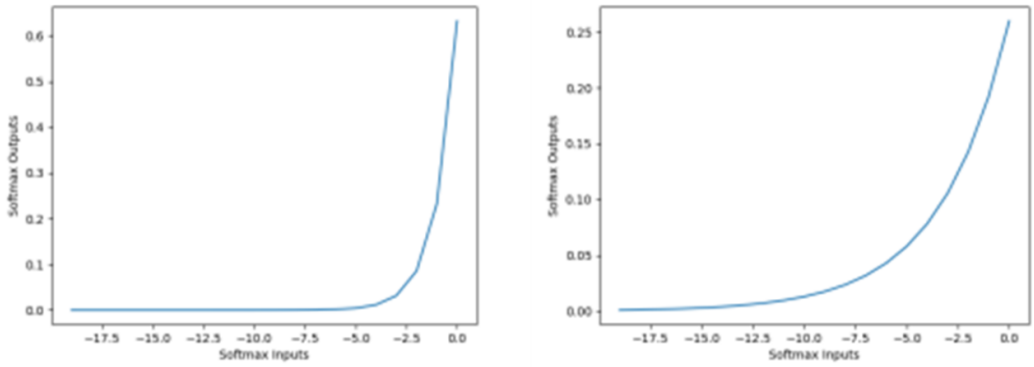

Fig. 3. Function Image.

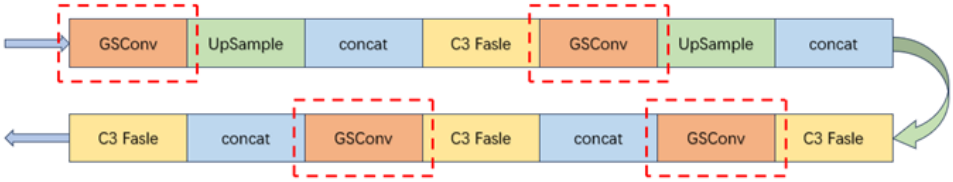

Fig. 4. Neck network after joining GSConv.

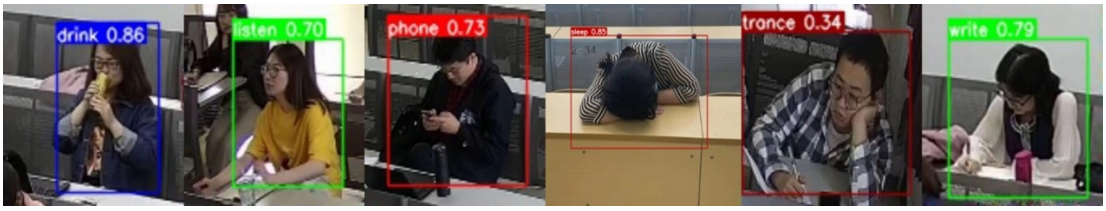

Fig. 5. Detection image.

GSCnov first downsamples the input and uses deep convolution to learn the feature relationships between different input channels, while point-by-point convolution.Then the features within each channel are learnt. Finally, the information generated by the standard convolution is infiltrated into the information generated by the depth-separable convolution through the Shuffle operation, which interpenetrates the information to provide a more comprehensive and enriched feature representation.

However, if all the standard convolutions of the model are replaced with GSCnov lightweight convolutions, it will make the network layers deeper, which will increase the resistance to data flow and may lead to an increase in inference time. However, by the time the feature map reaches the Neck stage, the channel dimension is maximal and the width dimension is minimal, eliminating the need for further transformations, thus reducing the computational burden and increasing the inference efficiency of the model. Meanwhile the feature map has the largest channel dimension at the Neck stage, and GSConv can better utilise the relationship between these channels to obtain richer and more useful features. As shown in Fig. 4, the structure of the Neck network Neck with the addition of GSCnov is shown.

2.4. Detection frame optimisation

Due to the specificity of the classroom behaviour detection application scenario, the density and similarity of the targets to be detected are high, and the interference is strong, so the observability of the multi-target detection frame needs to be improved. Therefore, in the model of the prediction head network to change the detection frame RGB value, the detection frame detection flexibility on the optimisation. According to the nature of this paper, the student behaviour is divided into three sets, respectively, the positive set, writing (write), listening (listen); neutral set, drink-

ing (drink); negative set, distraction (trance), sleep (sleep), play mobile phone (phone). Different coloured detection boxes were assigned to the three sets. A green detection box was assigned to the positive set to represent that the student was actively participating in the class. The neutral set was given a blue detection box, representing that the student's situation was unknown. The negative collection was given a red detection box, which is a warning that the student is disengaging from the class. Examples of the detection boxes are shown in Fig. 5, which allows for quick differentiation of student behaviours in the classroom based on the colour of the detection box, significantly improving observability.

3. Results and discussion

Analysis of experimental results. The experiments in this paper were conducted on a Windows 11 operating system with an AMD Ryzen 7 5800H CPU (3.20 GHz) and 16 GB of RAM. The graphics processing unit is an NVIDIA GeForce RTX 3060 Laptop GPU with 6 GB of video memory. The compilation language was Python 3.9, combined with the PyTorch 1.1 deep learning framework for model construction and training. CUDA version 11.6 was used to accelerate the deep learning algorithms.

3.1. Data set

The dataset used in this paper is the camera live capture data, the dataset contains 299 sequences, the target object abnormal behaviour contains drink (drink), listen to the class (listen), play mobile phone (phone), sleep (sleep), distraction (trance) and writing (write) six categories. As the behaviour of trance and listening is extremely similar, in the data on the monitoring of a posture for a long time an expression of the students defined as trance, in a certain period of time the student has the action behaviour is determined as listening, so the network model in addition to behavioural recognition and expression and demeanor recognition. In the pre-processing stage, the training set, validation set and test set are divided in the ratio of 8:1:1. The labelling is carried out by labelimg software.

3.2. Assessment of indicators

In this paper, the model performance is evaluated using Mean Average Precision Recall, Precision and mean Average Precision (mAP) and Frames Per Second (FPS). The calculation formulas are shown in Eqs. (16) and (17).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

The P-R curve is a curve drawn with Precision of a category as the vertical axis and Recall as the horizontal axis. The area enclosed by the curve and the axis represents the average precision AP of the defects in the category. By calculating the AP of the defects in all categories and averaging them, we can get the average precision mean value mAP.The calculation formula is shown in Eqs. (18) and (19). Where P denotes Precision, R denotes Recall, N denotes the total number of categories, and AP_i is the corresponding defect category.

$$AP = \int_0^1 P(R)dR \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{19}$$

3.3. Ablation experiment

In order to verify the effectiveness of the algorithm five sets of controlled experiments were designed for analysis and testing. The results of the experiments are shown in Table 1, Experiment 1 takes no measures for the original YOLOv5, Experiment 2 adds the FCA attention mechanism, Experiment 3 adds the EIoU, Experiment 4 adds both of the above two improvements, and Experiment 5 adds the GSCnov lightweight convolution on the basis of Experiment 4.

From Table 1, it can be seen that the improved algorithm in this paper has obvious advantages, in terms of average accuracy Map is improved by 2.9% over the original model, which is a substantial improvement. In the amount of computation is reduced by 1.9%, the number of parameters is reduced by 4.3%, so that the model is more lightweight, and the detection speed FPS is improved by 4.5.

As shown in Fig. 6, it is a comparison chart of the detection effect of the intercepted test set, Fig. 6(a) the group of images is the original YOLOv5, and Fig. 6(b) the group of images is the improved algorithm of this paper.

It can be seen that the improved model has an obvious advantage in confidence level. For Fig. 6(a4) the original network shows misdetection, Fig. 6(b4) the improved network results are improved. In Fig. 6(b6), the improved network recognises both phone and write, which is consistent with the actual situation. The observability of the detection frames is also improved, with the negative behaviours represented by red, i.e. classroom anomalies, the neutral behaviours represented by blue, and the positive behaviours represented by green being more visible and clearer.

3.4. Comparison experiment

In order to explore the performance of the model continued to do six sets of comparison experiments, from Table 2 can

Table 1. Comparative experiments.

Model	AP/%					
	Drink	Listen	Phone	Sleep	Trance	Write
YOLOv5s	90.1	68.7	57.3	97.5	84.5	83.2
YOLOv5s+FCA	89.4	78.6	79.3	98.2	69.5	81.8
YOLOv5s+E	86.3	82.8	57.9	98.5	92.0	79.2
YOLOv5s+FCA+E	89.6	79.4	70.2	99.1	72.3	90.5
YOLOv5s+FCA+E+G	90.6	70.7	74.6	98.8	71.1	92.9
	↑ 0.5%	↑ 2%	↑ 17.3%	↑ 1.3%	↓ 13.4%	↑ 9.7%

Model	mAP/%	Quantity of participants/M	Computational volume/GFLOPs	Detection speed / FPS
YOLOv5s	80.2	7.0	15.8	59.8
YOLOv5s+FCA	82.9	7.2	16.0	52.2
YOLOv5s+E	82.8	7.0	15.8	60.2
YOLOv5s+FCA+E	83.5	7.2	16.0	52.1
YOLOv5s+FCA+E+G	83.1	6.7	15.5	65.3
	↑ 2.9%	↓ 0.3%	↓ 0.3%	↓ 5.5%

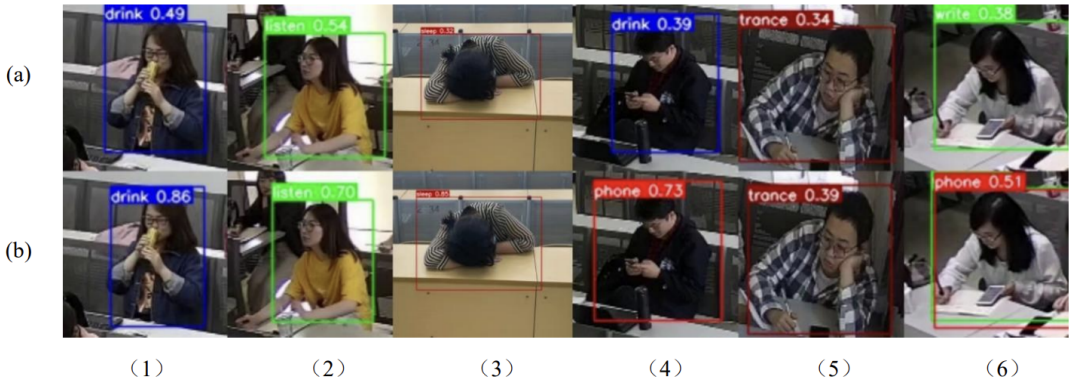

Fig. 6. Comparison of detection images.

Table 2. Performance comparison of mainstream models.

Model	AP/%						mAP	FPS
	Drink	Listen	Phone	Sleep	Trance	Write		
FasterR-CNN	88.9	66.2	71.8	93.1	80.3	90.0	81.7	22.7
SSD	88.7	72.2	60.7	95.3	74.4	82.6	79.0	67.1
YOLOv3	78.5	60.7	55.2	91.4	54.8	81.9	70.4	47.0
YOLOv4	88.8	63.2	52.0	92.5	76.3	86.7	76.6	52.2
YOLOv5s	90.1	68.7	57.3	97.5	84.5	83.2	80.2	59.9
Ours	**90.6**	**70.7**	**74.6**	**98.8**	**71.1**	**92.9**	**83.1**	**65.3**
	↑ **0.5%**	↑ **2%**	↑ **17.3%**	↑ **1.3%**	↓ **13.4%**	↑ **9.7%**	↑ **2.9%**	↑ **5.4%**

be seen that the algorithm improved in this paper and the current mainstream algorithms compared to the more advantageous in terms of comprehensive performance. Compared to the two-stage algorithm FasterR-CNN [27, 28], the accuracy is 1.4% ahead, and the detection speed is 42.6% ahead. compared to the previous generation of YOLOv4 and YOLOv3 [29], the detection speed and accuracy are far ahead. Compared to another one-stage algorithm SSD [30], although slightly lower in detection speed, the accuracy is 4.1% ahead, which is more advantageous.

4. Conclusion

In this paper, an improved classroom abnormal behaviour recognition algorithm based on YOLOv5 is proposed, firstly, a coordinate attention mechanism based on fractional order is introduced in the backbone network Backbone, so that the model pays more attention to the region of interest and improves the detection accuracy; secondly,

the EIoU loss function used is put into use for training optimization, which improves the regression accuracy of the detection frame, and further improves the detection accuracy; at the same time, a GSCnov lightweight convolution replaces the standard convolution in the neck network to reduce the amount of computation and the number of parameters, and improve the detection speed; finally, for the complexity of the application scenarios and the problem of high similarity and interference of the target to be detected in the classroom abnormal behaviour detection, we classify the student's behaviours into three sets by the nature of the behaviours, which are the positive set, the neutral set, and the negative set, respectively, and realise the observability of the abnormal behaviour detection in the classroom. Optimisation of classroom abnormal behaviour detection. Experiments show that the improved algorithm in this paper has significant improvement in both accuracy and speed. In future work, we can continue to explore the lightweight convolution with more advantageous performance, try to improve the detection speed without losing the accuracy, and perhaps change the idea to explore more suitable data enhancement methods to improve the detection performance of the model without changing the premise of the model.

References

[1] L. Shi and X. Di, (2023) *"A recognition method of learning behaviour in English online classroom based on feature data mining"* **International Journal of Reasoning-based Intelligent Systems *15***: 8–14. DOI: 10.1504/IJRIS.2023.128375.

[2] T. Guo, X. Bai, X. Tian, S. Firmin, and F. Xia, (2022) *"Educational anomaly analytics: features, methods, and challenges"* **Frontiers in big Data *4***: 811840. DOI: 10.3389/fdata.2021.811840.

[3] Y. Liu, H. Chen, and A. Thoff, (2020) *"Research on evaluation method of students' classroom performance based on artificial intelligence"* **International Journal of Continuing Engineering Education and Life Long Learning *30***: 476–491. DOI: 10.1504/IJCEELL.2020.110925.

[4] Y. Xie, S. Zhang, and Y. Liu, (2021) *"Abnormal Behavior Recognition in Classroom Pose Estimation of College Students Based on Spatiotemporal Representation Learning"* **Traitement du Signal *38***: 89–95. DOI: 10.18280/ts.380109.

[5] X. Zhang, (2022) *"A Gaussian High-Dimensional Random Matrix-Based Method for Detecting Abnormal Student Behaviour in Chinese Language Classrooms"* **Mathematical Problems in Engineering *2022***: 6957097. DOI: 10.1155/2022/6957097.

[6] S. Zhang, H. Liu, C. Sun, X. Wu, P. Wen, F. Yu, and J. Zhang, (2023) *"MSTA-SlowFast: A student behavior detector for classroom environments"* **Sensors *23***: 5205. DOI: 10.3390/s23115205.

[7] M. A. E. Abbas and S. Hameed, (2022) *"A Systematic Review of Deep Learning Based Online Exam Proctoring Systems for Abnormal Student Behaviour Detection"* **International Journal of Scientific Research in Science, Engineering and Technology *9***: 192. DOI: 10.32628/IJSRSET229428.

[8] C. Pabba and P. Kumar, (2022) *"An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition"* **Expert Systems *39***: e12839. DOI: 10.1111/exsy.12839.

[9] J. Zhang, Z. Zhang, L. Guan, and H. Hu. *Research on Classroom Behavior Recognition and Detection Method Based on Deep Learning*. 2024. DOI: 10.1109/cvidl62147.2024.10604092.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, 779–788. DOI: 10.48550/arXiv.1506.02640.

[11] J. Redmon and A. Farhadi. "YOLO9000: better, faster, stronger". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 7263–7271. DOI: 10.1109/CVPR.2017.690.

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

[13] J. Wen, Y. Qin, and S. Hu. "Abnormal behavior identification of examinees based on improved YOLOv5". In: *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022)*. ***12604***. 2022, 946–953. DOI: 10.1117/12.2674630.

[14] Z. Zhang, D. Ao, L. Zhou, X. Yuan, and M. Luo. "Laboratory behavior detection method based on improved Yolov5 model". In: *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*. 2021, 1–6. DOI: 10.1109/ICCSI53130.2021.9736251.

[15] F. Lei, F. Tang, and S. Li, (2022) *"Underwater target detection algorithm based on improved YOLOv5"* **Journal of Marine Science and Engineering *10***: 310. DOI: 10.3390/jmse10030310.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, 580–587. DOI: 10.1109/CVPR.2014.81.

[17] Q. Zheng, M. Yang, X. Tian, X. Wang, and D. Wang, (2020) *"Rethinking the Role of Activation Functions in Deep Convolutional Neural Networks for Image Classification"* **engineering letters *28***: 80.

[18] J. Bai, J. Dai, Z. Wang, and S. Yang, (2022) *"A detection method of the rescue targets in the marine casualty based on improved YOLOv5s"* **Frontiers in Neurorobotics *16***: DOI: 10.3389/fnbot.2022.1053124.

[19] C. Chen, F. Wang, Y. Cai, S. Yi, and B. Zhang, (2023) *"An improved YOLOv5s-based Agaricus bisporus detection algorithm"* **Agronomy *13***: 1871. DOI: 10.3390/agronomy13071871.

[20] M. Gong, D. Wang, X. Zhao, H. Guo, D. Luo, and M. Song. "A review of non-maximum suppression algorithms for deep learning target detection". In: *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*. ***11763***. SPIE, 2021, 821–828.

[21] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, (2022) *"Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles"* **arXiv**: DOI: 10.48550/arXiv.2206.02424.

[22] Q. Hou, D. Zhou, and J. Feng. "Coordinate attention for efficient mobile network design". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, 13713–13722. DOI: 10.1109/CVPR46437.2021.01350.

[23] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. "Distance-IoU loss: Faster and better learning for bounding box regression". In: *Proceedings of the AAAI conference on artificial intelligence*. ***34***. 07. 2020, 12993–13000. DOI: 10.1609/aaai.v34i07.6999.

[24] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, (2022) *"Focal and efficient IOU loss for accurate bounding box regression"* **Neurocomputing *506***: 146–157. DOI: 10.48550/arXiv.2101.08158.

[25] Z. Y. Khan and Z. Niu, (2021) *"CNN with depthwise separable convolutions and combined kernels for rating prediction"* **Expert Systems with Applications *170***: 114528. DOI: 10.1016/j.eswa.2020.114528.

[26] H. Srivastava and K. Sarawadekar. "A depthwise separable convolution architecture for CNN accelerator". In: *2020 IEEE Applied Signal Processing Conference (ASPCON)*. 2020, 1–5. DOI: 10.1109/ASPCON49795.2020.9276672.

[27] R. Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, 1440–1448. DOI: 10.1109/ICCV.2015.1691.

[28] S. Ren, K. He, R. Girshick, and J. Sun, (2016) *"Faster R-CNN: Towards real-time object detection with region proposal networks"* **IEEE transactions on pattern analysis and machine intelligence *39***: 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[29] J. Redmon and A. Farhadi, (2018) *"YOLOv3: An Incremental Improvement"* **ArXiv *abs/1804.02767***: DOI: 10.48550/arXiv.1804.02767.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. 2016, 21–37. DOI: 10.48550/arXiv.1512.02325.