Behavior Analysis Of Students In Preschool Mathematics Teaching Based On Deep Learning

Guangning Qin

School of Music and Dance, Zhengzhou University of Science and Technology, Zhengzhou 450064 China

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

Received: Apr. 19, 2025; Accepted: Jun. 06, 2025

Students behaviors can directly reflect the quality of the classroom. Analyzing and evaluating classroom behaviors through artificial intelligence and deep learning is conducive to improving teaching quality. The traditional methods for identifying students classroom behaviors involve that teachers directly observe students states or analyze them through surveillance videos after class. These methods are time-consuming, labor-intensive, and have a low recognition rate, making it difficult to reflect the problems existing in the classroom and exams in real time. To solve this problem, this paper proposes a novel students classroom behaviors based on YOLOv8 deep learning model. Combining the channel attention mechanism with deep convolution, a dynamic channel attention convolution (DCAConv) is proposed, which can dynamically adjust the channel weights and capture key features more sensitively. It introduces multi-scale convolutional attention (MSCA) to maximize the ability of mining multi-scale convolutional features through element multiplication, and enhance the attention to spatial details. Meanwhile, a multi-scale context fusion (MSCF) module is constructed. Through convolution and self-attention mechanism, multi-scale feature fusion is enhanced. Adding a small target detection layer and extracting local features from larger-sized feature maps significantly improves the ability to recognize the behaviors of students in the back row. The experimental results show that the average recognition accuracy rate of the proposed behavior recognition method for various parts of the human body can reach up to 83.7% at most. The recognition rates for various behaviors in simple and crowded scenarios reach more than 92.1% and 86.3% respectively, and it can effectively recognize various behaviors in the classroom.

Keywords: Behavior analysis, YOLOv8, deep learning, channel attention mechanism, multi-scale context fusion

http://dx.doi.org/10.6180/jase.202603_29(3).0009

1. Introduction

Today, with the rapid development of educational technology, the analysis of students classroom behaviors has become crucial for improving the quality of education [1]. However, traditional classroom observation and assessment methods have obvious limitations due to their significant consumption of time and human resources. Human behavior recognition is the most challenging and attractive research topic in the field of computer vision, and it has strong application values in various cross-disciplinary fields, such as video understanding, intelligent monitoring, robotics, human-computer interaction, industrial automation, healthcare and smart education [2]. A great deal of work has been done in the field of behavior recognition, but many challenges still exist. The student behavior recognition method mainly records and analyzes the specific behaviors of students through video surveillance. The specific behaviors of students include raising hands gestures [3], looking around [4], dozing behaviors, etc,. The hand-designed visual feature modeling method can identify students behaviors, as well as evaluate teaching

quality and students' learning attitudes. This method is time-consuming, labor-intensive and has a low accuracy rate. Therefore, student behavior recognition requires an accurate intelligent system. Deep learning methods can identify the key points (human postures) of the skeleton in videos to complete behavior recognition with a relatively high recognition accuracy rate, and have become a research hotspot in the field of computer vision [5].

Human behavior recognition based on key points of the skeleton is usually divided into two methods: top-down and bottom-up. The top-down method first uses the object detection algorithm to obtain the single-person detection box, and then uses the single-person posture estimation (SPPE) algorithm to find the key points of the human skeleton. Through the key points of the skeleton, specific behaviors are further identified. The bottom-up method has the opposite idea to the top-down method. It first detects the key points of all the people in the image, and then matches the key points with the target people through the algorithm. Bottom-up method is difficult to utilize the global context information. For this problem, Cao et al. [6] proposed the real-time multi-person two-dimensional pose recognition method (OpenPose). Li et al. [7] proposed a human behavior recognition method based on OpenPose and recurrent neural networks, which had a relatively high overall accuracy on public activity datasets. Lu et al. [8] proposed a child behavior recognition method based on convolutional neural networks and OpenPose. Based on this method, 13 kinds of behaviors were classified, and the accuracy rate reached 82.3%. Most research results show that the top-down method has a higher recognition accuracy than the bottom-up method. However, the object detection algorithm is prone to errors in positioning and recognition in crowded or occluding scenes such as classrooms, resulting in incorrect recognition of single-person poses. Therefore, Fang et al. [9] proposed the regional multi-person pose-estimation (RMPE) framework, namely the AlphaPose framework. This framework could effectively handle inaccurate detection boxes and redundant detection results, thereby improving the recognition accuracy. The above methods have problems of low recognition rate and poor real-time performance in complex occlusion and crowded scenarios with multiple people, and are not applicable to classroom video surveillance scenarios [10].

In the research of the smart classroom field, the recognition of students classroom behaviors has always been a core issue that attracts much attention from the academic community. With the rapid development of computer storage and deep learning technologies, more and more scholars have introduced deep learning technologies into the task of student classroom behavior recognition to improve the accuracy and efficiency of recognition. Yongqing and Dan [11] produced five datasets of student behaviors, namely listening, sleeping, raising hands, answering and writing, and proposed an improved SSD model, combined with the k-means clustering algorithm for student behavior recognition.

With the emergence of the YOLO series of object detection algorithms, Pabba et al. [12] deeply analyzed students behaviors by extracting the global features of human poses using the OpenPose algorithm and combining the local features of interactive objects extracted by the YOLOv3 algorithm. Jia and He [13] further improved the accuracy of student behavior detection through a human skeleton behavior recognition model that integrated the global attention mechanism and the spatio-temporal graph convolutional network. Furthermore, Xie et al. [14] utilized the YOLOv5 object detection model and the CA (coordinate attention) mechanism to accurately identify various behaviors of students in the classroom. The aforementioned research provided a direction for the production of behavior datasets and the selection of behavior recognition networks for students classroom behavior recognition.

Usually, the cameras in smart classrooms are located on the upper left or right of the classroom, resulting in student targets of different scales in the video, significant size differences between students in the front and back rows, and mutual occlusion among students in the back rows. Therefore, the model needs to have multi-scale awareness, pay attention to different behaviors at different positions, and be able to make full use of the information around the target to infer the information of the occluded part. The main contributions of this article can be summarized as follows.

1. This paper proposes to construct the dynamic channel attention convolution (DCAConv) by combining the CA attention mechanism, enabling the model to capture the key input features more sensitively. Meanwhile, the MSCA attention mechanism is introduced at the end of the YOLOv8 backbone network. By using a simple element multiplication operation, the multi-scale convolutional features are maximized to enhance the attention to spatial details.

2. A multi-scale context fusion (MSCF) module is proposed by combining the gathering-and-distribute (GD) mechanism of information. The MSCF module is operated through convolution and self-attention mechanisms, enhancing the fusion ability of multi-scale features. For the small targets, a small target detec-

tion layer is added to enhance the extraction of local features of larger-sized feature maps.

2. Materials and methods

The proposed behavior recognition model in this paper optimizes the structure of the original YOLOv8 model [15]. The improved model performs better than the previous model in small object detection, and at the same time achieves a balance between model lightweight and performance. And it has been verified on multiple datasets. The network structure diagram proposed in this paper is shown in Fig. 1.

The new model is mainly composed of four components: input, backbone, head and output. At the input end, Mosaic data augmentation is adopted. Meanwhile, data augmentation methods such as adaptive image scaling and stitching are employed to enhance the generalization ability of the model. In this paper, dynamic channel attention convolution (DCAConv) is constructed based on the channel attention (CA) mechanism, which is used to replace some Conv modules in the backbone network to improve the ability to capture the key information in the input features. At the end of the backbone network, a multi-scale convolutional attention (MSCA) module is introduced to enhance the attention and generalization performance of the model.

Meanwhile, in this paper, a multi-scale context fusion module (MSCF) is proposed by integrating the GD mechanism, and the neck structure of YOLOv8 is improved by using the MSCF module. The MSCF module integrates multi-level features and injects them into a higher-level network, significantly enhancing the information fusion ability of the head network and the feature expression ability of the model. In order to enhance the model's detection ability for small targets, a small target detection layer is added at P2 of the head network. The position information and feature details of the targets are further extracted, reducing the false detection and missed detection rates of the students and improving the detection accuracy of the students with small targets in the back row.

Finally, in order to solve the problem of the mismatch between the number of difficult samples and simple samples in the data set, slide loss is introduced to improve the loss function of YOLOv8. This improvement optimizes the processing methods of different difficulties samples by dynamically adjusting the weights of the loss function.

2.1. DCAConv module

Convolution operation is the basic component for building convolutional neural networks. It uses sliding windows

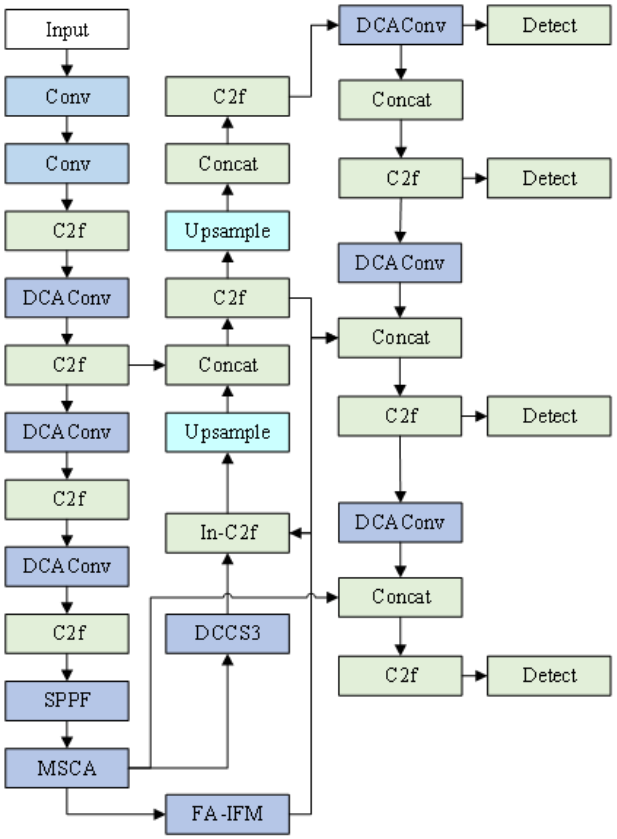

Fig. 1. Proposed behaviour recognition.

with shared parameters to extract feature information, overcoming the problems of multiple parameters and high computational overhead in fully connected layers when building neural networks.

Suppose the original image is a $n \times n$ matrix and the size of the convolution kernel is $f \times f$, then the size of the matrix obtained after the convolution operation is $(n-f+1) \times (n-f+1)$. The convolution kernel slides on the input image and calculates the weighted sum, thereby obtaining the corresponding pixel value of the output image. However, the standard convolution operation has a limitation, that is, it cannot capture the information differences brought by different positions, which limits the performance of convolutional neural networks to a certain extent.

The CA (coordinate attention) mechanism concatenates the channel attention into two one-dimensional feature encoding and aggregates the features in different directions. Meanwhile, the generated feature maps are encoded respectively to form a pair of direction-sensitive and location-sensitive feature maps, which can be complementary to the input feature maps to enhance the representation of the target of interest. The advantage of tandem lies in captur-

ing long-range dependencies in one spatial direction while retaining precise positional information in another spatial direction. The introduction of this mechanism is conducive to improving the understanding and representation ability of neural networks for input information.

To solve the problem that traditional convolution cannot capture the information differences brought by different positions, this paper combines the improved CA with convolution and proposes the dynamic channel attention convolution (DCAConv) to replace some of the traditional convolutions in YOLOv8. The network structure of DCAConv is shown in Fig. 2. DCAConv scales the attention weights of CA channels by calculating dynamic weights and combines them with traditional convolutions. This design resolves the limitations of traditional convolution in capturing information at different positions.

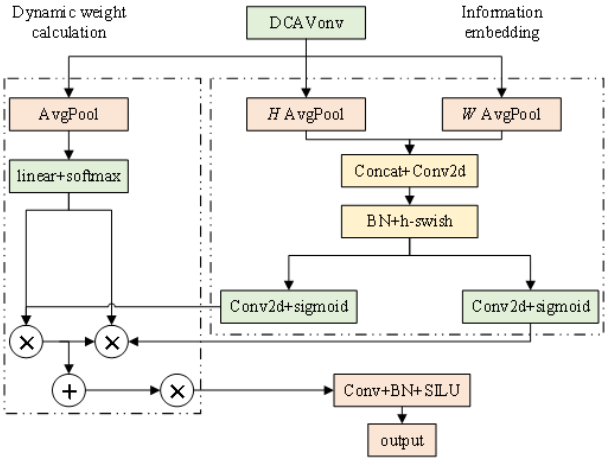

Fig. 2. DCAConv network structure.

DCAConv adaptively allocates the attention weights in the horizontal and vertical directions by introducing a dynamic weight allocation mechanism, and performs weighted fusion of the attention in these two directions, enabling the network to emphasize the features in different directions more flexibly, thereby improving the feature extraction ability and enhancing the richness of feature expression. The dynamic weight allocation mechanism can better capture the important feature information in different directions, while the weighted fusion method enables the network to integrate the feature information in different directions, improving the accuracy and effectiveness of feature extraction. Although DCAConv introduces dynamic weight calculation, the parameter increment and computational cost can be ignored, while significantly improving the network performance.

1. Coordinate information embedding

The input feature map with the size of $C \times H \times W$ is pooled in the X and Y directions to generate the global spatial information z^h and z^w with the sizes of $C \times H \times 1$ and $C \times 1 \times W$ respectively.

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < w} x(h, i) \tag{1}$$

$$z_C^w(w) = \frac{1}{H} \sum_{0 \le j < w} x(j, w) \tag{2}$$

Where W represents the width value of the input feature, and H represents the height value of the input feature. $z_c^h(h)$ represents the pooled output of the c-th channel with a height of $h.z_c^w(w)$ represents the pooled output of the c-th channel with a width of w. $x(h, i)$ and $x(j, w)$ represent the eigenvalues of the corresponding coordinates of the feature map.

Eq. (3) performs dimension transformation on the feature map z^h and then concatenates it with z^w, the aim is to preform information fusion and transformation of the channel dimensions.

$$F = \delta\left(BN\left(F_1\left(\left[z^h, z^w\right]\right)\right)\right) \tag{3}$$

Then, for the concatenated feature map $\left[z^h, z^w\right]$, it performs the F_1 operation (1×1 convolution), BN operation (normalization), the δ operation (sigmoid activation) in sequence to generate the feature map $F \in R^{C/r \times (H+W) \times 1}$. This operation is to fuse the information vectors extracted from the horizontal and vertical directions to integrate the information in these two directions and enhance the feature expression. Then, the processed feature map F is segmented into f^h and f^w along the channel dimension. Then, 1×1 convolution is used respectively for dimension increase. Finally, combined with the sigmoid activation function, the attention weight vectors $q^w \in R^{C \times 1 \times W}$ and $q^h \in R^{C \times H \times 1}$ in two directions are obtained, as shown in Eqs. (4) and (5).

$$q^h = \sigma\left(F\left(f^h\right)\right) \tag{4}$$

$$q^w = \sigma\left(F\left(f^w\right)\right) \tag{5}$$

2. Calculation of dynamic weights

$$Y^W, Y^H = q^w \times Y[0], q^h \times Y[1] \tag{6}$$

$$Y1 = X \times \left(Y^W + Y^H\right) \tag{7}$$

Eqs. (6) and (7) respectively perform global average pooling on the input feature maps, flatten them, and obtain the final dynamic weights $Y[0] \in R^{1\times1\times1}$ and $Y[1] \in R^{1\times1\times1}$ through the linear layer and the Softmax function. Then it multiplies and adds the dynamic weights with q^h and q^w to obtain $Y^H \in R^{C\times H\times1}$ and $Y^W \in R^{C\times1\times W}$. This operation scales the channel attention weights based on dynamic weights, enabling the network to capture input features more flexibly. Finally, it multiplies the obtained channel attention weights with the input feature map to obtain the output $Y1 \in R^{C\times H\times W}$.

3. Output

Finally, convolution, normalization, and SILU activation are performed successively on the above output result $Y1$ to obtain the final output Y, as shown in Eq. (8).

$$Y = \text{SILU}(BN(F(Y1))) \tag{8}$$

DCAConv generates dynamic weights through global average pooling and uses these dynamic weights to adjust the features in the CA mechanism. Such a design enables the network to adapt more flexibly to the changes of input data, accurately capture the key information in the input features, improve the robustness and generalization ability of the model. At the same time, it solves the problem that traditional convolution cannot capture the differences in information at different positions.

2.2. MSCA

In order to enhance the feature extraction ability of the YOLOv8 backbone, MSCA is introduced at the end of the backbone network in this paper. MSCA can effectively capture and utilize the multi-scale information of the input features, making the model more accurate when dealing with targets of different sizes and complexities. MSCA enhances the expressive ability of feature maps by generating and fusing multi-dimensional channel weights, enabling the model to better focus on important channel information and thereby improving the overall performance. Meanwhile, MSCA further enhances the expressive ability of the original feature maps by weighting them, and improves the robustness of the model in the face of complex scenes and occlusions.

Through the generation of attention for multi-scale information capture channels by MSCA, the details and key features of the input image are better reflected, the fusion of multi-dimensional features is improved, and the network is more flexible and accurate when processing targets of different dimensions. At the same time, it also improves the robustness of the detection and significantly enhances the feature expression ability of YOLOv8. MSCA consists of three parts: depth-separable convolution (for obtaining local information), multi-branch depth-separable convolution (for capturing multi-scale context), and 1×1 convolution (for modeling the relationship between different channels). Taking the output of 1×1 convolution directly as the attention weight and the output weight of MSCA. MSCA can be described as:

$$M = Conv_{5\times5}(Input) \tag{9}$$

$$M_1 = DWConv_{7\times1}\left(DWConv_{1\times11}(M)\right) \tag{10}$$

$$M_2 = DWConv_{11\times1}\left(DWConv_{1\times11}(M)\right) \tag{11}$$

$$M_3 = DW\ Conv_{21\times1}\left(DWConv_{1\times21}(M)\right) \tag{12}$$

$$Output = Conv_{1\times1}\left(M + \sum_{i=1}^{3} M_i\right) \otimes Input \tag{13}$$

Here, $Input$ represents the input, $DWConv_{1\times i}$ and $DWConv_{i\times1}$ represent depth-separable convolution. M_1, M_2, and M_3 represent three branches, and each branch performs convolution processing on the input using convolution kernels of different sizes. In each branch, two depth-separable convolution are adopted to approximate the standard-depth convolution with a large kernel. The kernel sizes of the three branches are set to 7,11 , and 21 respectively. Finally, the results of M and $\sum_{i=1}^{3} M_i$ are processed through 1×1 convolution kernels to establish the connections between different channels. The processing result is used as a weight to weight the Input and output the Output. The MSCA utilizes convolution kernels of different scales and depth-separable convolution to maximize the mining of feature information at different scales to enhance the attention to spatial details. The introduction of MSCA has enhanced the detection ability of the YOLOv8 network for student behavior recognition.

2.3. Multi-scale context fusion module (MSCF)

The neck structure of the YOLO series is inspired by the feature pyramid network (FPN), achieving multi-scale feature fusion by combining multiple branches together. However, FPN only fuses the features of adjacent layers, and the features of other layers can only be recursively fused through adjacent layers. This approach limits the network

performance. To solve this problem, PANet introduces a bottom-up path to make the information fusion between different levels more thorough and introduces additional connections to increase the information flow between different levels. EfficientDet proposed the Repeatable Module (BiFPN), which improves the quantity and quality of feature maps by repeatedly applying FPN, thereby enhancing the performance of object detection.

In order to further enhance the multi-scale feature fusion ability, inspired by the GD mechanism, this paper constructs a multi-scale context fusion (MSCF) module. The MSCF module is implemented through convolution and self-attention mechanism operations, enhancing the fusion ability of multi-scale features of the model. Compared with FPN, the MSCF module achieves efficient information exchange by globally fusing multi-level features and injecting global information into a higher level. In order to further enhance the multi-scale feature fusion ability, inspired by the GD mechanism, this paper constructs a MSCF module. The MSCF module is implemented through convolution and self-attention mechanism operations, enhancing the fusion ability of multi-scale features of the model. Compared with FPN, the MSCF module achieves efficient information exchange by globally fusing multi-level features and injecting global information into a higher level.

The implementation module of MSCF includes three small modules: feature fusion module (FA_IFM), information fusion module (DCCS3) and information injection module (In_C2f). Among them, the FA-IFM module is responsible for fusing and further extracting the features from layers P2 to P5. The DCCS3 module handles the fusion and extraction of features from layers P3 to P5. Finally, through the In_C2f module, the features of the FA_IFM and DCCS3 output modules are fused. The combination of MSCF and YOLOv8 makes better use of multi-scale feature information, improves the utilization efficiency of multi-scale features by the object detection model, and enhances the network performance.

3. Results and discussion

3.1. Dataset and evaluation metrics

This paper first evaluates the proposed behaviour recognition method in this paper on the MSCOCO Key-points dataset and the MPI dataset. The MSCOCO Key-points dataset contains approximately 150000 training samples and 80000 test samples. The MPII dataset includes 25000 labeled images over 40000 people, and occluded body parts are also included in the test set. Based on the evaluation index of MSCOCO, this paper uses average precision (AP) to evaluate the experimental results. At present, the most commonly used is the OKS (Object Key-point Similarity) metric, which is similar to IoU in object detection algorithms. To calculate the true value and predict the similarity of human key points, the calculation formula of OKS is shown as Eq. (14):

$$OKS = \frac{\sum_i \exp\left\{-d_{pi}^2/2S_p^2\sigma_i^2\right\}\delta\left\{v_{pi}=1\right\}}{\sum_i \delta\left(v_{pi}=1\right)} \tag{14}$$

Where p represents the actual id of the human body. i is the id of the key point. d_{pi} is the Euclidean distance between the true value and the predicted key point. S_p^2 represents the area occupied by the human body in the ground truth. σ_i represents the normalization factor of the i-th key point. The larger σ_i denotes that the key point is more difficult to label. v_{pi} represents whether the i-th key point of the p-th person is visible. This paper uses mAP as the evaluation index. mAP represents the average value of AP under the conditions of OKS $= [0.50 - 0.95]$, OKS $= 0.5$ and OKS $= 0.75$. The higher OKS denotes the greater threshold, the more stringent requirements, and the corresponding AP accuracy is lower.

In this paper, experiments are conducted using Anaconda3.0 and the PyTorch framework on a 64-bit notebook with memory 16 GB , equipped with a 2.6 GHz processor and a GTX 3060 graphics card. The resolution of the input image is set to 256×192 pixels. The learning rate is 1×10^{-4} within 100 epochs, changed to 1×10^{-5} after 100 epochs, the batch size is set to 64, and the RMSprop optimizer is used for optimization.

3.2. Comparison experiments

In this paper, the test results of different recognition algorithms are compared on the dataset. Experiments are conducted on the MSCOCO dataset and the MPII dataset. According to the evaluation index of MSCOCO, the results are evaluated using mAP . The comparison method includes WBKB [16], SCNN [17] and SDYOLO [18].

Test results with different methods on the MSCOCO and MPII datasets are shown in Tables 1 and 2, respectively. mAP@0.50 : 0.95 indicates an average accuracy from 0.50 to 0.95 . On the MSCOCO dataset, the performance of the proposed method in this paper is superior to that of WBKB, SCNN and SDYOLO. On the MPII dataset, the detection accuracy of the proposed method in the Head, Shoulder, Hip, Knee and Ankle parts has reached the highest, with an average accuracy rate of 83.7%, it is superior to other methods.

Table 1. Test results with different methods on the MSCOCO dataset/%.

Method	mAP@0.50:0.95	mAP@0.50	mAP@0.75
WBKB	62.9	86.0	68.6
SCNN	68.1	89.1	74.2
SDYOLO	74.4	90.3	80.2
Proposed	75.2	94.0	83.2

Table 2. Test results with different methods on the MPII dataset (mAP@0.50:0.95)/%.

Method	head	shoulder	elbow	wrist	hip	knee	ankle	Average
WBKB	92.3	88.7	78.8	67.9	76.5	70.0	62.8	76.7
SCNN	93.2	90.4	80.0	70.9	77.3	72.7	65.8	78.6
SDYOLO	92.4	91.6	85.1	77.5	81.4	81.0	73.5	83.2
Proposed	93.0	92.3	83.5	76.5	82.8	81.8	76.7	83.7

Table 3. Results of ablation experiments.

Method	DCAConv	MSCA	MSCF	slide	small	mAP@0.50:0.95	mAP@0.50	mAP@0.75
YOLOv8n	×	×	×	×	×	0.676	0.862	0.872
YOLOv8n_DC	√	×	×	×	×	0.689	0.874	0.848
YOLOv8n_MSCA	×	√	×	×	×	0.692	0.883	0.858
YOLOv8n_MSCF	×	×	√	×	×	0.710	0.877	0.856
YOLOv8n Small	×	×	×	√	×	0.713	0.889	0.868
YOLOv8n_Slide	×	×	×	×	√	0.688	0.871	0.846
DMS-YO LOv8	√	√	√	√	√	0.740	0.908	0.883

3.3. Ablation experiment

To verify the effectiveness of the proposed algorithm, this section designs the YOLOv8n, YOLOv8n_DC, YOLOv8n_MSCA, YOLOv8n_MSCF, YOLOv8n_Slide, YOLOv8n_Small and DS-YOLOv8 ablation experiments. Here, YOLOv8n_DC is to replace part of the Conv of YOLOv8 with the improved DCAConv in this paper. YOLOv8n_MSCA is implemented by adding a multi-scale convolutional attention (MSCA) module at the tail of the backbone network of YOLOv8. YOLOv8n_MSCF combines the MSCF module and YOLOv8. YOLOv8n_Slide replaces the original loss function with slide loss on the basis of YOLOv8n. YOLOv8n_Small only adds feature maps to the P2 layer of the backbone network in the original YOLOv8n and adds detection heads for small targets. DMS-YOLOv8 is the proposed algorithm in this paper, which includes DCAConv, MSCA, MSCF, small target detection head and slide loss.

This series of experimental designs fully considers the influence of different improvement factors. By comparing the experimental results, the contribution degree of each factor to the model performance can be obtained, thereby understanding the superiority of the DMS-YOLOv8 model more comprehensively. The comparison results of the ablation experiments are shown in Table 3. For the behavior recognition under the students classroom behavior tasks, the proposed algorithm shows significant improvement at each stage.

It can be seen from Table 3 that the YOLOv8n_DC based on the CA attention mechanism has improved by 1.2% and 1.3% respectively on mAP@0.50 and mAP@0.50 : 0.95 compared with the basic model YOLOv8n. This indicates that DCAConv can capture the key information in the input features more flexibly by dynamically adjusting the channel weight allocation. The YOLOv8n_MSCA model with the MSCA attention mechanism improves by 2.1% and 1.6% respectively compared with the basic model YOLOv8n on mAP@0.50 and mAP@0.50 : 0.95, indicating that the introduction of the MSCA attention mechanism improves the problem of difficult extraction of targets at different scales and fuzzy background features. The YOLOv8n_MSCF model constructed by the MSCF module is 1.5% and 3.4% higher than the basic model YOLOv8n on mAP@ 0.50 and mAP@0.50:0.95 respectively. It indicates that the MSCF module can better integrate features at different levels, and at the same time, significantly enhance the information fusion ability of the neck through global fusion of multi-level features.

4. Conclusions

Based on the YOLOv8 algorithm, this study proposes a deep learning-based YOLOv8 algorithm, aiming to effi-

ciently identify students classroom behaviors. The dynamic channel attention convolution (DCAConv) is introduced to dynamically adjust the weight distribution between channels. It integrates the MSCA attention mechanism and effectively combines multi-scale convolution features through element multiplication to enhance the focus on spatial information. The MSCF module integrating the GD mechanism aggregates the multi-scale information of the context and improves the efficiency of multi-scale feature utilization. Furthermore, the algorithm particularly adds a small target detection layer, significantly improving the recognition rate of the behaviors of students in the back row. Finally, the slide loss function is adopted to deal with the sample mismatch problem in the dataset. The experimental results fully confirm the outstanding performance of the proposed method in the tasks of student classroom behavior recognition. Although the accuracy of the proposed method in this paper is relatively high, there is still room for improvement in terms of accuracy rate. Furthermore, real-time correlation of students facial information and behavioral information remains a challenge.

References

[1] M. Imran and N. Almusharraf, (2024) *"Google Gemini as a next generation AI educational tool: a review of emerging educational technology"* **Smart Learning Environments *11***(1): 22. DOI: 10.1186/s40561-024-00310-z.

[2] A. Wilson, R. Kask, and L. W. Ming, (2024) *"Exploring circular digital economy strategies for sustainable environmental, economic, and educational technology"* **International Transactions on Education Technology (ITEE) *2***(2): 129–139. DOI: 10.33050/itee.v2i2.579.

[3] M. Jaboob, M. Hazaimeh, and A. M. Al-Ansi, (2025) *"Integration of generative AI techniques and applications in student behavior and cognitive achievement in Arab higher education"* **International journal of human–computer interaction *41***(1): 353–366. DOI: 10.1080/10447318.2023.2300016.

[4] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao. "Yolov9: Learning what you want to learn using programmable gradient information". In: *European conference on computer vision*. Springer. 2024, 26286–26296. DOI: 10.1109/CVPR52733.2024.02484.

[5] J. Yu, H. Li, S.-L. Yin, and S. Karim, (2020) *"Dynamic gesture recognition based on deep learning in human-to-computer interfaces"* **Journal of Applied Science and Engineering *23***(1): 31–38. DOI: 10.6180/jase.202003_23(1).0004.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, (2019) *"Openpose: Realtime multi-person 2d pose estimation using part affinity fields"* **IEEE transactions on pattern analysis and machine intelligence *43***(1): 172–186. DOI: 0.1109/TPAMI.2019.2929257.

[7] Q. Li, L. Xu, and X. Yang, (2022) *"2D multi-person pose estimation combined with face detection"* **International Journal of Pattern Recognition and Artificial Intelligence *36***(02): 2256002. DOI: 10.1142/S021800142256002X.

[8] C.-T. Lu, Y.-C. Liu, and Y.-C. Pan, (2024) *"An intelligent playback control system adapted by body movements and facial expressions recognized by OpenPose and CNN"* **Multimedia Tools and Applications *83***(10): 31139–31160. DOI: 10.1007/s11042-023-16880-y.

[9] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. "Rmpe: Regional multi-person pose estimation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, 2334–2343. DOI: 10.1109/ICCV.2017.256.

[10] S. Yin, H. Li, A. A. Laghari, L. Teng, T. R. Gadekallu, and A. Almadhor, (2024) *"FLSN-MVO: Edge Computing and Privacy Protection Based on Federated Learning Siamese Network With Multi-Verse Optimization Algorithm for Industry 5.0"* **IEEE Open Journal of the Communications Society**: DOI: 10.1109/OJCOMS.2024.3520562.

[11] C. Yongqing and L. Dan, (2024) *"Optimization of Student Behavior Detection Algorithm Based on Improved SSD Algorithm."* **International Journal of Advanced Computer Science & Applications *15***(5): DOI: 10.14569/ijacsa.2024.0150512.

[12] C. Pabba, V. Bhardwaj, and P. Kumar, (2024) *"A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom"* **Multimedia Tools and Applications *83***(12): 36975–37005. DOI: 10.1007/s11042-023-16388-5.

[13] Q. Jia and J. He, (2024) *"Student Behavior Recognition in Classroom Based on Deep Learning"* **Applied Sciences *14***(17): 7981. DOI: 10.3390/app14177981.

[14] F. Xie, B. Lin, and Y. Liu, (2022) *"Research on the coordinate attention mechanism fuse in a YOLOv5 deep learning detector for the SAR ship detection task"* **Sensors *22***(9): 3370. DOI: 10.3390/s22093370.

[15] X. Wang, H. Gao, Z. Jia, and Z. Li, (2023) *"BL-YOLOv8: An improved road defect detection model based on YOLOv8"* **Sensors *23***(20): 8361. DOI: 10.3390/s23208361.

[16] J. Luo, (2024) *"Dynamical behavior analysis and soliton solutions of the generalized Whitham–Broer–Kaup–Boussineq–Kupershmidt equations"* **Results in Physics *60***: 107667. DOI: 10.1016/j.rinp.2024.107667.

[17] M. Dang, G. Liu, Q. Xu, K. Li, D. Wang, and L. He, (2024) *"Multi-object behavior recognition based on object detection for dense crowds"* **Expert Systems with Applications *248***: 123397. DOI: 10.1016/j.eswa.2024.123397.

[18] J. Yu, G. Wang, X. Li, Z. Du, W. Xu, M. Akhter, and D. Li, (2025) *"SDYOLO-Tracker: An efficient multi-fish hypoxic behavior recognition and tracking method"* **Computers and Electronics in Agriculture *232***: 110079. DOI: 10.1016/j.compag.2025.110079.