Journal of Electronic Research and Application, 2026, Volume 10, Issue 10
https://bbwpublisher.com/journal/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Design and Implementation of a Classroom Gesture Recognition System Based on YOLOv11

Na Zhang[1]*, Jialin Ai[2]

[1]Shanwei Polytechnic, Shanwei, Guangdong 516600, China
[2]South China Normal University, Shanwei, Guangdong 516600, China

Corresponding author: Na Zhang, 18846181658@163.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: To solve the problems of feature redundancy, insufficient bounding box localization accuracy, and complex background interference in classroom gesture recognition, the SCMB-YOLOv11 algorithm based on YOLOv11 is proposed in this paper. It combines feature refinement, attention focusing, and localization optimization. SCConv is used to reconstruct the space and channel features of the Backbone network and reduce redundancy at the source; thus, it can improve the quality of feature representation. BSAM is added before the Head, and through channel-space joint attention of CBAM and dynamic sparse self-attention of BiFormer, local details and global context are integrated. MPDIoU loss is employed in place of CIoU, and a distance constraint is introduced between the corners of the predicted frame and the actual frame during regression to improve the convergence speed of regression and positioning accuracy. Experiments on 8 classes of classroom interactive gestures in the HaGRID dataset show that the algorithm achieves an mAP@0.5 of 95.8%, which is 5.4 percentage points higher than the original YOLOv11, and it has good robustness and real-time performance in complex classroom scenes.

Keywords: Classroom gesture recognition; YOLOv11; SCConv; BSAM attention mechanism; MPDIoU

Online publication: November 3, 2026

1. Introduction

As a natural and intuitive form of nonverbal interaction, gestures have considerable application value in fields such as smart classrooms, virtual teaching, and educational assistance. In class, a student raising a hand to ask a question, a teacher using the “OK” gesture to check understanding, or a thumbs-up to encourage students can each convey an instructional signal. However, there are still some problems with gesture recognition in the classroom, such as a complex background (desks, chairs, books, and overlapping hands of multiple students), changes in lighting, and a large number of different sizes and positions of gestures. This often makes it difficult to balance the accuracy and real-time performance of the recognition system. In recent years, deep-learning object detectors, especially models in the YOLO series, have been widely

used for gesture recognition because they balance speed and accuracy. Zhou et al. proposed CSSA-YOLO, which uses cross-scale spatiotemporal attention for fine-grained classroom behavior recognition [1]. Gu et al. introduced BSAM, combining bi-level routing attention with spatial attention, and reported gains in occluded scenes [2]. Xiong et al. studied natural-language-driven teacher gesture recognition by combining speech and visual features [3]. Liu et al. developed LGRDet, a lightweight gesture detector that recognizes gestures with low computational cost [4]. Lv et al. added global-context attention to YOLOv11 to improve traffic-police gesture recognition [5]. Although progress has been made in the above-mentioned work, most existing work improves individual modules, with less systematic study of how modules work together across the full model. Recognition of multiple interactive gesture classes in complex classrooms still leaves room for improvement. We therefore use YOLOv11 as the baseline and propose SCMB-YOLOv11. Its coordinated use of SCConv for feature refinement, BSAM for attention focusing, and MPDIoU for localization optimization is designed to handle complex classroom scenes while maintaining accuracy and efficiency.

2. YOLOv11 algorithm improvements

2.1. YOLOv11 algorithm

Released by Ultralytics in 2024, YOLOv11 is the latest model in the YOLO series [6]. It builds on YOLOv8 and introduces several architectural innovations. Its network has three parts: the Backbone, Neck, and Head.

The Backbone uses an enhanced CSPDarknet architecture [7]. Its core C3k2 (Cross Stage Partial with kernel size 2) module replaces the C2f module used in YOLOv8. C3k2 has self-adaptive feature extraction at all levels of the network and uses many kinds of convolutional kernels and cross-stage partial connections. At the end of the Backbone, SPPF (Spatial Pyramid Pooling - Fast) and C2PSA (Convolutional block with Parallel Spatial Attention) modules are also integrated. C2PSA uses parallel spatial attention to strengthen the model's response and perception of key regions.

The Neck uses a PAN-FPN structure [8]. Its top-down and bottom-up paths aggregate features at different scales and fuse multiscale information.

The Head uses a decoupled structure that separates classification from regression. It optimizes these tasks with binary cross-entropy loss and CIoU loss, respectively.

2.2. SCMB-YOLOv11 algorithm

Gesture recognition in classrooms faces substantial feature redundancy, uneven attention allocation, and insufficient bounding-box localization accuracy. Therefore, this paper proposes the SCMB-YOLOv11 algorithm and shows its general structure in **Figure 1**. The main improvements of this model can be summarized in the following three points. First, SCConv replaces all the standard convolutions in the Backbone to reduce spatial and channel redundancy at the feature extraction stage and obtain good feature representations for the next steps. Second, before the Head, BSAM attention is used to jointly focus on the three aspects of channels, spatial regions, and global context, and thus compensate for the long-range dependency information that may have been lost in the feature refinement stage. Finally, MPDIoU loss is employed instead of CIoU to improve the convergence properties of bounding box regression at the level of the training objective. The three parts are not independent modules but work together to improve the recognition ability of gestures in complex classroom scenes by focusing on the quality of features, distribution of attention, and regression objectives.

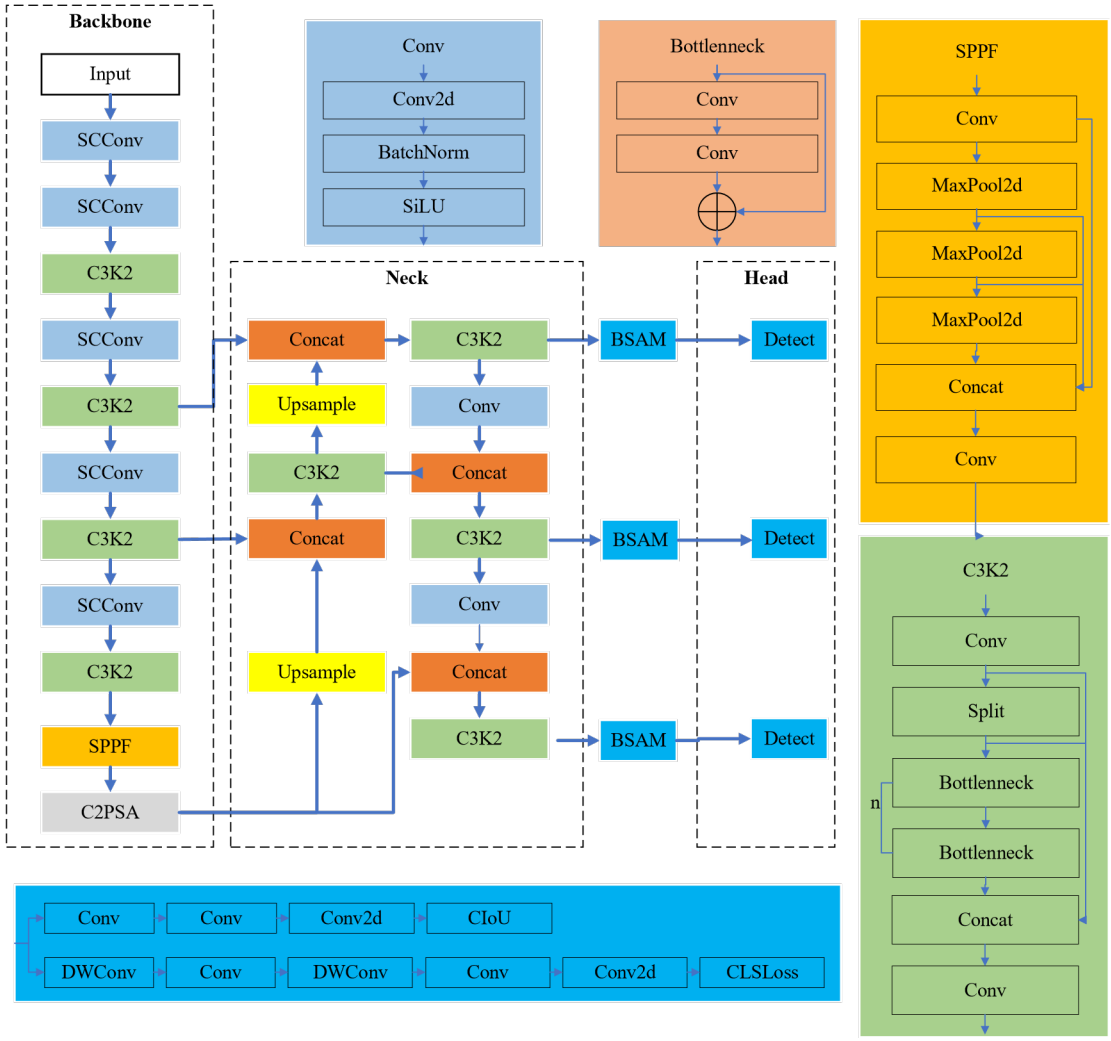

Figure 1. Architecture of the SCMB-YOLOv11 network

2.3. SCConv reconstruction convolution

SCConv (Spatial and Channel Reconstruction Convolution) consists of two sequential submodules—Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU) [9]. The structures of SCConv, SRU, and CRU are illustrated in **Figures 2**, **3**, and **4**, respectively. For the input feature map $X \epsilon R^{N \times C \times H \times W}$, the SRU first suppresses spatial redundancy through a separation-reconstruction strategy. It divides the input feature map into two groups along the channel dimension and calculates the spatial importance weights for each group using a Group Normalization layer. Denote the feature map of the Group i by X_i. The spatially refined feature is then obtained as shown in equation (1).

$$X_w^{(i)} = \gamma_i \cdot GN(X_i) + \beta_i \tag{1}$$

where γ_i and β_i are learnable scaling and shifting parameters. The spatial importance weight w_i is computed using the Sigmoid function, as given in equation (2):

$$w_i = \sigma(W_i \cdot GN(X_i)) \tag{2}$$

Then, CRU reduces channel redundancy through a split-transform-fusion strategy. The spatially refined

feature X_w is divided along the channel dimension into two parts, X_{up} and X_{low}, which are transformed separately, as shown in equation (3):

$$Y_1=M^{P1}(X_{up}) \ , \ Y_2=M^{P2}(X_{low}) \tag{3}$$

where M^{P1} and M^{P2} denotes transformation operations with different receptive fields. The final output is obtained through weighted fusion as shown in equation (4):

$$Y=\beta 1 \cdot Y1+\beta 2 \cdot Y2 \tag{4}$$

where β_1 and β_2 is the fusion weight normalized by Softmax.

In the scene of the classroom, the hand region is relatively small compared to the whole picture. The standard convolution operation generates a large amount of redundant spatial and channel information in the feature extraction stage; thus, it is computationally expensive and may reduce the representation power of important gesture features. Therefore, all the standard Conv modules in the YOLOv11 Backbone are replaced by SCConv modules in this paper. SCConv has a Spatial Reconstruction Unit (SRU) to reduce spatial redundancy in the background area (e.g., desks and blackboards), and a Channel Reconstruction Unit (CRU) is employed to compress channel redundancy related to skin tones and clothing color, thereby improving both the quality and representation efficiency of feature maps at the source of feature extraction.

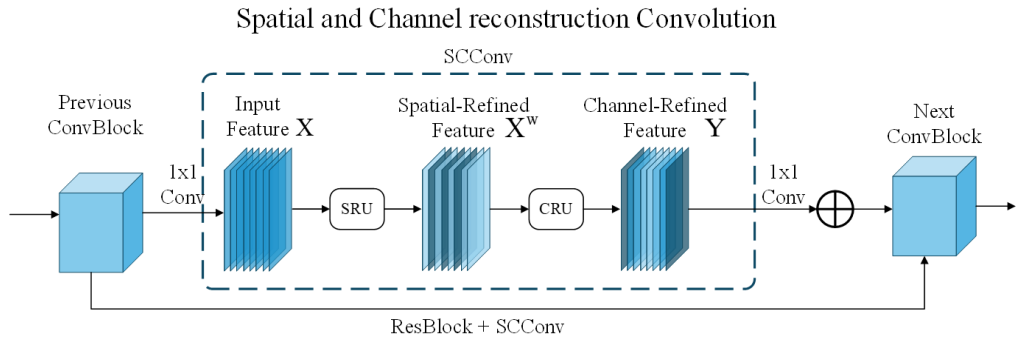

Figure 2. Structure of SCConv

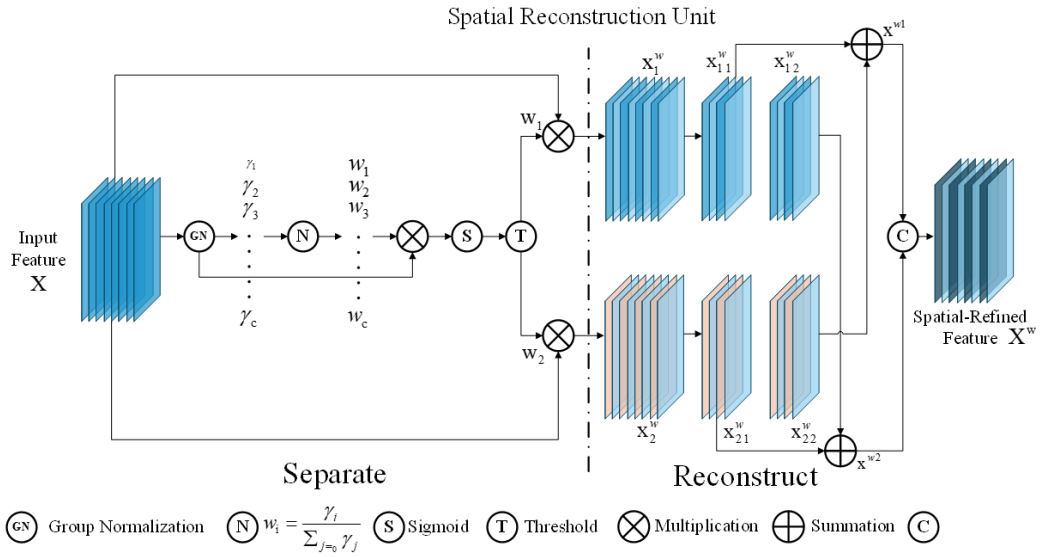

Figure 3. Structure of SRU

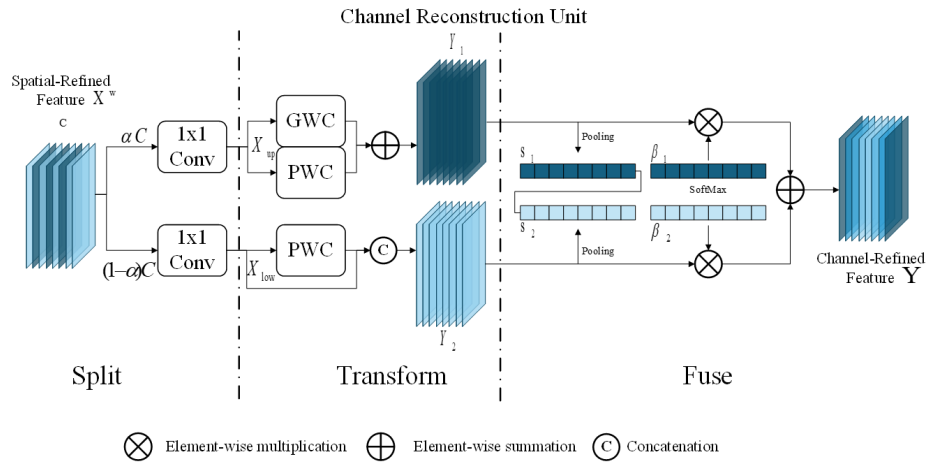

Figure 4. Structure of CRU

2.4. BSAM

BSAM (BiFormer-based Spatial Attention Mechanism) is a combination of the two attention mechanisms proposed by Cao et al., and it replaces the channel attention module in CBAM with a bi-level routing attention BiFormer module [10]. BSAM is a lightweight, versatile module that can be integrated into Convolutional Neural Network (CNN) architectures with little additional overhead; its structure is shown in **Figure 5**. CBAM recalibrates features along the channel and spatial dimensions through its Channel Attention Module (CAM) and Spatial Attention Module (SAM), as shown in equations (5) and (6) [11].

$$F' = M_c(F) \otimes F \tag{5}$$

$$F' = M_s(F) \otimes F \tag{6}$$

where M_c and M_s represent the channel and spatial attention maps, respectively, and $\otimes$ denotes element-wise multiplication.

BiFormer uses a bi-level routing strategy to select relevant key-value pairs dynamically for attention computation [12]. This reduces the computational complexity of standard Transformer self-attention while retaining its ability to model long-range dependencies; its structure is shown in **Figure 6**. BSAM takes the output of CBAM as input to BiFormer, achieving an organic integration of local detail attention and global context capture, as shown in equation (7):

$$F' = \text{Biformer}\left(M_s\left(M_c(F) \otimes F\right) \otimes M_c(F) \otimes F\right) \tag{7}$$

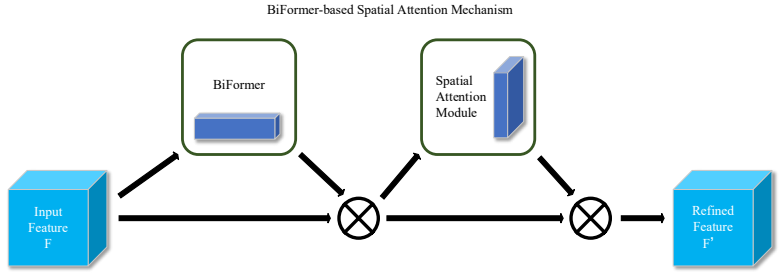

Figure 5. Structure of BSAM

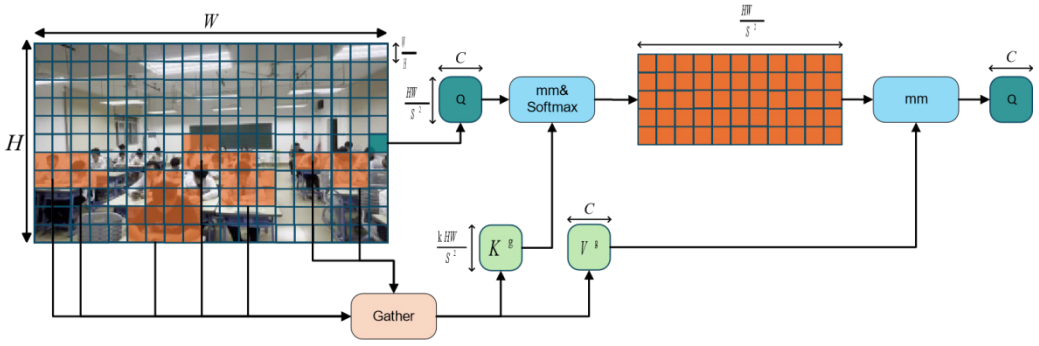

Figure 6. Structure of BiFormer

Although the feature maps reconstructed by SCConv in the Backbone network have been purified in both the spatial and channel directions, the compression mechanism of SCConv may have lost some of the long-range feature correlations. Therefore, in this paper, BSAM is placed before the YOLOv11 Head. The CBAM branch strengthens the representation of local gesture details, such as differences in finger shape, while the BiFormer branch restores global-context information that may have been compressed. Together, the branches refine local details and compensate for global context.

2.5. MPDIoU loss function

The original YOLOv11 uses CIoU as the bounding box regression loss. When classroom gestures are partially occluded, the predicted box may have a center point close to the ground-truth center and a similar aspect ratio, causing the penalty term in CIoU to approach 0. As a result, the loss degenerates into IoU, rendering the regression constraints ineffective and leading to reduced localization accuracy. MPDIoU (Intersection over Union with Minimum Points Distance) measures localization quality by directly minimizing the distance between corner points of the predicted and ground truth bounding boxes [13]. Unlike traditional CIoU, which relies on center point distance and aspect ratio, MPDIoU employs two pairs of key points—the top-left and bottom-right corners—to fully describe the geometric relationship of the bounding box, making it more robust to occlusions and better suited for detecting small objects, thus being more appropriate for classroom gesture detection scenes. Denote the top-left and bottom-right coordinates of the predicted box B_{prd} by $\left(x_1^{prd},y_1^{prd}\right)$ and $\left(x_2^{prd},y_2^{prd}\right)$, respectively, and those of the ground-truth box B_{gt} by $\left(x_1^{gt},y_1^{gt}\right)$ and $\left(x_2^{gt},y_2^{gt}\right)$, respectively. Then, MPDIoU is defined as given in equation (8):

$$\text{MPDIoU}=\text{IoU}-\frac{\text{d}_1^2}{\text{w}^2+\text{h}^2}-\frac{\text{d}_2^2}{\text{w}^2+\text{h}^2} \tag{8}$$

where IoU denotes the intersection-over-union ratio, i.e., the ratio of the intersection area of the predicted and ground-truth boxes to their union area; w and h are the width and height of the minimum bounding rectangle; and d_1^2 and d_2^2 are computed as given in equations (9) and (10), respectively:

$$\text{d}_1^2=\left(\text{x}_1^{\text{prd}}-\text{x}_1^{\text{gt}}\right)^2+\left(\text{y}_1^{\text{prd}}-\text{y}_1^{\text{gt}}\right)^2 \tag{9}$$

$$\text{d}_2^2=\left(\text{x}_2^{\text{prd}}-\text{x}_2^{\text{gt}}\right)^2+\left(\text{y}_2^{\text{prd}}-\text{y}_2^{\text{gt}}\right)^2 \tag{10}$$

The final boundary loss is given by equation (11) as follows:

$$L_{MPDIoU}=1\text{-}MPDIoU \tag{11}$$

Hands only cover a small part of the image in the scene of a classroom, so gesture recognition is a typical problem of small-object detection. CIoU usually has a problem of gradient instability in small-object regression, whereas MPDIoU provides a more direct and stable gradient signal by explicitly constraining the distance between predicted and ground-truth box corner points. It is only a problem of coordinate algebra; therefore, convergence efficiency can be enhanced. The three improvements in SCMS-YOLOv11 form a collaborative mechanism: SCConv reduces the redundancy of space and channels at the feature extraction stage to get high-quality features for the subsequent modules; BSAM enhances local details and addresses the lack of global context based on the purification of features; and MPDIoU improves bounding box regression by changing the training objective. The performance gains achieved by the proposed improvements demonstrate the effectiveness of the progressive and complementary interaction among feature refinement, attention focusing, and localization optimization.

3. Experiment and results analysis

3.1. Gesture recognition dataset

The public HaGRID (Hand Gesture Recognition Image Dataset) is used in this paper [14]. The original HaGRID dataset has more than 550,000 FullHD RGB images and about 18 gesture categories. The study selected eight categories relevant to classroom interaction: one (asking a question), like (approval), dislike (disagreement or not understanding), palm (stop), call (emphasis or direction), ok (acknowledgment), mute (pause or silence), and peace (victory or encouragement). As instructional signals in the classroom, these gestures have specific meanings and are suitable for building a classroom gesture recognition system. The data were divided into training, validation, and test sets in an 8:1:1 ratio. The training set was further augmented using techniques such as Mosaic stitching, random flipping (horizontal/vertical), brightness/contrast adjustment, and Gaussian blur, increasing its size by threefold to a total of 37,440 images, thereby enhancing the model's adaptability to complex classroom scenes.

3.2 Evaluation metrics

The four metrics of this paper are Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP). The calculation formulas are given by equations (12), (13), (14), and (15) in the following section.

$$P=\frac{TP}{TP+FP}*100\% \tag{12}$$

$$R=\frac{TP}{TP+FN}*100\% \tag{13}$$

$$AP=\int_0^1 P(R)dR \tag{14}$$

$$mAP=\frac{1}{n}\sum_{i=1}^{n}AP_i \tag{15}$$

Where TP denotes the number of true positive samples, FP denotes the number of false positive samples, and FN denotes the number of false negative samples. AP_i represents the average precision of the Class i under the P-R curve, and n is the number of gesture categories.

3.3. Experimental environment configuration

Table 1 shows the experimental environment and training parameter settings of this paper.

Table 1. Experimental environment table

Name	Parameters
CPU	Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz
GPU	NVIDIA A800-SXM4-80G
Operating System	Windows Server 2022 Standard
Deep learning framework	Pytorch
Epoch	200
Batch Size	32
Learning rate	0.01
Weight decay	0.005

3.4. Comparative experiments of different attention mechanisms

To validate the advantages of the BSAM over other mainstream attention mechanisms in classroom gesture recognition tasks, the framework of YOLOv11+SCConv+MPDIoU is taken as the baseline, and the five attention mechanisms SE, ECA, CA, CBAM, and BSAM were tested. The results are shown in **Table 2**. According to the above results, pure channel-based attention mechanisms (SE and ECA) do not improve performance, and SE is even slightly worse. This may be because discriminative information about gestures in classroom scenes is not only in the channel dimension but also heavily dependent on spatial location. CA, which incorporates positional information into channel attention, achieves better performance than pure channel attention. CBAM further introduces a spatial attention branch, boosting mAP to 94.7%. BSAM adds a BiFormer global self-attention branch to CBAM, raising mAP to 95.8%, 1.1 percentage points above CBAM. These results demonstrate the complementary value of global contextual information for gesture recognition in classroom scenes.

Table 2. Results for different attention mechanisms

Types of attention mechanisms	P (%)	R (%)	mAP0.5 (%)
/	93.8	92.7	94.0
SE	93.9	92.5	93.8
ECA	94.1	92.8	94.1
CA	94.2	93.0	94.3
CBAM	94.5	93.3	94.7
BSAM	95.3	94.2	95.8

3.5. Algorithm comparison experiment

To comprehensively evaluate the performance of SCMB-YOLOv11, this study compares it with mainstream YOLO models, recent classroom gesture recognition methods, and lightweight detectors. The results are shown in **Table 3**. Based on the above comparison, it can be seen that SCMB-YOLOv11 is 4.8, 5.4, and 5.4 percentage points higher than the baseline model YOLOv11 in precision (P), recall (R), and mAP@0.5, respectively; thus, the proposed improvements have been effective. Compared to the dedicated gesture model MediaPipe Hands, SCMB-YOLOv11 outperforms it by 3.0, 3.4, and 4.2 percentage points in P, R, and mAP@0.5, addressing the insufficient robustness of specialized models in complex classroom scenes. Compared with the two-stage detector Faster R-CNN, the enhanced model achieves a mAP more than 13 percentage points higher. Overall, the proposed improvements prove effective.

Table 3. Comparison of experimental results for the different algorithm models

Model Name	P (%)	R (%)	mAP0.5 (%)	FPS
Faster R-CNN	83.1	81.5	82.3	18
YOLOv8	90.1	88.3	89.2	95
YOLOv11	90.5	88.8	90.4	98
MediaPipe Hands	92.3	90.8	91.6	60
SCMB-YOLOv11	95.3	94.2	95.8	87

3.6. Ablation experiment

To evaluate the individual effectiveness and synergistic enhancement of each enhanced module, this study uses the original YOLOv11 as the Baseline and designs seven ablation experiments. The results are presented in **Table 4**. As shown in the validation results for the individual modules in **Table 4**, SCConv raises mAP by 2.3 percentage points, indicating that feature redundancy is a significant issue in classroom scenes and that spatial and channel reconstruction can refine gesture features. BSAM is 1.9 percentage points higher than the previous one, and a combination of CBAM and BiFormer has been used to consider local details and global context. The improvement of the MPDIoU loss function is 1.6 percentage points, and it has been shown that corner distance constraints are effective for bounding box regression of small objects. The combined performance of SCConv+BSAM is 3.9 percentage points better, and that of SCConv+MPDIoU is 3.6 percentage points better; both exceed the sum of the gains obtained by the individual modules. With all three modules, mAP rises by 5.4 percentage points, and both precision and recall exceed 94%. These results show that these three improvements form an effective synergistic enhancement mechanism.

Table 4. Ablation experiment results

YOLOv11	SCConv	BSAM	MPDIoU	P (%)	R (%)	mAP@0.5 (%)
√	-	-	-	90.5	88.8	90.4
√	√	-	-	92.3	91.0	92.7
√	-	√	-	91.9	90.5	92.3
√	-	-	√	91.6	90.2	92.0
√	√	√	-	94.1	93.0	94.3
√	√	-	√	93.8	92.7	94.0
√	√	√	√	95.3	94.2	95.8

4. Conclusion

SCMB-YOLOv11 is introduced in this paper to solve the problems of feature redundancy, background interference, and localization errors in classroom gesture recognition. The three parts of the algorithm are feature refinement, attention focusing, and localization optimization. SCConv reduces spatial and channel redundancy; BSAM combines local information with global context; MPDIoU improves the convergence of bounding box regression. Based on the experiments performed on the HaGRID dataset for the eight categories of classroom gestures, the mAP@0.5 reached 95.8%, which is an increase of 5.4 percentage points over YOLOv11, and it can also operate in real time. However, the model is prone to misclassify or miss gestures under extreme lighting or severe occlusion, and it currently handles only static gestures. Future work will introduce a temporal attention mechanism for dynamic gesture recognition, and further compress the model size to facilitate deployment on edge devices.

Funding

Guangdong Provincial Ordinary Higher Education Institutions Youth Innovation Talent Program (Project No.: 2025KQNCX259).
Shanwei Polytechnic Research Project (Project Nos.: 2025XJXM003, 2024XJXM026).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Zhou L, Liu X, Guan X, et al., 2025, CSSA-YOLO: Cross-Scale Spatiotemporal Attention Network for Fine-Grained Behavior Recognition in Classroom Environments. Sensors, 25(10): 3132.

[2] Gu Y, Niu K, 2024, Research on Classroom Behavior Analysis Based on YOLO Target Detection Model. In 2024 4th International Symposium on Computer Technology and Information Science (ISCTIS), IEEE.

[3] Xiong Y, Chen S, Cai T, et al., 2025, Natural Language-Driven Teacher Gesture Recognition. 18th International Conference on Educational Data Mining (EDM), Palermo, Italy.

[4] Liu H, Wan Y, Zhang M, et al., 2025, LGRDet: A Light Object Detection Network for Gesture Recognition. Concurrency and Computation: Practice and Experience, 37(25–26): e70350.

[5] Lv C, Sun Z, 2025, Research on Traffic Police Gesture Recognition Algorithm Based on Improved YOLOv11. In 2025 WRC Symposium on Advanced Robotics and Automation (WRC SARA) (pp. 228–233).

[6] Khanam R, Hussain M, 2024, YOLOv11: An Overview of the Key Architectural Enhancements, arXiv, https://doi.org/10.48550/arXiv.2410.17725

[7] Zhao B, Zhao J, Song R, et al., 2025, Enhanced YOLO11 for Lightweight and Accurate Drone-Based Maritime Search and Rescue Object Detection. PLoS ONE, 20(7): e0321920.

[8] Jocher G, Qiu J, 2024, Ultralytics YOLO11[CP/OL]. https://github.com/ultralytics/ultralytics

[9] Li J, Wen Y, He L, 2023, SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6153–6162).

[10] Cao J, Luo F, Ding W, 2026, BS-YOLO: A Small Object Detection Algorithm Based on BSAM

Attention Mechanism and SCConv. Computer Engineering, 52(3): 119–127. https://doi.org/10.19678/j.issn.1000-3428.0070159

[11] Woo S, Park J, Lee JY, et al., 2018, CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–19), Springer.

[12] Zhu L, Wang X, 2023, BiFormer: Vision Transformer with Bi-Level Routing Attention. arXiv. https://doi.org/10.48550/arXiv.2303.08810

[13] Ma S, Xu Y, 2023, MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. arXiv. https://doi.org/10.48550/arXiv.2307.07662

[14] Kapitanov A, Kvanchiani K, 2024, HaGRID – Hand Gesture Recognition Image Dataset. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 4572–4581), IEEE. https://doi.org/10.1109/WACV57701.2024.00451

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.