Ukrainian Journal of Physical Optics
ISSUE 4
2026

OPTICS-ORIENTED STUDENT CLASSROOM BEHAVIOR RECOGNITION USING AN IMPROVED YOLOV11N-RD MODEL FOR OPTICAL SURVEILLANCE IMAGING

MIN PENG [1], YUJIA FU [2], ZUFANG YANG [1], XIAOKUN YANG [3], XIANGQING WANG [1]*
DONGFEI WANG [2]

[1] Wuhan Institute of Technology & Business, Wuhan, 430065, Hubei Province, China
[2] School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China
[3] School of Electronic and Information Engineering, Nanchang Institute of Technology, Nanchang 330044, China
*Corresponding author: wxqing@fynu.edu.cn

Received: 21.10.2026

Abstract. Students' classroom behavior is an important indicator for teaching evaluation. Optical classroom surveillance provides a non-contact, wide-coverage way to observe student behavior, but reliable recognition from such imagery remains challenging because student behaviors are often small-scale, partially occluded, and affected by complex illumination, background clutter, perspective distortion, and optical feature loss during network downsampling. To address these problems, this paper proposes an optics-oriented improved algorithm, YOLOv11n-RD, for student classroom behavior recognition from optical surveillance images. First, a receptive-field convolutional block attention module (RFCBAMConv) is used to improve C3k2, enhancing multi-scale optical feature extraction and enabling the model to focus on local optical cues associated with small student behaviors. Second, a dynamic detection head (DyHead) replaces the original detection head to integrate scale-aware, spatially aware, and task-aware attention, thereby improving target detection accuracy and reducing optical feature loss in complex classroom scenes. Real classroom optical images are used for training and testing. Experimental results show that the improved model can more accurately detect students' classroom behavior from optical surveillance imagery. The proposed model achieves an mAP@0.5 of 97.5%, representing an 8.4 percentage-point improvement over the baseline YOLOv11n, demonstrating enhanced robustness for optical classroom behavior recognition under complex imaging conditions.

Keywords: optical classroom monitoring, student behavior recognition, YOLOv11, RFCBAMConv, DyHead, optical imaging
UDC: 004.932:004.85:37.091.3
DOI: 10.3116/16091833/Ukr.J.Phys.Opt.2026.04207
This work is licensed under the Creative Commons Attribution International

1. Introduction

Recent advances in artificial intelligence have accelerated the use of deep learning in computer vision, particularly for object detection and image recognition. In parallel, the continuing development of educational informatization and intelligent classrooms has created a strong demand for technologies that can improve teaching effectiveness and classroom management [1]. Classroom behavior is a key indicator of students' learning status and classroom participation; accurate recognition and analysis of such behavior can therefore provide useful evidence for teaching evaluation. However, classroom behaviors are diverse and often subtle, which complicates classroom order management and may affect the implementation of teaching activities [2]. Systematic evaluation of student classroom behavior is thus important for optimizing classroom teaching.

Traditional classroom behavior recognition relies mainly on manual observation and recording by teachers. Such approaches are inefficient, subjective, limited in coverage, and

poorly suited to modern educational management. As smart classrooms and surveillance equipment become more common, computer-vision-based automated behavior recognition has become a practical alternative [3]. Deep learning algorithms can analyze classroom surveillance video and estimate student concentration, and convolutional neural network-based object detection has shown strong potential because of its efficiency and accuracy.

Object detection aims to locate and recognize specific objects in images or videos. Before 2012, detection methods depended largely on handcrafted features such as the histogram of oriented gradients [4] and scale-invariant feature transform (SIFT), combined with classifiers such as support vector machines. In 2014, Girshick et al. [5] introduced R- convolutional neural networks (R-CNN) (regions with CNN features), bringing CNNs into object detection and initiating the two-stage detection paradigm. R-CNN improved accuracy through region proposals and feature extraction, but its computational efficiency was limited. Fast R-CNN [6] and Faster R-CNN [7] subsequently improved training efficiency and detection speed by sharing convolutional features and introducing region proposal networks. The original R-CNN required independent feature extraction for each region proposal, incurring substantial computational and memory costs; Fast R-CNN reduced this cost by sharing convolutional features, and Faster R-CNN further enabled end-to-end region proposal generation.

To address these issues, Redmon et al. [8] proposed the one-stage "You Only Look Once" (YOLO) algorithm in 2016. YOLO directly predicts bounding boxes and class probabilities without region proposals, greatly improving detection speed and meeting real-time requirements. Subsequent versions, including YOLOv3, YOLOv4, and YOLOv5, further improved accuracy, training efficiency, and small-object detection [9-11]. Nevertheless, the YOLO series still has limitations in small-object detection. In classroom behavior recognition, student behaviors often appear as small objects and are vulnerable to occlusion and complex background interference, which creates significant detection challenges.

This paper proposes an improved YOLOv11 algorithm to address these limitations. By optimizing the network structure and introducing multi-scale feature fusion and attention mechanisms, the method improves small-target detection, thereby enhancing classroom behavior recognition accuracy. The proposed model is designed for complex classroom scenes and aims to provide reliable technical support for educational intelligence under optical surveillance imaging. This integration links local optical feature enhancement within C3k2 (a feature-extraction building block) with adaptive multi-scale prediction through DyHead, improving student behavior recognition in complex classroom surveillance images.

2. Improved YOLOv11n algorithm

YOLOv11 [12] is an object detection algorithm released by Ultralytics on September 30, 2024. Compared with previous YOLO series models, YOLOv11 has achieved significant improvements in accuracy, speed, and computational efficiency. Compared with YOLOv8, YOLOv11 replaces C2f with C3k2 and introduces a C2PSA self-attention module after the spatial pyramid pooling-fast (SPPF) module to enhance feature extraction. The neck network adopts the path aggregation network-feature pyramid network structure, which integrates shallow spatial information and deep semantic information, improving the model's object localization capability and enhancing the interaction and utilization of spatial information at different levels. Additionally, the two Conv layers in the classification detection head are replaced with depthwise convolution, reducing the model's parameter count and computational overhead.

Although YOLOv11 has excellent detection performance, small-target detection still has room for improvement. To address missed detections, false detections, and optical feature loss in small-target detection, this paper introduces the following modifications to YOLOv11n. First, we use RFCBAMConv, which incorporates receptive-field attention and the convolutional block attention module (CBAM), to improve C3k2 and enhance optical feature extraction at different scales. Secondly, in order to deal with the problem of feature information loss caused by multiple downsampling, the detection head of the original model is replaced with a dynamic detection head with a self-attention mechanism, which integrates scale perception, spatial perception, and task perception into one framework, significantly improving the representation ability of the target detection head and the small-target detection effect. The structure diagram of the improved model network YOLOv11n-RD is shown in Fig. 1.

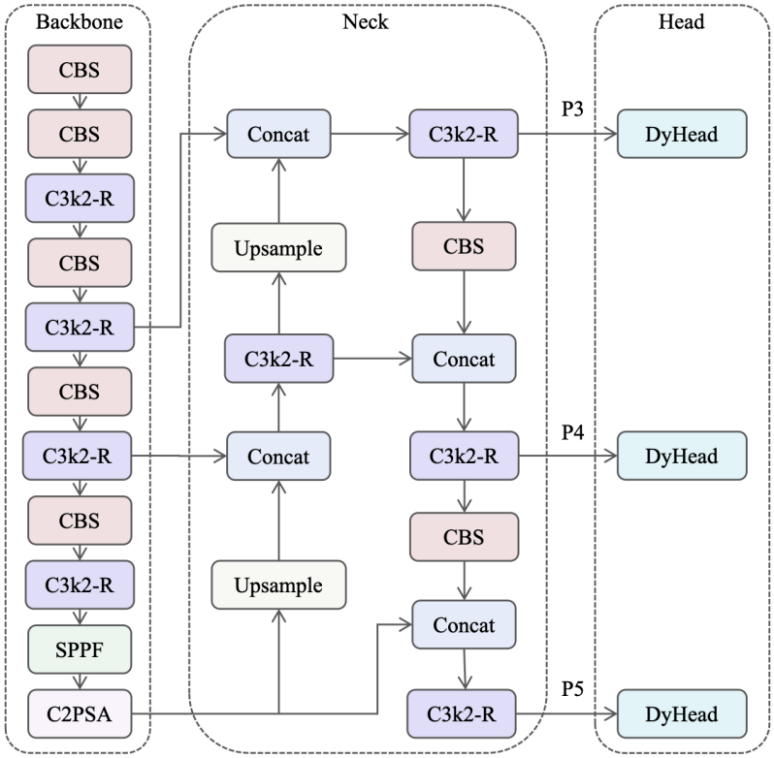


Fig. 1. Architecture of the proposed YOLOv11n-RD model. (CBS - convolution–batch normalization–sigmoid linear unit; Concat - feature concatenation; DyHead - dynamic detection head. C3k2-R denotes a C3k2 cross-stage partial block modified with RFCBAMConv, and C2PSA denotes a cross-stage partial block with self-attention. P_3, P_4, and P_5 denote feature levels with strides of 8, 16, and 32, respectively. The suffix *n* denotes the nano variant, and RD denotes the integration of RFCBAMConv and DyHead).

2.1. Improved C3k2 module

Small objects typically have less pixel information and are easily submerged by background interference. Traditional detection models often miss small targets due to limitations in their receptive field design. Specifically, first, the fixed-scale receptive field struggles to adapt to multi-scale target changes; second, the parameter-sharing mechanism of convolution results in insufficient sensitivity to object location. To address these problems, Liu et al. [13] proposed the receptive field block (RFB) mechanism, which improves the limitations of traditional convolution operations by simulating the multi-scale characteristics of the receptive field in the human visual system. The RFB employs a multi-branch structure with

convolutions of different kernel sizes and dilation rates to construct multi-scale receptive fields within a single module. This design enables the network to capture contextual information at different spatial scales and better represent local and surrounding features. Such multi-scale feature extraction is particularly beneficial for detecting small targets in complex classroom scenes.

The C3k2 module in the original YOLOv11n model uses standard convolution operations, which have a large and fixed receptive field range that lacks dynamic adjustment capabilities. This design fails to flexibly adapt to the scale variations of the target, limiting the ability to capture the subtle differences of small objects at various positions. To address this, we incorporate the existing RFCBAMConv module [14] into the C3k2 module of YOLOv11n to enhance local optical feature extraction, and its structure is shown in Fig. 2. RFCBAMConv incorporates receptive-field feature processing and channel and spatial attention derived from the CBAM [15]. In the modified C3k2 module, RFCBAMConv is used for the input projection before channel splitting and the fusion projection after feature concatenation. Both operations use a kernel size of 1 and a stride of 1, preserving the spatial dimensions. The original channel splitting, internal feature processing, and concatenation structure are retained. When a C3k unit is used inside C3k2, its projection layers also employ RFCBAMConv. Channel attention is generated through global average pooling and two fully connected layers with 16 hidden units. Spatial attention is obtained from the channel-wise maximum and average feature maps using a 3×3 convolution followed by a sigmoid activation. These operations adaptively weight local optical features within the C3k2 structure.

CBAM focuses on the importance of feature channels and the relevance of spatial positions through two sub-modules: channel attention and spatial attention. At the same

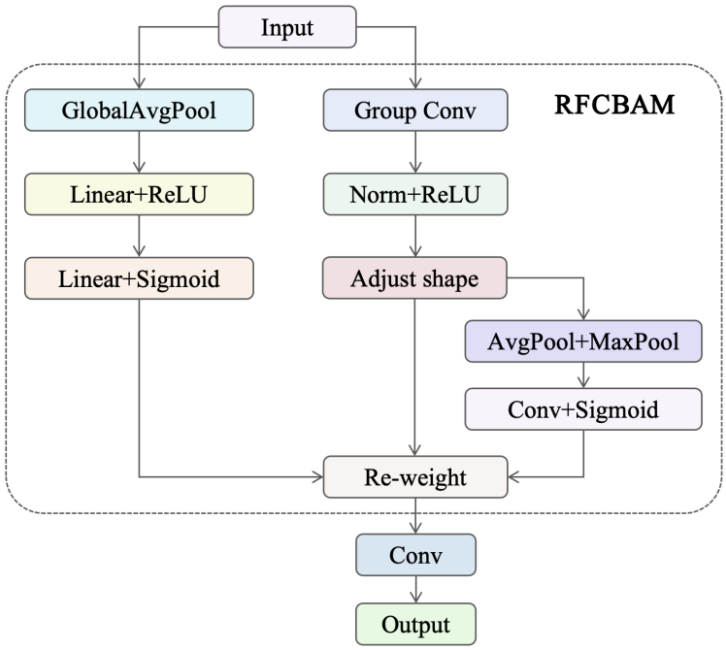


Fig. 2. Architecture of the RFCBAMConv module. (RFCBAM denotes convolutional block attention applied to receptive-field features. GlobalAvgPool - global average pooling; Group Conv - grouped convolution; Linear - fully connected layer; Norm - batch normalization; ReLU - rectified linear unit; AvgPool - channel-wise average pooling; MaxPool - channel-wise maximum pooling. Adjust shape denotes feature rearrangement, and Re-weight denotes element-wise feature weighting).

time, CBAM's channel attention increases the weight of feature channels related to small targets, while its spatial attention emphasizes the spatial position of small targets. This dual mechanism effectively suppresses background noise interference on small targets, enhances feature relevance, and significantly improves small-target recognizability. Receptive-field feature processing provides contextual information, while channel and spatial attention highlight behavior-related optical features and suppress background interference. This synergy enables RFCBAMConv to focus on local features more precisely while retaining the ability to perceive global context, significantly improving the robustness of feature representation. Therefore, RFCBAMConv combines a receptive field and an attention mechanism, making the model more sensitive to small targets and reducing the missed detection rate, especially in dense scenes or long-distance target detection.

The improved C3k2 module is significant for detecting students' classroom behavior because it introduces the RFCBAMConv structure. Since students' behavior patterns in the classroom are typically complex, there are different learning attitudes, levels of classroom participation, and behavioral details, which are often difficult to capture using traditional models. Therefore, RFCBAMConv combines multi-scale receptive fields and attention mechanisms, enabling the model to focus more effectively on key behavioral features within small ranges and avoid interference from background noise. This improves recognition accuracy, especially in classroom environments with complex backgrounds or a large number of students. This improvement has profound implications for classroom behavior monitoring systems in education, providing administrators with more accurate analyses of student engagement and helping teachers identify and intervene in students' classroom behavior in a timely manner.

2.2. Dynamic detection head

As the latest version of the YOLO series, YOLOv11 uses the classic single-stage detector design in its original detection head, directly classifying and regressing bounding boxes from multi-scale feature maps produced by FPN. Despite the model's strong real-time performance and accuracy, it still has limitations. First, the original detection head cannot dynamically adjust its feature-extraction strategy to changes in target scale, making it difficult for the model to effectively capture detailed features across scale variations. Specifically, for small targets, static convolution may overlook local details, while for large targets, the detection accuracy may decrease due to excessive attention to local features. Secondly, in complex scenes, it is difficult for the detection head to adaptively adjust the focus area according to the target distribution or background characteristics, and it may not be able to distinguish subtle movements such as "raising hands" and "lowering heads" in classroom behaviors, resulting in high false detection and missed detection rates. In addition, in multi-target dense scenes, the original detection head cannot respond to key features, reducing the robustness of classification and bounding box localization.

Because the original detection head struggles to detect small objects, this paper introduces a dynamic detection head [16] to replace it. The three attention mechanisms are conceptually illustrated in Fig. 3. The core of DyHead is to improve the model performance through an adaptive mechanism and unify the object detection head through an attention mechanism. By combining scale-aware, spatial-aware, and task-aware attention, DyHead enhances the detection head's feature representation.

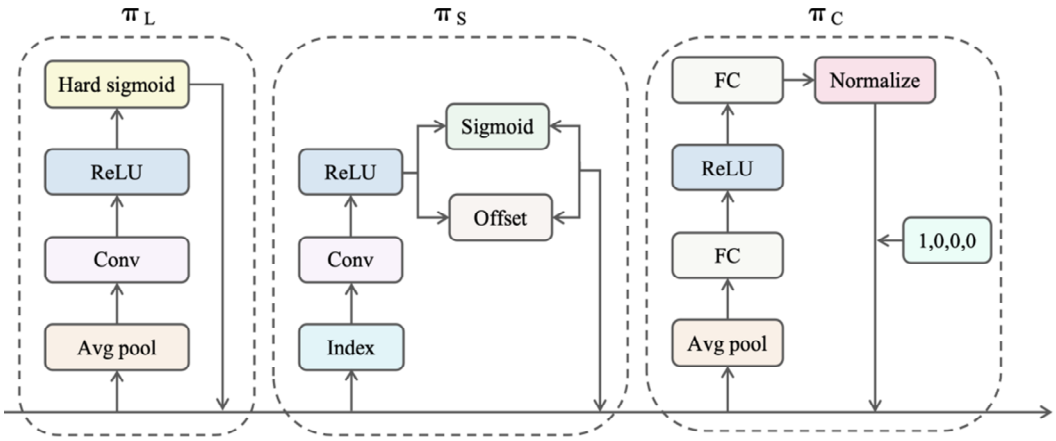


Fig. 3. Conceptual illustration of the three attention mechanisms in DyHead. (DyHead - dynamic detection head; Avg pool - average pooling; ReLU - rectified linear unit; FC - fully connected layer. *L*, *S*, and *C* denote feature level, spatial position, and channel, respectively, corresponding to scale-aware, spatial-aware, and task-aware attention).

In YOLOv11, the multi-scale features output by the neck network are described along three dimensions: feature level (*L*), spatial position (*S*), and channel (*C*). DyHead deploys attention mechanisms in the three dimensions of scale perception, spatial perception, and task perception, respectively, and introduces scale-aware attention in the detection head, so that the weights of each feature level are adaptively adjusted according to the input, thereby achieving a unified detection head design and enhancing feature representation capabilities. Specifically, given the multi-scale feature representation *F*, the general attention formulation is expressed as follows:

$$W(F)=\pi_C\left(\pi_S\left(\pi_L(F)\cdot F\right)\cdot F\right)\cdot F \tag{1}$$

where $W(F)$ denotes the attention-refined features. The functions π_L, π_S, and π_C represent scale-aware, spatial-aware, and task-aware attention, respectively. Eq. (1) summarizes the DyHead attention framework, while the operations used in the proposed network are described below.

For a 640×640 input image, DyHead receives three optical feature maps from P_3, P_4, and P_5, with dimensions of B×64×80×80, B×64×40×40, and B×64×20×20, respectively, where B is the batch size. Thus, L denotes the number of feature levels (3), S denotes the number of spatial positions at each level (height×width), and C denotes the number of channels (64).

Fig. 3 illustrates the three attention mechanisms conceptually. In our implementation, spatial feature alignment is followed by scale-weighted fusion and task-aware activation. A 3×3 convolution directly generates the offset and modulation parameters without an intervening ReLU, and a sigmoid is applied only to the modulation channels. The task-aware branch uses dynamic ReLU, with its activation parameters obtained through hard-sigmoid and subsequent scaling and shifting.

Spatial-aware attention uses a 3×3 modulated deformable convolution with 18 sampling-offset channels and nine sigmoid-activated modulation channels, followed by GroupNorm with 16 groups. Features from adjacent levels are aligned to the target resolution using stride-2 convolution for the finer branch and nearest-neighbor interpolation for the coarser branch.

Scale-aware attention generates a weight for each aligned branch through global average pooling, a 1×1 convolution, ReLU, and hard-sigmoid. The weights are broadcast across the spatial and channel dimensions, and the weighted features are averaged over the participating branches. Task-aware attention uses dynamic ReLU, which takes the maximum of two channel-dependent affine transformations. Its parameters are generated through global average pooling and two fully connected layers with 16 hidden units and 256 output values. Three DyConv blocks are applied successively, each processing all three feature levels. The refined optical feature maps retain their input spatial resolutions and 64 channels before subsequent classification and localization.

The introduction of the DyHead in YOLOv11 is highly significant for detecting student behavior in classrooms. In real classroom environments, students' micro-expressions, gestures, and subtle movements (such as raising a hand, nodding, or lowering the head) are often difficult to detect because they occupy little space and are easily obscured by background clutter. The DyHead structure uses multi-level feature fusion and adaptive attention mechanisms to better capture small-scale behavioral features. Specifically, the scale-aware mechanism enables the model to dynamically adjust feature extraction based on scale, preserving detailed information. In addition, the spatial perception mechanism enables the detection head to distinguish adjacent individuals in a classroom environment with dense students, even when students are closely packed, thereby mitigating occlusion-induced confusion. The weight-sharing mechanism also helps the model balance between classification and positioning tasks, ensuring that behavioral features can be learned and recognized with higher accuracy.

Therefore, by incorporating DyHead into the improved network, we can better integrate features of different scales and improve the detection of small student behaviors in optical classroom images. The modified C3k2 module enhances local optical features, while DyHead adaptively processes the resulting representations for classification and localization. This improvement provides technical support for optical classroom behavior monitoring and the analysis of student participation. By providing more reliable and refined behavioral data analysis, educators can more effectively evaluate students' classroom participation and engagement, thereby optimizing teaching strategies.

3. Experiment

3.1. Experimental environment and parameter configuration

The experimental environment and hyperparameter settings used in this experiment are shown in Tables 1 and 2, respectively.

Table 1. Experimental environment.

Environment Configuration	Information
GPU	NVIDIA GeForce RTX 4090D
CPU	Intel(R) Xeon(R) Platinum 8474C
Memory	80 GB
Operating System	Windows11
Python	3.10
PyTorch	2.1.0
Cuda	12.1

Table 2. Hyperparameter settings.

Name	Numeric
Image size	640×640×3
Epochs	500
Batch size	16
Initial learning rate (lr0)	0.01
Final learning rate fraction (lrf)	0.01
Confidence threshold for the predicted box	0.35

3.2. Dataset

The self-built dataset comprises 6351 optical classroom surveillance images extracted from 40 sessions recorded in 10 physical classrooms. Frames are sampled at five-second intervals, and blurred or redundant images are removed. Labelme is used to annotate six behaviors: hand-raising, reading, writing, using the phone, bowing the head, and leaning on the table.

The data are partitioned at the physical-classroom level. Eight classrooms provide 5646 images for model development and final training, while two other classrooms provide 705 held-out test images. Internal model development uses 4941 training images from six classrooms and 705 validation images from two other classrooms. After hyperparameters and training duration are determined using internal validation, each model is reinitialized and trained on all 5646 development images. The final-epoch checkpoint is then evaluated on the held-out test set, which is not used for parameter optimization or model selection.

3.3. Model evaluation indicators

The performance of the proposed model is evaluated using four commonly used object detection metrics [17], namely, precision (*P*), recall (*R*), average precision (*AP*), and mean average precision (*mAP*). Precision (*P*) measures the proportion of correctly predicted positive samples among all samples predicted as positive, whereas recall measures the proportion of correctly predicted positive samples among all actual positive samples. Average precision summarizes the precision–recall performance for each category, while mean average precision is the mean of the AP values across all categories. The calculation formula is as follows:

$$P = \frac{TP}{FP + TP}, \tag{2}$$

$$R = \frac{TP}{TP + FN}, \tag{3}$$

$$AP_i = \int_0^1 P_i(R)dR, \tag{4}$$

$$mAP = \frac{\sum AP}{N} \tag{5}$$

In the formula, *TP* represents the number of correctly detected targets, *FP* represents the number of incorrectly detected targets, *FN* represents the number of targets that actually exist but have not been detected, N is the total number of categories of the target to be detected, and AP_i represents the *AP* value of the *i*-th category.

3.4. Comparative experiment

To evaluate the effectiveness of the improved YOLOv11n-RD algorithm for student behavior recognition, this paper conducted experiments on a self-built classroom image dataset and selected multiple mainstream target detection models as baselines for performance comparison. The results are shown in Table 3. To assess training variability, we trained YOLOv11n and YOLOv11n-RD independently using five predefined random seeds (2025-2029) on the primary partition of the optical classroom surveillance dataset. All runs follow the internal-validation and final-training procedure described in Section 3.2. For these two models, the mean *mAP* differences and their approximate 95% paired t confidence intervals are calculated using seed-matched results and reported in the text.

As shown in the Table 3, the improved YOLOv11n-RD algorithm performs well in detection. On the primary partition, the mean differences between YOLOv11n-RD and YOLOv11n are 8.4 percentage points for *mAP*@0.5 and 9.8 percentage points for *mAP*@0.5:0.95. The corresponding approximate 95% confidence intervals are [7.3, 9.5] and [8.5, 11.1] percentage points, respectively. These intervals quantify training-seed variability conditional on this partition. In addition, compared with YOLOv5s, YOLOv8s, and YOLOv10s, the proposed model achieves higher detection accuracy on the evaluated optical classroom surveillance dataset.

Table 3. Comparative experiment results.

Model	**P/%**	**R/%**	***mAP*@0.5:0.95/%**	***mAP*@0.5/%**
YOLOv5s	77.7	69.5	58.5	78.3
YOLOv8s	80.2	72.3	61.2	80.7
YOLOv10n	83.3	75.5	64.9	83.5
YOLOv10s	86.9	78.2	68.5	86.7
YOLOv11n	88.5	91.5	69.8	89.1
YOLOv11n-RD	94.3	93.8	79.6	97.5

3.5. Ablation experiment

To fully verify the improvement effect of each improved module on model performance, this paper conducted an ablation experiment on the proposed improved model using a self-built dataset. The experiment is based on the YOLOv11n model and uses a step-by-step integration method to quantify each module's effect on detection accuracy. The specific experimental process is as follows: First, the C3k2 module in the original network is improved using the RFCBAMConv structure to enhance the model's ability to extract multi-scale features; secondly, the dynamic detection head DyHead is used to replace the original detection head to improve the model's dynamic perception of target location and category. During the experiment, the impact of each improved module on the experimental results was tested separately to clarify its contribution. The results of the ablation experiment are shown in Table 4.

The ablation results are visualized, and the original curves are smoothed for clarity. Fig. 4 compares the *mAP*@0.5, *mAP*@0.5:0.95, precision, and recall of the four models before and after improvement.

Table 4. Ablation experiment results.

Baseline	RFCBAMConv	DyHead	*P*, %	*R*, %	*mAP*@0.5:0.95, %	*mAP*@0.5, %
YOLOv11n			88.5	91.5	69.8	89.1
	√		91.4	94.5	73.4	91.6
		√	91.9	89.4	75.2	95.4
	√	√	94.3	93.8	79.6	97.5

Note: In the RFCBAMConv and DyHead columns, "√" indicates that the corresponding module is included, and a blank cell indicates that it is not included.

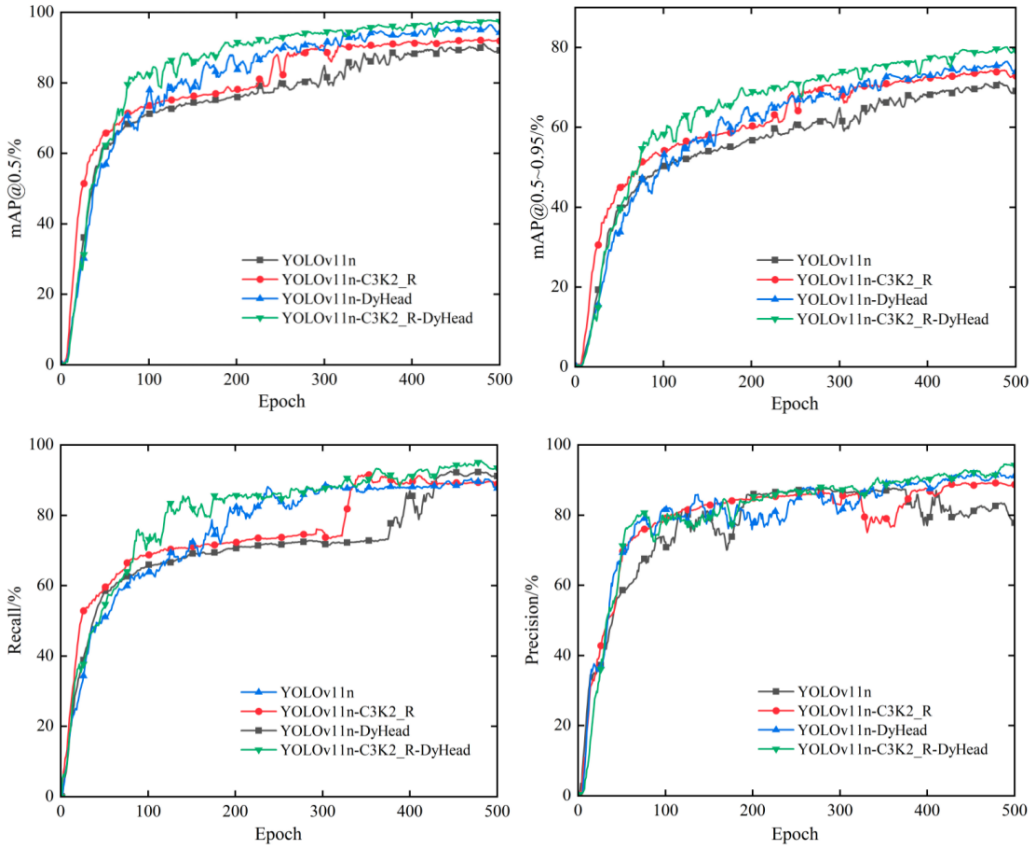


Fig. 4. Comparison of precision, recall, *mAP*@0.5, and *mAP*@0.5:0.95 among the baseline and improved models.

Ablation experiments show that, on the self-built dataset, the three improved models outperform the original YOLOv11n model in identifying student classroom behaviors. First, after using the RFCBAMConv structure to improve the C3k2 module in the original network, *mAP*@0.5 is improved by 2.5 percentage points, and *mAP*@0.5:0.95 is improved by 3.6 percentage points. This shows that the RFCBAMConv module effectively alleviates the information loss problem of traditional convolution in multi-scale feature fusion by adaptively adjusting the receptive field and channel attention mechanism, and also significantly enhances the sensitivity and adaptability to targets of different scales, especially in complex classroom scenes, and can better capture the detailed features of student behavior. Second, after replacing the original detection head with the dynamic detection head DyHead, the model's *mAP*@0.5

and *mAP*@0.5:0.95 improve by 6.3 percentage points and 5.4 percentage points, respectively. This shows that DyHead not only reduces the loss of detailed information but also enhances the ability to capture multi-scale, fine-grained features and effectively handles small targets and occlusion, thereby improving the model's detection stability in complex scenes. Finally, when RFCBAMConv and DyHead are integrated into YOLOv11n, the resulting YOLOv11n-RD model increases *mAP*@0.5 and *mAP*@0.5:0.95 by 8.4 and 9.8 percentage points over the baseline, respectively. Compared with DyHead alone, the combined model further improves *mAP*@0.5 and *mAP*@0.5:0.95 by 2.1 and 4.4 percentage points, respectively. These results support the effectiveness of combining local optical feature enhancement with adaptive multi-scale detection for classroom behavior recognition.

Therefore, the YOLOv11n-RD model proposed in this paper significantly improves small-target detection performance, better identifies and detects students' classroom behaviors, reflects students' learning status, and is suitable for behavior detection in real classroom scenarios. The proposed model not only accurately identifies typical classroom behaviors such as hand-raising and reading, but also adapts well to complex classroom scenarios involving multi-person interactions and partial occlusions, providing reliable technical support for quantifying student engagement distribution and building intelligent teaching evaluation systems.

3.6. Visualization

The difference in the effect of the original YOLOv11n model and the improved YOLOv11n-RD model on student classroom behavior detection in the same classroom scene is shown in Fig. 5. From the comparison chart, it can be seen that the YOLOv11n-RD model has stronger detection capabilities than the YOLOv11n model in classroom scenes with dense students and complex backgrounds, and can accurately identify and locate student behavior targets, significantly improving the detection accuracy of student behavior. This result verifies the effectiveness of the improved algorithm proposed in this paper for handling real classroom scenes and complex background tasks. By introducing the receptive field attention mechanism and dynamic detection head structure, the model improves mAP@0.5:0.95 by 9.8 percentage points on the optical classroom surveillance dataset. In particular, the model significantly enhances its ability to capture the features of typical classroom behaviors such as raising hands, lowering heads, and using mobile phones.



Fig. 5. Detection results of YOLOv11n (left) and YOLOv11n-RD (right) on the same classroom image.

4. Conclusions

In this study, an optics-oriented YOLOv11n-RD algorithm was proposed to improve the accuracy and robustness of student classroom behavior recognition from optical surveillance imagery. Existing methods often suffer from false detections and missed detections because

classroom optical images contain small behavioral targets, partial occlusion, complex illumination, background interference, and perspective-induced scale variation. To address the C3k2 module's limitations in extracting small-target optical features, we introduced the RFCBAMConv module to adaptively adjust the receptive field and improve the model's perception of optical targets at different scales. In addition, DyHead was introduced to replace the original detection head; this structure integrates scale-aware, spatial-aware, and task-aware attention, further enhancing the representation of small-target optical features and reducing information loss caused by downsampling.

Comparative experiments on a self-built optical classroom image dataset show that the proposed YOLOv11n-RD algorithm achieves superior detection performance compared with the baseline and other benchmark models. The results indicate that the proposed method more effectively captures behavior-related optical features in complex classroom scenes, including hand-raising, reading, writing, phone use, head lowering, and leaning on the table, while remaining robust under dense-student and partially occluded optical imaging conditions. Overall, the YOLOv11n-RD model provides reliable technical support for optical-imaging-based classroom behavior monitoring, student engagement quantification, and intelligent teaching evaluation. Future work may further incorporate optical calibration, illumination normalization, low-light enhancement, and multispectral or polarization imaging to extend the method to more challenging optical classroom environments.

Declaration of competing interests. The authors declare no conflicts of interest.

Funding and acknowledgements. This research work was supported by the Special Fund of Advantageous and Characteristic disciplines (Group) of Hubei Province; the Fundamental Research Funds for the Municipal Universities of Beijing (Ee202531); the research project on digital education in Beijing (BDEC2024ZX042); the National Natural Science Foundation of China (62472040); the Jiangxi Provincial Natural Science Foundation (20242BAB20041); and the Nanchang Institute of Technology Special Project of the Research Project Support Program for Doctoral Faculty (NLZX2613).

Authors' contributions. Min Peng: writing – original draft; Yujia Fu: methodology, formal analysis; Zufang Yang: formal analysis; Xiaokun Yang: software, investigation; Xiangqing Wang: sources; writing – review and editing; Dongfei Wang: supervision, writing – review and editing.

References

1. Zhou, J., Ran, F., Li, G., Peng, J., Li, K., & Wang, Z. (2022). Classroom learning status assessment based on deep learning. *Mathematical Problems in Engineering*, Article 7049458.
2. Wang, Z., Li, L., Zeng, C., & Yao, J. (2023). Student learning behavior recognition incorporating data augmentation with learning feature representation in smart classrooms. *Sensors, 23*(19), Article 8190.
3. Abozeid, A., Alrashdi, I., & Al-Makhlasawy, R. M. (2026). Intelligent recognition of students' behavior for smart learning environments. *Scientific Reports*, 16, Article 5674.
4. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 886–893).
5. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 580–587).
6. Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1440–1448).

7. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.
8. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779–788)..
9. Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement* (arXiv:1804.02767). arXiv.
10. Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal speed and accuracy of object detection* (arXiv:2004.10934). arXiv.
11. Ultralytics. (2021). YOLOv5: A state-of-the-art real-time object detection system [Computer software]. GitHub. https://github.com/ultralytics/yolov5
12. Khanam, R., & Hussain, M. (2024). *YOLOv11: An overview of the key architectural enhancements* (arXiv:2410.17725). arXiv.
13. Liu, S., Huang, D., & Wang, Y. (2018). Receptive field block net for accurate and fast object detection. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), In *Computer Vision – ECCV 2018* (pp. 385–400). Springer.
14. Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., & Song, Y. (2023). *RFAConv: Innovating spatial attention and standard convolutional operation* (arXiv:2304.03198). arXiv.
15. Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), In *Computer Vision – ECCV 2018* (pp. 3–19). Springer.
16. Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7369–7378).
17. Li, F., Jiang, Z., Zhou, S., Deng, Y., & Bi, Y. (2022). Spilled load detection based on lightweight YOLOv4 trained with easily accessible synthetic dataset. *Computers and Electrical Engineering, 100*, Article 107944.

Peng, M., Fu, Yu., Yang, Z., Yang, X., Wang, X., Wang, D. (2026). Optics-Oriented Student Classroom Behavior Recognition Using an Improved YOLOv11n-RD Model for Optical Surveillance Imaging. *Ukrainian Journal of Physical Optics*, *27*(4), 04207 – 04219.
doi: 10.3116/16091833/Ukr.J.Phys.Opt.2026.04207

Анотація. *Поведінка учнів у класі є важливим показником для оцінювання навчального процесу. Оптичний нагляд у класі забезпечує безконтактний спосіб спостереження за поведінкою учнів, що охоплює велику площу. Однак надійне розпізнавання на основі таких оптичних зображень залишається складним завданням, оскільки дії учнів часто є дрібномасштабними, частково перекритими іншими об'єктами, а також зазнають впливу складного освітлення, захащеного фону, перспективних спотворень і втрати оптичних ознак під час зменшення розмірності даних у нейронній мережі. Для розв'язання цих проблем у статті пропонується вдосконалений алгоритм YOLOv11n-RD, орієнтований на оптичні дані, призначений для розпізнавання поведінки учнів у класі за допомогою зображень із систем спостереження. По-перше, для вдосконалення модуля C3k2 застосовано блок уваги на основі згорткових шарів із рецептивним полем (RFCBAMConv). Це покращує вилучення оптичних ознак у різних масштабах і дозволяє моделі зосереджуватися на локальних візуальних деталях, пов'язаних із дрібними діями учнів. По-друге, замість стандартного блоку виявлення об'єктів використано динамічний блок (DyHead), який інтегрує механізми уваги, що враховують масштаб, просторове розташування та специфіку завдання. Це підвищує точність виявлення об'єктів і зменшує втрату оптичних ознак у складних умовах класного приміщення. Для навчання та тестування використовувалися реальні оптичні зображення, отримані в класах. Експериментальні результати показують, що вдосконалена модель здатна точніше розпізнавати поведінку учнів на основі зображень із систем спостереження. Запропонована модель досягає показника mAP@0.5 на рівні 97,5%, що на 8,4 відсоткових пункти перевищує результат базової моделі YOLOv11n. Це свідчить про підвищену стійкість моделі до складних умов фільмування при розпізнаванні поведінки учнів.*

Ключові слова: *оптичний нагляд у класі, розпізнавання поведінки учнів, YOLOv11, RFCBAMConv, DyHead, оптичні зображення*