IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2024, accepted 1 March 2024, date of publication 7 March 2024, date of current version 13 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3373536

RESEARCH ARTICLE

YOLOv8n_BT: Research on Classroom Learning Behavior Recognition Algorithm Based on Improved YOLOv8n

QINGTANG LIU[1,2], (Member, IEEE), RUYI JIANG[1,2], QI XU[1,2], DENG WANG[1,2], ZHIQIANG SANG[1,2], XINYU JIANG[1,2], AND LINJING WU[1]

[1]Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, Hubei 430000, China
[2]Hubei Research Center for Educational Informatization, Central China Normal University, Wuhan, Hubei 430000, China

Corresponding author: Qingtang Liu (liuqtang@mail.ccnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62277021, in part by the Fundamental Research Funds for the Central Universities under Grant CCNU22JC011, and in part by the National Teacher Development Collaborative Innovation Experimental Base Construction Research Program under Grant CCNUTEIII 2021-12.

ABSTRACT Classroom learning behavior recognition can provide effective technical support for teaching and learning. However, in natural classroom teaching scenarios, classroom learning behaviors are often missed or falsely detected due to character occlusion and the small object. To tackle the above issues, this study proposed an improved classroom learning behavior recognition algorithm (YOLOv8n_BT) based on YOLOv8n. On the one hand, for the occlusion problem of classroom learning behaviors, this study incorporated the BRA into the Backbone to better capture feature information; on the other hand, for the small object problem of classroom learning behaviors for back-row-students, this study expanded a Tiny Object Detection Layer (TODL) to detect small targets better. Experiments show that the BRA and the TODL can significantly improve the model performance. The YOLOv8n_BT model, which incorporated both the BRA and the TODL into the YOLOv8n(baseline) model simultaneously, has the most significant performance improvement. Compared with the YOLOv8n(baseline), the YOLOv8n_BT model improved by 3.0%, 6.7%, 5.0%, 3.6%, and 9.0% on P, R, F1, mAP50, and mAP50-90, respectively. The detection performance of YOLOv8n_BT also outperforms other state-of-the-arts.

INDEX TERMS YOLOv8, BRA mechanism, learning behavior recognition, target detection, occluded targets, small targets.

I. INTRODUCTION

The performance of the student's learning behavior is crucial to instruction and assessment [1], [2]. Assessment of classroom learning behaviors usually includes manual and automated measures [3]. Manual measurement mainly includes self-reporting, interviews, and observation. However, they have the problems of retrospective bias, high subjectivity, and low efficiency [4]. With the development of artificial intelligence in the education community, using smart technology to track and detect classroom learning behaviors has become a new trend. Classroom learning behavior recognition results can be utilized to analyze and visualize behavior statistics, teaching patterns, etc. This is important for conducting learning situation analysis, learning diagnostics, and achieving comprehensive, process-oriented, "multidimensional" assessments. Classroom learning behavior analysis is a data-driven mechanism. It promotes the integration of objective and quantitative assessment, improving the accuracy of classroom instruction assessment.

Currently, deep learning is progressively improving the automatic measurement of classroom behavior. Automatic classroom behavior measurement methods have become a research hotspot in the field of education informatization [6], [7] due to their highly automated, real-time, and efficient features [5]. Object detection algorithms based on deep

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

learning have become mainstream, such as YOLO technology [8], SSD technology [9], and fasterRCNN [10] technology. However, classroom learning behavior recognition is a complex issue. The difficulties are (1) classroom learning behaviors are often obscured; (2) The learning behaviors for back-row-students are small objects. The object detection methods are not friendly to recognizing occluded objects and small objects in natural classrooms [11]. As a result, classroom learning behaviors are often missed or falsely detected. This leads to low recall and precision in identifying classroom learning behaviors, ultimately failing to provide a comprehensive and accurate response to classroom learning engagement. Therefore, for the problem of missed and false detection of classroom learning behaviors due to occlusion problems and small object problems for classroom learning behaviors, improving the recall and precision of classroom learning behavior recognition is a pressing issue in current education.

Among the existing algorithms for object detection, YOLO is popular due to its perfect speed-accuracy balance. YOLOv8 is the next major update to YOLOv5, open-sourced by Ultralytics, on January 10, 2023. The YOLOv8 model exhibits a faster and more accurate performance, thereby delivering enhanced technical support for classroom learning behavior recognition. Therefore, this study creates a dataset of learning behavior detection for elementary students in natural classroom scenarios. To address the above issues, this study proposed an improved classroom learning behavior recognition algorithm YOLOv8n_BT using YOLOv8n as the baseline model. The main contributions of this study are as follows:

(1) Constructing a classroom learning behavior dataset. A small-scale classroom learning behavior dataset was constructed based on regular elementary classroom videos;

(2) Incorporating Bi-Level Routing Attention(BRA) [12]. Aiming at the occlusion problem of classroom learning behaviors, the BRA was incorporated into the Backbone to preserve fine-grained detail information and better capture global information and rich contextual information;

(3) Adding a Tiny Object Detection Layer(TODL). To address the problem of small objects for classroom learning behaviors, a TODL was added to improve the performance of capturing small object feature information;

(4) Proposing an improved classroom learning behavior recognition algorithm, YOLOv8n_BT. On the self-constructed dataset of this study, the P-value, R-value, and mAP (50-90) of YOLOv8n_BT are improved by 3%, 6.7%, and 9.0%, respectively, compared with the YOLOv8n (baseline). The improved model(YOLOv8n_BT) effectively addresses the problem of missed and falsely detected classroom learning behaviors.

The rest of the paper is organized as follows. Section II describes the related work of this study, reviewing object detection, occluded object detection, and small object detection. Section III describes the overall framework and implementation details of the YOLOv8n_BT model proposed in this study. Section IV shows specific experiments and experimental results. Section V presents conclusions and future work.

II. RELATED RESEARCH WORK

A. OBJECT DETECTION

Object detection task involves recognizing both the position of objects (localization) and categorizing each object (classification) within a given image [13]. Object detection broadly consists of traditional and deep learning-based algorithms [14]. The latter can be categorized into single-stage and two-stage detection algorithms based on the prediction stage required by the detector [15]. Single-stage algorithms perform relatively poor accuracy [16] but can significantly improve the computational speed. With the development of deep learning, single-stage algorithms have achieved comparable accuracy to two-stage algorithms. However, their performance on small object detection should be improved [17], [18], especially in severe occlusion situations. SSD [9] and YOLO [8] are the main single-stage algorithms. The YOLO (You Only Look Once) algorithm family is trained end-to-end to improve accuracy and have good compatibility. With the advent of the YOLO architectural successor, the detection accuracy of YOLO is improving significantly. Sometimes, the detection accuracy of YOLO is better than that of two-stage algorithms [19]. The YOLO is adopted in various fields. For example, Bie et al. [20] proposed an improved lightweight YOLOv5 algorithm (YOLOv5n-L) that can be applied to mobile terminal devices to achieve real-time accurate detection of vehicle targets. Yang et al. [21] used the YOLO neural network for end-to-end prediction of the travel area of agricultural machinery. YOLOv8 is the latest version of YOLO series at present, which was released on January 10, 2023. Aiming at the features and shortcomings of YOLOv8, many researches have improved and applied YOLOv8 by combining the features and needs of their respective research fields. Finally, better experimental results were achieved. For example, to cope with dense fish populations and underwater plants that obscure them, Li et al. [22] integrated an innovative module in Real-time Detection Transformer (RT-DETR) into YOLOv8 and applied repulsion loss; aiming at the large-scale changes of different forms of traffic signs and the rapid speed of vehicles, Zhang et al. [23] implemented multi-scale traffic sign detection based on YOLOv8 by introducing the attention module and RFB module and improving the loss function; in response to the blurriness of UAV-collected images and the large number of small target objects, Wang et al. [24] introduced a small target detection structure (STC) and the global attention GAM into YOLOv8.

The YOLO series is also widely used in the field of educational research. Chen and Guan [25] proposed an improved YOLOV4 behavior detection algorithm to recognize the behavior of teachers and students based on classroom teaching scenarios. Kumari et al. [26] used YOLOV4 to detect mobile eye-tracking data in a student laboratory session. Xu et al. [27] introduced the simAM attention module into YOLOv8 to detect students' cognitive-behavioral engagement. However, the above studies do not take into account the challenges of traditional classroom target detection, such as severe occlusion and small scale of students in the back rows, which lead to high miss and false detection rates for student behavior recognition. In conclusion, the YOLO algorithm family is widely used in various industries due to its good speed-accuracy balance.However, YOLO needs to be further adapted and improved by combining the characteristics of application scenarios and research needs in the traditional classroom.

B. SMALL OBJECT DETECTION

In deep learning, current object detection algorithms have achieved good results on medium and large targets. However, the performance of small target detection is not satisfactory. Because small targets have a small percentage of area in the image, which makes it difficult for the model to obtain adequate feature information [28]. There are four challenges for small object detection: insufficient feature information, limited contextual information, uneven distribution of categories, and insufficient positive instances. To address the above challenges, the existing solutions mainly include super-resolution techniques, context-based information, multi-scale representation learning, data augment, and loss function-based [29]. Liu et al. [30] proposed an algorithm to generate clear, high-resolution faces directly from blurred small faces using GAN. References [11], [31], [32], and [33] incorporated attention mechanisms into models to acquire contextual detail information, alleviating the problem of small target detection in their respective research areas. To address the limitations of visual tracking due to target scale changes, Gu et al. [34] proposed a novel parallel Transformer network architecture based on the attention mechanism. References [35] and [36] added detection heads for small targets to capture multi-scale information, significantly improving the model's performance in detecting small targets. References [37] and [38] improved the model's small object detection performance by optimizing the loss function. Data augment (e.g., geometric transformations, color transformations, random occlusion, etc.) is also an effective way to improve the robustness of the model for small object detection. For example, Gao et al. [39] increases the training set by random flipping, Bochkovskiy et al. [40] uses a mosaic enhancement technique for small object detection in images, and Zhang et al. [41] uses a flipping mosaic algorithm to enhance the network's perception of small targets. The results show that all the above methods are beneficial in improving the detection performance for small objects in images.

C. OCCLUDED OBJECT DETECTION

Occlusion is the other challenge for object detection in real-world scenarios [42], [43], [44]. The difficulties in dealing with occluded objects are: (1) Occlusion interferes with feature extraction. (2) Occlusion causes overlapping prediction frames, which can be wrongly filtered out by non-maximal suppression (NMS), leading to missed detections. (3) Complex occlusion datasets make it difficult for the model to have strong robustness. To address the above difficulties, optimization mainly includes data augment, objective structure improvement, loss function improvement, and non-maximal suppression improvement. Yun et al. [45] performed image augment by cutting and pasting masking blocks on the training image, so that the negative effect of uninformative pixels can be avoided during the training process, making the training more effective. References [46], [47], and [48] incorporated different attention mechanisms in models to capture global and rich contextual information. The visible part is fully utilized for detection, thereby effectively reducing the effect of occlusions. To cope with the occlusion problem of target tracking, Yuan et al. [49] developed an Aligned Spatial-Temporal Memory network-based Tracking method (ASTMT), and Gu et al. [50]proposed a novel shared-encoder dual-pipeline Transformer architecture. Compared to the Common Mean Square Error Loss Function, L1 Loss Function, and L2 Loss Function, more new loss strategies are proven to be useful, such as IoU Loss [51], Focal Loss [52], GIoU Loss [53], DIoU Loss [54], EIoU Loss [55] and so on. Wang et al. [56] proposed a repulsion loss specifically designed for crowded scenes. This loss function plays a good optimization role in pedestrian-dense occlusion detection. Tan et al. [57] introduced soft non-maximum suppression to minimize the occurrence of missed targets due to occlusion. Guo et al. [58] optimized the NMS algorithm by linear attenuation confidence score to improve the detection accuracy of occluded vehicles. It has been shown that all the above methods help to improve the detection performance of occluded targets in images.

YOLO is the leading target detector due to its perfect speed-accuracy balance [59]. However, the YOLO family is not friendly to detecting small objects and occluded objects, because YOLO lacks shallow network information and does not have full access to global and contextual information [11]. To overcome YOLO's limitations and meet research needs, studies have optimized the model, successfully applying improved versions in various fields [60], [61], [62]. In summary, referring to the above optimization schemes of occlusion target detection and small target detection, this study carried out model improvement based on YOLOv8n. The improved model is expected to increase the recall and precision of classroom learning behavior recognition and thus reflect classroom learning more comprehensively and accurately.

III. METHODOLOGIES

A. BASELINE MODEL YOLOv8N

YOLO is the leading target detector due to its good speed-accuracy balance. YOLOv8 is the update to YOLOv5, open-sourced by Ultralytics, on January 10, 2023. It provides the most advanced object detection performance. Compared with YOLOv5, YOLOv8 improvements are mainly as follows:

(1) The C3 module in YOLOv5 is replaced with the C2f module to achieve further lightweight;

(2) YOLOv8 abandons the previous Anchor-Base and uses the Anchor-Free idea;

(3) YOLOv8 uses Decoupled-Head to decouple the classification and detection processes;

(4) It uses the sample matching method of TaskAlignedAssigner [63]. There is no Objectness loss branch compared to the YOLOv5.

The YOLOv8 consists of four layers: Input, Backbone, Neck, and Head. It includes five architectures of YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. The architectures are suitable for datasets of different sizes. Because the dataset in this study is relatively small, we chose YOLOv8n as the baseline experimental model. The YOLOv8n architecture is shown in Fig.1.

B. OVERALL FRAMEWORK OF IMPROVED MODELING: YOLOv8N_BT

The difficulties of learning behavior recognition in natural classroom scenarios are: (1) Classroom learning behaviors are often obscured. There are many students, desks, chairs, and books in natural lecture scenarios, resulting in classroom learning behaviors often occluded. The occlusion of behaviors interferes with the model for feature extraction, which makes it difficult for the model to have strong robustness; (2) Classroom learning behaviors for back-row-students are small objects. Students in the front or back rows have different proportions of pixels in the image [64]. Back-row-students have smaller pixels and have more serious occlusion problems [65]. As a result, the pixel scales of classroom learning behaviors are inconsistent, especially the classroom learning behaviors of back-row-students have insufficient pixels. So, it is difficult for the model to extract effective feature information. Aiming at the above-mentioned problems, this study proposed an improved algorithm based on YOLOv8n: YOLOv8n_BT. The structure of the YOLOv8n_BT model is shown in Fig. 2, and the part of the dashed module is the module added to this study. The functions of the improved model are as follows:

(1) Introducing the BRA. The BRA can better capture fine-grained, global, and rich contextual information. It provides high accuracy and computational efficiency in intensive prediction tasks [12]. The BRA was introduced in layer 7 in Backbone to address the occlusion problem for classroom learning behaviors.

(2) Adding a TODL. The TODL can better extract feature information for relatively smaller objects. It is specifically designed to recognize the classroom learning behavior of back-row-students. This study added an upsampling layer and a downsampling layer to the 17th -22nd layers of the Neck and a tiny object detection layer to the Head. This is to solve the small object problem for classroom learning behaviors.

C. SPECIFIC IMPROVEMENT MEASURES

1) BI-LEVEL ROUTING ATTENTION

The BRA was introduced to address the occlusion problem of classroom learning behaviors. The BRA is a dynamic, query-aware sparse attention mechanism. The key idea of BRA is to filter out the least relevant key-value pair vectors at the coarse region level, retaining only a small fraction of the routing regions. Then, a fine-grained token-to-token attention mechanism is applied to concatenating these routing regions [12]. The results show that the BRA can effectively optimize small targets and dense occlusion in computer vision tasks such as object detection and semantic segmentation [12], especially in intensive prediction tasks. Therefore, this study incorporated the BRA to YOLOv8n for the occlusion problem of classroom learning behaviors. To fit our data better, this study has done subsidiary experiments. We introduced BRA into different model layers to improve the model performance. Based on the experimental results (see Section IV-C1), the BRA was finally added to the seventh layer of the model, as shown in Layer 7 of Fig. 2. The 40 × 40 network feature maps outputted from Layer 6 are fed into the BRA. The BRA is shown in Fig. 3.

First, region partition and input projection. Input feature map X with dimension H×W×C, $X \in \mathbb{R}^{H\times W\times C}$, divide this feature map into S×S non-overlapping regions, each including $\frac{HW}{S^2}$ feature vectors, i.e., turn X into X^r, $X^r \in \mathbb{R}^{S^2\times\frac{HW}{S^2}\times C}$, and then obtain the query, key, value tensor,Q, K, $V \in \mathbb{R}^{S^2\times\frac{HW}{s^2}\times C}$, by linear projections:

$$Q = X^r W^q\,,\quad \mathrm{K} = X^r W^k \quad \mathrm{v} = X^r W^V \tag{1}$$

where W^q,W^k,$W^v \in \mathbb{R}^{C\times C}$ are projection weights for the query, key, value, respectively.

Second, a directed graph is constructed through the adjacency matrix, and the region-to-region routing of the directed graph is used to find the attending relationship corresponding to different key-value pairs. Specifically, (1) the average values of Q and K in each region are computed to obtain Q^r, $K^r \in \mathbb{R}^{S^2\times C}$; (2) the adjacency matrix A^r of the region-to-region affinity graph is computed by matrix multiplication between Q^r and the transposed K^r, $A^r \in \mathbb{R}^{S^2\times S^2}$, and the entries in the adjacency matrix A^r measure the degree to which the two regions are semantically related; (3) only the first k connections of each region are kept to prune the relevance graph, specifically, deriving a routing index matrix I^r, $I^r \in \mathbb{N}^{S^2\times k}$,which keeps the indexes of the first k connections row by row, and the i^{th} row of I^r contains the indexes of the first k most relevant regions in the ith region:

$$\mathbf{A}^r = \mathbf{Q}^r(\mathbf{K}^r)^T \tag{2}$$

$$\mathbf{I}^r = \mathrm{topkIndex}(\mathbf{A}^r) \tag{3}$$

Finally, fine-grained token-to-token attention is applied using the region-to-region routing index matrix I^r. The key and value tensors are first gathered, then attention is applied to the gathered key-value pairs:

$$\mathbf{K}^g = gather\,(\mathbf{K}, \mathbf{I}^r) \tag{4}$$

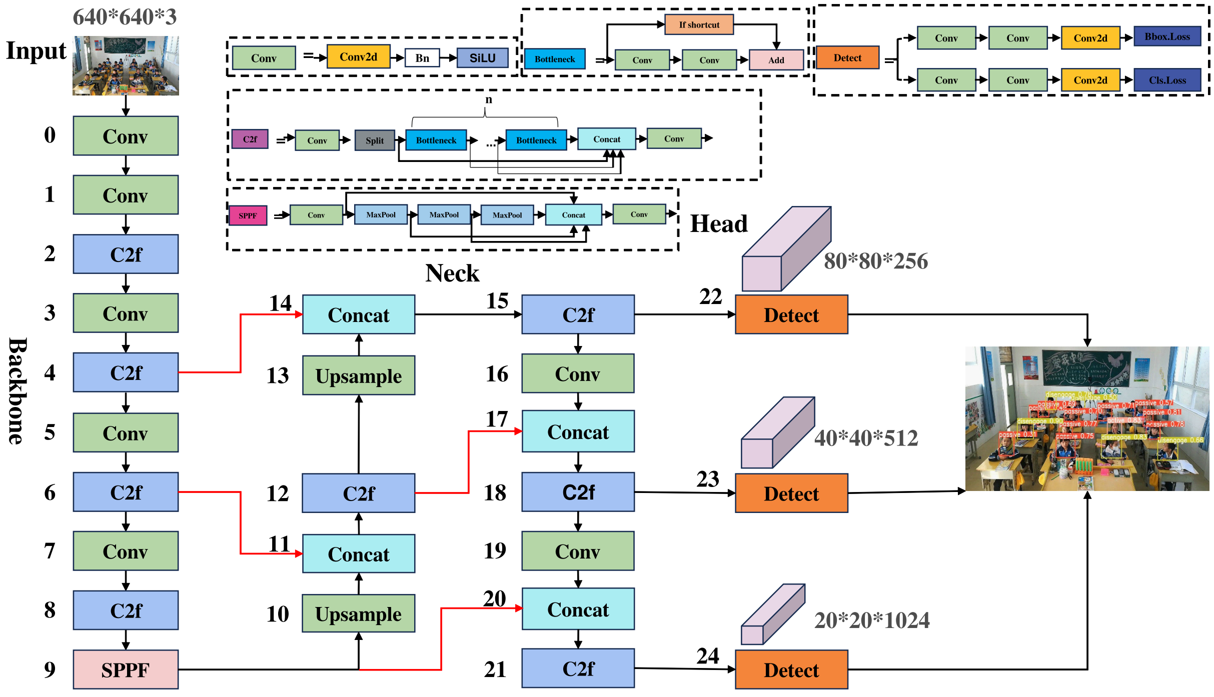


FIGURE 1. Model structure of YOLOv8n.

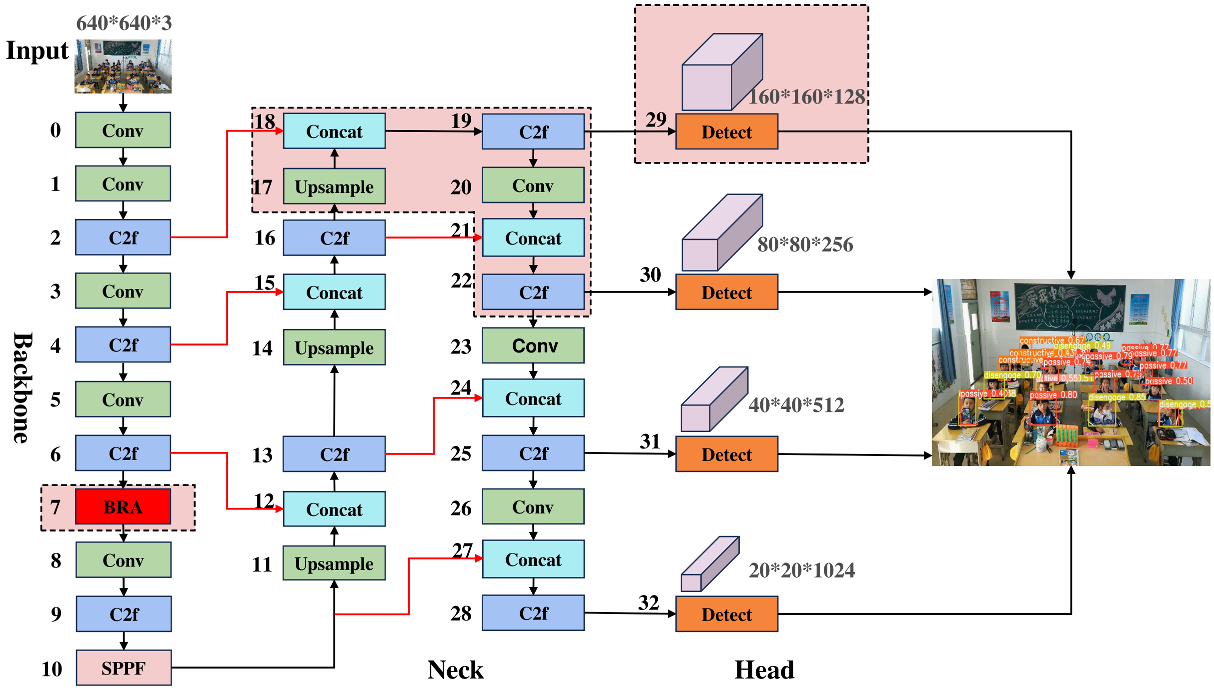


FIGURE 2. Model structure of YOLOv8n_BT.

$$\mathbf{V}^{g} = gather\left(\mathbf{V}, \mathbf{I}^{r}\right) \tag{5}$$

$$\mathbf{O} = Attention(\mathbf{Q}, \mathbf{K}^{g}, \mathbf{V}^{g}) + LCE(\mathbf{V}) \tag{6}$$

where $\mathbf{K}^{g}$ and $\mathbf{V}^{g}$ are gathered key and value tensor, and LCE ($\mathbf{V}$) is a local context augmentation term, parameterized as a function by deep convolution with convolution kernel size set to 5.

This attention mechanism saves the number of parameters and computation by gathering key-value pairs in the first k relevant windows and utilizing sparsity operations to skip the computation of the least relevant regions directly. The final 40 × 40 network feature map with identified vital information is then output and fed into layer 8 for feature learning, as shown in Fig. 2.

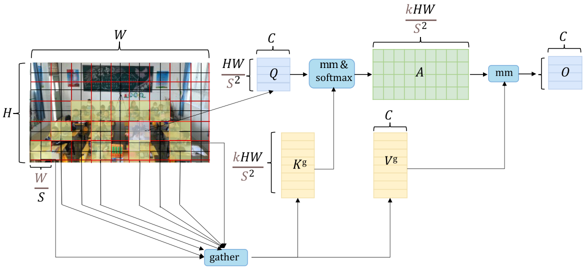


FIGURE 3. Bi-Level routing attention.

2) A TINY OBJECT DETECTION LAYER

The TODL was added to solve the small object problem for classroom learning behaviors. The head detection layer of the YOLOv8n(baseline) has only three detection layers. They detect small, medium, and large targets, respectively. However, the YOLOv8n(baseline) model often suffers from missed detection or poor detection accuracy for occlusion groups of inconsistent size, especially for small target objects [35]. It has been shown that adding a smaller object detection layer improves the capture information of different dimensions and facilitates small object detection [32], [66]. Classroom learning behaviors of back-row-students have a small proportion of pixels in the image, leading to the fact that classroom learning behaviors are often missed or falsely detected. To address this problem, this study added a TODL to YOLOv8n, as shown in the dashed part of Neck and Head in Fig. 2. Detailed parameters of the network structure of the TODL are shown in Table 1.

TABLE 1. Parameters of the tiny target detection layer network.

layers	module	Input		Output	
		input layer	feature map	output layer	feature map
17th	Upsample	16th	80×80×256	18th	160×160×256
18th	Concat	2nd 17th	160×160×128 160×160×256	19th	160×160×384
19th	C2f	18th	160×160×384	20th 29th	160×160×128
20th	Conv	19th	160×160×128	21st	80×80×512
21st	Concat	16th 20th	80×80×256 80×80×512	22nd	80×80×768
22nd	C2f	21st	80×80×768	23rd 30th	80×80×256
29th	Detect	19th	160×160×128	-	-

An upsampling module is added to layers 17-19. The network feature map of size 80 × 80 output from layer 16 of the model is fed into layer 17 for up-sampling. The up-sampling will magnify the network feature map by a factor of 2, resulting in a 160 × 160 network feature map. The enlarged image allows the model to learn detailed information about small targets, which improves the model's robustness in detecting the students' learning behaviors in the back row. At layer 18, the output feature map of layer 17 is feature-connected to the output feature map of layer 2 for feature fusion and connecting process information. In layer 19, the 160 × 160 network feature map is output through the C2f module to perform convolution and feature learning. The network feature map output from layer 19 is delivered to the Detect layer (layer 29), external to layer 19. This Detect layer (Layer 29, the TODL) is used to detect relatively more minor targets.

A downsampling module is added to layers 20-22. At layer 20, the network feature map output from layer 19 is reduced by half with a step size of 2 through a 3 × 3 convolution kernel to obtain an 80 × 80 network feature map; at layer 21, the network feature map from layer 20 is spliced with the network feature map from layer 16 to fuse the features. At layer 22, the spliced network feature map is convolved by the C2f module to output an 80 × 80 network feature map. The network feature maps output from layer 22 is conveyed to the Detect layer (layer 30). The detect layer (layer 30) is used to detect small targets. Similarly, the feature maps of layers 13 and 24 are spliced and then convolved to obtain a 40 × 40 network feature map delivered to the Detect layer (layer 31). The detect layer (layer 31) is used to detect medium targets. The feature maps of layers 10 and 27 are spliced, and then a 20 × 20 network feature map is obtained by convolution and delivered to the Detect layer (layer 32). The detect layer (layer 32) is used to detect large targets.

IV. RESULTS AND DISCUSSION

A. DATASETS

1) DATA SOURCES

This study created a dataset of elementary students' classroom learning behaviors based on nature classroom videos. Before the data collection, video capture was agreed upon by the teacher and students. Image acquisition was accomplished by securing the camera to the front center of the classroom via a tripod. In the end, we obtained four videos with a duration of 40 minutes in the format of mp4. We selected one of the videos (19 students, 10 girls, and 9 boys, respectively) and extracted one frame of image every 3 seconds. Finally, we obtained 925 images for a total of 17,575 classroom learning behavior samples.

2) DATA ANNOTATION

This study adopt the ICAP (Interactive-Constructive-Active-Passive) framework to classify classroom learning behaviors into passive, active, constructive, interactive, and disengaged. The annotation details are in Table 2. The ICAP is a coding framework that distinguishes cognitive engagement based on patterns of explicit behaviors. The ICAP framework organically combines explicit behaviors with implicit cognitive states, providing a theoretical basis for calculating students' cognitive engagement through their learning behaviors [67], [68]. Compared with previous studies that coded behaviors such as looking up, looking down, looking left, looking right, and lying down, the ICAP framework can be directly coded for the student's learning engagement states, which allows for a more direct characterization of the students' learning states. We used an open-source software annotation

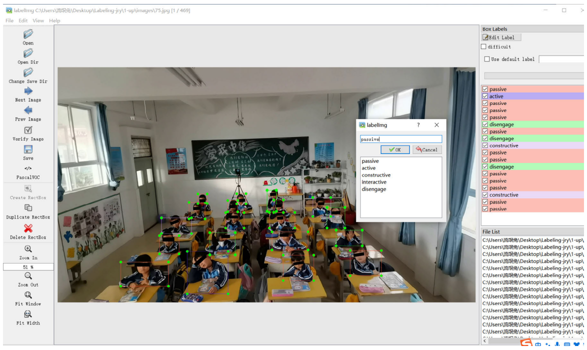

labelImg C:\Users\流氓兔\Desktop\Labeling-jry\1-up\images\75.jpg [1 / 469]
File Edit View Help
Open
Open Dir
Change Save Dir
Next Image
Prev Image
Verify Image
Save
PascalVOC
Create RectBox
Duplicate RectBox
Delete RectBox
Zoom In
51 %
Zoom Out
Fit Window
Fit Width
labelImg
passive
OK
Cancel
passive
active
constructive
interactive
disengage
Box Labels
Edit Label
difficult
Use default label
passive
active
passive
passive
passive
disengage
passive
disengage
constructive
passive
passive
disengage
passive
passive
passive
constructive
passive
passive
File List
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\
C:\Users\流氓兔\Desktop\Labeling-jry\1-up\

FIGURE 4. LableImage annotation tool interface and annotation examples.

tool (LableImage) to annotate the images. The interface and annotation examples of the LableImage annotation tool are shown in Fig. 4. The software annotation tool can mark each student's location (coordinates of the annotation box) and classroom learning behavior category. Each student in each image is labeled with a classroom learning behavior category and corresponding location coordinates. Eventually, the labeling information of all students in each image is saved as an XML file named the same as the image. Three graduate students labeled data. The data with inconsistent coding was determined by consensus among the three coders. In this study, the labeled 17,575 data were divided into training and validation sets according to the ratio of 8:2, with 14,060 data in the training set and 3,515 data in the validation set.

B. TRAINING ENVIRONMENT AND EVALUATION INDICATORS

1) TRAINING ENVIRONMENT

The experimental environment and training parameters for training in this study are shown in Table 3.

2) EVALUATION INDICATORS

The evaluation metrics used in this study include precision (P), recall (R), F1 value, mAP50, mAP50-95, and FPS. Precision (P) indicates how many of the predicted positive samples are truly positive; recall (R) indicates how many positive classes in the samples are predicted correctly, and the F1 value is a combination of precision (P) and recall (R).

For each category, the AP is the average of the precision calculated at different confidence thresholds. Precision is measured by calculating the overlap between detected and real targets (usually using IoU, intersection-union ratio). The AP is calculated based on the precision-recall curve (PR curve). The mAP is the value obtained by averaging the APs of all the categories. The mAP is a critical metric for evaluating overall object detection system performance, and it is the average precision of the model when dealing with multiple categories. When calculating the mAP, a confidence threshold is usually used to determine the positive samples. The mAP50 is the mAP value calculated at a confidence threshold of 50%. The mAP50-95 is a more comprehensive evaluation metric, which calculates the mAP values for confidence thresholds ranging from 50% to 95%. The mAP50-95 can better evaluate the robustness and performance of the model. FPS, the number of images that can be processed within a second, is used to evaluate the speed of object detection. The calculation formulas are shown in Eqs. (7)-(12):

$$P = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{8}$$

$$F_1 = \frac{2TP}{P + R} \times 100\% \tag{9}$$

$$AP = \int_0^1 P(R)\, dR \tag{10}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{11}$$

$$FPS = \frac{N}{t} \tag{12}$$

where TP refers to the number of samples that are positive and are predicted to be positive, FP refers to the number of samples that are negative but are predicted to be positive, FN refers to the number of samples that are positive but are predicted to be negative, n is the number of target classes detected, APi is the AP of the i^{th} target class, N is the number of detected images, and t is the detection time.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) COMPARATIVE EXPERIMENTS ON THE LOCATION OF ATTENTION MECHANISMS

This study added the BRA to different layers of the Backbone and Neck modules of YOLOv8n, respectively. The positional ablation experiments of the attention mechanisms were conducted to evaluate the effect of the added positions of the BRA on the performance of the YOLOv8n algorithm. The positions of the BRA are added, as shown in Fig. 5, a-f. The red dashed modules are the BRA mechanisms added separately in different layers for this study. The Concat layer splices the output feature of the previous layer with those of the other layers to achieve feature fusion. The red arrows represent the output features of other layers. The experimental results are shown in Table 4, where the bold parts are the optimal results.

As seen from Table 4, compared to the YOLOv8n (baseline), adding BRA to different locations in the Backbone and Neck of YOLOv8n significantly improved the model's performance. This demonstrates the effectiveness of the BRA. Among them, when the BRA was integrated into the 7th layer of the YOLOv8n model, as depicted in Fig. 5b, it resulted in the most significant enhancement. This change led to a 3.9% increase in the precision, a 6.7% increase in the recall, a 4.1% increase in the F1 value, a 3.3% increase in mAP50, and an 8.1% increase in mAP50-90. Therefore,

TABLE 2. Detailed rules for classroom learning behavior coding.

Category	Detailed Rules for Data Annotation
passive	Reading textbooks, teachers, peers (when peers answer questions), presentation materials (PowerPoint, etc.)
active	Draw lines on the content, search for information, point to the learning content, and hof posture.
constructive	Raise hands to ask questions, take notes, propose new ideas, and ask peers.
interactive	Rotate conversations with teachers, peers, and initiate conversations on devices.
disengage	Viewing irrelevant content, looking around, not engaging in designated activities.

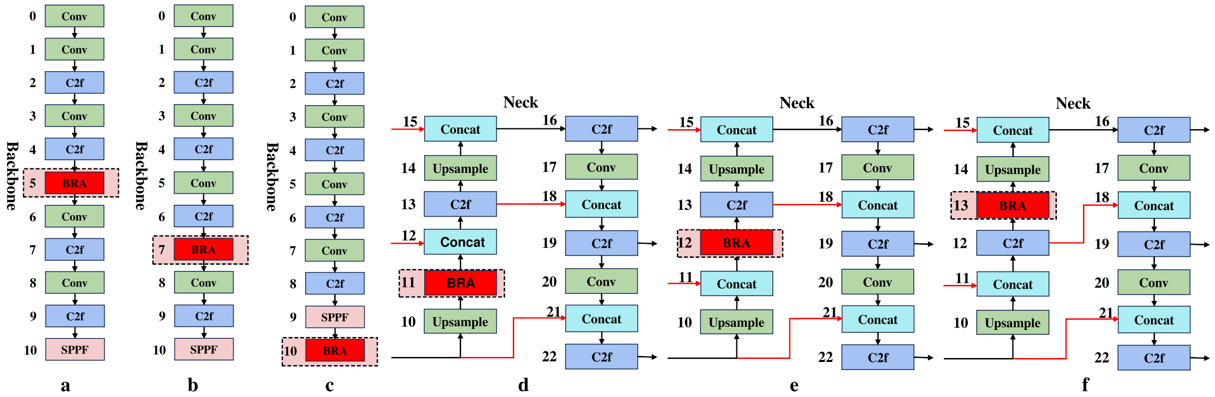


FIGURE 5. Location of BRA attention mechanism additions.

TABLE 3. Experimental environment.

Category	Name	Parameter
Hardware	CPU	Intel(R) Core(TM) i9-10920X CPU @ 3.50GHz
	GPU	NVIDIA GeForce RTX 3090 Ti
configfiguration	Memory	64G
Software	Operating System	Windows10
	Frameworks	Pytorch
configfiguration	Python	3.7.16
	CUDA	11.3
	Input image	640*640
	Epoch	300
Hyperparameter	Batch Size	32
	Initial learning rate	0.01
setting	Final learning rate	0.01
	momentum	0.937
	Weight decay	0.0005

TABLE 4. Experimental results of different integration positions in BRA.

Exp	BRA position	P （%）	R （%）	F1 （%）	mAP50 （%）	mAP50-90 （%）
1	YOLOv8n(baseline)	92.4	85.3	88.7	93.7	63.2
2	Backbone (*a*)	96.5	90.6	93.5	97.0	70.6
3	**Backbone** (*b*)	**96.3**	**90.7**	**93.4**	**97.0**	**71.3**
4	Backbone (*c*)	95.3	89.9	92.5	96.5	69.4
5	Neck (*d*)	95.5	89.4	92.3	96.0	69.1
6	Neck (*e*)	94.6	89.7	92.1	96.3	69.3
7	Neck (*f*)	94.9	90.3	92.5	96.3	69.1

based on the experimental results, this study incorporated the BRA into layer 7 of YOLOv8n.

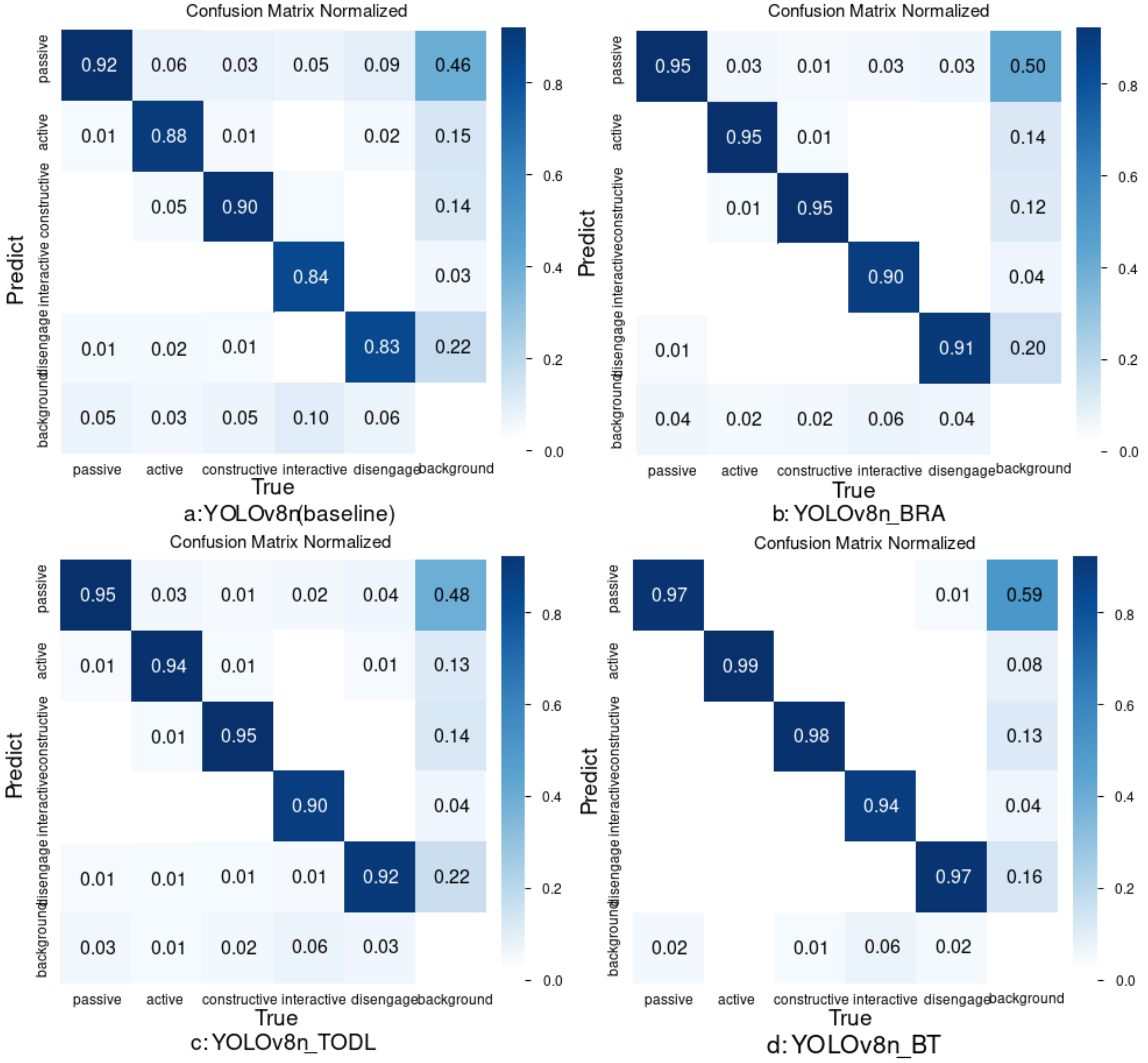
2) ABLATION EXPERIMENT

We conducted ablation experiments for both the BRA and the TODL. This was to verify the optimization effects of each improvement module comprehensively and to further evaluate the contributions of these enhancement techniques to the YOLOv8n algorithm. Experiment 1 (YOLOv8n) is the baseline model for this study. Experiment 2 (YOLOv8n_BRA) only incorporates the BRA in layer 7 of the YOLOv8n(baseline). Experiment 3 (YOLOv8n_TODL) only adds the TODL to the YOLOv8n(baseline). Experiment 4 (YOLOv8n_BT) adds both the BRA and the TODL to the YOLOv8n(baseline) at the same time. The experimental results are shown in Table 5, where the bolded parts are the optimal results. The optimization results of the model based on YOLOv8n (baseline) are shown in parentheses.

TABLE 5. Results of ablation experiment.

Exp	Model	P （%）	R （%）	F1 （%）	mAP50 （%）	mAP50-90 （%）	FPS
1	YOLOv8n (*baseline*)	92.4	85.3	88.7	93.7	63.2	312.5
2	YOLOv8n_BRA	96.3 (3.9)	90.7 (5.4)	93.4 (4.7)	97.0 (3.3)	71.3 (8.1)	322.6 (10.1)
3	YOLOv8n_TODL	95.2 (2.8)	91.6 (6.3)	93.4 (4.7)	97.2 (3.5)	71.5 (8.3)	312.5 (0)
4	**YOLOv8n_BT**	**95.4** **(3.0)**	**92.0** **(6.7)**	**93.7** **(5.0)**	**97.3** **(3.6)**	**72.2** **(9.0)**	**312.5** **(0)**

As seen from Table 5, the model's performance is significantly improved by adding the BRA and the TODL in the YOLOv8n, respectively, as well as by adding the BRA and the TODL simultaneously. The specific improvement effects are shown in the bracketed data in Table 5. The results of Experiment 2 show that incorporating the BRA into the YOLOv8n significantly improves the model's detection performance and speed. This result is consistent with the research result [12]. Zhu et al. [12] found that BRA could improve the detection performance and computational efficiency of the model due to its working principle. The results of Experiment 3 reveal that the TODL can significantly improve the detection performance of the model. The



a:YOLOv8n(baseline)

b: YOLOv8n_BRA

c: YOLOv8n_TODL

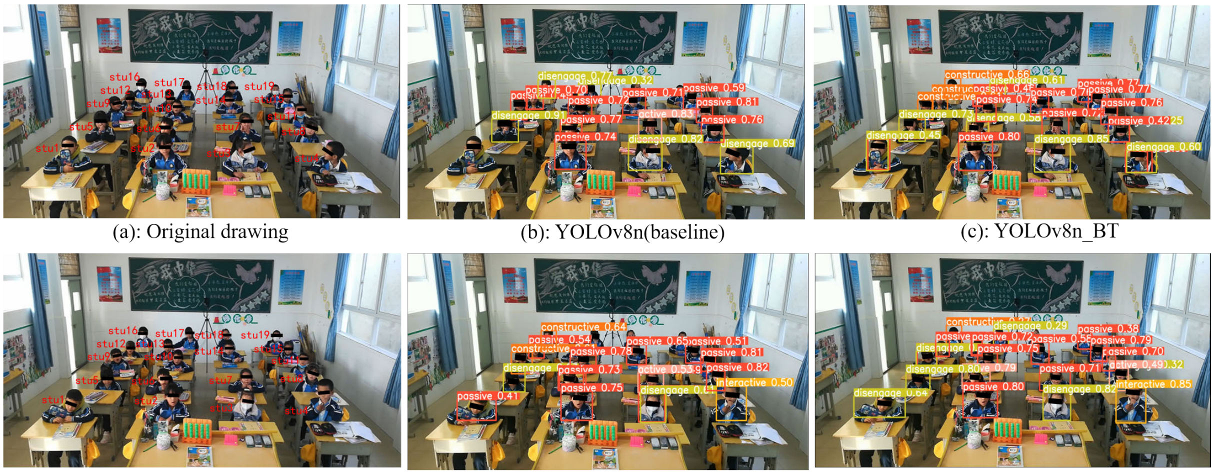
d: YOLOv8n_BT

FIGURE 6. Confusion matrix for the BRA and the TODL ablation experiments.

detection speed is relatively equal to the YOLOv8n. The results of Experiment 4 show that incorporating both the BRA and the TODL into the YOLOv8n can significantly improve the model performance without reducing detection speed.

Comparing Experiment 2 and 3, YOLOv8n_BRA and YOLOv8n_TODL improve the precision by 3.9% and 2.8% and the recall by 5.4% and 6.3%, respectively. This indicates that the BRA improves the precision of the model more than the TODL. In contrast, the TODL improves the recall of the model more than the BRA. Comparing Experiment 1, Experiment 2, Experiment 3, and Experiment 4, the study's results indicate that YOLOv8n_BT outperforms YOLOv8n, YOLOv8n_BRA, and YOLOv8n_TODL in terms of detection performance for all evaluation metrics, except for the precision, which is 0.9% lower than that of YOLOv8n_BRA. The detection rate of YOLOv8n_BT is 312.5 FPS (i.e., it can detect 312 frames per second). It can fully satisfy the requirement of real-time detection for daily video of 24-30 frames per second. When P, R, F1, mAP50, mAP50-90, and FPS metrics are considered simultaneously, the YOLOv8n_BT model has the best boost. For the missed detection and false detection problems of classroom learning behaviors, which are intended to be solved in this study, the recall and precision of YOLOv8n_BT are improved by 6.7% and 3.0%, respectively. It can be seen that YOLOv8n_BT can effectively improve the model's recall and precision and solve the problem of missed detection and false detection for classroom learning behaviors.

The confusion matrices of YOLOv8n_BT, YOLOv8n_BRA, YOLOv8n_TODL and YOLOv8n(baseline) are shown in Fig. 6. Each column of the confusion matrix represents the predicted category, and each row represents the real attributed



(a): Original drawing (b): YOLOv8n(baseline) (c): YOLOv8n_BT

(d): Original drawing2 (e): YOLOv8n (baseline) (f): YOLOv8n_BT

FIGURE 7. Comparison of recognition cases.

category of the data. The confusion matrix is mainly used to compare predicted and actual values. Table 6 displays the improvements in classroom learning behavior detection performance for YOLOv8n_BRA, YOLOv8n_TODL, and YOLOv8n_BT compared to the YOLOv8n(baseline). The bolded portion of Table 6 shows the optimal improvement effect. Fig. 6 and Table 6 show that the BRA and the TODL benefit the model's feature learning for various classroom learning behaviors. They contribute to enhancing the model's performance in detecting these behaviors. YOLOv8n_BT achieves the best recognition performance for different classroom learning behaviors, with detection performance exceeding 0.94 and 5% or more improvements for all categories of these behaviors. For the category of "disengage," YOLOv8n_BT's improvement effect is even 14%. The results above show that the YOLOv8n_BT model has an excellent fitting effect, high stability, and detection performance for all classroom learning behaviors.

TABLE 6. Confusion matrix enhancement results for each type of classroom learning behavior.

Category	YOLOv8n_BRA	YOLOv8n_TODL	**YOLOv8n_BT**
passive	3%	3%	**5%**
active	7%	6%	**11%**
constructive	5%	5%	**8%**
interactive	6%	6%	**10%**
disengage	8%	9%	**14%**

3) COMPARATIVE EXPERIMENTS OF DIFFERENT MODELS

To further prove the superiority of the YOLOv8n_BT algorithm, the self-constructed dataset of this study is trained on other classical object detection models, FasterRCNN [10], YOLOv5_s, YOLOv5_m, YOLOv5_l, YOLOv7_tiny, YOLOv7 and YOLOv7_X. The testing results are compared with those of YOLOv8n_BT. The experimental results are shown in Table 7, where the bolded parts are the optimal results. Compared with other classical models, YOLOv8n_BT also has the best detection performance.

TABLE 7. Comparative experimental results of different models.

Model	P（%）	R（%）	F1（%）	mAP50（%）	mAP50-90（%）
fasterRCNN	50.7	55.0	52.8	50.3	29.2
YOLOv5_s	72.2	66.6	69.3	69.8	37.2
YOLOv5_m	75.5	66.5	70.7	70.5	38.8
YOLOv5_l	77.4	64.6	70.4	69.6	38.9
YOLOv7_tiny	64.5	60.8	62.6	63.6	32.9
YOLOv7	65.1	56.6	60.6	61.0	31.1
YOLOv7_X	54.1	54.6	54.3	56.6	28.4
YOLOv8n	92.4	85.3	88.7	93.7	63.2
YOLOv8n_BT	**95.4**	**92.0**	**93.7**	**97.3**	**72.2**

4) APPLICATION COMPARISON

The YOLOv8n and YOLOv8n_BT models were applied to another classroom video to verify the feasibility of the improved model. The classroom video is a 40-minute recording including 19 students (9 girls and 10 boys). This classroom video is another lesson recording of an elementary math class from the same school as the training set. Fig.7 shows two representative classroom learning behavior recognition results for YOLOv8n and YOLOv8n_BT. In order, the three columns of images are the original image with the students' serial numbers, the YOLOv8n recognition results, and the YOLOv8n_BT recognition results. In the first case, as shown in Fig.7a, YOLOv8n missed the classroom learning behaviors of 4 students, stu1, stu13, stu18, and stu19 as shown in Fig.7b, and YOLOv8n_BT missed the learning behavior of only one student, stu18 as shown in Fig.7c. In the second case, as shown in Fig.7d, YOLOv8n missed the classroom learning behaviors of 4 students, stu13, stu17, stu18, and stu19 (as shown in Figure 7e), and YOLOv8n_BT missed the learning behavior of only one student, stu18 as shown in Fig.7f. It can be seen that YOLOv8n is not friendly to the classroom learning behaviors of the occluded and back-row students. The learning behaviors of the back-row-students are often missed, as well as the learning behaviors of stu1. The YOLOv8n_BT model is better than YOLOv8n in small and densely occluded objects and has strong robustness. YOLOv8n_BT can improve the detection performance of classroom learning behaviors and more comprehensively and precisely respond to students' classroom learning engagement. This verifies the feasibility of YOLOv8n_BT model improvement.

5) COMPARISON OF MODELS ON PUBLIC DATASETS

To verify the generalization ability of the improved network proposed in this paper (YOLOv8n_BT), we compare the improved network's object detection performance with other more networks such as YOLOv8n(baseline) on public data sets such as Pascal VOC 2012 and COCO. The experimental results for Pascal VOC2012 are shown in Table 8 and for COCO in Table 9. YOLOv8n_BT has the best overall detection performance in the Pascal VOC 2012 and COCO public datasets. Because these public datasets cover information beyond the classroom scenario, this indicates that YOLOv8n_BT has a robust generalization ability and is not limited to applications in classroom scenarios.

TABLE 8. The experimental result of the Pascal VOC 2012 dataset.

Model	R（%）	F1（%）	mAP50（%）	mAP50-90（%）
YOLOv8n [69] (baseline)	55.1	-	62.2	45.9
YOLOv8n+SimAm+WIoUv1 [69]	57.0	-	63.8	46.8
YOLOv8n+SimAm+WIoUv2 [69]	53.6	-	62.8	45.6
YOLOv8n+SimAm+WIoUv3 [69]	54.5	-	**63.6**	46.4
YOLOv8n_BT	**57.9**	**62.3**	64.2	**48.0**

TABLE 9. The experimental result of the COCO dataset.

Model	mAP50-90（%）
YOLOv5_small [70]	36.7
YOLOv7_tiny_SiLU [71]	38.7
YOLOv8n(baseline) [72]	37.3
YOLOv8n_BT	**39.1**

V. CONCLUSION

To solve the missed detection and false detection of classroom learning behaviors due to the target occlusion and the small target in the natural classroom, this study proposed an improved YOLOv8n_BT based on YOLOv8n. The YOLOv8n_BT improvement consists of two main aspects. On the one hand, for the occlusion problem of classroom learning behaviors, it incorporated the BRA in the Backbone to better capture fine-grained, global, and rich contextual information. On the other hand, for the small target problem of classroom learning behaviors for back-row-students, it added a TODL to better capture feature information about small objects. The results show that YOLOv8n_BRA, YOLOv8n_TODL, and YOLOv8n_BT can significantly improve the model performance, among which the YOLOv8n_BT model performs best. Compared to YOLOv8n(baseline), YOLOv8n_BT improves the detection performance of P, R, F1, mAP50, and mAP50-90 by 3.0%, 6.7%, 5.0%, 3.6%, and 9.0%, respectively, and the detection speed (FPS = 312.5) is relatively flat. This detection speed can still be detected in real-time. YOLOv8n_BT outperforms the current classical models, such as fasterRCNN, YOLOv8n, YOLOv7 and YOLOv5. In summary, YOLOv8n_BT can significantly improve the model's detection performance and solve missed and false detection issues in classroom learning behavior recognition. This enables the automated recognition of classroom learning behaviors to provide more comprehensive and precise data for teachers' instruction and classroom assessment.

In future work, first, we will reduce the network parameters while ensuring the detection performance. We hope to create a lightweight network to meet the lightweight requirements of mobile or embedded devices. Second, we will apply YOLOv8n_BT to other classroom datasets with different grades and subjects to validate the validity and generalization of YOLOv8n_BT. Third, we will develop a system with YOLOv8n_BT to recognize and analyze natural classroom learning behaviors.

REFERENCES

[1] J. Zhou, F. Ran, G. Li, J. Peng, K. Li, and Z. Wang, "Classroom learning status assessment based on deep learning," *Math. Problems Eng.*, vol. 2022, pp. 1–9, Apr. 2022.

[2] H. Liu, Y. Liu, R. Zhang, and X. Wu, "Student behavior recognition from heterogeneous view perception in class based on 3-D multiscale residual dense network for the analysis of case teaching," *Frontiers Neurorobotics*, vol. 15, Jul. 2021, Art. no. 675827.

[3] M. Hu, Y. Wei, M. Li, H. Yao, W. Deng, M. Tong, and Q. Liu, "Bimodal learning engagement recognition from videos in the classroom," *Sensors*, vol. 22, no. 16, p. 5932, Aug. 2022.

[4] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1012–1027, Nov. 2021.

[5] B. Wu, C. Wang, W. Huang, D. Huang, and H. Peng, "Recognition of student classroom behaviors based on moving target detection," *Traitement du Signal*, vol. 38, no. 1, pp. 215–220, Feb. 2021.

[6] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, p. 5314, Aug. 2021.

[7] A. Jisi and S. Yin, "A new feature fusion network for student behavior recognition in education," *J. Appl. Sci. Eng.*, vol. 24, no. 2, pp. 133–140, 2021.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Comput. Vis.–ECCV 14th Eur. Conf.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–12.

[11] Y. Liu, G. He, Z. Wang, W. Li, and H. Huang, "NRT-YOLO: Improved YOLOv5 based on nested residual transformer for tiny remote sensing object detection," *Sensors*, vol. 22, no. 13, p. 4953, Jun. 2022.

[12] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, "BiFormer: Vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10323–10333.

[13] S. Wang, M. Xu, Y. Sun, G. Jiang, Y. Weng, X. Liu, G. Zhao, H. Fan, J. Li, C. Zou, Y. Xie, L. Huang, and B. Chen, "Improved single shot detection using DenseNet for tiny target detection," *Concurrency Computation, Pract. Exper.*, vol. 35, no. 2, Jan. 2023, Art. no. e7491.

[14] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[15] L. Yang, Y. Xu, S. Wang, C. Yuan, Z. Zhang, B. Li, and W. Hu, "PDNet: Toward better one-stage object detection with prediction decoupling," *IEEE Trans. Image Process.*, vol. 31, pp. 5121–5133, 2022.

[16] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[17] L. Jiao, C. Kang, S. Dong, P. Chen, G. Li, and R. Wang, "An attention-based feature pyramid network for single-stage small object detection," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18529–18544, May 2023.

[18] Y. Li, H. Zheng, Z. Yan, and L. Chen, "Detail preservation and feature refinement for object detection," *Neurocomputing*, vol. 359, pp. 209–218, Sep. 2019.

[19] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," *Multimedia Tools Appl.*, vol. 82, no. 6, pp. 9243–9275, Mar. 2023.

[20] M. Bie, Y. Liu, G. Li, J. Hong, and J. Li, "Real-time vehicle detection algorithm based on a lightweight you-only-look-once (YOLOv5n-L) approach," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119108.

[21] Y. Yang, Y. Zhou, X. Yue, G. Zhang, X. Wen, B. Ma, L. Xu, and L. Chen, "Real-time detection of crop rows in maize fields based on autonomous extraction of ROI," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118826.

[22] E. Li, Q. Wang, J. Zhang, W. Zhang, H. Mo, and Y. Wu, "Fish detection under occlusion using modified you only look once V8 integrating real-time detection transformer features," *Appl. Sci.*, vol. 13, no. 23, p. 12645, Nov. 2023.

[23] L. J. Zhang, J. J. Fang, Y. X. Liu, H. Feng Le, Z. Q. Rao, and J. X. Zhao, "CR-YOLOv8: Multiscale object detection in traffic sign images," *IEEE Access*, vol. 12, pp. 219–228, 2024.

[24] F. Wang, H. Wang, Z. Qin, and J. Tang, "UAV target detection algorithm based on improved YOLOv8," *IEEE Access*, vol. 11, pp. 116534–116544, 2023.

[25] H. Chen and J. Guan, "Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology," *Electronics*, vol. 11, no. 23, p. 3998, Dec. 2022.

[26] N. Kumari, V. Ruf, S. Mukhametov, A. Schmidt, J. Kuhn, and S. Küchemann, "Mobile eye-tracking data analysis using object detection via YOLO v4," *Sensors*, vol. 21, no. 22, p. 7668, Nov. 2021.

[27] Q. Xu, Y. Wei, J. Gao, H. Yao, and Q. Liu, "ICAPD framework and simAM-YOLOv8n for student cognitive engagement detection in classroom," *IEEE Access*, vol. 11, pp. 136063–136076, 2023.

[28] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 2, pp. 936–953, Feb. 2022.

[29] K. Tong and Y. Wu, "Deep learning-based detection from the perspective of small or tiny objects: A survey," *Image Vis. Comput.*, vol. 123, Jul. 2022, Art. no. 104471.

[30] D. Liu, Z.-Q. Zhao, and W.-D. Tian, "TFPGAN: Tiny face detection with prior information and GAN," in *Proc. Intell. Comput. Methodologies, 16th Int. Conf. (ICIC)*, Bari, Italy. Cham, Switzerland: Springer, Oct. 2020, pp. 62–73.

[31] J. Zhang, J. Zhang, K. Zhou, Y. Zhang, H. Chen, and X. Yan, "An improved YOLOv5-based underwater object-detection framework," *Sensors*, vol. 23, no. 7, p. 3693, Apr. 2023.

[32] M. Kim, J. Jeong, and S. Kim, "ECAP-YOLO: Efficient channel attention pyramid YOLO for small object detection in aerial image," *Remote Sens.*, vol. 13, no. 23, p. 4851, Nov. 2021.

[33] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "EANTrack: An efficient attention network for visual tracking," *IEEE Trans. Autom. Sci. Eng.*, vol. 99, pp. 1–18, 2004.

[34] F. Gu, J. Lu, and C. Cai, "RPformer: A robust parallel transformer for visual tracking in complex scenes," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[35] J. Li, C. Liu, X. Lu, and B. Wu, "CME-YOLOv5: An efficient object detection network for densely spaced fish and small targets," *Water*, vol. 14, no. 15, p. 2412, Aug. 2022.

[36] Y. Hu, Y. Dai, and Z. Wang, "Real-time detection of tiny objects based on a weighted bi-directional FPN," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2022, pp. 3–14.

[37] J. Yao, X. Fan, B. Li, and W. Qin, "Adverse weather target detection algorithm based on adaptive color levels and improved YOLOv5," *Sensors*, vol. 22, no. 21, p. 8577, Nov. 2022.

[38] Z. Cao, T. Liao, W. Song, Z. Chen, and C. Li, "Detecting the shuttlecock for a badminton robot: A YOLO based approach," *Exp. Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113833.

[39] C. Gao, W. Tang, L. Jin, and Y. Jun, "Exploring effective methods to improve the performance of tiny object detection," in *Proc. Comput. Vis.–ECCV Workshops*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 331–336.

[40] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[41] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang, and X. Guo, "Real-time vehicle detection based on improved YOLO v5," *Sustainability*, vol. 14, no. 19, p. 12274, Sep. 2022.

[42] K. Saleh, S. Szénási, and Z. Vámossy, "Generative adversarial network for overcoming occlusion in images: A survey," *Algorithms*, vol. 16, no. 3, p. 175, Mar. 2023.

[43] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[44] R. Huan, J. Zhang, C. Xie, R. Liang, and P. Chen, "MLFFCSP: A new anti-occlusion pedestrian detection network with multi-level feature fusion for small targets," *Multimedia Tools Appl.*, vol. 82, no. 19, pp. 29405–29430, Aug. 2023.

[45] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6023–6032.

[46] S. Chen, X. Zou, X. Zhou, Y. Xiang, and M. Wu, "Study on fusion clustering and improved YOLOv5 algorithm based on multiple occlusion of camellia oleifera fruit," *Comput. Electron. Agricult.*, vol. 206, Mar. 2023, Art. no. 107706.

[47] Q. Gao, Z. He, X. Jia, Y. Xie, and X. Han, "Lightweight high-precision pedestrian tracking algorithm in complex occlusion scenarios," *KSII Trans. Internet Inf. Syst.*, vol. 17, no. 3, pp. 1–21, 2023.

[48] X. Chen, Y. Jia, X. Tong, and Z. Li, "Research on pedestrian detection and DeepSort tracking in front of intelligent vehicle based on deep learning," *Sustainability*, vol. 14, no. 15, p. 9281, Jul. 2022.

[49] D. Yuan, X. Shu, Q. Liu, and Z. He, "Aligned spatial–temporal memory network for thermal infrared target tracking," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 3, pp. 1224–1228, Mar. 2023.

[50] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "Repformer: A robust shared-encoder dual-pipeline transformer for visual tracking," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20581–20603, Oct. 2023.

[51] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.

[52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[53] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 658–666.

[54] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000.

[55] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.

[56] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[57] L. Tan, X. Lv, X. Lian, and G. Wang, "YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107261.

[58] J. Guo, L. Liu, F. Xu, and B. Zheng, "Airport scene aircraft detection method based on YOLO v3," *Laser Optoelectronics Prog.*, vol. 56, no. 19, 2019, Art. no. 191003.

[59] Y. H. Shao, D. Zhang, and H. Y. Chu, "A review of YOLO object detection based on deep learning," *J. Electron. Inf. Technol.*, vol. 44, no. 10, pp. 3697–3708, Oct. 2022.

[60] Y. Tang, H. Zhou, H. Wang, and Y. Zhang, "Fruit detection and positioning technology for a camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118573.

[61] C. Qi, J. Gao, S. Pearson, H. Harman, K. Chen, and L. Shu, "Tea chrysanthemum detection under unstructured environments using the TC-YOLO model," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116473.

[62] L. Li, G. Shi, and T. Jiang, "Fish detection method based on improved YOLOv5," *Aquaculture Int.*, vol. 31, no. 5, pp. 2513–2530, Oct. 2023.

[63] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.

[64] L. Tang, T. Xie, Y. Yang, and H. Wang, "Classroom behavior detection based on improved YOLOv5 algorithm combining multi-scale feature fusion and attention mechanism," *Appl. Sci.*, vol. 12, no. 13, p. 6790, Jul. 2022.

[65] H. Liu, W. Ao, and J. Hong, "Student abnormal behavior recognition in classroom video based on deep learning," in *Proc. 5th Int. Conf. Electron. Inf. Technol. Comput. Eng.*, 2021, pp. 664–671.

[66] W. Yang, X. Ma, W. Hu, and P. Tang, "Lightweight blueberry fruit recognition based on multi-scale and attention fusion NCBAM," *Agronomy*, vol. 12, no. 10, p. 2354, Sep. 2022.

[67] M. T. H. Chi and R. Wylie, "The ICAP framework: Linking cognitive engagement to active learning outcomes," *Educ. Psychologist*, vol. 49, no. 4, pp. 219–243, Oct. 2014.

[68] M. T. H. Chi, J. Adams, E. B. Bogusch, C. Bruchok, S. Kang, M. Lancaster, R. Levy, N. Li, K. L. McEldoon, G. S. Stump, R. Wylie, D. Xu, and D. L. Yaghmourian, "Translating the ICAP theory of cognitive engagement into practice," *Cognit. Sci.*, vol. 42, no. 6, pp. 1777–1832, Aug. 2018.

[69] Q. Liu, W. Huang, X. Duan, J. Wei, T. Hu, J. Yu, and J. Huang, "DSW-YOLOv8n: A new underwater target detection algorithm based on improved YOLOv8n," *Electronics*, vol. 12, no. 18, p. 3892, Sep. 2023.

[70] J. Wang, H. Dai, T. Chen, H. Liu, X. Zhang, Q. Zhong, and R. Lu, "Toward surface defect detection in electronics manufacturing by an accurate and lightweight YOLO-style object detector," *Sci. Rep.*, vol. 13, no. 1, p. 7062, May 2023.

[71] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[72] G. Jocher. (2023). *YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics

QINGTANG LIU (Member, IEEE) was born in Hubei, China, in 1969. He received the Ph.D. degree in electronic information engineering from Huazhong University of Science and Technology, Wuhan, in 2005.

He is currently a Professor with Central China Normal University. His current research interests include learning analytics technology and digital learning. He is also a member of ISO/IEC JTC1 SC36, AVS Standard Organization, the National Beacon Committee Education Technology Sub-Technical Committee (CELTSC), and ACM.

RUYI JIANG was born in Neijiang, Sichuan, China, in 2000. She received the bachelor's degree in educational technology from Sichuan Normal University, Sichuan, China, in 2022. She is currently pursuing the master's degree with Central China Normal University, Wuhan, China. Her research interest includes multimodal learning analysis.

QI XU was born in Jingmen, Hubei, China, in 1997. She received the master's degree in educational technology from Hubei Normal University, Huangshi, China, in 2022. She is currently pursuing the Ph.D. degree with Central China Normal University, Wuhan, China. Her research interest includes educational technology.

DENG WANG was born in Hubei, China, in 1999. He received the B.S. degree in educational technology form South-Central Minzu University, in 2022. He is currently pursuing the master's degree with Central China Normal University, Wuhan, China. His current research interest includes multimodal learning analysis.

ZHIQIANG SANG was born in Yunnan, China, in 1977. He received the bachelor's degree in industrial automation engineering from Yunnan University of Technology, in 1999, and the master's degree in computer software and applications from Kunming University of Science and Technology, in 2011. He is currently pursuing the Ph.D. degree with Central China Normal University, Wuhan, China. His research interest includes learning analysis.

XINYU JIANG was born in Hubei, China, in 1999. She received the M.S. degree in educational technology from Hubei University, Hubei, in 2023. She is currently pursuing the D.S. degree with Central China Normal University, Hubei. Her current research interests include machine learning and classroom teaching behavior detection.

LINJING WU was born in Hubei, China in 1987. She received the Bachelor of Science degree from Hubei University, China, in June 2007, and the Ph.D. degree in science from the Education Information Technology Research Center, Central China Normal University, China, in June 2013.

She is currently teaching with Central China Normal University, as an Associate Professor and a Doctoral Supervisor. She has published over 60 papers in domestic and foreign journals and international academic conferences and applied for four national invention patents and obtained four software copyrights, involving learning analysis and data mining, resource management and support services, user feature modeling, and personalized resource recommendation. Her research interests include data mining, artificial intelligence, and educational applications.