

Received 23 October 2023, accepted 21 November 2023, date of publication 28 November 2023,
date of current version 8 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337435

RESEARCH ARTICLE

ICAPD Framework and simAM-YOLOv8n for Student Cognitive Engagement Detection in Classroom

QI XU^{1,2}, YANTAO WEI^{1,2}, JIE GAO^{1,2}, HUANG YAO^{1,2},
AND QINGTANG LIU^{1,2}, (Member, IEEE)

¹Hubei Research Center for Educational Informationization, Central China Normal University, Wuhan 430079, China

²Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China

Corresponding author: Yantao Wei (yantaowei@ccnu.edu.cn)

This work was supported in part by the National Collaborative Innovation Experimental Base Construction Project for Teacher Development of Central China Normal University under Grant CCNUTEIII-2021-19, in part by the National Natural Science Foundation of China under Grant 62277029 and Grant 62277021, in part by the Humanities and Social Sciences of China Ministry of Education (MOE) under Grant 20YJC880100, in part by the Knowledge Innovation Program of Wuhan–Basic Research under Grant 2022010801010274, and in part by the Fundamental Research Funds for the Central Universities under Grant CCNU22JC011.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Central China Normal University under Approval No. CNU-IRB-202305004a.


ABSTRACT Research has shown that cognitive engagement plays a key role in effective learning, resulting in extensive efforts have been devoted to measuring it. Whereas most of the literature explores manual methods to measure cognitive engagement, the research on automatic detection of cognitive engagement levels in real classrooms is very limited. Automatic detection of cognitive engagement has been a problem for a long time due to the lack of behavior annotation guidance and effective detection algorithms. For the first challenge, a theory of cognitive engagement called Interactive-Constructive-Active-Passive-Disengage (ICAPD) is proposed in this paper. ICAPD links visual behaviors with cognitive engagement in the classroom. According to the ICAPD framework, a cognitive engagement dataset is constructed to train the detection model. To tackle the second challenge, the simAM-based You Only Look Once version 8 Nano (simAM-YOLOv8n) model is designed, simAM-YOLOv8n utilizes the simAM attention module to strengthen feature extraction and detect different levels of cognitive engagement precisely and efficiently. The experimental results on the self-built dataset have demonstrated the effectiveness of the proposed theory framework and detection algorithm, indicating that the proposed methodology could be used to detect real-time cognitive engagement in the classroom scenario. This work has the potential to help teachers to carry out learning analysis and instructional adjustments.

INDEX TERMS Cognitive engagement detection, attention mechanism, ICAP, YOLO.

I. INTRODUCTION

Cognitive engagement refers to students' effort, persistence, resilience, concentration, paying attention, and contributing to class in the context of instruction [1], [2]. Research has demonstrated that high-level engagement often leads to better learning achievements [3], as it encourages students to gain a deeper understanding of topics. Measuring cognitive engagement can provide insights for improving curriculum

and instructional design [4]. Therefore, measuring cognitive engagement is a hot topic in the field of education. Currently, the popular tools for measuring cognitive engagement include self-reports, observational checklists, and automated detection. However, both the self-reporting and observational checklist require a great deal of time and effort from both the students and the observers. In the past few years, automatic detection methods, which extract features automatically and do not interrupt the learning of students, have attracted more and more attention. Among these automatic detection methods, computer vision-based engagement detection is

The associate editor coordinating the review of this manuscript and approving it for publication was Daniela Cristina Momete .

the most popular one, which can offer many ways to detect engagement, such as body behavior, facial expression, head posture, etc [5], [6]. For example, Levordashka et al. [7] used head posture features (i.e., head yaw, pitch, and roll) to detect cognitive engagement. The advantages of the computer vision-based methods, such as real-time, non-invasive, and low-cost, make them suitable for cognitive engagement detection in the classroom [4]. However, cognitive engagement detection is still difficult due to the lack of a representation framework, which well links visual behaviors with cognitive engagement levels. According to Fredricks et al. [1], behaviors are the carrier and representation of cognitive engagement. Consequently, we can assess the cognitive engagement level by observing student behaviors. However, most cognitive engagement detection methods use the behavior of students in a coarse-grained and shallow way. For example, head posture (moving or not moving), eye movement (staring or not staring), and facial expression (laughing or not laughing) are used to discriminate engagement and disengagement. These simple representations and binary judgments cannot characterize the internal complexity of cognitive engagement. The difficulty of assessing students' cognitive engagement in the classroom consists of defining the different levels and the visible behaviors of it. To this end, a more concrete and fine-grained behavioral representation of cognitive engagement is needed. The Interactive, Constructive, Active, Passive (ICAP) framework provides a possible solution [8], [9]. This framework links students' visual behaviors with levels of cognitive engagement, such as looking at teachers as passive behavior, pointing to learning materials as active behavior, etc. Goldberg et al. [10] used it to map students' engagement into visible head pose, gaze, and facial expressions. However, the ICAP framework is designed for classroom observation. It can not provide enough information for engagement annotation, which is the essential step for training the cognitive engagement detection model. Consequently, designing a framework for cognitive engagement annotation will be beneficial for computer vision-based detection.

Besides behavior annotation, designing an effective detection model is another important issue in computer vision-based cognitive engagement detection. Some studies first extract the individual's behavioral features and then detect engagement from single-person images [11], [12], which requires that each individual be equipped with a camera to record their behaviors. For example, Li et al. [12] use the OpenFace system to extract facial behavior features, and Naïve Bayes, k-NN, Decision tree, Random forest, and Support Vector Machine were used to detect engagement. Another way is to first locate the individual and then detect engagement from multi-person images by the object detection model [13], [14], which is suitable for real classroom scenarios. In this case, only one camera is needed to record the behaviors of all students. The

widely used object detection models are Fast-Region-based Convolutional Neural Network (Fast-R-CNN) [15], Single Shot MultiBox Detector (SSD) [16] and You Only Look Once (YOLO) [17] etc. Among them, the YOLO series are popular due to their good performance and the convenience of single-stage detection [14]. These advantages of YOLO make it suitable for cognitive engagement detection in the classroom [18]. However, due to the low resolution, occlusion, and complexity of behaviors, the YOLO-based engagement detection method also faces challenges. For example, the behaviors of the students in the back row are difficult to detect, and behaviors belonging to different levels of cognitive engagement may be similar in appearance, which makes it difficult to detect engagement accurately.

Previous studies have shown that the attention mechanism is an effective way to enhance deep learning models [19], [20], [21], [22], [23], [24], which can adjust the weight according to the given task. Many attention modules have been proposed in the past few years, such as Efficient Multi-Scale Attention (EMA) [25], Global attention mechanism (GAM) [26], Normalization-based Attention Module (NAM) [27], Simple, Parameter-Free Attention Module (simAM) [28], Efficient channel attention (ECA) [29], Selective kernel networks (SKnets) [30], Convolutional block attention module (CBAM) [31] and Squeeze-and-Excitation (SE) [32], etc. For example, channel attention (CA) can achieve good performance in pedestrian detection [33], and the SE is effective in shadow detection [34]. Adding the CBAM in the initial stage of the feature extraction network of YOLOv5s has been shown to improve the performance of student behavior detection [35]. Recently, YOLOv8 series methods have been proposed, offering cutting-edge performance in terms of accuracy and speed. However, in classroom scenes, many students' engagement levels need to be detected, the background is complex, and the occlusion problem is common, which greatly affects the accuracy of YOLOv8. In contrast, the attention mechanism extracts features that significantly improve the accuracy, which has always led to good results in the improvement of YOLO models. Consequently, an attention-based YOLOv8 model is proposed in this paper to improve the performance in the dense scenes.

The specific research questions of the paper include two: First, how can cognitive engagement be directly annotated by students' overt behavior? Second, how to effectively improve the detection performance of cognitive engagement? To our knowledge, computer vision-based cognitive engagement detection in the classroom has rarely been studied due to the difficulty of annotation and detection. To this end, this paper establishes a representation framework for linking students' overt behavior with their cognitive engagement levels in the classroom, which provides support for the automatic detection of cognitive engagement. Aiming to deal with the problems of dense classroom scenarios, we designed the simAM-based YOLOv8n (simAM-YOLOv8n) model to

obtain a better detection performance. The main contributions of this paper can be summarized as follows:

- 1) The ICAPD framework is proposed as an annotation guide for computer vision-based cognitive engagement detection in the real classroom. It maps the overt behaviors to five different levels of cognitive engagement.
- 2) The ICAPD dataset of student cognitive engagement is built in the real classroom according to the ICAPD framework. Unlike experiment-induced behaviors, it obtains students' visual behavior data non-invasively, providing data for training the automatic detection model in real classrooms.
- 3) The simAM-YOLOv8n is proposed to solve the problems in dense scenes for cognitive engagement detection. It adds the simAM attention module in the Backbone network and the Head network of the YOLOv8n to improve the cognitive engagement detection performance in the real classroom.

II. RELATED WORKS

A. ICAP FRAMEWORK

The ICAP framework was proposed by Chi et al. [36], and the main contents of it are shown in Table 1. It consists of a taxonomy that differentiates cognitive engagement levels based on overt behaviors. By definition, knowledge-change processes are dynamic processes that students engage in while learning new information. Four broad knowledge-change processes can be associated with cognitive engagement levels. However, it cannot be used directly for engagement detection. One reason is that the ICAP framework does not consider the state of disengagement and relevant behaviors. Magana et al. [37] used the ICAP framework to encode students' level of interaction. Their use of video data and inclusion of the disengagement dimension aligns well with our ideas. Importantly, this study has inspired us in two significant ways: 1) there is room for improvement within the vision-based ICAP framework, and 2) cognitively disengaged and disruptive behaviors can demonstrate no engagement. According to the engagement structure [38], disengagement may lead to dropping class [39], [40]. Consequently, it is necessary to annotate and detect disengagement [41], [42], such as playing with others, looking around, crying, etc., which is helpful for teachers to strengthen the management of abnormal learning cases. Another reason is that the behaviors in the ICAP framework include the students' speaking, actions, work products, etc. However, the computer vision methods can only work in the description of the behavior in the image-dependent ICAP framework rather than text- and product-related clues. Therefore, it is better to improve the ICAP framework by focusing on visible non-verbal behavior, rather than considering verbal behaviors.

B. OVERVIEW OF YOLOv8

YOLO version 8 (YOLOv8) [43], the latest YOLO version for real-time object detection, includes YOLOv8 nano (YOLOv8n), YOLOv8 small (YOLOv8s), YOLOv8 medium

TABLE 1. ICAP framework.

Classes of cognitive engagement	Knowledge-change processes
Passive: Learners orient toward and receive information from the instructional materials without overtly doing anything else related to learning.	Store: New information is stored in an isolated way.
Active: Learners' engagement with instructional materials if some form of overt motoric action or physical manipulation is undertaken.	Integrate: New information activates relevant prior knowledge; while storing, new information is integrated with activated prior knowledge.
Constructive: Learners generate or produce additional externalized outputs or products beyond what was provided in the learning materials.	Infer: New information is integrated with activated prior knowledge, and new knowledge is inferred from activated and integrated knowledge.
Interactive: Dialogues, examples include a learner talking with another person who can be a peer, a teacher, a parent, or a computer agent.	Co-Infer: Each learner infers new knowledge from activated and integrated knowledge and iteratively infers knowledge with new inputs from conversational partner(s).

(YOLOv8m), YOLOv8 large (YOLOv8l) and YOLOv8 extra-large (YOLOv8x) suitable for datasets of different sizes. This model family belongs to one-stage object detection models that process an entire image in a single forward pass of a convolutional neural network. It involves detecting objects in an image or video frame and drawing bounding boxes around them. The detected objects are classified into different classes based on their features. Because our database is small, we chose YOLOv8n as the basic network framework. As shown in Figure 1, YOLOv8n divides the network structure into three parts: Backbone, Neck, and Head. First, the input image is sent to the Backbone network to complete feature extraction. Then, the fusion of features with different scales is completed in the Neck network. Finally, the Head network predicts the bounding box, class, and confidence through the output feature maps of three scales.

The YOLOv8n is fast in detection but suffers from low accuracy of small objects and complex behaviors in the classroom. For example, the Active class contains many complex behaviors, as shown in Figure 2. Through multiple image down-sampling, the similar behaviors adjacent to the aggregation areas are mapped to the deep feature map and aggregated into one point, which makes the objects indistinguishable. Based on this, we consider adding the attention module to extract more effective behavioral features. The Head network in YOLOv8n is updated to the current mainstream decoupling head structure, separating the classification and detection heads and changing from Anchor-Based to Anchor-Free. Since students are small targets, the small target detection layer in the Head network needs to be carefully designed. Finally, YOLOv8n uses Binary Cross Entropy (BCE) Loss as the classification loss and Distribution Focal Loss (DFL) + Complete IOU (CIU) Loss as the regression loss.

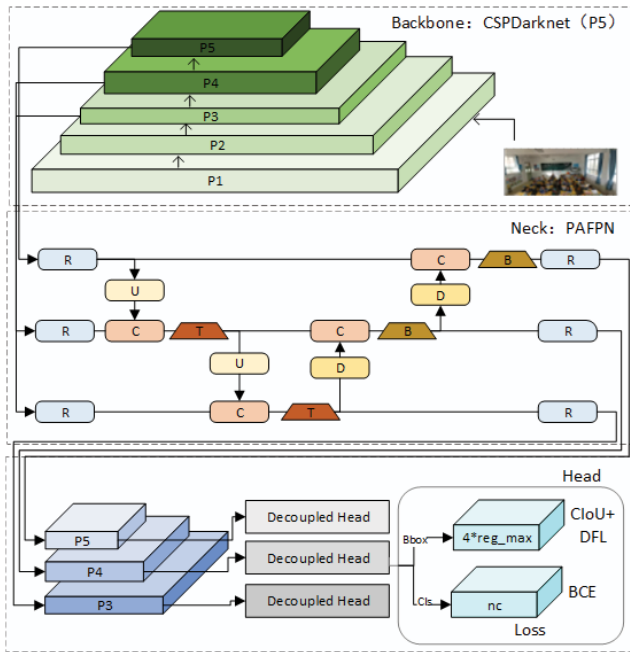


FIGURE 1. The structure of YOLOv8n network. The Backbone, Neck, and Head networks are from top to bottom.



FIGURE 2. Examples in Active class. According to ICAP, it may include students pointing to materials, gestures expressing digital symbols, looking for information, picking up stationery, manipulating objects, etc.

III. THE PROPOSED METHOD

As shown in Figure 3, the proposed method consists of three parts. The first is a behavior-based representation framework for cognitive engagement annotation. The second part is dataset construction for the detection model training, where we built the ICAPD dataset under the guidance of the proposed representation framework. Lastly, the third part is the attention-based YOLOv8 model for cognitive engagement detection.

A. THE ICAPD FRAMEWORK

1) COGNITIVE ENGAGEMENT LEVELS

The Interactive, Constructive, Active, Passive, Disengage (ICAPD) framework is a variant of the ICAP framework [36]. ICAPD framework defines cognitive engagement on the basis of students’ visual behaviors and proposes that engagement behaviors can be categorized and differentiated into one of five modes. It is designed for the student cognitive engagement annotation in the classroom. The ICAPD framework is shown in Table 2, which mainly provides a guide to judge the level of cognitive engagement from visual behaviors. It is different from the original ICAP framework in two aspects. First, considering the engagement boundary [39], [40], the ICAPD framework includes the Disengage class. Therefore, there are five classes of cognitive engagement in the ICAPD framework: (1) Interactive, (2) Constructive, (3) Active,

(4) Passive, and (5) Disengage. This order represents the whole process of students from high engagement to high disengagement. Second, the ICAPD framework defines the specific visual behaviors corresponding to different levels of cognitive engagement. It provides guidance on data annotation to train an effective detection model.

2) VISUAL BEHAVIORS

The following will introduce the content of external visual behaviors in detail. In terms of visual behaviors, first, it combines the positional relationship between the gesture and the face to enrich the behaviors of the Active class. The reason for this is that the hand-over-face gestures are important engagement cues [44], [45]. The behaviors like stroking, tapping, and touching facial regions - especially with the index finger - are all associated with cognitive mental states, namely thinking [46]. This state is highly consistent with the Active class because passive behavior does not require students to do anything and active behavior points out the minimum level of students’ visible behavior. Second, some non-task action behaviors (motor activity not associated with the assigned academic task; e.g., leaving the seat to throw a piece of paper in the trash can), non-task audio-visual behaviors (utterances not associated with the academic task; e.g., talking to a peer about something other than the current assignment, humming), and non-task negative behaviors (passive nonengagement; e.g., looking out the window) are classified into the Disengaged class, inspired by the Behavior Observation of Student in Schools (BOSS) framework [47]. The BOSS was designed to assess student academic behaviors in the classroom environment, which is consistent with our consideration of cognitive engagement annotation in the real classroom scene. Third, it includes some Back-Channels into the Interactive class [48] because these behaviors occur in the dialogue, they conform to the behavior mapping connotation of this class, example Non-verbal back-channels include head nods and shakes signaling to the initiator, listens or desires to continue the conversation, applaud to others’ answers, etc. It must be acknowledged that the same visual behavior may correspond to different classes of cognitive engagement. Determining students’ true cognitive engagement accurately is challenging with just a single video frame. The automated detection models can only process images or videos. So, we make the following distinctions when annotating vision-based constructive, interactive, and disengage:

- Most behaviors accompanied by standing belong to interactive. So, the subject of “talk with others” is the standing state. This is because students only stand when they are allowed to interact in the recording situation. However, when a student stands to explain to others, it requires additional textual annotating to be classed as constructive. We annotate this as “interactive” based on visual cues - while imperfect - it can offer a reasonable alternative. When a student is instructed to stand, he/she

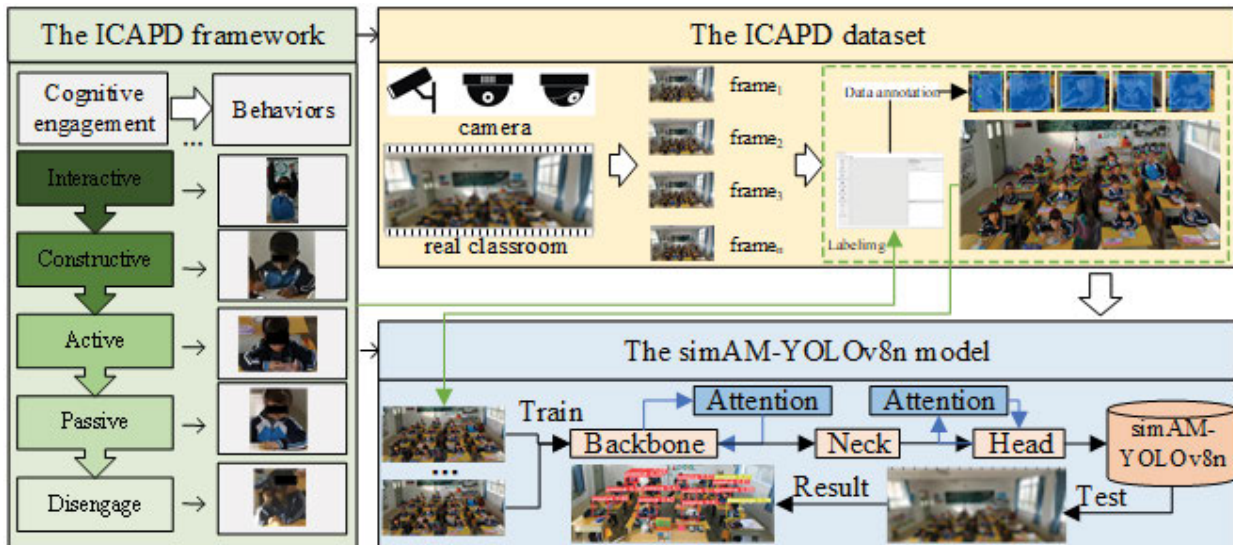


FIGURE 3. The proposed method of student cognitive engagement detection. The first step is to design an ICAPD framework for data annotation. Then we construct the ICAPD dataset by using this ICAPD framework. Finally, we develop the simAM-YOLOv8n model for detecting cognitive engagement.

TABLE 2. The proposed ICAPD framework. The first column is five different levels of cognitive engagement, the second column is the behavior description corresponding to different cognitive engagement levels, the third column is the specific overt behaviors for the given engagement levels, and the fourth column is the behavior example images, which is used as a guidance for the annotation.

Classes	Connotation	Behaviors	Examples
Disengage	Behaviors that are unrelated to learning activities.	Yawn, drink water, lie on the desk, sleep, cry, play with toys, look out the window, leave the seat without permission, etc.	
Passive	Behaviors of being oriented toward and receiving information from the materials without overtly doing anything else related to learning.	Read a text silently, watch a video, or listen to a lecture without undertaking any other visible activities, etc.	
Active	Some form of overt action is undertaken without providing any new information.	Point to or gesture at the materials, measure objects, rotate objects, underline sentences, Take out the tool, scratch the head, hands on the face or head, etc.	
Constructive	Produce externalized ideas containing information that goes beyond what is provided in the materials or instruction.	Explain to others, take notes, pose problems, ask questions, raise hands, draw on paper, etc.	
Interactive	Interactions between two peers (or a small group), often through dialogues.	Stand up, talk with others, applaud to mates, clap hands, handshaking, pat others, etc.	

might yawn, look out the window, etc. In such cases, it can be annotated as disengage. In other words, we rely on visual cues for annotating whenever possible. We do not consider situations that require referencing text or other data.

- Writing-related behaviors belong to constructive. the subject of “Explain to others” can be the writing state. Because the written content can serve as a good self-explanation [49], [50].

- Behaviors where students look around belong to disengage. Students are not allowed to sit and interact with others in the recording situation. Sitting accompanied by turning around and discussing topics related to course materials is not permitted by the teacher, so it would be annotated as disengage.

Overall, the ICAPD framework covers most of the possible visual body behaviors of students in the classroom learning process, such as listening to a lecture, pointing to or gesturing

at parts of what learners are reading or solving, asking questions, standing up to talk to the teacher and other behaviors without engagement. According to the examples in Table 2 and Figure 2, it can be found that complex behavioral features in the Active class are easily confused with those in other classes, which brings great challenges to the application of automatic detection models.

B. THE ICAPD DATASET

1) DATA COLLECTION

The classroom videos utilized in the study were recorded in real-world lessons at a primary school, and approved by the ethics committee of Central China Normal University (Approval CNU-IRB-202305004a). We paid attention to the following ethical issues in our study. First, participants are free to opt in or out of the study at any point in time. Second, participants know the study's purpose, benefits, and risks before they agree or decline to join. Third, sensitive personally identifiable data (e.g., full name) is not collected. Fourth, we anonymize identifiable related data (e.g., eye) so that it can't be linked to other data by anyone else. A total of forty students from grades one to three were videotaped during three lessons across two subjects. The duration of each recording was 40 minutes 48 seconds, 44 minutes 34 seconds, 38 minutes 32 seconds, a total of 123 minutes 54 seconds. All recordings were captured by non-invasive cameras (30 frames per second) with the resolution 1920×1080 . The camera is placed in the front center of the classroom. Its angle of placement considers two aspects: 1) capturing each attending student as comprehensively as possible, and 2) minimizing occlusion between students. Each recording lasts around 40 minutes and contains students' spontaneous behavior. To extract features from video, frame-based sampling is used. Students' behaviors in consecutive frames do not change much, which brings a lot of redundant information. So, automatic frame sampling is performed to get suitable frames for cognitive engagement detection. We extract the images from the video stream with a fixed step (e.g., every 3 seconds) and a total of 2600 sample data are generated.

2) BEHAVIOR ANNOTATION

Annotating behavior data is the first step in computer vision-based detection of student cognitive engagement. As shown in Figure 4, the frames are sampled uniformly to convert dynamic videos into static images of 480×270 pixels, and the data format of images is jpg. Then the obtained images are annotated by the LabelImg tool. A single student's location and behavior in the image are accurately marked with a bounding box based on the ICAPD framework. Only the smallest rectangular box around each student's location is marked to make the bounding box contain as few background areas as possible. The contents of the annotation include folder, filename, path, source, size, and multiple objects (name, pose, bandbox, etc.).

We saved the labeled results to an XML file, as shown in Figure 4(e). As shown in Figure 4(f). Two human raters manually annotated cognitive engagement in all 3 videos. In the pre-annotation phase, The Kappa value of the two raters for annotations was 0.60, indicating moderate inter-rater reliability. Then an annotation meeting is held to negotiate inconsistent annotations and revise the annotations of all videos together. The instances in the ICAPD dataset are quite small, and the annotations in each image are intensive. This presents great challenges in extracting the students' behavior features for automatic detection.

After annotation guided by the ICAPD framework, the obtained ICAPD dataset contains 2600 images with 32716 objects in total, and the format of the dataset is produced according to the COCO dataset. There are 22145, 2665, 2649, 764, and 4493 samples in the Passive, Active, Constructive, Interactive, and Disengage classes. We combined 1724 and 426 images in the train and validation sets to train the model and report results on 450 test images. The ratio of training set, validation set, and test set is 6: 2: 2.

C. THE simAM-YOLOv8n MODEL

1) ATTENTION MODULE: simAM

The simAM [28] is a conceptually simple but very effective attention module for convolutional neural networks, as shown in Figure 5. The simAM infers 3-D attention weights for the feature map in a layer without adding parameters to the original networks. It is based on some well-known neuroscience theories and optimizes an energy function to find the importance of each neuron. The energy function is defined as shown in Equation (1):

$$e(w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} ((-1 - (w_t x_i + b_t))^2 + (1 - (w_t x_i + b_t))^2) + \lambda w_t^2 \quad (1)$$

Among them, w_t and b_t are the parameters of the energy function, M is a constant term, λw_t^2 is a regular term, x_i is the input feature map, y is the output of the new feature map. Most operators are selected based on the solution to the defined energy function, avoiding too much effort for structure tuning.

The simAM is a channel and spatial attention module. By extracting features in two dimensions, it can extract some important features about the small objects and the complex behaviors as much as possible. This may make the simAM effective for extracting the deep behavioral channel attention and width-length of the annotation box's spatial attention in the active class of the ICAPD dataset. The detection of the added Disengage class in the proposed ICAPD framework may also be effective through simAM, because this class contains many behaviors that lead to high complexity. Overall, simAM may solve this task because the behaviors involved in the proposed ICAPD framework are numerous and complex.

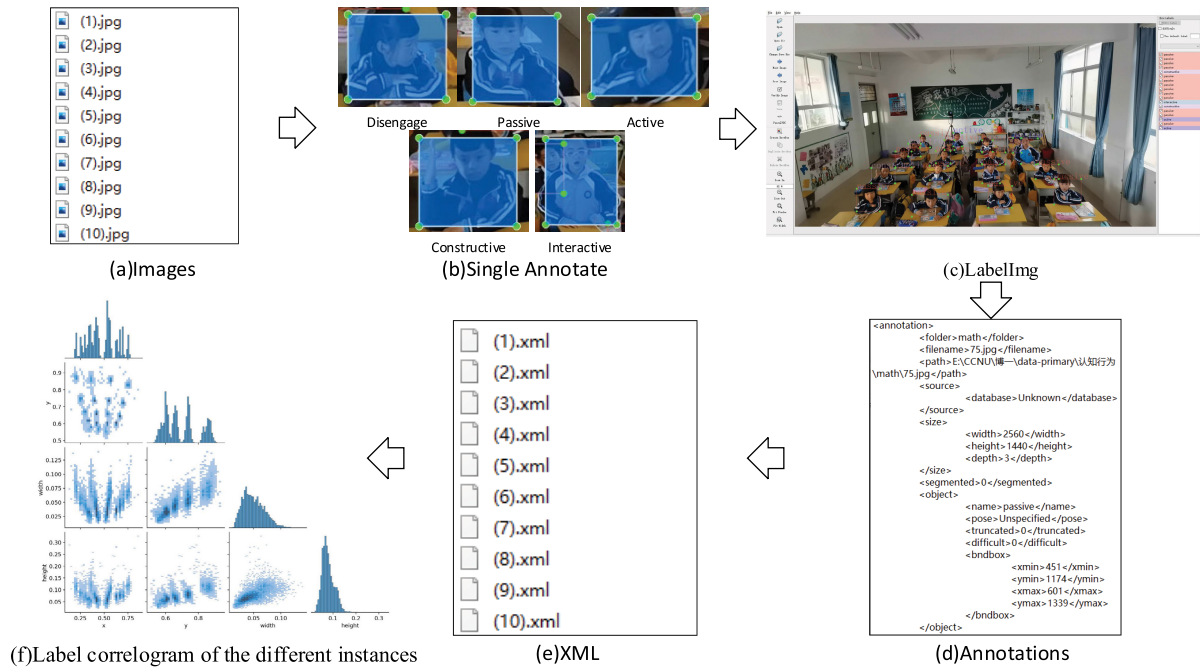


FIGURE 4. Student cognitive engagement annotation process. The image in (a) is formed by automatically sampling video frames. (b) denotes the process of annotating a single person from a whole image in (a). (c) shows the process of labeling a whole image in the Labeling tool. (d) shows the annotated content in a whole image provided by the Labeling tool. (e) represents XML files saved after annotating multiple video frames. (f) represents the distribution of all video annotations on the width and height of the image.

2) THE simAM-YOLOv8n MODEL

The simAM-YOLOv8n model is the improved YOLOv8n model with an attention module. The idea of the attention mechanism is based on the mechanism used by humans in cognitive science to process information. This mechanism can automatically learn and calculate the contribution of input data to output data. To accurately detect the complex behaviors in the different classes and help teachers understand students' cognitive engagement status, we use the simAM in the feature extraction stage and after the feature fusion of the YOLOv8n. Accurately, it is added to the last feature extraction stage in the Backbone network and the small target detection stage in the Head network, as shown in Figure 5, which helps the network learn more valuable features.

The parameter of the simAM-YOLOv8n is shown in Table 3. The added simAM module is on layer 9 and layer 17. The output $20 \times 20 \times 512$ feature map of layer 8 is input into the simAM module. The attention is extracted along the channel dimension and the spatial dimension. Then the attention feature extracted by the simAM is transmitted to the SPPF of YOLOv8n, which avoids the problem of image distortion caused by the attention calculation of the image area. Then, the output feature map $40 \times 40 \times 512$ of layer 16 is input into the simAM module, and the features of the small target detection enhanced by the simAM are transmitted to the Conv of YOLOv8n to help the network perform subsequent splicing and detection calculations. Finally, the size of the final output feature map is still $256 \times 256 \times 5$ through a convolutional layer with 256 convolution kernels (convolution kernel size is 1×1).

TABLE 3. Model summary of the proposed simAM-YOLOv8n, where simAM modules are at layers 9 and 17.

Layer	Params	Module	Arguments
0	464	Conv	[3, 16, 3, 2]
1	4672	Conv	[16, 32, 3, 2]
2	7360	C2f	[32, 32, 1, True]
3	18560	Conv	[32, 64, 3, 2]
4	49664	C2f	[64, 64, 2, True]
5	73984	Conv	[64, 128, 3, 2]
6	197632	C2f	[128, 128, 2, True]
7	295424	Conv	[128, 256, 3, 2]
8	460288	C2f	[256, 256, 1, True]
9	0	simAM	[256, True]
10	164608	SPPF	[256, 256, 5]
11	0	Upsample	[None, 2, 'nearest']
12	0	Concat	[1]
13	148224	C2f	[384, 128, 1]
14	0	Upsample	[None, 2, 'nearest']
15	0	Concat	[1]
16	37248	C2f	[192, 64, 1]
17	0	simAM	[64]
18	36992	Conv	[64, 64, 3, 2]
19	0	Concat	[1]
20	123648	C2f	[192, 128, 1]
21	147712	Conv	[128, 128, 3, 2]
22	0	Concat	[1]
23	493056	C2f	[384, 256, 1]
24	2598943	Detect	[5, [64, 128, 256]]

We use BCE Loss as the classification loss and DFL Loss + CIOU Loss as the regression loss for the improved network structure. Since the cognitive engagement detection task is a multi-classification problem, the binary representations of multiple class labels are written together to form a one-hot vector.

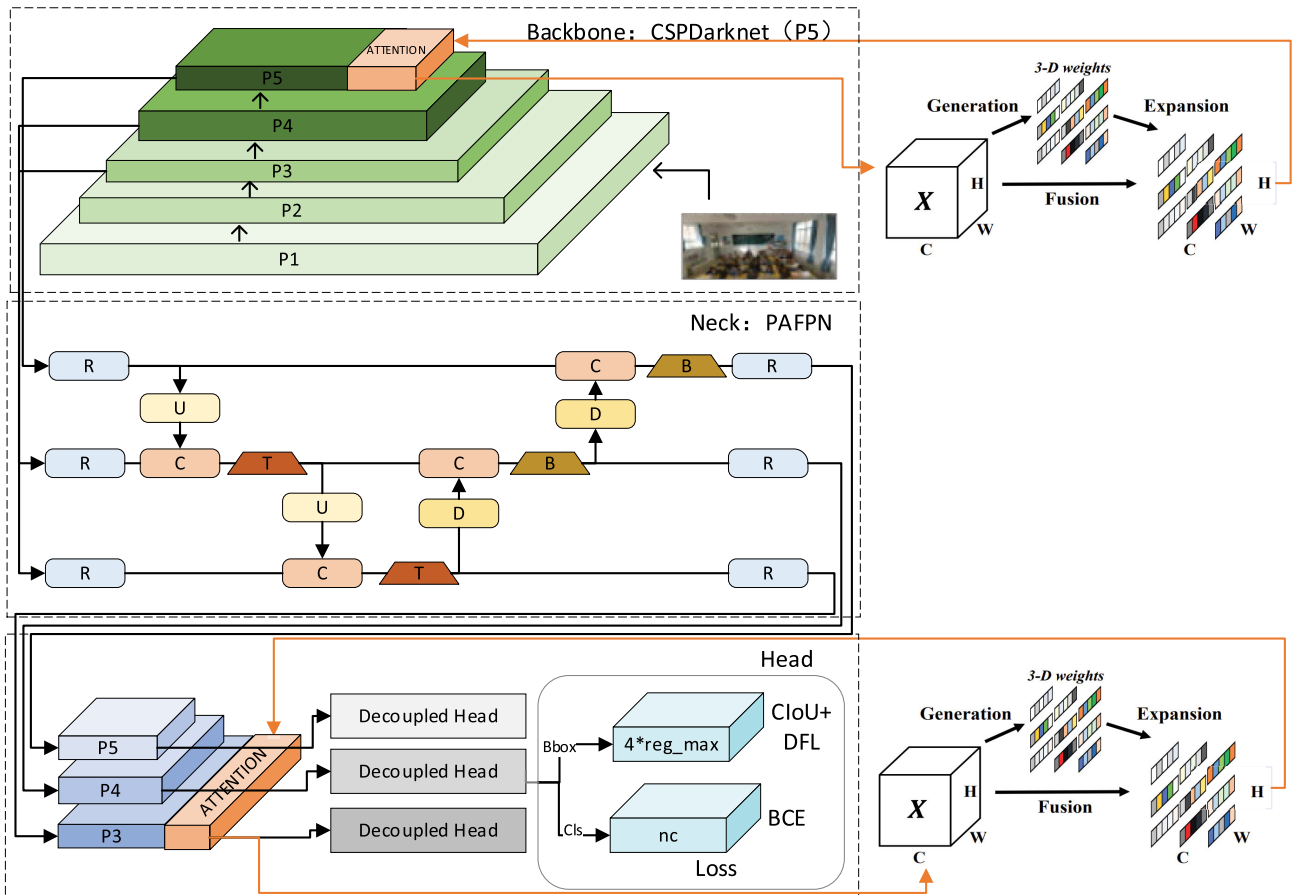


FIGURE 5. The structure of simAM-YOLOv8n model. The orange rectangles in the Backbone and Head networks represent the added simAM attention modules.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL SETTING

We implemented the experiments in Python-3.9 torch-1.13 on a server with 32 GB of RAM and NVIDIA GeForce RTX 3080 10240MB GPU. The code is based on Reis et al. [43]. The images were to get $3 \times 640 \times 640$ ($C \times H \times W$) tensors as inputs to the simAM-YOLOv8n model. The $3 \times 224 \times 224$ dimensions are the standard dimension for input to the model. In this paper, Accuracy (ACC), Precision (P), Recall (R), F1-Score (F1), Mean Average Precision 50 (mAP50), Mean Average Precision 50-90 (mAP50-95), and Average Accuracy (AA) are used to evaluate the detection performance of the different algorithms. Because there is no similar ICAPD-related dataset, the proposed simAM-YOLOv8n can only be directly trained on the ICAPD dataset, so we conduct experiments on the ICAPD dataset to verify the effectiveness of the proposed method. We adopted a stochastic gradient descent (SGD) optimization algorithm, and the initial learning rate and momentum parameters were set to 0.01 and 0.937, respectively. The number of iterations of the model training is 200 iterations. The batch size was set to 16, the number of workers was set to 0, training was performed using Automatic Mixed Precision, the patience was set to 50, label smoothing was set to 0,

and the nominal batch size was set to 64. In addition, we used some strategies to learn more features: first, we increased the weight of the bounding box loss. It adjusted the weight of the bounding box loss from 7.5 to 8.5 to obtain more features related to behavioral annotation boxes. This provides more effective features for cognitive engagement prediction. Second, we dynamically allocated image weights. It controlled the probability of sampling images based on the number of behavioral categories to learn more complex behavioral image features. Third, we dynamically adjusted the learning rate. It used the cosine annealing algorithm to control the learning rate and the speed at which the simAM-YOLOv8n learns more behavioral deep features.

B. TRAINING PROCEDURES

Firstly, the ICAPD dataset is divided into 16 batches, and each batch is input into the model for training. Then, the images in each batch are fed through the simMA-YOLOv8n model for forward propagation, resulting in predicted bounding boxes and class probabilities for each student. Next, the predicted bounding boxes and class probabilities are compared with the ground truth labels to calculate the losses. Subsequently, the model weights are updated through backpropagation using the gradients of the loss functions to minimize the losses. This

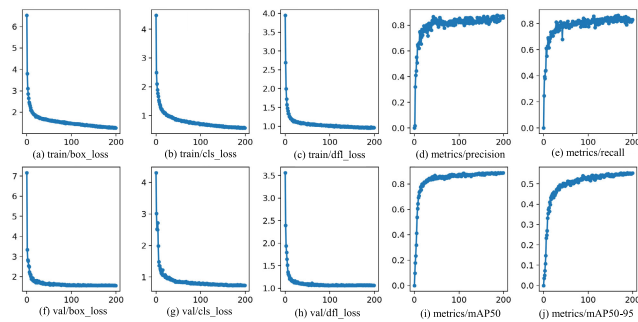


FIGURE 6. The simAM-YOLOv8n training loss functions and performance curves on ICAPD dataset. The first three columns represent the loss curves, and the last two represent the performance curves.

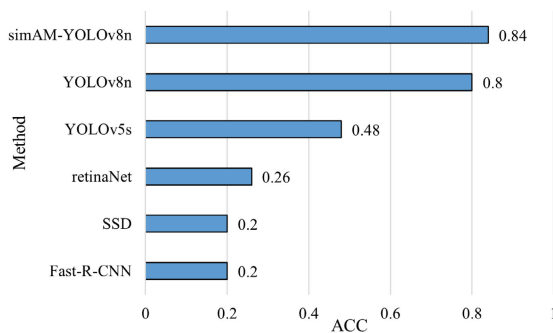


FIGURE 7. Comparison of the ACC with existing methods on ICAPD dataset.

process is repeated until the predefined number of training epochs, which is set to 200, or until a stopping criterion is met. As shown in Figure 6, the loss values of the proposed simAM-YOLOv8n keep decreasing with the increase of epochs. When the number of epochs approaches 200, the loss values decrease to a relatively small value. The P and R values increase with the increase of epochs and finally reach a relatively stable level. This shows that under the joint efforts of the optimization algorithm, the empirical risk and error of the model gradually decrease, and the optimization degree of the network gradually increases. Finally, the optimal weight parameters of the model are saved, and the detailed evaluation of the automatic detection of student cognitive engagement is carried out on the test set.

C. EXPERIMENTAL ANALYSIS AND COMPARISON

1) COMPARISON WITH BASELINES AND STATE-OF-THE-ARTS

This work focuses on solving the accuracy problem in detecting cognitive engagement classes. Especially in the Active and Disengage classes, teachers need to guide teaching through the detection results of these classes. We select the popular object detection models for comparison, which are Fast-R-CNN [15], retinaNet [51], SSD [16], YOLOv5s [52] and YOLOv8n [43]. We compared the ACC of different models in automatically detecting student cognitive engagement, as shown in Figure 7 and Table 4.

It finds that YOLO series models perform better than other object detection models. The performance of YOLOv8n

TABLE 4. Comparison of the ACC of methods in the five classes. The D, P, A, C and I letters represent the Disengage, Passive, Active, Constructive, and Interactive classes. The thickened value indicates the best improvement effect.

Method	D	P	A	C	I
Fast-R-CNN	0.17	0.6	0.19	0.02	0
retinaNet	0.18	0.62	0.21	0.26	0.04
SSD	0.17	0.52	0.15	0.14	0
YOLOv5s	0.41	0.82	0.31	0.61	0.25
YOLOv8n	0.72	0.93	0.63	0.86	0.87
simAM-YOLOv8n	0.76	0.95	0.71	0.85	0.93

is much better than that of YOLOv5s, indicating that the updated version of YOLO has better performance. The proposed simAM-YOLOv8n achieves an ACC of 84%, outperforming the YOLOv8n, which obtains an ACC of 80%. The proposed approach with the simAM attention mechanism outperformed any other state-of-the-art method on the ICAPD dataset in terms of overall ACC value. All methods can achieve good detection results in the Passive and Constructive classes. Because the behavioral features in these classes are not complicated. The poor performance of Fast-R-CNN and SSD in the Interactive class is due to the data imbalance on the ICAPD dataset. The proportion of these behaviors will not be too large, which is consistent with the characteristics of classroom spontaneity. Compared with other object detection models, the proposed model has strong comparability in each class, which verifies that adding attention mechanisms helps improve the model's performance. Compared with YOLOv8n, the proposed simAM-YOLOv8n maintains a comparable ACC value in some classes, but has a higher ACC value in the more confusing Active and Disengage classes, which is more conducive to solving classroom teaching problems. Because the teacher always pays more attention to the students who are not engaged or initiate active behaviors. Therefore, the proposed simAM-YOLOv8n has advantages in terms of overall accuracy and has strong comparability in different classes on the ICAPD dataset.

Furthermore, results of the YOLOv8n and the simAM-YOLOv8n on two example frames are shown in Figure 8, where the first row is the input image, and the second and third rows are the detection results of YOLOv8n and simAM-YOLOv8n, respectively. In the first column, the YOLOv8n detects the location of most students but mis-detects some active behaviors. The active behaviors, such as the third column of the second row, the second column of the third row, and the first column of the first row in the input image, are more accurately detected by the simAM-YOLOv8n. The proposed simAM-YOLOv8n model can detect the location and cognitive engagement of the students wearing masks. Moreover, the simAM-YOLOv8n detects smaller objects in the back row, which verifies the effectiveness of the attention mechanism on dense and small object detection. From the case of the second column, it can be observed that the behaviors in the front row and

TABLE 5. Comparison of the P, R, F1, and ACC with existing cognitive engagement detection methods. The thickened value indicates the best improvement effect. Metrics for which values were not provided in the original paper have been marked as "none."

Method	Data	P	R	F1	ACC	Annotation
BERT-CNN	text	0.79	0.78	0.78	0.79	active, constructive and interactive
GAN	audio	none	none	none	0.34	valence, arousal
SVM	video	0.83	none	none	0.75	engaged, and less engaged
GMM	video	none	none	none	0.68	left-slide, right-slide, and unfocused
simAM-YOLOv8n	video	0.86	0.85	0.86	0.84	disengage, passive, active, constructive and interactive

the small objects in the back row (e.g., the fifth row of the second column) are more accurately detected by the simAM-YOLOv8n. We can also find that YOLOv8n is not very accurate. For example, there is an active behavior that is mistaken for passive behavior on the first row of the first column and a passive behavior is mistaken for an active behavior on the first row of the third column. Overall, the above results show that the ICAPD framework can support computer vision-based cognitive engagement detection, and the simAM-YOLOv8n model has a better performance than YOLOv8n in detecting cognitive engagement with low-resolution images and complex behaviors.

2) COMPARISON WITH EXISTING COGNITIVE ENGAGEMENT DETECTION METHODS

To validate the superiority of the proposed ICADP framework and simAM-YOLOv8n method in cognitive engagement detection, we conducted comparisons with Bidirectional Encoder Representation from Transformers - Convolutional Neural Networks (BERT-CNN) based on text [53], Generative Adversarial Networks (GAN) based on audio [5], Support Vector Machine (SVM) based on video [12], Gaussian Mixture Models (GMM) based on video [54]. The evaluation metrics include Precision (P), Recall (R), F1-score (F1), Accuracy (ACC), and Annotation, as shown in the Table 5. From the Annotation perspective, the proposed ICAPD framework offers finer-grained annotations for cognitive engagement, allowing for a more comprehensive representation of engagement levels. In terms of other metrics, the proposed simAM-YOLOv8n outperforms other models. Overall, our ICAPD framework and simAM-YOLOv8n model are better suited for cognitive engagement detection tasks.

D. DISCUSSION

1) EFFECT OF DIFFERENT ATTENTION MECHANISMS

To further verify the advantages of combining the simAM with the YOLOv8n, we compared the proposed simAM-YOLOv8n with different variants of attention-based YOLOv8n. These compared models were YOLOv8n with the SE [32] (SE-YOLOv8n), YOLOv8n with the ECA [29] (ECA-YOLOv8n), YOLOv8n with the SKnets [30] (SKnets-YOLOv8n) and YOLOv8n with the EMA [25]

(EMA-YOLOv8n). These attention modules were added twice in the YOLOv8n, the same as the position added by the simAM module. In addition, baseline YOLOv8n was also used for comparison. The experimental environment and the training parameters were the same for all models. The confusion matrices and AA values of different methods are shown in Figures 9 and 10.

It can be found that the proposed simAM-YOLOv8n achieves the best performance in all classes. However, other attention-based methods only exceed the YOLOv8n in some classes, which is not friendly to the overall accuracy. Except for the proposed simAM-YOLOv8n, none of the compared models can get more than 70% accuracy on Active class. The reason for this may be the high inter-class variance in the Active class (As shown in Figure 2). Furthermore, the behaviors belonging to the Disengage class should be prevented or solved with early identification. So there is a high demand for higher detection accuracy on the Disengage class. The proposed simAM-YOLOv8n can achieve a comparable result of 76%, which exceeds most attention-based models. Although SKnets-YOLOv8n exceeds the proposed simAM-YOLOv8n in the Disengage class, it has severe performance degradation in other classes. Overall, the proposed simAM-YOLOv8n can exceed or equal YOLOv8n in each class and can achieve the best results in the Active class containing the most complex behaviors. So the proposed method can provide assistance for teachers to adapt teaching.

2) EFFECT OF DIFFERENT POSITIONS OF ATTENTION MODULE

To further explore the effect of adding attention mechanism modules at different positions of YOLOv8n on the performance of the algorithm, the simAM module was added to the P3, P4, and P5 in the Backbone network (as shown in Figure 1), i.e. added between layers 4 and 5, between layers 6 and 7, and in layer 9 (as shown in Table 3). These new models were named BP3-simAM-YOLOv8n, BP4-simAM-YOLOv8n, and BP5-simAM-YOLOv8n, respectively. After adding the simAM module to the P3, P4, and P5 in the Head network (as shown in Figure 1), i.e. added in layer 17, between layers 20 and 21, between layers 23 and 24 (as shown in Table 3), the models were named HP3-simAM-YOLOv8n, HP4-simAM-YOLOv8n, and HP5-simAM-YOLOv8n, respectively. We named the proposed model of adding simAM module both behind the P5 in the Backbone network (i.e. layer 9 as shown in Table 3) and behind the P3 in the Head network (i.e. layer 17 as shown in Table 3) as simAM-YOLOv8n. The above methods were tested in turn. The experimental results are shown in Table 6.

The experimental results show that in terms of the ACC and P, the BP5-simAM-YOLOv8n model performs the best when compared with BP3-simAM-YOLOv8n, BP4-simAM-YOLOv8n and BP5-simAM-YOLOv8n. The ACC and R values of the HP3-simAM-YOLOv8n perform best when comparing the models of simAM added to the Head network



FIGURE 8. Comparison of the YOLOv8n and simAM-YOLOv8n in student cognitive engagement detection results. The first line is the original image, the second line is the detection result of YOLOv8n, and the third line is the detection result of simAM-YOLOv8n. The first and second columns represent two different cases.

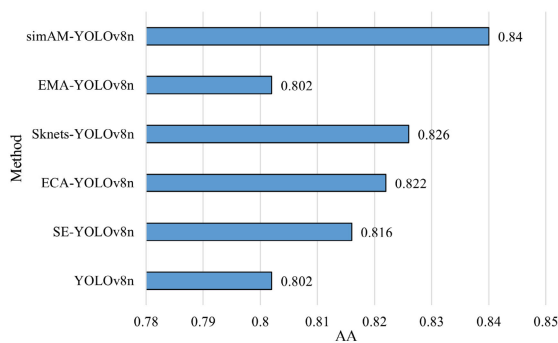


FIGURE 9. Comparison of the AAs of different methods in the ICAPD dataset. The AA is the mean of class-specific accuracy values.

of YOLOv8n. Since the ACC values are a useful reference for evaluating the classification performance of cognitive engagement, the simAM-YOLOv8n considers the advantages of both complex behavior feature extraction in the Backbone network and small target detection in the Head network, making the algorithm achieve the best performance in terms of ACC value. The proposed simAM-YOLOv8n does not emphasize the values of mAP50 and mAP50-90 because we pay more attention to the classification of cognitive engagement rather than the detection of students.

3) ICAPD FRAMEWORK AND THE ROLE OF THE TEACHER

In our recording situation, there is no prescribed standard for what the students should do. The role of the teacher is fundamentally that of an instructional guide. Because in an authentic learning environment (such as a classroom), teachers cannot tell which knowledge-change processes students are thinking, how can this challenge be resolved so teachers can know how students are engaged? One possible way is to map what students do while engaged with instruction that is visible behaviorally to the invisible knowledge-change processes they are undertaking. Accordingly, a taxonomy was proposed that specified roughly four distinguishable classes of student overt behaviors. The benefit of our ICAPD is that this operational definition can guide teachers to evaluate their own design of student activities, to determine what class of student outputs their designed activities elicit (i.e., they can compare the anticipated student response with the presented instructional materials). Subsequently, it remains essential for teachers to utilize the results of our detection according to their specific needs. For example, ICAPD's easily defined rubric, requiring only a comparison of what is generated to what is presented instructionally, can teach teachers how to design a deep question: A deeper question can simply be one that elicits generative responses from

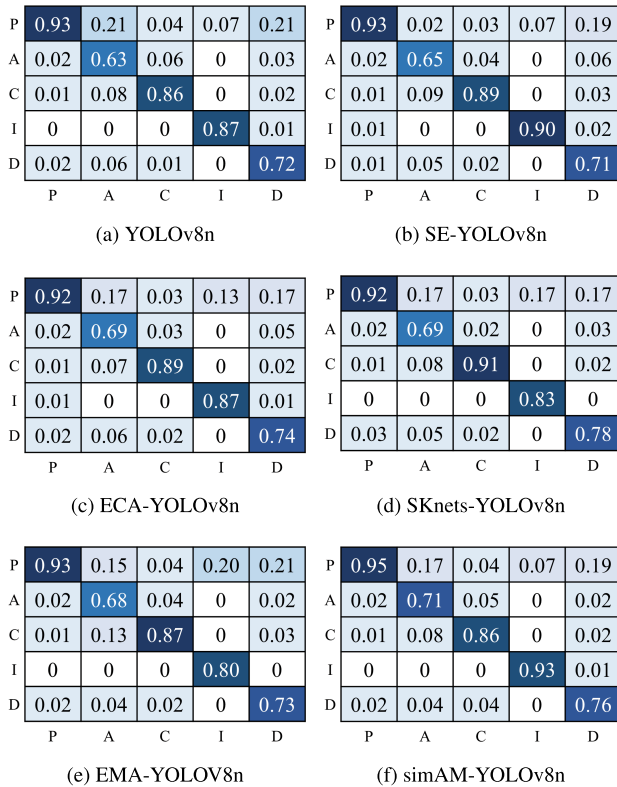


FIGURE 10. Confusion matrices of different attention-based methods on ICAPD dataset. The horizontal axis represents the real class, and the vertical axis represents the predicted class. The P, A, C, I, and D represent the Passive, Active, Constructive, Interactive, and Disengage classes. The values on the diagonal represent the proportion of classes correctly classified. Note that the confusion matrix shows the confusion only between P, A, C, I, and D classes.

TABLE 6. Comparison of the overall performance of models based on different positions on ICAPD dataset. The thickened value indicates the best improvement effect.

Method	P	R	F1	ACC	mAP50	mAP50-95
BP3-simAM-YOLOv8n	0.843	0.808	0.825	0.814	0.882	0.537
BP4-simAM-YOLOv8n	0.836	0.823	0.829	0.790	0.867	0.535
BP5-simAM-YOLOv8n	0.844	0.794	0.818	0.824	0.871	0.531
HP3-simAM-YOLOv8n	0.811	0.848	0.829	0.820	0.876	0.543
HP4-simAM-YOLOv8n	0.834	0.808	0.821	0.792	0.869	0.531
HP5-simAM-YOLOv8n	0.854	0.826	0.840	0.812	0.882	0.542
simAM-YOLOv8n	0.861	0.846	0.853	0.840	0.888	0.555

students, containing information that has not been presented instructionally. We hope to prescribe what teachers should do in instruction based on what the ICAPD mode a teacher’s activity elicits in students’ engagement.

4) INFLUENCE OF THE PROPOSED METHOD IN CLASSROOM The ICAPD framework can provide guidance for annotating visual behaviors. It offers a more comprehensive set of annotations for cognitive engagement than the ICAP. The

simAM-YOLOv8n model can provide outputs that closely approximate genuine cognitive engagement results. It delivers more precise detection results compared to YOLOv8n. Furthermore, the proposed method provides a promising starting point for reducing the effort involved in manual video inspection and annotation, which in turn would facilitate the analysis of larger numbers of individuals and longer videotaped lessons. The results can be combined with other empirical experiences to enhance student engagement [55], [56], [57].

5) LIMITATIONS OF OUR STUDY

This study proposes the ICAPD framework and simAM-YOLOv8n for student cognitive engagement detection in the classroom. The experimental results verify the usefulness of it, which provides a new perspective for learning analytics. However, this study has several limitations that should be acknowledged. Firstly, there was inadequate consideration of parental consent and the understanding of the study by the children. Secondly, accurately determining students’ true cognitive engagement with a single video proved to be challenging. This limitation may impact the accuracy and reliability of the findings related to students’ cognitive engagement. Thirdly, the study did not sufficiently consider the role of instructional teachers in determining students’ cognitive engagement. Teachers play a crucial role in facilitating and guiding students’ learning experiences, and their influence on students’ engagement should not be overlooked. Lastly, the performance of the simAM-YOLOv8n model in handling complex classroom scenarios with significant occlusions was not thoroughly evaluated. This limitation raises concerns about the generalizability and applicability of the model in other real-world settings where occlusions and complex classroom dynamics are common. Acknowledging these limitations is essential for a comprehensive understanding of the study’s findings and to guide future research in addressing these concerns to enhance the validity and reliability of the research in this area.

V. CONCLUSION

This paper proposes the ICAPD framework to represent cognitive engagement in the classroom. This framework defines five different levels: Disengage, Passive, Active, Constructive, and Interactive. These higher-order levels are mapped to lower-order explicit learning behaviors. Based on the ICAPD framework, an ICAPD dataset was constructed for cognitive engagement detection in the classroom. Furthermore, the simAM-YOLOv8n model was proposed to detect the students’ cognitive engagement more effectively. Experimental results on the self-built dataset have validated the effectiveness of our method. Therefore, future research will focus more on the impact of boundary conditions in data collection and the role of instructional teachers. The use of multimodal and multi-video data to detect students’ cognitive engagement will also be considered. Additionally,

algorithmic structures tailored to complex scenarios will be explored to optimize the model's performance. These efforts will help to address the limitations of the current study and improve the validity and reliability of future research in this area.

ACKNOWLEDGMENT

The authors would like to thank Zhiqiang Sang for collecting the classroom videos and Ruyi Jiang for annotating the data.

REFERENCES

- J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Rev. Educ. Res.*, vol. 74, no. 1, pp. 59–109, Mar. 2004.
- C. A. Wolters and D. J. Taylor, "A self-regulated learning perspective on student engagement," in *Handbook of Research on Student Engagement*. Boston, MA, USA: Springer, 2012, pp. 635–651.
- H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Social Behav. Personality, Int. J.*, vol. 46, no. 3, pp. 517–528, Mar. 2018.
- 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, Apr. 2023.
- P. Guhan, N. Awasthi, a. K. McDonald, K. Bussell, D. Manocha, G. Reeves, and A. Bera, "Developing an effective and automated patient engagement estimator for telehealth: A machine learning approach," 2020, *arXiv:2011.08690*.
- S. Li, "Measuring cognitive engagement: An overview of measurement instruments and techniques," *Int. J. Psychol. Educ. Stud.*, vol. 8, no. 3, pp. 63–76, Oct. 2022.
- A. Levordashka, D. Stanton Fraser, and I. D. Gilchrist, "Measuring real-time cognitive engagement in remote audiences," *Sci. Rep.*, vol. 13, no. 1, p. 10516, Jun. 2023.
- 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.
- C. Antonietti, M.-L. Schmitz, T. Consoli, A. Cattaneo, P. Gonon, and D. Petko, "Development and validation of the ICAP technology scale to measure how teachers integrate technology into learning activities," *Comput. Educ.*, vol. 192, Jan. 2023, Art. no. 104648.
- P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educ. Psychol. Rev.*, vol. 33, no. 1, pp. 27–49, Mar. 2021.
- N. Bosch and S. K. D'Mello, "Automatic detection of mind wandering from video in the lab and in the classroom," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 974–988, Oct. 2021.
- S. Li, S. P. Lajoie, J. Zheng, H. Wu, and H. Cheng, "Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving," *Comput. Educ.*, vol. 163, Apr. 2021, Art. no. 104114.
- J. Mo, R. Zhu, H. Yuan, Z. Shou, and L. Chen, "Student behavior recognition based on multitask learning," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 19091–19108, May 2023.
- 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.
- R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 21–37.
- 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.
- J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 and beyond," 2023, *arXiv:2304.00501*.
- V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
- J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*.
- D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
- K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1462–1471.
- M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.
- D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
- Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.
- Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," 2021, *arXiv:2111.12419*.
- L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.
- Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
- S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- Z. Xu, J. Li, Y. Meng, and X. Zhang, "CAP-YOLO: Channel attention based pruning YOLO for coal mine real-time intelligent monitoring," *Sensors*, vol. 22, no. 12, p. 4331, Jun. 2022.
- A. Kumar, "SEAT-YOLO: A squeeze-excite and spatial attentive you only look once architecture for shadow detection," *Optik*, vol. 273, Feb. 2023, Art. no. 170513.
- L. Li, M. Liu, L. Sun, Y. Li, and N. Li, "ET-YOLOv5s: Toward deep identification of students' in-class behaviors," *IEEE Access*, vol. 10, pp. 44200–44211, 2022.
- 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.
- A. J. Magana, T. Amuah, S. Aggrawal, and D. A. Patel, "Teamwork dynamics in the context of large-size software development courses," *Int. J. STEM Educ.*, vol. 10, no. 1, pp. 1–16, Sep. 2023.
- N. Bergdahl, J. Nouri, U. Fors, and O. Knutsson, "Engagement, disengagement and performance when learning with technologies in upper secondary school," *Comput. Educ.*, vol. 149, May 2020, Art. no. 103783.
- K. Salmela-Aro, K. Upadyaya, K. Hakkarainen, K. Lonka, and K. Alho, "The dark side of internet use: Two longitudinal studies of excessive internet use, depressive symptoms, school burnout and engagement among Finnish early and late adolescents," *J. Youth Adolescence*, vol. 46, no. 2, pp. 343–357, Feb. 2017.
- M.-T. Wang, J. Fredricks, F. Ye, T. Hofkens, and J. S. Linn, "Conceptualization and assessment of adolescents' engagement and disengagement in school," *Eur. J. Psychol. Assessment*, vol. 35, no. 4, pp. 592–606, Jul. 2019.
- A. Abedi and S. S. Khan, "Detecting disengagement in virtual learning as an anomaly using temporal convolutional network autoencoder," 2022, *arXiv:2211.06870*.
- N. Naik and M. A. Mehta, "An improved method to recognize hand-over-face gesture based facial emotion using convolutional neural network," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2020, pp. 1–6.

- [43] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with YOLOv8," 2023, *arXiv:2305.09972*.
- [44] M. Mahmoud and P. Robinson, "Interpreting hand-over-face gestures," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Cham, Switzerland: Springer, 2011, pp. 248–255.
- [45] M. Mahmoud, T. Baltrušaitis, and P. Robinson, "Automatic analysis of naturalistic hand-over-face gestures," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, pp. 1–18, Aug. 2016.
- [46] B. Pease and A. Pease, *The Definitive Book of Body Language: The Hidden Meaning Behind People's Gestures and Expressions*. New York, NY, USA: Bantam, 2008.
- [47] E. Shapiro, *Direct Observation: Manual for the Behavioral Observation of Students in Schools (BOSS)*. Bloomington, MN, USA: Pearson, 2004.
- [48] H. Salam, O. Belkhat, H. Gunes, and M. Chetouani, "Automatic context-aware inference of engagement in HMI: A survey," *IEEE Trans. Affect. Comput.*, early access, May 26, 2023, doi: [10.1109/TAFFC.2023.3278707](https://doi.org/10.1109/TAFFC.2023.3278707).
- [49] D. Ramadhanti and D. P. Yanda, "Students' metacognitive awareness and its impact on writing skill," *Int. J. Lang. Educ.*, vol. 5, no. 3, pp. 193–206, 2021.
- [50] L. E. Margulieux and R. Catrambone, "Scaffolding problem solving with learners' own self explanations of subgoals," *J. Comput. Higher Educ.*, vol. 33, no. 2, pp. 499–523, Aug. 2021.
- [51] 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. (ICCV)*, Oct. 2017, pp. 2999–3007.
- [52] G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon, and Y. Defretin, "ultralytics/YOLOv5: V5. 0-YOLOv5-P6 1280 models, AWS, supervise. ly and YouTube integrations," Zenodo, CA, USA, Tech. Rep. 6, 2021.
- [53] S. Liu, S. Liu, Z. Liu, X. Peng, and Z. Yang, "Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement," *Comput. Educ.*, vol. 181, May 2022, Art. no. 104461.
- [54] Y. Liu, J. Chen, M. Zhang, and C. Rao, "Student engagement study based on multi-cue detection and recognition in an intelligent learning environment," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28749–28775, Nov. 2018.
- [55] C. Bryson and L. Hand, "The role of engagement in inspiring teaching and learning," *Innov. Educ. Teaching Int.*, vol. 44, no. 4, pp. 349–362, Nov. 2007.
- [56] S. J. Mann, "Alternative perspectives on the student experience: Alienation and engagement," *Stud. Higher Educ.*, vol. 26, no. 1, pp. 7–19, Mar. 2001.
- [57] Z. Zhang, P. Lin, S. Ma, and T. Xu, "An improved YOLOv5s algorithm for emotion detection," in *Proc. 5th Int. Conf. Pattern Recognit. Artif. Intell. (PRAI)*, Aug. 2022, pp. 1002–1006.



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.



YANTAO WEI received the Ph.D. degree from the School of Artificial Intelligence and Automation, Institute of Image Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently an Associate Professor with the Faculty of Artificial Intelligence in Education, Central China Normal University, China, where he is also the Vice Director of the Hubei Research Center for Educational Informationization. His publications have appeared in international journals, such as *IEEE TRANSACTIONS ON CYBERNETICS*, *Pattern Recognition*, *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, *IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT*, *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS*, and *Neurocomputing*. His research interests include educational artificial intelligence, computer vision, and machine learning.



JIE GAO was born in Lanzhou, China, in 1998. She is currently pursuing the master's degree with Central China Normal University, Wuhan, China. Her research interests include intelligent education, emotion evolution analysis, and deep learning.



HUANG YAO received the Ph.D. (Eng.) degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2011. He is currently an Associate Professor with the Faculty of Artificial Intelligence in Education, Central China Normal University, China. His research interests include deep learning, image processing, and computer vision.



QINGTANG LIU (Member, IEEE) received the Ph.D. degree in electronic information engineering from the Huazhong University of Science and Technology, Wuhan, in 2005. He is a Professor with Central China Normal University. His current major research interests include learning analytics technology and digital learning. He is a member of ISO/IEC JTC1 SC36, AVS Standard Organization, and the National Beacon Committee Education Technology Sub-Technical Committee (CELTSC); and an ACM Member.

...