

SBCP-YOLO-R3D: Student Behavior Recognition and Visualization Framework Using Improved YOLO and R3D for Class Video

Chunyan Yu,^{1,3} Qin Ding,² and Yuchen Bai¹

¹School of Computer and Information Engineering, Chuzhou University, Chuzhou, 239000, Anhui, China

²School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan, 232001, Anhui, China

³School of Education Science, Nanjing Normal University, Nanjing, 210023, Jiangsu, China

(Received 28 October 2024; Revised 29 November 2024; Accepted 04 December 2024; Published online 02 February 2025)

Abstract: Real-time recognition and visualization of students' behaviors in face-to-face classrooms serve as pivotal indicators of learning engagement. However, current methods exhibit limitations in both real-time performance and accuracy. Additionally, in-depth studies have not been extensively conducted to evaluate learning status more conveniently by utilizing computer vision techniques. To address these issues, a novel Student Behavior Recognition and Dynamic Class Portraits Construction framework, named SBCP-YOLO-R3D, incorporating the StB-YOLO and R3D methods, has been proposed to detect student behaviors and construct class portraits. The developed framework comprises two layers: the StB-YOLO detection layer and the R3D classification layer. In the StB-YOLO detection layer, the Lightweight-SEAM (LW-SEAM) is incorporated into YOLOv5 to enhance the recognition of occluded students, by capturing contextual information and enhancing occlusion-related features. Moreover, a Double-SlideLoss function is devised, employing adaptive weighting mechanisms to strike an optimal balance between simple and challenging samples. In the R3D classification layer, the results generated by StB-YOLO are then processed using R3D to produce class portraits. Experiments conducted on the StuAct and SCB-DATASET3-S datasets demonstrate the effectiveness of the StB-YOLO. Compared with the baseline model, StB-YOLO increases the mAP by 3.1%.

Keywords: learning portraits; occlusion attention; student behavior recognition; YOLO

I. INTRODUCTION

Classroom teaching is the primary form of educational activities and a significant subject of educational research. Student behavior in the classroom serves as an external manifestation of their level of concentration, containing complex and valuable information. Real-time analysis and understanding of students' classroom behavior positively impact teachers' ability to promptly adjust teaching strategies and enhance teaching effectiveness. With the improvement of educational facilities, many campuses have implemented routine classroom recording systems. These systems, by installing cameras in classrooms, can comprehensively record students' classroom behavior. In universities and colleges, a course typically spans over a dozen weeks, with each week's sessions occurring at a fixed time. The video data, which captures each learning process, provides rich material for data analysis. However, since the cameras are primarily used to record the overall classroom situation, they often have low resolution and the insufficient lighting in some classrooms results in poor data quality.

Deep learning has achieved a series of advancements in the field of image and video processing. Many researchers have applied deep learning and computer vision techniques to analyze student engagement in classrooms. For instance, Zhao *et al.* [1] employed YOLOv7 along with a multi-head self-attention mechanism to enhance feature extraction for detecting student and teacher behaviors in classrooms. Li *et al.* [2] proposed an algorithm based

on attention-based relational reasoning and relational feature fusion to simulate the relationships between individuals and objects in the classroom. Zhou *et al.* [3] combined YOLOv8, MTCNN, CovPoolFER, and OpenPifPaf to recognize behaviors. Although existing methods have achieved remarkable results, several challenges remain in this context:

- **Scale variation.** Students who are closer to the surveillance camera appear larger size, while students who are further away have smaller sizes. On the one hand, small targets lack visual features, making it difficult to recognize them accurately. On the other hand, significant scale differences can cause the model to over-attend to targets of certain sizes and under-perform targets of other sizes.
- **Occlusion.** Occlusion among students in classroom poses a significant challenge in behavior recognition. Severe occlusion can result in substantial loss of critical features, ultimately compromising the algorithm's precision in identifying occluded students.
- **Imbalanced samples.** While certain behaviors, such as looking up to listen and reading/writing, are prevalent in classrooms, others like standing up to answer and resting on the desk are infrequent. This tends to skew the algorithm's focus toward more common behaviors, thereby affecting its accuracy in identifying exceptional behaviors that warrant special consideration.

Additionally, available datasets are lacking due to privacy concerns. However, we discover Student Classroom Behavior (SCB) dataset [4], which consists of images of students in

Corresponding author: Chunyan Yu (e-mail: yuchy@chzu.edu.cn).



Fig. 1. Sample images are presented from the StuAct and SCB-DATASET3-S datasets. (a) The Stu-Act dataset and (b) the SCB-DATASET3-S dataset.

classrooms collected from the Internet, as shown in Fig. 1(a). The dataset categorizes student behaviors into three classes: hand-raising, reading, and writing.

In this paper, we originate a dataset of student behaviors in real-world college classroom scenarios, as depicted in Fig. 1(b). Subsequently, we propose a Student Behavior Recognition and Dynamic Class Portrait Construction framework, SBCP-YOLO-R3D. It comprises two layers: StB-YOLO detection layer for identifying student behaviors and R3D classification layer for structuring class portraits. Our main contributions are summarized as follows:

1. We construct a student behavior dataset consisting of classroom surveillance videos with data privacy compliance. We annotate each student with seven distinct behavioral categories.
2. We employ Lightweight Separated and Enhancement Attention Module (LW-SEAM) to address the occlusion issue. Existing SEAM compensates for information loss in occluded regions. However, SEAM contains a large number of parameters, which is not conducive to edge devices. Thus, we develop a more lightweight version, LW-SEAM, to reduce the model size.
3. We propose a Double-SlideLoss function to address the issue of imbalanced samples. ClsSlideLoss is a novel classification loss function designed to address the issue of sample imbalance in classification tasks. It utilizes adaptive weighting for few and hard samples.
4. We propose a novel Student Behavior Recognition and Dynamic Class Portraits Construction framework using the StB-YOLO and R3D (SBCP-YOLO-R3D) methods.

The rest of the paper is organized as follows. Section II discusses related works. Section III presents the design of our proposed model. Section IV reports the experimental results, and Section V summarizes our work.

II. RELATED WORKS

Student behavior recognition has attracted significant research interest in recent years. Researchers commonly employ object

detection methods, pose estimation techniques, and spatio-temporal behavior recognition methods to accomplish this task.

A. OBJECT DETECTION AND ATTENTION MECHANISM

Convolutional Neural Network (CNN)-based algorithms can be broadly classified into two-stage detectors and one-stage detectors. Two-stage detectors such as Faster R-CNN [5] tend to exhibit slower inference speeds. In contrast, one-stage detectors, such as YOLO, directly predict bounding boxes, leading to faster computation speeds.

Many Transformer-based object detection models emerged in recent years, for instance, DETR [6], Deformable DETR [7], and YOLOS [8]. A multimodal fusion-based Transformer was proposed in [9] to enhance the performance of expression recognition in conversation. However, the extensive dot-product operations in the Transformer model significantly hindered the inference speed of the algorithm in vision tasks. Despite some sparse attention methods such as Swin Transformer [10] and BiFormer [11] were proposed, the computational speed of these models remained slower than many purely CNN-based architectures.

Attention mechanism enables the model to focus on the salient parts of the input data while ignoring irrelevant information. GAM-Attention [12] took into account the interaction between channels and spatial dimensions. NLNet [13] employed the self-attention mechanism to model long-range dependencies. Zhang *et al.* [14] employed an attention mechanism across channels to represent various occlusion patterns. Xie *et al.* [15] utilized an attention network to emphasize visible pedestrian regions while suppressing occluded regions.

B. CLASSROOM BEHAVIOR RECOGNITION BASED ON DEEP LEARNING METHODS

Skeleton and pose-based methods. Lin *et al.* [16] proposed an error correction scheme based on pose estimation and person detection technology. Similarly, Pabba *et al.* [17] extracted both skeleton features, facial action units, and head pose features and combined these features to jointly predict student behavior.

Object detection-based methods. Rashmi *et al.* [18] employed YOLOv3 [19] to identify five types of student behaviors, leveraging image template matching to reduce the number of image frames, thereby enhancing video processing speed. Zhao *et al.* [1] integrated Efficient Transformer Block (ETB) and Efficient Convolution Aggregation Block (ECAB) modules into YOLO to extract image features.

Spatio-temporal-based methods. To enhance the recognition of interactions between people and objects, Li *et al.* [2] independently utilized Deep Convolutional Neural Network (3DCNN) and Faster R-CNN separately to extract scene features and Region Of Interest (ROI) from classroom videos. Albert *et al.* [20] posited that students' behavioral changes require real-time observation, hence utilizing 3DCNN to identify eight types of student behaviors.

Although some achievements have been made in the recognition of students' classroom behaviors, current research still faces some limitations. The changes in students' postures during class are subtle, thus resulting in minor behavioral differences between frames. Spatio-temporal-based behavior recognition needs to redundantly process many similar frames. The skeleton-based method requires extensive manual annotations. While existing object detection algorithms possess the ability to classification, occlusion and uneven sample distribution in classroom

environments often result in poor recognition and classification performance.

III. METHOD

This section introduces the Student Behavior Recognition and Dynamic Class Portraits Construction framework using the StB-YOLO and R3D (SBCP-YOLO-R3D) methods. The framework is divided into two layers: StB-YOLO detection layer and R3D classification layer, as depicted in Fig. 2. StB-YOLO detection layer is employed to detect student behaviors, while the role of the R3D classification layer is to categorize the learning state throughout the entire class swiftly. In R3D classification layer, the detection results are encoded using different colors to obtain behavior-color-coded charts (BCCC) for each image. The input data for R3D consists of instantaneous BCCC, which are processed by R3D to obtain classroom portraits. As depicted in Table I, there are four types of class portraits. The explanations of the colors in the examples can be found in Fig. 2.

A. StB-YOLO DETECTION LAYER

We approach the task of student classroom behavior recognition as an object detection problem and utilize StB-YOLO to handle this task. The StB-YOLO model, as shown in Fig. 3, consists of a backbone, neck, and head. We employ the backbone network of YOLOv5 to extract image features. The P2 layer is also integrated

into the neck, aiming to enhance the detection performance for multi-scale objects. We incorporate the LW-SEAM into the neck to enhance the model’s occlusion detection capability. As studied in [21], the features focused on during training differ between classification and regression tasks, and the utilization of separate branches for computation favors performance enhancement. Therefore, a decoupled head is utilized to independently predict the position and classification of targets.

Double-SlideLoss. Binary Cross-Entropy loss has lower recognition accuracy for minority behaviors such as lying on the desk or using a mobile phone. Additionally, some behaviors appear visually similar and prone to confusion. Therefore, this paper automatically assigns weights to hard samples by SlideLoss [22]. The weighting function of SlideLoss can be expressed as Eq. 1. SlideLoss views the distinction between easy and hard samples as the IoU (Intersection over Union) size. However, to address the issue of high IoU yet misclassification, we designed a modified version of SlideLoss specifically for classification losses, named ClsSlideLoss. Inspired by the design of SlideLoss, ClsSlideLoss utilizes the classification accuracy of all prediction boxes as the threshold β , considering samples with accuracies below β as negative and above β as positive. The weighting function of ClsSlideLoss can be expressed as Eq. 2:

$$F(x) = \begin{cases} 1, & \text{if } x \leq \mu - 0.1 \\ e^{1-\mu}, & \text{if } \mu - 0.1 < x < \mu \\ e^{1-x}, & \text{if } x \geq \mu \end{cases} \quad (1)$$

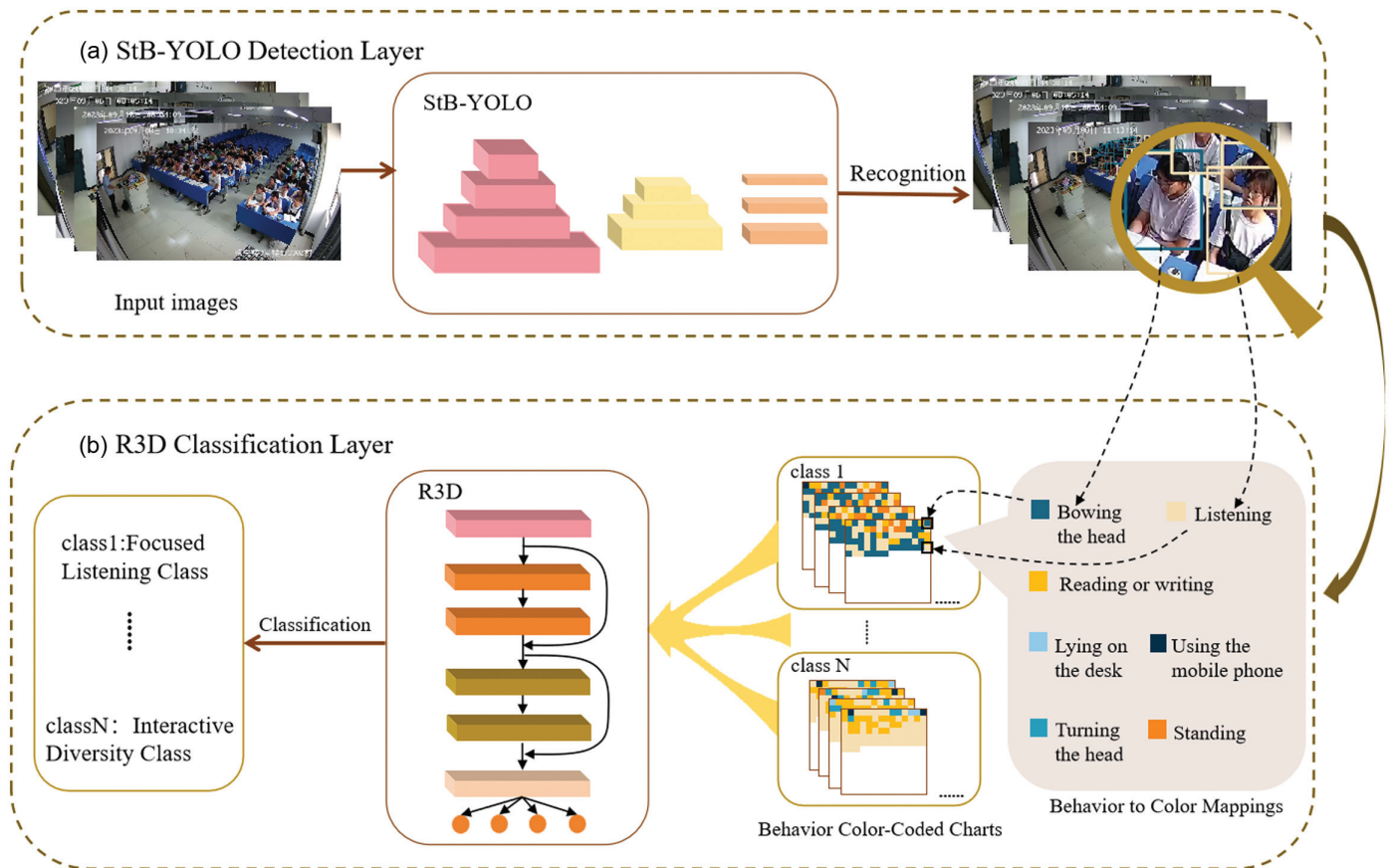


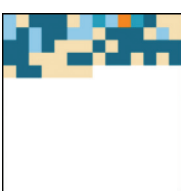



Fig. 2. Model for student behavior and dynamic class portraits using YOLO and R3D (SBCP-YOLO-R3D). It consists of two parts, (a) Stab-YOLO Detection Layer and (b) R3D classification layer.

Table I. Class portraits

Category	Detailed description	Example
Focused Listening Class	The majority of students are fully attentive during the lecture, while some are engaged in reading or taking notes. It is rare or absent for other classroom behaviors	
Interactive Diversity Class	A proportion of students is fully attentive during the lecture, while another proportion is looking down. Other classroom behaviors are rarely observed or absent.	
Free Activity Class	The majority of students are looking down, listening to the lecture, and using mobile phones or resting their heads on the desks. Only a handful of students are taking notes or reading.	
Silent Learning Class	The vast majority of students are looking down, with a small percentage of students using mobile phones, resting their heads on the desks, or reading and writing. Other classroom behaviors are rarely observed or absent.	

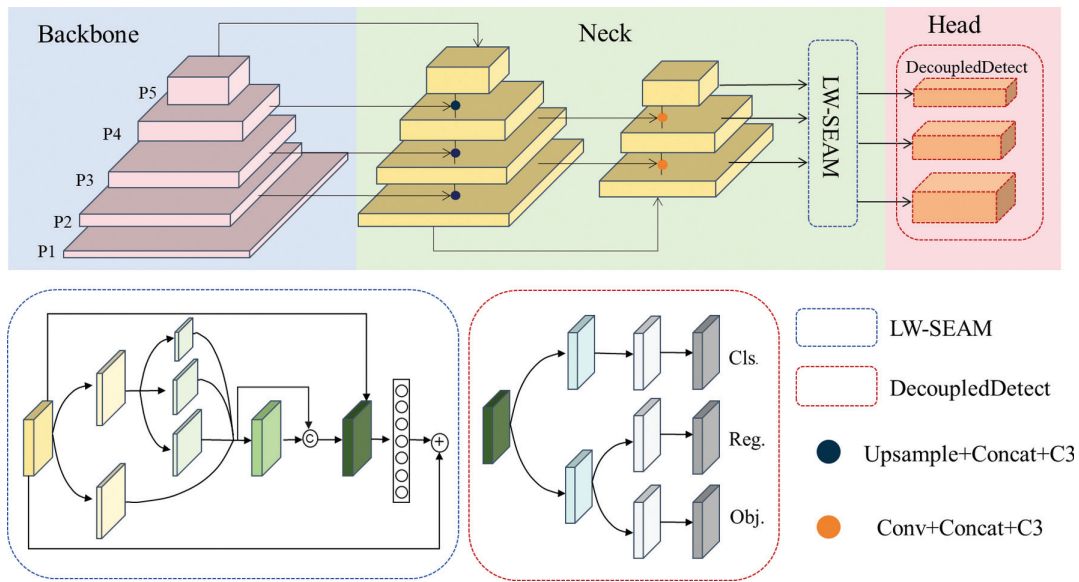


Fig. 3. An overview of the proposed StB-YOLO framework. The first row illustrates the StB-YOLO model, while the second row depicts the individual structures of the LW-SEAM and the decoupled head.

$$F(x) = \begin{cases} e^{1-\beta}, & \text{if } x < \beta \\ e^{1-x}, & \text{if } x \geq \beta \end{cases} \quad (2)$$

Lightweight Occlusion Attention. The structure of LW-SEAM is illustrated in the upper left side of Fig. 4. It initially comprises three Channel and Spatial Mixing Modules (CSMMs) with different

kernel sizes. The Patch Embedding Module (PEM) within CSMM extracts multi-scale features through convolutions with three different kernel sizes. However, large convolution kernels introduce a significant number of parameters. Thus, PEM initially utilizes 1×1 convolution kernels to adjust the input feature channels, resulting in two feature maps, where the height and width remain unchanged

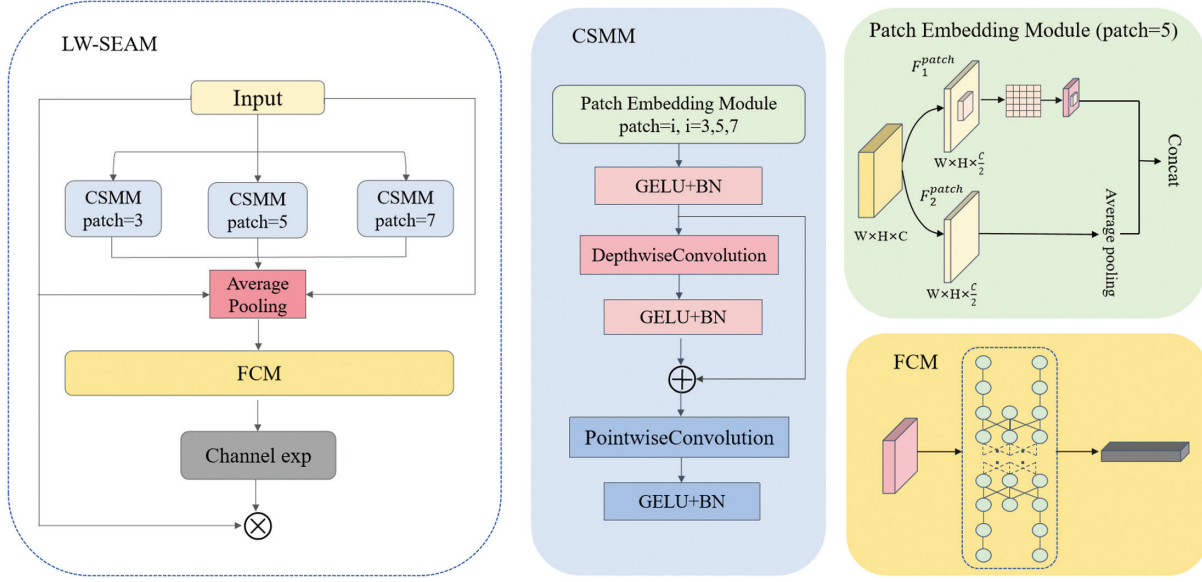


Fig. 4. Overview of LW-SEAM. The left part is the overall structure of LW-SEAM, and the middle part is the CMSS. The right part represents lightweight multi-scale feature extraction and fully connected module (FCM).

while the number of channels is reduced to half. This uniform multi-branch structure reduces parameters while effectively preventing overfitting. Subsequently, multi-scale features are obtained and average-pooled to the same size. Finally, the two feature maps are concatenated along the channel dimension. The computation of PEM is as follows:

$$PEM_{patch}(F) = f^{patch \times patch}(f^{1 \times 1}(F) \oplus AP(f^{1 \times 1}(F))) \quad (3)$$

where $F \in R^{C \times H \times W}$ denotes input feature map, $f^{patch \times patch}$ denotes the convolution operation with a kernel size of $patch \times patch$, AP denotes average pooling, and \oplus denotes concat.

After PEM, depthwise separable convolutions are employed to extract features. While depthwise separable convolutions can learn the importance of different channels and reduce the number of parameters, it neglects the information relationships between channels. To compensate for this loss, pointwise convolutions are subsequently used, and they constitute the entire CMSS.

Subsequently, average pooling is applied to capture global contextual information in the spatial dimension. A two-layer fully connected network is used to fuse the information of each channel, enabling the network to strengthen the connections between channels. The channel exp is employed to expand the range from $[0,1]$ to $[1,e]$ to make the results more tolerant of positional errors. Finally, the output is multiplied by the original features as attention coefficients.

B. R3D CLASSIFICATION LAYER

The R3D incorporates the strengths of residual learning and 3D CNNs, enabling efficient processing of video data and extraction of both temporal and spatial features.

The BCCC, which transforms student behaviors into visual indicator charts, is processed using R3D to acquire class portraits. R3D can not only extract behavioral features of all students spatially but also capture their changes over time.

C. ASSUMPTIONS AND PROBLEM FORMULAS

Given a class with m students, let $S = \{S_1, S_2, \dots, S_m\}$ represent the students in this class, and given a lesson with n key frames, let $I = \{I_1, I_2, \dots, I_n\}$ represent the set of n images. Let C be the set of z types of student behaviors, then $C = \{C_1, C_2, \dots, C_z\}$. Assuming the correct behavior of student S_i among the image I_j is denoted by y_{ij} , whereas the model's predicted behavior is denoted by \hat{y}_{ij} . Define $G_{i,j,k}$ and $\hat{G}_{i,j,k}$ as a Boolean variable that are used to indicate whether the answers to and correspond to behavior, respectively. They are represented by Eq. 4 and Eq. 5:

$$G_{i,j,k} = \begin{cases} 1, & \text{if } y_{ij} = C_k \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

$$\hat{G}_{i,j,k} = \begin{cases} 1, & \text{if } \hat{y}_{ij} = C_k \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Based on this, the prediction results of S_i on image I_j are $TP_{i,j,k}$, $TN_{i,j,k}$, $FP_{i,j,k}$, $FN_{i,j,k}$. Their formula is presented as follows:

$$TP_{i,j,k} = \begin{cases} 1, & \text{if } \hat{G}_{i,j,k} \cdot G_{i,j,k} = 1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$TN_{i,j,k} = \begin{cases} 1, & \text{if } (1 - \hat{G}_{i,j,k}) \cdot (1 - G_{i,j,k}) = 1 \\ G & \text{otherwise} \end{cases} \quad (7)$$

$$FP_{i,j,k} = \begin{cases} 1, & \text{if } \hat{G}_{i,j,k} \cdot (1 - G_{i,j,k}) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

$$FN_{i,j,k} = \begin{cases} 1, & \text{if } (1 - \hat{G}_{i,j,k}) \cdot G_{i,j,k} = 1 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Therefore, TP are represented by Eq. 10. Similarly, for TN, FP, and FN:

$$TP = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^z TP_{i,j,k} \quad (10)$$

To validate and compare the performance of our model, we use Precision (P), Recall (R), Mean Average Precision (mAP), Frames Per Second (FPS), and Params as evaluation metrics. The definitions of Precision, Recall, mAP, and F1 are as follows:

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$AP = \int_0^1 P(R) dR \quad (13)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (14)$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (15)$$

FPS measures the inference speed of a model and is an essential parameter for assessing the real-time performance of an algorithm. The number of parameters affects model complexity, generalization ability, and training speed.

IV. EXPERIMENTS

This section provides details on the StuAct dataset and evaluates our model on SCB-DATASET3-S and StuAct using various metrics. Subsequently, we compare and analyze the performance of the R3D and ViVit algorithms in the Dynamic Class Portraits task.

A. StuAct DATASET

StuAct dataset has been collected from surveillance videos of 33 classes at University C, with a student population ranging from 40 to 120 per course. These data cover four grades and eight schools across the university, totaling 771 hours. The raw surveillance videos downloaded from the academic administration system, with data privacy compliance, contain both in-class and pre-class/post-class periods. To extract frames with differentiation and filter out data outside the in-class periods, we employ a frame difference method based on local maxima to extract key frames from the videos. The specific steps are as follows: (1) for each video, we

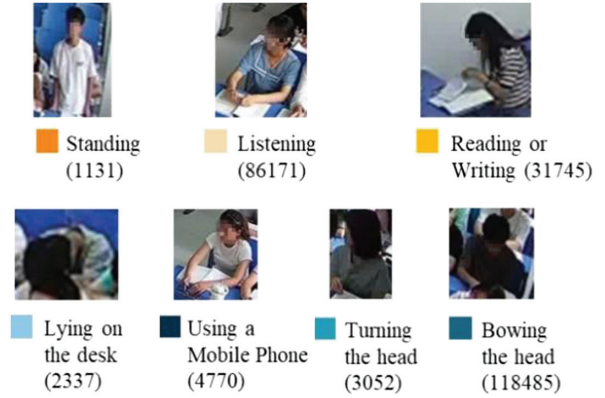


Fig. 5. Examples and quantities of behaviors. The numbers in brackets are the number of samples, and the color of each behavior visualization is below the image.

calculate the frame-to-frame difference every 24 frames and plot the variation of this difference over time. (2) We identify the range of frame-to-frame differences during in-class periods and compute the average. (3) Frames with differences below the average are filtered out. Finally, the StuAct dataset contains 4,721 images, each with a size of 1920×1080 pixels. We classify the student's classroom behaviors into seven categories. They are standing, reading/writing, using the mobile phone, bowing the head, lying on the desks, listening and turning the head. As shown in Fig. 5, the numbers in brackets are the number of samples and the color of each behavior visualization is below the image. All students involved in the video surveillance used in this experiment have been informed and have given their consent for their images to be used for research purposes. Furthermore, all images presented in the paper have undergone facial blurring to safeguard privacy.

B. EXPERIMENTAL RESULTS ON StuAct and SCB DATASETS

To validate the contribution of the Double-SlideLoss and LW-SEAM proposed in our paper for improving model performance, we used YOLOv5s as the baseline and applied the Double-SlideLoss and LW-SEAM to its network structure. The results can be seen in Table II. The SCB-DATASET3-S dataset does not exhibit significant data imbalance issues; thus, the application of Double-SlideLoss yields nearly identical results compared to the baseline. LW-SEAM results in an increase in mAP by 0.8%. On the StuAct dataset, after employing Double-SlideLoss, the P increases by 3.1%, the R increases by 1%, and the mAP improved by 0.8%. With the inclusion of LW-SEAM, the R increases by 0.5% and the

Table II. The comparison before and after adding different modules

Datasets	Baseline	Double-SlideLoss	LW-SEAM	P	R	mAP_50
SCB-DATASET3-S [4]	✓			73.3	67.8	72.9
	✓	✓		72.5	68.1	73
	✓		✓	71.3	70	73.7
StuAct	✓			67.2	61.7	63.2
	✓	✓		70.3	62.7	64
	✓		✓	70	62.2	63.9

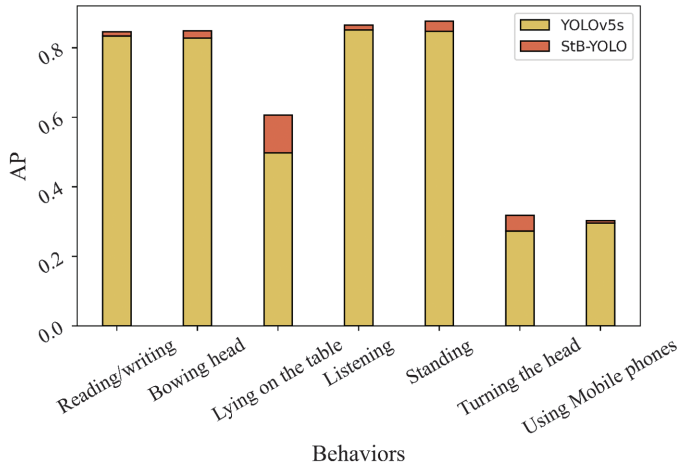
Table III. AP values of the before and after adding different modules in StuAct dataset

Model	Reading/ writing	Using mobile phone	Bowing head	Lying on the desk	Listening	Turning the head	Standing
Baseline	83.4	29.6	82.8	49.8	85.1	27.3	84.7
+ Double-SlideLoss	83.2	34.2	82.6	53	85	26.7	84
+ LW-SEAM	82.8	32.6	82.6	55.1	84.9	24.6	84.7
+ Double-SlideLoss & LW-SEAM	83.8	33.5	82.8	53.5	85.3	24.7	86.6

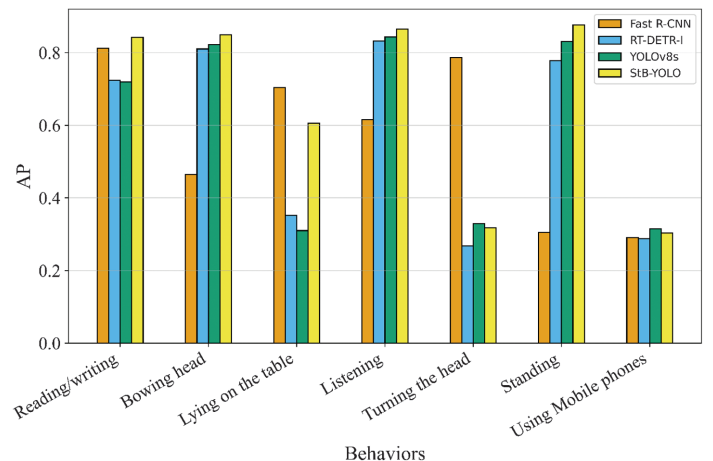
Table IV. Comparisons study on StuAct dataset

Model	Backbone	Input size	mAP_50 (%)	F1	Params (M)	FPS	GFLOPs
Faster R-CNN	ResNet50	600 × 600	56.87	0.56	136.812	10.188	\
RT-DETR-l	\	640 × 640	57.9	0.62	31.998	36.453	108.0
YOLOv8s	\	640 × 640	60.6	0.60	11.128	98.847	28.7
YOLOv8m	\	640 × 640	62.1	0.61	25.858	87.232	79.1
YOLOv8l	\	640 × 640	64.1	0.64	43.612	69.754	164.8
YOLOv5s	\	640 × 640	63.2	0.62	7.029	88.960	15.8
YOLOv5m	\	640 × 640	61.3	0.60	20.877	93.975	47.9
YOLOv5l	\	640 × 640	62.6	0.62	46.14	67.469	107.7
StB-YOLO	\	640 × 640	66.3	0.65	11.311	77.019	41.8

P value increases by 2.8%, and the mAP improves by 0.7%. Table III presents the AP for each behavior category. As evident from the table, the recognition accuracy is relatively high for categories such as reading/writing (83.4%), bowing head (82.8%), listening (85.1%), and standing (84.7%). However, the recognition accuracy is notably lower for using the mobile phone (29.6%), lying on the desk (49.8%), and turning the head (27.3%). This can be attributed to two primary factors: first, these three behaviors are considered as abnormal in the classroom context, with fewer samples compared to other categories. Second, using a phone and lying on the table exhibit high similarity to bowing the head, while turning the head shares similar characteristics with listening, resulting in the model's poor performance in identifying these samples. After using Double-Slide Loss and LW-SEAM, the AP increases significantly.

**Fig. 6.** The AP values of seven categories of baseline and the proposed approach in StuAct dataset.

We conduct a comparative analysis between StB-YOLO and several object detection methods, including RT-DETR [23], YOLOv8 [24], YOLOv5, and Faster-R-CNN [5]. The results of comparison on the StuAct dataset are presented in Table IV. StB-YOLO introduces only a small increase in parameters compared to YOLOv8s, while the mAP is much higher than YOLOv8s. The AP values of each behavior for the proposed model are compared with those of the baseline. Fig. 6 illustrates the comparison of performance between the baseline model and our approach across various categories. As shown in Fig. 6, our model achieves higher AP values for each behavior category compared to the baseline model, with significant improvements in the categories of lying on the desk, turning the head, and standing up to answer questions. Figure 7 provides a detailed exhibition of the performance of each algorithm within each category. According to Fig. 7,

**Fig. 7.** The AP values for each behavior of models in StuAct dataset.

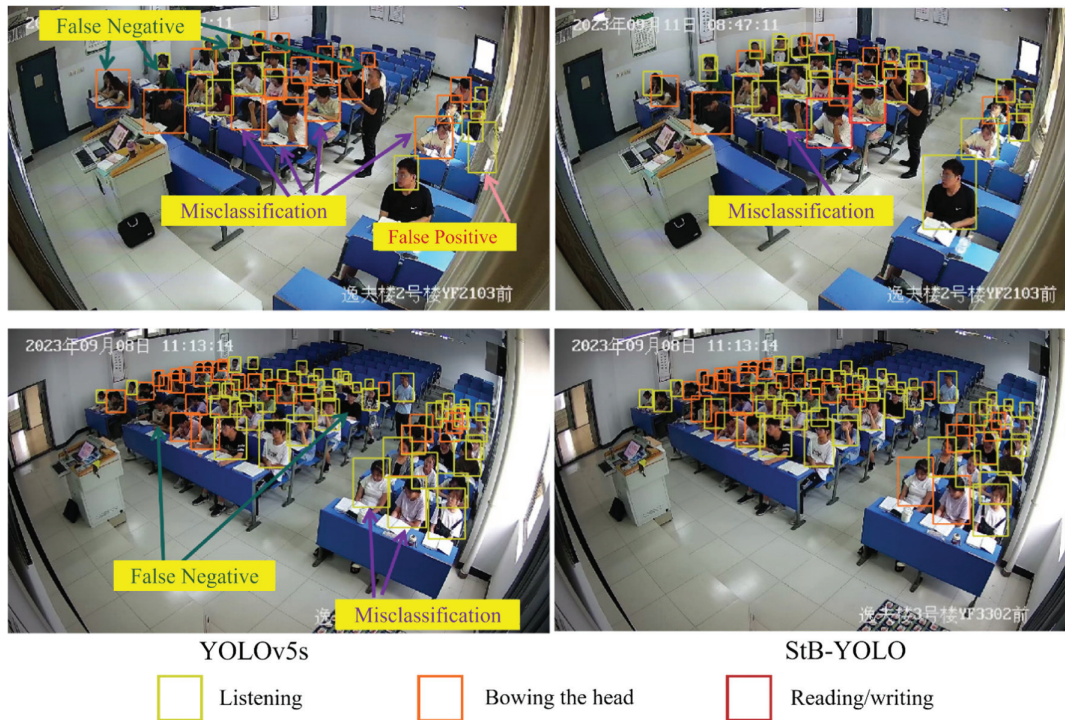


Fig. 8. The comparison between YOLOv5 and our method.

Fast R-CNN demonstrates superior performance in turning the head and lying on the desk category compared to other algorithms.

Figure 8 visually demonstrates the student behavior recognition results of YOLOv5 and our model. As can be observed from the figure, YOLOv5 tends to misclassify background and misclassifies challenging behavior samples. StB-YOLO achieves superior performance in multi-scale detection tasks. Although the behaviors of bowing the head and reading/writing are highly similar, ours can distinguish between them more accurately. In addition, StB-YOLO demonstrates stronger anti-interference capabilities in complex backgrounds, focusing its attention on the foreground regions.

C. EXPERIMENTAL RESULTS ON CLASS PORTRAITS

Classroom portraits can assist teachers in quickly comprehending classroom dynamics. To this end, we represent students’ in-class behaviors as 8×8 pixel patches, where each patch’s color corresponds to a specific behavior. We then depict the behaviors of all students in a given frame onto a 112×112 pixel image, which we term the “Behavior Color-Coded Charts (BCCC).” Analogously, time-series-based BCCC for the entire class session can be obtained. This paper utilizes the R3D and ViViT [25] for classroom portraits. As seen in Fig. 9, the R3D algorithm achieves an accuracy of over 85%, significantly outperforming ViViT with a lower loss.

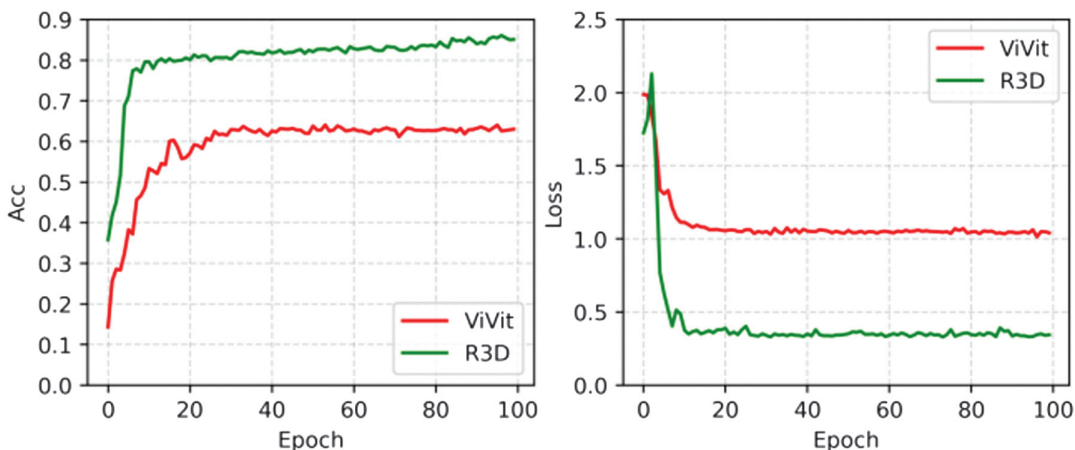


Fig. 9. Accuracy and loss in R3D and ViViT.



Fig. 10. Visualization of student behavior and classroom status portrait.

D. RESULT ON SBCP-YOLO-R3D

Figure 10 shows the outputs of SBCP-YOLO-R3D, which include visualizations of student behaviors and class portraits. The input of SBCP-YOLO-R3D can be a video or a set of images. Teachers can quickly acquire information about student behaviors and classroom status.

V. CONCLUSION

In this paper, we introduced the StuAct dataset. We delved into solutions for identifying hard samples first. Second, we employed LW-SEAM for occluded feature extraction modifications. Finally, we developed a scheme for Student Behavior Recognition and Dynamic Class Portrait Construction using the StB-YOLO and R3D.

Despite this study having yielded some conclusions, there are still limitations. Our future research on student classroom behavior recognition and learning engagement focuses on the detection effect of abnormal behaviors.

DATA AVAILABILITY

StuAct dataset is available on request from the corresponding author.

FUNDING

The work was supported by science and technology innovation 2030—major project of “New Generation Artificial Intelligence” (2022ZD0115905), the Key Project of Anhui Provincial Scientific Research Planning Project (No. 2022AH040153) and Anhui University of Science and Technology Graduate Innovation Projects (2023cx2130).

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

- [1] J. Zhao, H. Zhu, and L. Niu, “Bitnet: a lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network,” *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 8, p. 101670, 2023.
- [2] Y. Li, X. Qi, A. K. J. Saudagar, A. M. Badshah, K. Muhammad, and S. Liu, “Student behavior recognition for interaction detection in the classroom environment,” *Image Vis. Comput.*, vol. 136, p. 104726, 2023.
- [3] H. Zhou, F. Jiang, J. Si, L. Xiong, and H. Lu, “Stuart: Individualized classroom observation of students with automatic behavior recognition and tracking,” in *ICASSP 2023–2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, IEEE, Piscataway, NJ, 2023, pp. 1–5.
- [4] F. Yang and T. Wang, “Scb-dataset3: A benchmark for detecting student classroom behavior,” *ArXiv*, vol. abs/2310.02522, 2023.
- [5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Eur. Conf. Comput. Vis.*, Springer, Berlin, German, 2020, pp. 213–229.
- [7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: deformable transformers for end-to-end object detection,” *arXiv preprint arXiv:2010.04159*, 2020.
- [8] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series in 2021,” *arXiv preprint arXiv:2107.08430*, 2021.

- [9] S. Zuo, et al., "Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 258, 2022. DOI: [10.1016/j.knosys.2022.109978](https://doi.org/10.1016/j.knosys.2022.109978).
- [10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, IEEE, Piscataway, NJ, 2021, pp. 10012–10022.
- [11] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: vision transformer with bi-level routing attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, IEEE Computer Society, Los Alamitos, CA, 2023, pp. 10323–10333.
- [12] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," *ArXiv*, vol. abs/2112.05561, 2021.
- [13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, IEEE Computer Society, Los Alamitos, CA, 2018, pp. 7794–7803.
- [14] S. Zhang, D. Chen, J. Yang, and B. Schiele, "Guided attention in CNNs for occluded pedestrian detection and re-identification," *Int. J. Comput. Vis.*, vol. 129, pp. 1875–1892, 2021.
- [15] J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3872–3884, 2020.
- [16] 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, pp. 5314, 2021.
- [17] C. Pabba, V. Bhardwaj, and P. Kumar, "A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 36 975–37 005, 2024.
- [18] M. Rashmi, T. Ashwin, and R. M. R. Guddeti, "Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 2907–2929, 2021.
- [19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [20] C. C. Yin Albert, et al., "Identifying and monitoring students' classroom learning behavior based on multisource information," *Mobile Inf. Syst.*, vol. 2022, pp. 1–8, 2022.
- [21] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, IEEE Computer Society, Los Alamitos, CA, 2020, pp. 11563–11572.
- [22] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang, "Yolo-facev2: a scale and occlusion aware face detector," *Pattern Recogn.*, vol. 155, p. 110714, 2024.
- [23] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, IEEE Computer Society, Los Alamitos, CA, 2024, pp. 16965–16974.
- [24] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, IEEE, Piscataway, NJ, 2024, pp. 1–6.
- [25] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: a video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, IEEE, Piscataway, NJ, 2021, pp. 6836–6846.