RESEARCH ARTICLE

BON VIEW PUBLISHING

SiT-YOLOv9: An Efficient Algorithm for Learning Behavior Detection in the Home Environment

Zhendan Sun[1,*] and Vladimir Y. Mariano[1]

[1]College of Computing and Information Technologies, National University, Philippines

Abstract: In the context of home-based learning, accurate identification of learning behaviors is essential for enhancing post-classroom learning efficiency. However, due to background interference and computational constraints in the TinyML terminal within home environments, CNN-based algorithms are susceptible to reduced performance and accuracy, leading to an increased false positive rate. To address this challenge, we propose a lightweight detection model called SiT-YOLOv9, which integrates MODNet, image enhancement, and other modules into the YOLOv9 model while also implementing moderate network pruning to effectively mitigate issues related to image noise and training sample computational power. Evaluation of the SiTBehaviors video dataset demonstrates that the SiT-YOLOv9 model achieves outstanding performance with a recognition accuracy of 0.948 (mAP50) at a high processing speed of 90.9 frames per second. When compared with original models such as YOLOv8, YOLOv10, and RT-DERT, our proposed model exhibits superior recognition accuracy of 0.948 mAP, and a processing speed of 0.2 ms.

Keywords: SiT-YOLOv9, SiTBehaviors dataset, home environment, learning behavior recognition, image enhancement

1. Introduction

To improve the learning quality and after-school learning efficiency of students in the compulsory education stage, the Chinese government introduced the "double reduction" policy in March 2022. This policy aims to reduce the high educational costs of extracurricular tutoring for students. It applies to all educational institutions in China [1]. In the requirements, the quality and effect of homework after class are proposed as important assessment requirements, which strongly causes anxiety among 160 million primary and secondary school students and their parents [2].

However, the home learning environment presents unique challenges that differ from the structured classroom setting. A classroom is a dedicated learning space that fosters concentration, aided by teacher supervision and peer presence, making it easier for students to focus on completing their homework [3]. In contrast, students in a home learning environment often struggle with distractions due to their limited self-management skills and motivation for tasks such as homework, review, and online learning. The reasons for these distractions include the temptation of digital devices like smartphones and social media [4, 5], interruptions from family activities or background noise, insufficient lighting or noisy environments, personal mood swings or health issues, unengaging or difficult learning materials, and physical states like hunger or fatigue [6].

Many studies have shown that distraction behavior has a considerable adverse effect on learning [7–11], resulting in an inability to concentrate and a decrease in learning efficiency. It will not only affect the effect of preview, review, and test after class but also squeeze out students' rest and extracurricular activities, leading to the accumulation of homework and a decline in academic performance [12], causing a chain reaction such as increased dependence on supplementary lessons and the rising cost of family education investment. In the family environment, the fine monitoring of students' learning behavior can provide an important data basis for personalized education support and intervention measures [12]. This can not only enhance students' self-management ability but also help them adjust their learning habits, thus improving their learning efficiency [13]. Therefore, to improve the accuracy and real-time performance of after-class learning behavior recognition monitoring to a new level, and make it move toward practical application, is the key problem we need to solve.

Most of the existing research focuses on classroom learning behavior recognition, Bui et al. [14] use machine vision algorithms such as image segmentation, feature extraction, object detection, and motion analysis to recognize learning behaviors by analyzing faces, and expressions [15], emotions and gestures. In terms of algorithms, Convolutional Neural Networks (CNNS) have baseline models that cover a variety of networks, For example, pyramid network, ACAM, MOC, BSN, DBG, BiFPN, HBR, and SOTA [16–20], especially the improved YOLO series networks, are popular due to their efficient real-time object detection capabilities. In addition, some researchers are using advanced algorithms such as Transformer and LLM to continuously improve the accuracy of recognition and the depth of analysis, which has become a new technology development trend.

_____
*Corresponding author: Zhendan Sun, College of Computing and Information Technologies, National University, Philippines. Email: sunz@students.national-u.edu.ph

In recent years, the number of datasets in the field of classroom behavior recognition has increased, especially in the field of classroom activities of teachers and students. For example, Hussein et al published the CK+ dataset which contains five categories of student expressions: surprised, tired, confused, happy, and neutral [21]. EduNet, on the other hand, focuses on classroom identification of teacher and student activities, including 20 different types of behaviors [22]. The BNU-LCSAD dataset records the behaviors of students in the classroom, such as raising their hands, reading, writing, using the phone, bowing their heads, and bending over [23]. SCB-Dataset provides important data support for the identification of distracted and negative behaviors between teachers and students in class [24]. These datasets provide rich video samples, which are of great value for learning behavior analysis. In contrast, there is no dataset for learning behavior recognition in the home environment.

In recent years, object detection algorithms based on CNN, KNN, CPM VGG-16, 3DCNN and LSTM have achieved significant results in classroom learning behavior recognition [25–28]. The target detection algorithm has achieved remarkable results in classroom learning behavior recognition. This is due to the fixed nature of the camera position in the classroom environment, the neat nature of the classroom seat layout, adequate lighting conditions, uniform uniforms, and correct sitting posture. In addition, the unimpeded school network, the reorganization of the data center calculation cases, and the easy operation of the above large algorithms. However, when these algorithms are applied to variable and complex home learning environments, their effectiveness is greatly reduced. This is mainly because the learning behavior is significantly different between the classroom and the family environments.

Firstly, in a home learning environment, it is necessary to detect and recognize a variety of complex behaviors such as sleeping, using a smartphone, reading, writing, touching one's hair, drinking, eating, holding a pen, staring at or adjusting glasses [29]. In contrast, classroom environments typically evaluate student focus by identifying behaviors like raising hands, turning heads, whispering, and standing up as shown in Figure 1 [24].

Secondly, the angles of behavior collection differ significantly between home and classroom settings. A home learning environment usually employs a single front-facing camera, capturing the learner's upper body movements and facial expressions. In classrooms, however, multiple cameras (front, rear, and side) are used to focus on the body movements of the entire student group.

Moreover, the diversity of home environments is reflected in varying clothing and background settings, whereas classroom environments are relatively uniform in terms of background and attire, facilitating more standardized behavior recognition.

Lastly, the instability of home internet connections and the privacy concerns associated with long-term fixed cameras make data collection in home environments more difficult, often relying on edge computing. In classroom environments, data collection is easier and more controlled, with remote computing based on centralized data centers being more convenient.

Therefore, designing a behavior recognition model suitable for home learning environments requires optimizing the model's structure to reduce complexity, allowing it to function on devices with lower computational power. Additionally, given the limited data available from home environments, the model must also exhibit higher accuracy, robustness, real-time performance, and faster processing speeds to ensure effective application.

In response to the above problems, this study uses matting, image enhancement, and YOLOv9 technology to develop a high real-time home learning behavior recognition model. The model effectively overcomes background noise interference in the family environment and has smaller parameters than similar models. Through a small amount of dataset training, it can get better recognition and understanding of learning behavior accuracy, performance, and robustness. The main contributions of this study are as follows:

1) In this study, a new lightweight detection model, SiT-YOLOv9, is developed.

Figure 1
Example of learning behavior in a home environment

front

computer classes

kindergarten

side

english classes

elementary school

back
I shooting angles

art classes
II different classes

middle school
III different learning stages

2) A 70Gb family learning behavior dataset SiTBehavior was constructed, covering 10 types of behaviors from notebooks, pads, and mobile phones.

3) A series of excellent models such as SiT-YOLOv9 and YOLOv9 original model, YOLOv8, CNN, and ResNet are evaluated, including key performance indicators such as parameter scale, running speed, and recognition accuracy.

4) Discussion of limitations and related recommendations for applying the method in home Settings.

2. Literature Review

In the past five years, many researchers have paid attention to issues related to classroom learning behavior recognition, mainly focusing on recognition methods and anti-interference research.

2.1. Recognition methods

In 2020, Zhu and Chen [30] innovationally proposed an automatic learning expression recognition method to reduce feature redundancy. By using the global shallow texture features with dimensionality reduction and the deep feature extraction technology of CNN, support vector machine (SVM) was used as the classifier. It shows good results on the self-built expression database CK database and the Chinese emotional picture system. Wang et al. [31] proposed a method of combining face and hand key point detection in multi-person scenes. This method uses OPENPOSE's dual keypoint detection technology to remove the fully connected layer and optimizes face detection by depth-wise separable convolution and SSD technology. They borrowed the design concept of reducing parameters and computing time and reconstructed the network architecture for real-time expression recognition, which fused the advanced techniques of Convolutional pose machine (CPM) and CMU. The architecture was composed of four layers of depth-wise separable convolution (DSC), each convolution layer was equipped with batch normalization and ReLU activation function, while the average pooling layer and softmax function were used. After a series of optimization, the number of parameters of the final model is reduced to nearly 60,000, which is 80 times less than that of the original CNN model, which greatly reduces the number of parameters and computing time, and provides an efficient solution for real-time behavior detection and expression recognition, The proposed scheme combines pose estimation and human detection technology to reduce false connections in skeleton data in complex situations. The experimental results show that compared with the skeleton-based scheme, the proposed scheme improves the average precision of the system by 15.15%, and the average recall rate by 12.15%. In addition, Wang et al. [32] integrated the feature pyramid into R-FCN, used ellipses to fit the degree of mouth opening and opening to refine the detection results, and solved the problem of yawning gesture detection. Experiments show that the mAP50 accuracy of our method reaches 0.90 Furthermore, to reduce the required computational resources, we pruned the channels without harming the performance, further reduced the number of network parameters, and improved the detection speed by a factor of 2.

Another branch is effectively supervised and managed classroom learning state recognition of students. Lu et al. [33] proposed an assistant teaching management system to monitor and manage students' learning status in class, especially to detect abnormal behaviors such as playing on mobile phone and sleeping. The core of the system uses the transfer learning technology of the VGG pre-trained network model to extract the characteristics of

student's abnormal classroom behaviors. The test results show that the average recognition accuracy of the system reaches 85.2775%, specifically, the recognition accuracy of sleeping normal and playing mobile phone behaviors is respectively 95.1510%,90.5490%, 70.1334%. In addition, this method can use background subtraction technology to identify multiple targets in the video and determine the abnormal behavior of students through continuous identification and statistical analysis, and finally, automatically generate classroom behavior analysis reports. Studies show that this method has a significant effect in extracting features, is suitable for complex environments, and has high applicability and strong anti-interference. In 2021, Chen et al. [34] proposed to use the intelligent detection double YOLO algorithm to identify the classroom learning state of students. The algorithm combines K-means clustering to analyze the target candidate box and improves the efficiency of classroom teaching evaluation and teaching management through result scoring and grading. The accuracy of the method reaches 86.62%, and the processing time of each frame of the image is only 0.2 seconds. Wang et al. [31] proposed an innovative method in 2022 to identify students' learning attitudes by analyzing classroom videos and detecting behaviors such as listening, turning, bowing, and raising hands in real time. This method combines the YOLOv3 and Openpose algorithms to extract global and local features of the human body, and the verification accuracy is as high as 95.45%. Lu et al. [33] also used the YOLOv3 deep learning method in 2022 and achieved 87.33% recognition accuracy in an English classroom environment.

2.2. Research on anti-interference

When learning behavior recognition is performed in a classroom environment, the two major challenges are facial occlusion between students and the recognition of small objects in images. These problems not only affect the accuracy of recognition but may also weaken the robustness of the system.

The non-controlled scene face recognition algorithm developed by Chen et al. [34] uses principal component sparse representation and low-rank decomposition algorithm to process photos. Experiments show that the algorithm can effectively deal with illumination changes and occlusion, and the accuracy of face recognition is as high as 92.4%. Liu et al. [35] designed the Rs-YOLOv3 model, which improved the capture ability of feature information and the efficiency of multi-layer multiplexing by adding RFB and SE-Res2net modules. The DIoU_Loss technology is used to improve the positioning accuracy and effectively improve the ability to identify students' abnormal behaviors in complex classroom environments. Experimental results show that these improvements make the accuracy of Rs-YOLOv3 reach 80.1%, which is 5.8% higher than that of traditional YOLOv3, and effectively reduces the misjudgment of recognition. Yu et al. [36] also innovated a student expression recognition model and intelligent teaching evaluation algorithm based on a deep attention mechanism. Their method uses the clipping and occlusion strategy to generate multi-channel face images, establishes a multi-channel deep attention network, and uses the self-attention mechanism to assign different weights. The loss function was constrained to optimize the weight allocation to realize student expression recognition under occlusion conditions. In addition, the algorithm also integrates the teaching evaluation algorithm of students 'expression and behavior state. The experimental results on FERplus public dataset and self-built classroom teaching video dataset show that the model achieves 87.34% accuracy in student expression recognition. The algorithm

proposed by Huang et al. [37] obtains the image stream of students through real-time object detection and tracking technology and uses the deep spatio-temporal residual convolutional neural network to analyze the spatio-temporal characteristics of student behavior, which effectively reduces the facial occlusion problem in the classroom. Experimental results on the classroom teaching video dataset show that the proposed model can achieve an accuracy of 88.5% for classroom behavior recognition. Longyu Tang et al. applied multi-scale feature fusion and attention mechanism to classroom behavior detection, improved YOLOv5 network, fused feature pyramid structure and weighted bidirectional feature pyramid network, and introduced spatial and channel convolutional attention mechanism to reduce facial feature loss. It greatly enhances the ability of the model to detect various behavior states of students. The experimental results show that the improved algorithm shows excellent detection effects in listening, bending, lying down, and standing scenes, with an average accuracy of 89.8% and a recall rate of 90.4%. Chen et al. [38] improved the classroom detection model based on YOLOv8. By combining Res2Net and YOLOv8 network, the C2f_Res2block module is innovatively proposed, and combined with MHSA and EMA technology, it is integrated into the YOLOv8 model. The experimental results show that compared with the original YOLOv8 model, the detection performance of the improved model is significantly improved. The average precision (mAP@0.5) is improved by 4.2%.

Wang et al. [39] proposed an algorithm to improve the accuracy of small object detection. It combines deformable DETR, Swin Transformer, and lightweight Feature Pyramid Network (FPN). The network can effectively process the multi-scale feature maps extracted by Swin Transformer, thus capturing more nuanced behavioral features. In addition, the CARAFE lightweight operator is integrated into the FPN structure to improve the detection accuracy. Experiments on real student classroom behavior datasets show that the proposed method improves the detection accuracy by 6.1% compared with the current most advanced method. Zhao et al. [40] developed BiTNet, a detection network for identifying objects with small head size. The network specially designed a small detection head to identify small-size objects, and effectively fused semantic and location information through a Bidirectional Feature Pyramid Network (BiFPN). At the same time, BiTNet integrates Efficient Converter Block (ETB) and Efficient Convolutional Aggregation Block (ECAB) and uses the self-attention mechanism to optimize the recognition ability of occluded objects. In addition, by combining the homogeneous multi-branch design with ECAB's single aggregation operation, BiTNet not only improves the recognition accuracy but also greatly reduces the computational cost. Experimental results on the classroom behavior dataset and Pascal VOC 2007 and VisDrone2021 confirm that BiTNet has good generalization ability and small object recognition ability.

Image interference factors, such as low-resolution images, background clutter, light, and Angle of view change, are important factors leading to the increase of misjudgment rate. A method of head behavior recognition based on depth map is proposed. It constructs a face model using 3D coordinate key points, and extracts features from the dynamic relationship between adjacent frames. Through depth information and a support vector machine (SVM) classifier, the proposed method effectively dealt with background clutter and illumination changes [40]. Experimental results show that the proposed outlier feature representation method shows the ability to effectively identify head behavior and the superiority of resisting the interference of background clutter and illumination change on the public dataset and the real behavior data of the classroom environment collected by Kinect. The anti-interference

features fusion network developed by Jisi and Yin [25] combines a spatial affine transformation network and convolutional neural network to deeply mine detailed features in the feature extraction stage. In classification, the spatio-temporal features are fused by weighted summation, and the softmax classifier is optimized, which significantly improves the accuracy of classification. In the experiments on the UCF101 dataset, the recognition accuracy of 1, 5, and 8 spatial affine transformation network (STN) structures is compared, reaching 89.2%, 91.5%, and 91.7%, respectively, which proves that the increase of the number of STN can significantly improve the recognition accuracy.

Studies in the past three years have shown that the YOLO series object detection network architecture has been widely recognized in the field of classroom learning behavior detection due to its classic and advanced nature. Many improved models have successfully improved the accuracy of detection by trying different anti-interference methods. However, research on learning behavior recognition in home Settings is still very limited. Therefore, this study is the first to explore the research of lightweight algorithms for this field under the condition of small datasets and low computing power terminals.
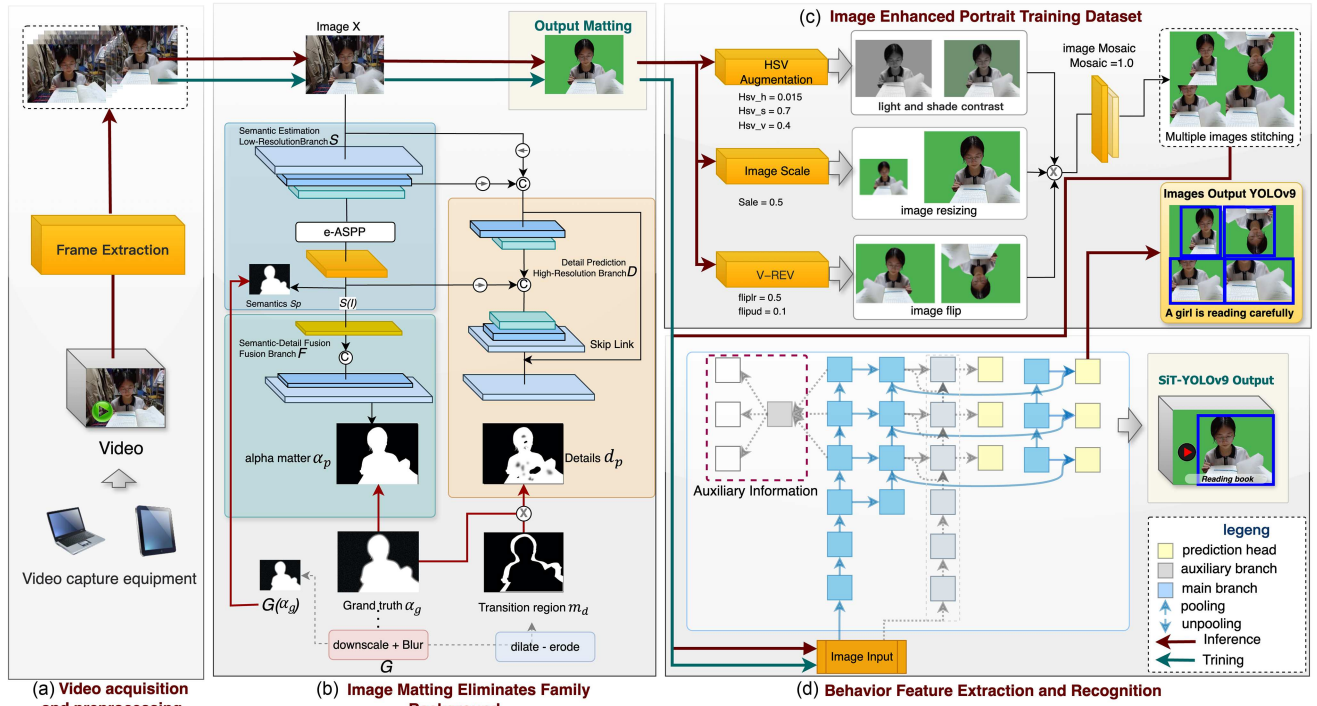
3. Theoretical Framework

In home learning environments, various challenges arise when trying to accurately recognize student behaviors, including background noise, complex behaviors, and limited data availability. These issues are compounded by the need for computational efficiency, especially on low-power devices such as TinyML platforms. To address these challenges, this study proposes an innovative solution, the SiT-YOLOv9 model. This model integrates the MODNet matting algorithm to reduce background noise and network computational scale [41], while improving the YOLOv9 model [42] with a pre-processing image enhancement module. The goal is to expand the dataset's scale for more precise identification of diverse student learning behaviors, providing an efficient and accurate solution that caters to the operational needs of low-power devices.

The design framework of the SiT-YOLOv9 model includes four core steps: video frame extraction, portrait matting, image enhancement, and feature recognition, with the conceptual architecture detailed in Figure 2.

Training process: (a) The mobile terminal collected the video and extracted frames into images (b) the images in the training process were passed through the MODnet facial image matting module to reduce the calculation of the subsequent network (c) the images after matting were passed through the HSV and other image enhancement modules to enhance the size of the dataset, (d) and then sent to the auxiliary reversible branch of YOLOv9 network to realize lossless feature extraction. Obtain the maximum accuracy of the model. Inference process: The extracted frame video from (a) is directly sent to the improved YOLOv9 network (d) that omits the Auxiliary branch, so as to realize the balance between inference speed and accuracy.

Video frame extraction is a standardized pre-processing procedure that captures video at a rate of 20 frames per minute from home TinyML mobile devices, storing the pre-processed video frames in cache. Background removal aims to address the issue of image interference in students' learning behaviors within home environments. By utilizing real-time portrait masking technology, the extracted video frames undergo background removal, akin to green screen technology, retaining only key objects such as the student's face, upper limbs, body, head, and handheld items. This approach not only reduces noise caused by complex backgrounds, movement

Figure 2
SiT-YOLOv9 network conceptual architecture

(a) **Video acquisition and preprocessing**

(b) **Image Matting Eliminates Family Background**

(c) **Image Enhanced Portrait Training Dataset**

(d) **Behavior Feature Extraction and Recognition**

of family members, and abnormal lighting in the home learning environment but also reduces the network scale through network pruning, effectively lowering computational load while enhancing system robustness, recognition accuracy, and speed. Considering the diversity of home learning environments, the complexity of student behaviors, and the initial shortage of image data, this study introduces an HSV portrait enhancement module during the training phase to increase the scale and quality of image data [43]. This module aims to enhance the diversity and expressiveness of the data by scaling and improving the quality of the training dataset. Through geometric transformation, color space enhancement, image blending, random erasure, feature space enhancement, adversarial training, generative adversarial networks, neural network style transfer, and meta-learning of the previous network's matting output, training accuracy is effectively improved while reducing the computational power consumption on TinyML devices. Feature extraction and recognition are divided into training and reference processes. The training process involves extracting features from the image mosaic portrait output and using these features to identify individual learning behavior patterns in the home environment, such as sleeping, reading, and using a smartphone. The introduction of programmable gradient information (PGI) and the generalized efficient layer aggregation network (GELAN) can address issues encountered in recognizing learning behaviors in home environments, such as unstable frames, object occlusion, small object recognition, and light changes, ensuring continuous tracking of learning behaviors under any circumstances [42].

4. Experiment

4.1. Experimental environment

The experimental training inference platform utilizes CUDA:0 (Tesla T4, 15110MiB) provided by Kaggle. The deep learning framework used is PyTorch 2.0.0, with the optimizer set to SGD

(Stochastic Gradient Descent). The initial learning rate is 0.1. The model will undergo 128 training epochs to ensure it has sufficient time to learn complex features from the data while avoiding overfitting. Input images are uniformly resized to 640 pixels, and the batch size is set to 9. These choices strike a balance between GPU utilization and memory constraints.

4.2. Evaluation metrics

In this study, the mean average precision (mAP) was used to evaluate model performance. Average precision (AP) assesses the model's performance on each class by considering both precision (P) and recall ®. mAP represents the average of AP across all classes and incorporates the intersection over union (IoU) between predicted bounding boxes and ground truth bounding boxes to reflect the accuracy of the bounding boxes. Commonly, mAP50 refers to the model's class recognition accuracy when the IoU threshold is set to 0.5, which is a crucial metric for measuring detection accuracy.

The frames per second (FPS) were also used to evaluate model speed. This included the time for pre-processing (e.g., padding to maintain aspect ratio, channel transformations from HWC to CHW format, and data dimensionality expansion), the time for inference (i.e., the time the pre-processed image takes to be input into the model and produce output results), and the time for postprocessing (i.e., converting the model's output results).

The important evaluation parameters equation is as follows:

$$_i = \int_0^1 P_i(R_i)dR_i \tag{1}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{2}$$

$$FPS = \frac{1000ms}{preprocess + inference + postprocess} \tag{3}$$

AP (Equation (1)) is used to measure the average precision under the Precision-Recall curve and mAP (Equation (2)) is the mean of average precision values measured for each class. For measuring the frame per second (FPS) speed we have used Equation (3).

4.3. SiTBehavior dataset

The SiTBehavior dataset is designed to support research in analyzing learning behaviors within home environments, particularly through the use of TinyML devices and mobile phones. This dataset comprises a total of 150 videos, featuring 10 distinct action categories that reflect typical behaviors observed during home learning activities. The categories include sleeping, reading a phone, reading a book, writing, touching hair, drinking water, eating, holding a pen, being in a trance, and touching glasses. Within these, reading and writing are classified as focused behaviors, while the remaining actions are considered indicators of distracted learning. See Figure 3 for details.

The dataset is constructed from a total of 1740 manually annotated images, sourced both from the internet and real home environments. These images are extracted as feature frames using TinyML devices, which are commonly available in home settings, particularly through mobile phones. To enhance the diversity and richness of the dataset, an additional 300 images were sourced via web searches. This supplemental data is crucial for capturing a wider array of behavior characteristics, making the dataset more robust and generalizable.

For effective machine learning model development, the SiT-Behavior dataset is divided into three subsets: the training set, validation set, and test set, following a 7:1:2 ratio. This division ensures a balanced and systematic approach to training, tuning, and evaluating machine learning models. The video content in the dataset is formatted with a resolution of 1280 × 1024 pixels and a frame rate of 30 frames per second. To ensure efficient storage without compromising quality, H.265 coding is employed, with a bit rate set at 800 Kbps. All videos are uniformly saved in MP4 format, which is widely supported and ideal for efficient video storage and playback.

To facilitate easy management and retrieval, the video files in the SiTBehavior dataset are named according to a specific convention: "Shooting Angle + Shooting Object or Behavior". This naming scheme allows for straightforward identification of the content and perspective of each video, aiding researchers in their analysis and application of the data. The SiTBehavior dataset is particularly useful for developing models that can analyze and predict learning behaviors in home environments, with potential applications in improving focus and productivity in educational settings.

4.4. Results analysis

To visually display the detection outcomes, we performed visualization processing using Figures 4, 5, and 6. In these figures, student behaviors are outlined by blue bounding boxes, and key performance metrics such as the mAP50 value and loss function value are illustrated through line charts. To evaluate the model's proficiency in behavior classification, the left section of Figure 7 compares the predicted values of the 10 types of learning behaviors with the actual values by generating a heat map based on the confusion matrix to evaluate the model's proficiency in behavior classification identification. The color intensity within each block in the heat map corresponds to recognition accuracy for respective categories; deeper colors indicate higher accuracy levels.

Additionally, the right segment of Figure 7 presents category instances via a bar chart and examines the relationship between boundary dimensions for behavior detection through a scatter plot. Object distribution across categories is represented using distinct colors. The experimental results show that the minimum recognition accuracy of the model is more than 82.8%, and the average mAP50 of all categories is as high as 94.8%.

Table 1 presents the mAP50 values of the original YOLOv9t model and the SiT-YOLOv9 model following image enhancement and object extraction by MODNet. The results demonstrate an enhanced accuracy in recognizing student behaviors, with an increase from 91.8% to 94.8%. Notably, the recognition precision has also shown significant improvement, particularly evident in reading behavior recognition which rose from 69.0% to 93.7%,

Figure 3
SiTBehavior dataset MODnet images and normal images of students' behavior

Figure 4
Home learning behavior test results of SiT-YOLOv9

Figure 5
Various metrics of the SiT-YOLOv9 model such as precision, loss function, etc.
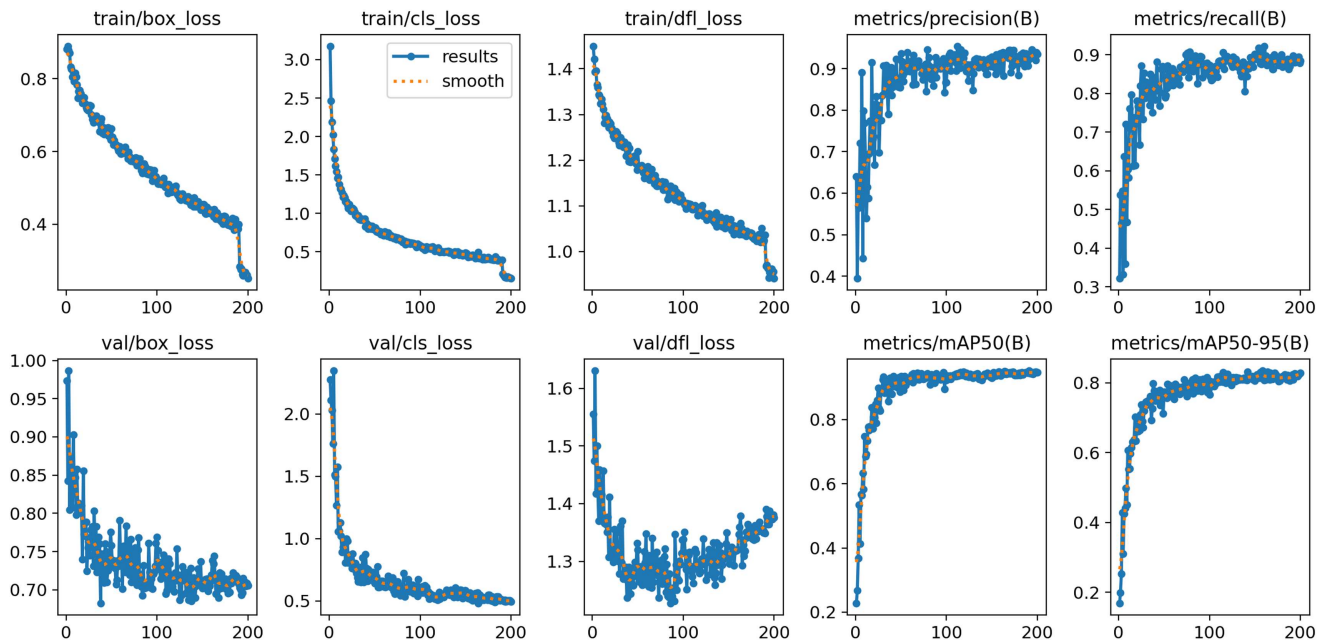
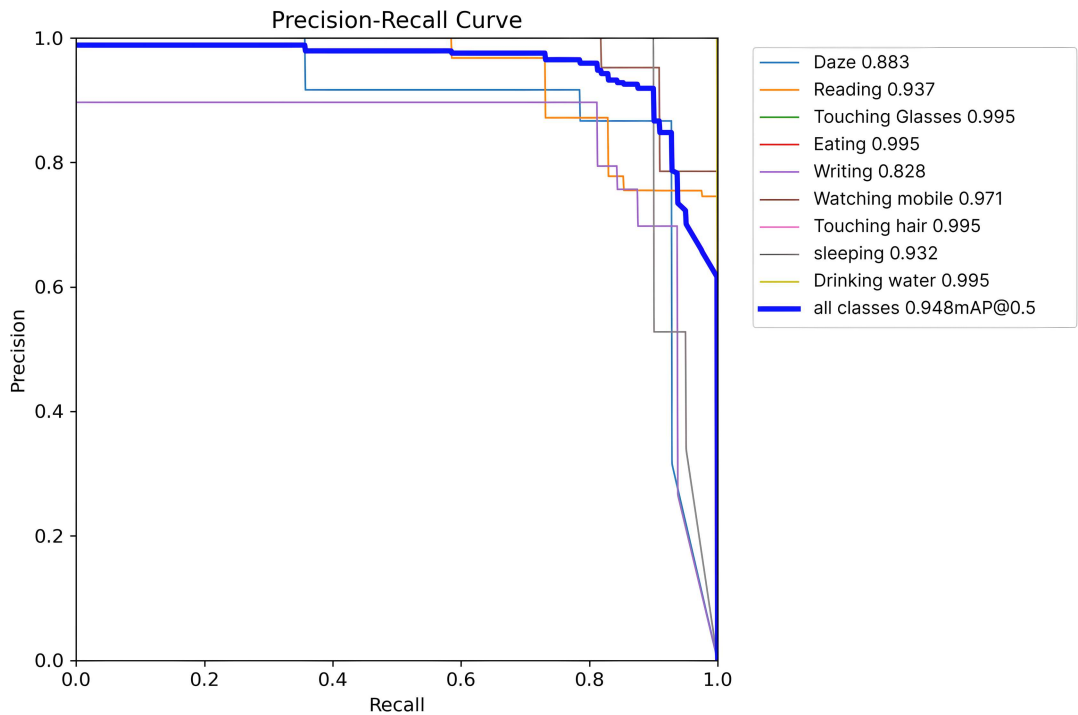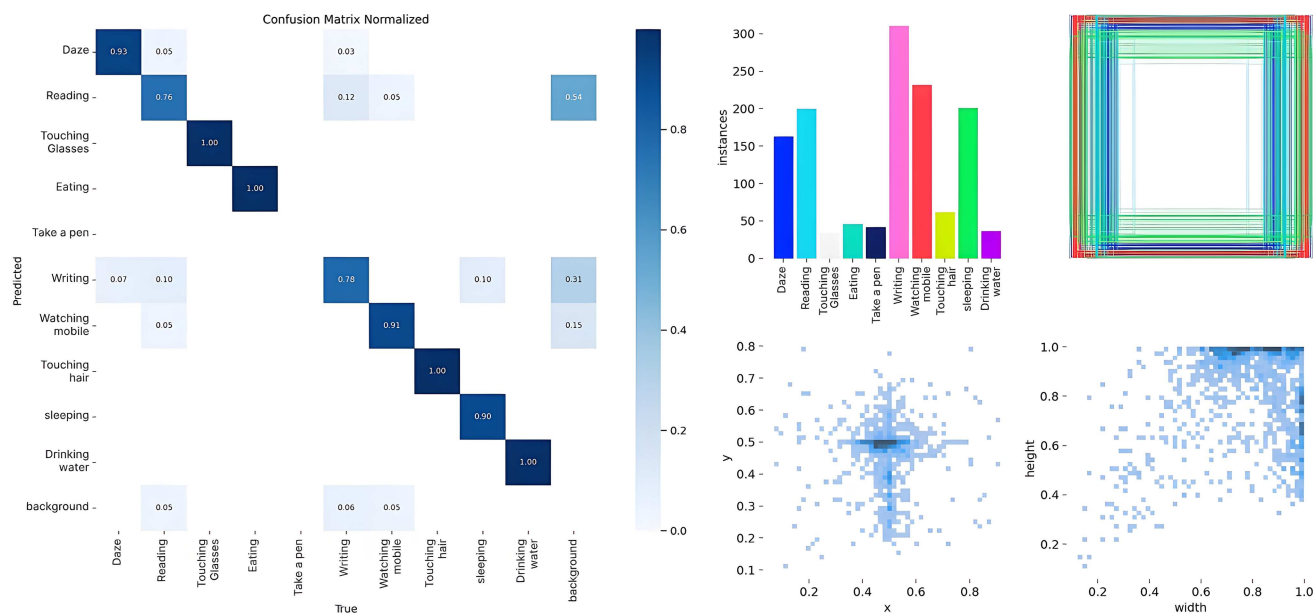**Figure 6**
PR curve of SiT-YOLOv9 model

Figure 7
SiT-YOLOv9 model classification accuracy and number of labels, labeling box location, etc.

marking a notable increase of 35.8%. This study highlights how enhancements to the YOLOv9 model contribute to more accurate recognition of various student behaviors.

The comparison data of the original YOLOv9t model and the improved SiT-YOLOv9 model after MODNet matting and image enhancement on inference speed indicators such as FPS (frames per second) are shown in Table 2. From the table, the FPS for recognizing student behavior increased from 80.6 to 90.9, a recognition speed improvement of 12.8%. It is particularly noteworthy that the post-processing time per image was reduced from 3.7 to 1.2 ms, making the processing speed three times faster than before. This significant improvement indicates that the optimization of the YOLOv9 model

Table 1
The mAP50 of various behaviors were detected with YOLOv9t and SiT-YOLOv9 models

Behavior	mAP50 (YOLOv9t)	mAP50 (SiT-YOLOv9)
All	0.918	0.948
Read the book	0.690	0.937
Drink water	0.995	0.995
Write	0.907	0.828
Touch glasses	0.828	0.995
Daze	0.964	0.883
Sleep	0.990	0.932
Eat food	0.995	0.995
Look at the phone	0.924	0.971
Touch hair	0.995	0.995

Table 2
Speed metrics for YOLOv9t model and SiT-YOLOv9 model

Speed metrics	YOLOv9t	SiT-YOLOv9
Preprocess (Unit: ms per image)	0.2	0.2
Inference (Unit: ms per image)	8.5	9.6
Postprocess (Unit: ms per image)	3.7	1.2
FPS	80.6	90.9

in this study can significantly enhance the speed and efficiency of processing and recognizing images of student behavior.

In the comparative analysis of model performance, this study observed that the recognition accuracy (mAP50 = 0.9006) of the YOLOv10n model is quite close to that of the YOLOv9 model (mAP50=0.9184). Notably, the YOLOv8n model exhibited a higher mAP50 value (0.9316), significantly outperforming YOLOv9 (0.9184). Based on this finding, the study further implemented the MODNet algorithm for image matting improvement on the YOLOv8n and YOLOv10n [44, 45] models and compared their student behavior recognition accuracy with more precise and larger models such as YOLOv8s, YOLOv9s, YOLOv10s, and RT-DERTl [46]. The results showed that the SiT-YOLOv9 model still maintains an advanced level among these models.

The accuracy values and parameter quantities of the ten models trained in the experiment are compared in Figure 8 and Table 3. Observations reveal that the improvements made in this study not only enhanced the recognition accuracy of the YOLOv8n and YOLOv10n models but also demonstrated the universal applicability of model improvements such as MODNet matting. In particular, the SiT-YOLOv8 model achieved the highest recognition accuracy, with its mAP50 value reaching 0.9591, improving the recognition accuracy from 93.16% to 95.91%. Following closely behind, the SiT-YOLOv9 model had a recognition accuracy of (mAP50 = 0.9480) and the smallest number of parameters (2,007,358), which is only about two-thirds of that of the SiT-YOLOv8 model (3,012,798).

Figure 9 and Table 4 present a comparison of the speed metrics, including FPS, for ten models derived from experimental training. By observing the percentage stacked bar chart of model inference time, it can be seen that the speed of model inference is primarily determined by the Inference and Postprocess times

Figure 8
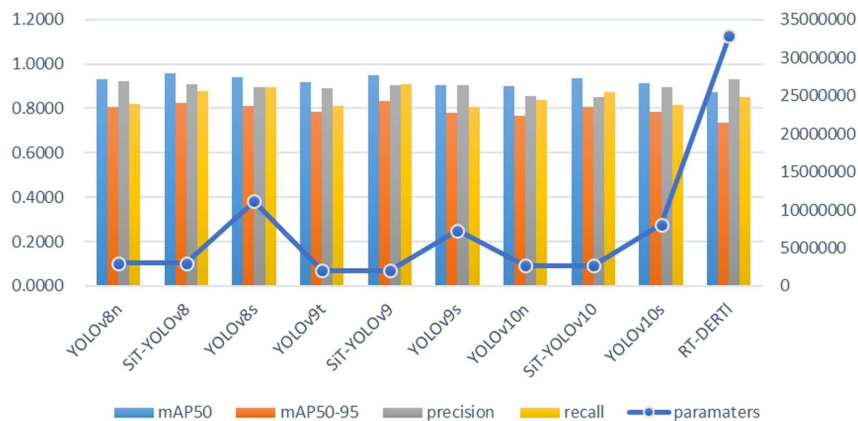**Comparison of accuracy values and parameters between models**

Table 3
Accuracy values and parameters for each model

Model	mAP50	mAP50-95	Precision	Recall	Parameters
YOLOv8n	0.9316	0.8057	0.9242	0.8191	3,012,798
SiT-YOLOv8	0.9591	0.8233	0.9084	0.8777	4,858,291
YOLOv8s	0.9385	0.8102	0.8961	0.8965	1,113,9470
YOLOv9t	0.9184	0.7851	0.8911	0.8101	2,007,358
SiT-YOLOv9	0.9480	0.8339	0.9047	0.9088	4,763,315
YOLOv9s	0.9043	0.7800	0.9027	0.8075	7,291,278
YOLOv10n	0.9006	0.7669	0.8563	0.8381	2,710,940
SiT-YOLOv10	0.9344	0.8082	0.8513	0.8755	4,556,433
YOLOv10s	0.9137	0.7863	0.8967	0.8175	8,074,092
RT-DERTl	0.8734	0.7375	0.9299	0.8522	32,826,626

Figure 9
Comparison of speed metrics such as FPS between models
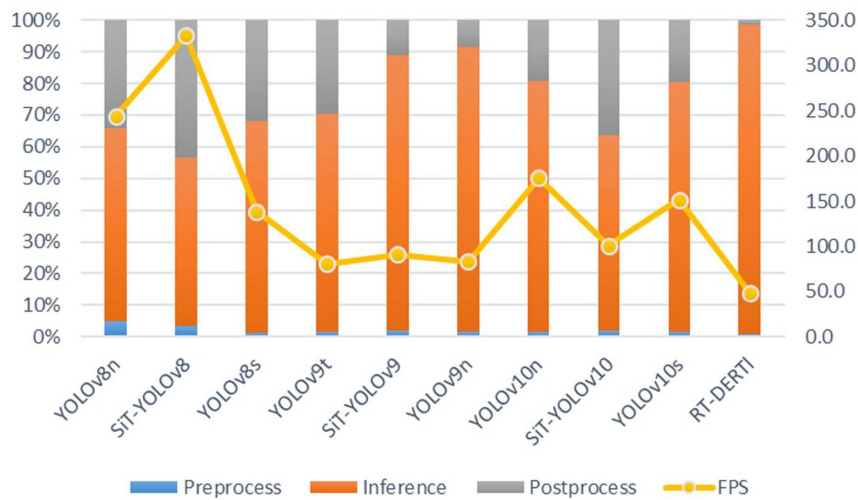
Table 4
Speed metrics such as FPS for each model

Model	Preprocess	Inference	Postprocess	FPS
YOLOv8n	0.2	2.5	1.4	243.9
SiT-YOLOv8	0.1	1.6	1.3	333.3
YOLOv8s	0.1	4.8	2.3	138.9
YOLOv9t	0.2	8.5	3.7	80.6
SiT-YOLOv9	0.2	9.6	1.2	90.9
YOLOv9s	0.2	10.8	1.0	83.3
YOLOv10n	0.1	4.5	1.1	175.4
SiT-YOLOv10	0.2	6.1	3.6	101.0
YOLOv10s	0.1	5.2	1.3	151.5
RT-DERTl	0.2	20.0	0.3	48.8

per image. Among these models, the SiT-YOLOv8 model exhibits the fastest recognition speed, with an FPS value reaching 333.3, which represents a significant improvement over the previous 243.9. This model ranks first in both recognition accuracy and speed. Meanwhile, the FPS of the SiT-YOLOv9 model has also increased compared to the original YOLOv9t model, and it has the smallest model size.

4. Conclusion

This paper innovatively proposes a behavior recognition model, SiT-YOLOv9. The model can accurately identify and analyze various learning behaviors in the home learning environment, with a particular focus on the operational requirements of low-power devices. To verify the performance of the model, we constructed a specialized video dataset for home learning behaviors, SiTBehavior. Through this dataset, the SiT-YOLOv9 model not only learned to recognize individual behaviors but also gained the ability to distinguish and understand different combinations of behaviors in complex scenes.

The SiT-YOLOv9 model effectively reduces background noise and computational scale of the network by integrating the MOD-Net matting algorithm and an original image enhancement module, significantly improving the accuracy of core steps such as background elimination, image enhancement, and feature recognition.

This improvement effectively solves interference problems caused by factors such as complex backgrounds, lighting changes, and family member movements.

Experimental results show that the SiT-YOLOv9 algorithm outperforms current mainstream algorithms like Yolov8, YOLOv9, Yolov10, and RT-DET on the SiTBehavior dataset, demonstrating excellent performance. The model's mean average precision (mAP50) reached 0.948, and the recognition speed reached 90.9 FPS, with a processing speed increase of 12.8%. This indicates that SiT-YOLOv9 has not only made breakthroughs in recognition accuracy but also achieved significant improvements in processing speed.

These achievements not only demonstrate the ability of SiT-YOLOv9 to quickly and accurately recognize student behaviors in home learning environments but also mean it can operate efficiently on resource-constrained mobile devices. This is of great significance for the popularization of intelligent learning tools. Overall, the contribution of this study lies in providing an efficient and accurate solution for recognizing home learning behaviors in the field of intelligent education.

Recommendations

However, we are also aware that some challenges need to be addressed. Firstly, it is suggested to further improve the recognition

accuracy, which requires a larger dataset of mobile phones, including the use of PAD, mobile phone, laptop, and other TinyML devices to collect video data from different resolutions, shooting angles, depth of field, and range of view. Secondly, the existing process of manually selecting feature frames is tedious and time-consuming, and the recognition tips are not detailed and specific enough. It is suggested to combine the image-to-text ability of the large language model to automatically extract the description information in the image, identify the learning state, behavior, atmosphere, and project of students in the home environment, and realize zero-shot generalization annotation.

Funding Support

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/danielsun1974/sitbehavior/settings.

Author Contribution Statement

Zhendan Sun: Conceptualization, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Vladimir Y. Mariano:** Methodology, Supervision, Project administration, Funding acquisition.

References

[1] Wang, J., Zhou, D., Li, X., Meng, D., & Zheng, Y. (2022). "Shuāng jiǎn" bèijǐng xià de jiātíng zuòyè: Wèntí huígù, zuòyòng jīzhì fēnxī yǔ tí zhì zēng xiào lùjìng xuǎnzé [Homework under the background of "Shuangjian": Problem review, mechanism analysis, path selection of improving quality and efficiency]. *Modern Distance Education, 1,* 57–63. https://doi.org/10.13927/j.cnki.yuan.20220105.001

[2] Fowler, K. S. (2000). *Understanding home schooling: The perceptions of three case study families of approach, process, and outcomes.* PhD Thesis, University of Virginia.

[3] Stefanovic, D., Drapsin, M., Nikolic, J., Scepanovic, D., Radjo, I., & Drid, P. (2011). Empirical study of student satisfaction in e-learning system environment. *Technics Technologies Education Management, 6*(4), 1152–1164.

[4] Yan, Y. (2018). Research on the influence of mobile media on college ideological and political education and countermeasures. In *International Conference on Education, Psychology, and Management Science,* 412–415. https://doi.org/10.25236/icepms.2018.091

[5] Chen, Q., & Yan, Z. (2016). Does multitasking with mobile phones affect learning? A review. *Computers in Human Behavior, 54,* 34–42. https://doi.org/10.1016/j.chb.2015.07.047

[6] Hodges, L. C. (2019). Becoming the distraction in the classroom. *The National Teaching & Learning Forum, 28*(5), 1–4. https://doi.org/10.1002/ntlf.30206

[7] Schmidt, S. J. (2020). Distracted learning: Big problem and golden opportunity. *Journal of Food Science Education, 19*(4), 278–291. https://doi.org/10.1111/1541-4329.12206

[8] Du, S., & He, Y. (2020). Problems and countermeasures of high school students' mathematics learning anxiety. *European Journal of Research in Social Sciences, 8*(2), 21–26.

[9] Craik, F. I. M. (2014). Effects of distraction on memory and cognition: A commentary. *Frontiers in Psychology*, *5*, 841. https://doi.org/10.3389/fpsyg.2014.00841

[10] Chun, Z., Hang, S., & Xiaoqing, G. (2021). Measurement and analysis of students' classroom learning behavior engagement based on computer vision technology. *Modern Educational Technology, 31*(6), 96–103.

[11] Chen, Q., Yan, Z., Moeyaert, M., & Bangert-Drowns, R. (2024). Mobile multitasking in learning: A meta-analysis of effects of mobilephone distraction on young adults' immediate recall. *Computers in Human Behavior, 162,* 108432. https://doi.org/10.1016/j.chb.2024.108432

[12] Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology, 94*(1), 88–106. https://psycnet.apa.org/doi/10.1037/0022-0663.94.1.88

[13] Berger, J. L., Kipfer, N., & Büchel, F. (2008). Effects of metacognitive intervention in low-performing vocational students. *Journal of Cognitive Education and Psychology, 7*(3), 337–367. https://doi.org/10.1891/194589508787724060

[14] Bui, N. A., Ngo, T. S., Lam, P. T., Le, P. C., Nguyen, H. T., Dat, N. C., ..., & van Dinh, T. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, *9*(22), 4729. https://doi.org/10.3390/app9224729

[15] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2,* 568–573. https://doi.org/10.1109/CVPR.2005.297

[16] Li, X., Yu, J., & Zhan, S. (2016). Spontaneous facial micro-expression detection based on deep learning. In *IEEE 13th International Conference on Signal Processing,* 1130–1134. https://doi.org/10.1109/ICSP.2016.7878004

[17] Ulutan, O., Rallapalli, S., Srivatsa, M., Torres, C., & Manjunath, B. S. (2020). Actor conditioned attention maps for video action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,* 516–525. https://doi.org/10.1109/WACV45572.2020.9093617

[18] Lin, T., Zhao, X., Su, H., Wang, C., & Yang, M. (2018). BSN: Boundary sensitive network for temporal action proposal generation. In *Computer Vision – ECCV 2018: 15th European Conference,* 3–21. https://doi.org/10.1007/978-3-030-01225-0_1

[19] Lin, C., Li, J., Wang, Y., Tai, Y., Luo, D., Cui, Z., ..., & Ji, R. (2020). Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(7), 11499–11506. https://doi.org/10.1609/aaai.v34i07.6815

[20] Wang, X., Wang, Y. F., & Wang, W. Y. (2018). Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 795–801. https://doi.org/10.18653/v1/N18-2125

[21] Hussein, H., Angelini, F., Naqvi, M., & Chambers, J. A. (2018). Deep-learning based facial expression recognition system evaluated on three spontaneous databases. In *9th International Symposium on Signal, Image, Video and Communications,* 270–275. https://doi.org/10.1109/ISIVC.2018.8709224

[22] Sharma, V., Gupta, M., Kumar, A., & Mishra, D. (2021). EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors, 21*(17), 5699. https://doi.org/10.3390/s21175699

[23] Sun, B., Zhao, K., Xiao, Y., He, J., Yu, L., Wu, Y., & Yan, H. (2019). BNU-LCSAD: A video database for classroom student action recognition. In *Optoelectronic Imaging and Multimedia Technology: Proceedings of SPIE, 11187,* 111871V. https://doi.org/10.1117/12.2539052

[24] Yang, F., & Wang, T. (2023). SCB-dataset3: A benchmark for detecting student classroom behavior. *arXiv Preprint: 2310.02522.* https://doi.org/10.48550/arXiv.2310.02522

[25] Jisi, A., & Yin, S. (2021). A new feature fusion network for student behavior recognition in education. *Journal of Applied Science and Engineering, 24*(2), 133–140. https://doi.org/10.6180/jase.202104_24(2).0002

[26] Zhao, Y., Yan, H., & Wang, Z. (2019). The advisable technology of key-point detection and expression recognition for an intelligent class system. *Journal of Physics: Conference Series, 1187*(5), 052011. https://doi.org/10.1088/1742-6596/1187/5/052011

[27] Ben Abdallah, T., Elleuch, I., & Guermazi, R. (2021). Student behavior recognition in classroom using deep transfer learning with VGG-16. *Procedia Computer Science, 192,* 951–960. https://doi.org/10.1016/j.procs.2021.08.098

[28] Yang, H.,Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., & Maybank, S. J. (2019). Asymmetric 3D convolutional neural networks for action recognition. *Pattern Recognition, 85,* 1–12. https://doi.org/10.1016/j.patcog.2018.07.028

[29] Si, J., Lin, J., Jiang, F., & Shen, R. (2019). Hand-raising gesture detection in real classrooms using improved R-FCN. *Neurocomputing, 359,* 69–76. https://doi.org/10.1016/j.neucom.2019.05.031

[30] Zhu, X., & Chen, Z. (2020). Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network. *The Visual Computer, 36*(4), 743–755. https://doi.org/10.1007/s00371-019-01660-3

[31] Wang, Z., Shen, C., Zhao, C., Liu, X., & Chen, J. (2022). Rónghé réntǐ zītài gūjì hé mùbiāo jiǎncè de xuéshēng kètáng xíngwéi shìbié [Recognition of classroom learning behaviors based on the fusion of human pose estimation and object detection]. *Journal of East China Normal University (Natural Science), 2022*(2), 55–66. https://doi.org/10.3969/j.issn.1000-5641.2022.02.007

[32] Wang, Z., Jiang, F., & Shen, R. (2019). An effective yawn behavior detection method in classroom. In *Neural Information Processing: 26th International Conference,* 430–441. https://doi.org/10.1007/978-3-030-36708-4_35

[33] Lu, M., Li, D., & Xu, F. (2022). Recognition of students' abnormal behaviors in English learning and analysis of psychological stress based on deep learning. *Frontiers in Psychology, 13,* 1025304. https://doi.org/10.3389/fpsyg.2022.1025304

[34] Chen, B., Zhu, J. N., & Dong, Y. Z. (2021). Expression recognition in sparse principal component combine low-rank decomposition architecture. *Journal of Computers, 32*(6), 195–205. https://doi.org/10.53106/199115992021123206017

[35] Liu, H., Ao, W., & Hong, J. (2021). Student abnormal behavior recognition in classroom video based on deep learning. In *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering,* 664–671. https://doi.org/10.1145/3501409.3501529

[36] Yu, W., Liang, M., Wang, X., Chen, Z., & Cao, X. (2022). Jīyú shēndù zhùyì lì wǎngluò de kètáng jiàoxué shìpín zhōngxuéshēng biǎoqíng shìbié yǔ zhìnéng jiàoxué pínggū [Student expression recognition and intelligent teaching evaluation in classroom teaching videos based on deep attention network]. *Journal of Computer Applications, 42*(3), 743–749. https://doi.org/10.11772/j.issn.1001-9081.2021040846

[37] Huang, Y., Liang, M., Wang, X., Chen, Z., & Cao, X. (2022). Jīyú shēndù shíkōng cán chà juàn jī shénjīng wǎngluò de kètáng jiàoxué shìpín zhōng duō rén kètáng xíngwéi shìbié [Multi-person classroom action recognition in classroom teaching videos based on deep spatiotemporal residual convolution neural network]. *Journal of Computer Applications, 42*(3), 736–742. https://doi.org/10.11772/j.issn.1001-9081.2021040845

[38] Chen, H., Zhou, G., & Jiang, H. (2023). Student behavior detection in the classroom based on improved YOLOv8. *Sensors, 23*(20), 8385. https://doi.org/10.3390/s23208385

[39] Wang, Z., Yao, J., Zeng, C., Li, L., & Tan, C. (2023). Students' classroom behavior detection system incorporating deformable DETR with swin transformer and light-weight feature pyramid network. *Systems, 11*(7), 372. https://doi.org/10.3390/systems11070372

[40] Zhao, J., Zhu, H., & Niu, L. (2023). BiTNet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network. *Journal of King Saud University-Computer and Information Sciences*, *35*(8), 101670. https://doi.org/10.1016/j.jksuci.2023.101670

[41] Ke, Z., Sun, J., Li, K., Yan, Q., & Lau, R. W. H. (2022). MODNet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(1), 1140–1147. https://doi.org/10.1609/aaai.v36i1.19999

[42] Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2025). YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision – ECCV 2024: 18th European Conference*, 1–21. https://doi.org/10.1007/978-3-031-72751-1_1

[43] Hu, J., Jiang, Q., Cong, R., Gao, W., & Shao, F. (2021). Two-branch deep neural network for underwater image enhancement in HSV color space. *IEEE Signal Processing Letters, 28,* 2152–2156. https://doi.org/10.1109/LSP.2021.3099746

[44] Safaldin, M., Zaghden, N., & Mejdoub, M. (2024). An improved YOLOv8 to detect moving objects. *IEEE Access,*

12, 59782–59806. https://doi.org/10.1109/ACCESS.2024.3393835

[45] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. In *38th Annual Conference on Neural Information Processing Systems,* 107984–108011.

[46] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ..., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605

[47] Tang, L., Xie, T., Yang, Y., & Wang, H. (2022). Classroom behavior detection based on improved YOLOv5 algorithm combining multi-scale feature fusion and attention mechanism. *Applied Sciences, 12*(13), 6790. https://doi.org/10.3390/app12136790