International Journal of Innovative
Computing, Information and Control
Volume **19**, Number **3**, June **2023**

ICIC International ©2023 ISSN 1349-4198
pp. 877–891

DETECTION OF STUDENTS' CLASSROOM CONCENTRATION BASED ON COMPONENT ATTENTION

Jianwen Mo, Rui Zhu, Hua Yuan* and Zhaoyu Shou

School of Information and Communication
Guilin University of Electronic Technology
No. 1, Jinji Road, Guilin 541001, P. R. China
{jwmo; guilinshou}@guet.edu.cn; 614824028@qq.com
*Corresponding author: yuanhua@guet.edu.cn

Received July 2022; revised November 2022

Abstract. *To detect students' concentration status in the classroom, a classroom concentration evaluation algorithm based on component attention is proposed to calculate students' concentration based on their classroom behaviors. First, the classroom videos are sampled and detected to obtain students' location information. Secondly, the multi-object video in the classroom is converted into a sequence of student single-object numbers (IDs) using a tracking assignment method. Finally, the student ID sequences are fed into a component attention-based behavior recognition network to obtain individual student concentration scores. A weighted fusion approach is designed to get concentration scores for all students and the overall classroom. In this paper, object detection and behavior classification datasets are created to analyze learning attentiveness, and transfer learning is used to improve the problem of insufficient sample size. After several experiments, it is shown that the detection accuracy of the component attention-based behavior recognition algorithm is more than 85%. Each student's concentration is fused using weighted fusion to obtain the curve of students' individual and overall classroom concentration over time.*
Keywords: Student behavior recognition, Object detection, Object tracking, Concentration

1. **Introduction.** As the basic form of teaching and learning, classroom teaching has always been the centerpiece of educational research. Students' classroom behavior is an important basis for evaluating students' learning status and the quality of classroom instruction, and it helps teachers to obtain an overall picture of classroom concentration. It helps teachers identify the results, optimize the content, and improve the quality of teaching. Students use the results to understand their classroom performance and regulate their classroom status. In traditional analysis methods, teachers or researchers use classroom observations and questionnaires to obtain information about students' classroom concentration. This human-driven detection method is difficult to avoid subjectivity and suffers from low efficiency and incomplete observation, which makes it difficult to extend to intelligent information-based classroom teaching. Therefore, teachers need an automated classroom learning status analysis method that coaches teachers to obtain the learning status of individual students in the classroom as well as the overall status of the classroom.

With the development of intelligent teaching systems, many computer vision-based methods for student classroom concentration detection have emerged. Pan et al. [1] used learning facial expressions and attention to classify the impact of student learning, constructed a learning focus migration model, and briefly described how to evaluate classroom teaching/learning effectiveness using students' learning affects (SLA) analysis. Guo [2]

DOI: 10.24507/ijicic.19.03.877

constructed a multi-task convolutional neural network to detect students' head-up rate to quantify classroom participation. Huang et al. [3] used head posture and facial expressions to distinguish students' classroom behaviors. A deep convolutional neural network and cascade based face feature point localization method were proposed to distinguish students' classroom behaviors by head pose and facial expressions. Pise et al. [4] proposed a temporal relational network (TRN) to estimate student engagement using a multilayer perceptron for classification work. Gupta et al. [5] proposed a maximum edge base face detection method using students' facial expressions for emotional content analysis, where the emotional content analysis includes the analysis of four different emotions of students, namely high positive emotion, low positive emotion, high negative emotion, and low negative emotion, and finally the student's concentration score is calculated by the four emotions. Ashwin and Guddeti [6] proposed a convolutional neural network based structure that uses nonverbal information such as facial expressions, gestures and body postures to analyze students' engagement and found a positive correlation between students' engagement and their test performance.

Most of these methods estimate student classroom engagement by extracting learned facial expressions or head postures, for example. However, there is still the problem of insufficient recognition accuracy, and these methods mainly object to the single learning situation and rarely object to the overall classroom concentration situation. Therefore, this paper proposes a component attention-based classroom concentration evaluation algorithm to evaluate student and overall classroom concentration by detecting students' classroom behaviors. The experimental results show that the proposed multi-participant classroom concentration evaluation algorithm can effectively identify classroom behaviors to get classroom concentration scores. To fully validate the proposed classroom concentration evaluation algorithm, a student classroom dataset is constructed in this paper for the training and testing of the algorithm. The main contributions and innovations of this paper are as follows.

1) Construct a dataset with object detection and behavior labeling to provide a database for subsequent student attention measurement in classroom teaching videos.

2) Propose a classroom behavior recognition algorithm based on component attention, and the overall detection accuracy is over 85%, combined with student object detection and tracking, to achieve student behavior recognition in classroom teaching videos.

3) Design a classroom concentration evaluation method to achieve single student concentration and overall classroom concentration evaluation in classroom teaching videos.

The main structure of this paper is as follows. Section 2 reviews the related work. Section 3 proposes an algorithm of classroom concentration evaluation based on component attention. Section 4 first introduces the experimental dataset, and then discusses the experimental results in detail. Section 5 summarizes the conclusion.

2. **Related Work.**

The popular areas of object detection, object tracking and behavior recognition are computer vision. This paper's research work is a combination of these three areas. Therefore, an overview of the work related to these fields is presented in this section.

2.1. **Object detection.**

Student target detection is the first step to achieve classroom concentration discrimination. Deep learning-based target detection algorithms are divided into single-step algorithms and two-step algorithms. The two-stage target detection algorithm needs to calculate the target candidate region and then perform target prediction. It has high localization and recognition accuracy, but the accuracy is slow. The single-stage target detection algorithm is a fast but low-accuracy end-to-end method.

Tan et al. [7] proposed the EfficientDet network, which performs quick multi-scale feature fusion through a weighted bi-directional pyramidal network (BiFPN), while proposing a composite scaling method with uniform scaling resolution, depth and width. Zhao et al. [8] proposed the M2Det algorithm to solve the target object through a mutil-scaled detection network of sizes and objects of different complexity. YOLOv4 [9] made further optimization based on YOLOv3 [10] to improve the speed and accuracy of the detection algorithm through CSPDarkNet53, Spatial Pyramid Pooling (SPP), Pyramid Attention Network (PAN) and data enhancement methods. And YOLOv5 [11], introduced in the same year, enhances the target detection algorithm's speed and accuracy through data enhancement and adaptive anchor frames. In [12] YOLOvX improved the detection speed and accuracy with operations such as decoupled head, anchor-free, and advanced label assigning strategy based on YOLOv3.

2.2. **Object tracking.** After the student target detection is completed, the listening status of each student needs to be continuously tracked by target tracking. In a classroom scenario, if the classroom behavior of a particular student needs to be analyzed, a real-time target tracking algorithm is just the right boost. Bewley et al. [13] proposed a simple online real-time target tracking (Sort) algorithm by combining the Kalman filtering algorithm and Hungarian algorithm; Wojke and Bewley [14] proposed a deep simple online real-time tracking (DeepSort) algorithm; Wang et al. [15] proposed the walk toward real-time multi-target tracking (JDE) algorithm by embedding the appearance model into the target detection model so that only one depth model is required for target tracking; Zhang et al. [16] proposed a fair multi-target tracking (FairMOT) that combines the target detection task with the pedestrian re-identification (Re-ID) task thereby achieving target tracking.

2.3. **Action recognition.** Two types of behavior recognition methods exist in computer vision: still-image based and video based. Static image based behavior recognition does not contain temporal information and is more difficult than video based recognition approaches. To compensate, still-image action recognition approaches rely on cues such as human pose, human-object and scene interactions. By sharing the underlying convolutional layers, Li et al. [17] proposed to explore body structure information with behavior recognition information using a deep model. Zheng et al. [18] proposed a spatial attention-based action mask network, which adds a spatial attention layer to the neural network to create specific action masks for images and generates semantic frames containing specific actions by a region selection strategy. Yan et al. [19] proposed a multi-branch attention network to extract global and local contextual information. For the lack of sufficient information in images, a migration learning approach on a large amount of action recognition data is combined with an attention mechanism to extract more distinguishing features for behavior classification. And for video-based human action recognition, motion information is the key factor and it is more discriminative than non-temporal recognition. Carreira and Zisserman [20] combined 3D convolution and dual-flow methods to recognize behaviors and proposed an inflated convolutional network algorithm for behavior recognition. Feichtenhofer et al. [21] proposed a slow-fast network that takes two branches to extract behavioral features, a slow branch and a fast branch, which are responsible for acquiring spatial and temporal information. The following year Feichtenhofer [22] discarded the two-branch approach. They proposed an expanded 3D convolutional network by extending the model depth and width, adjusting the model's image resolution and other parameters to obtain excellent performance even with very small computational effort.

In this paper, we apply the algorithm of single-person behavior recognition to the multi-object classroom teaching scene. We add the algorithm of target detection and tracking and single-person behavior recognition to solve the task of students' classroom behavior

recognition in the classroom teaching scene. For example, Yang et al. [24] designed a feature fusion-based YOLOv3 algorithm to detect the location and category of people and helmets in surveillance videos. Dang et al. [25] proposed an improved version of deep_sort_yolov3 target tracking architecture to achieve less identity exchange and faster computing speed.

3. **Methods.** The overall workflow of the algorithm proposed in this paper is shown in Figure 1. Firstly, the classroom video is input into the detection module, and the location information of students is obtained through sampling, detection and localization. Secondly, using the tracking assignment module, the multi-target videos in the classroom are converted into ID sequences of single targets of students. Finally, in the output module, the ID sequences are separately input to a component-based behavior recognition network to obtain individual student behavior score sequences, and a weighted fusion is used to obtain all student behavior scores at this moment. The analysis results include individual student listening concentration and overall classroom listening concentration for a certain period of time.

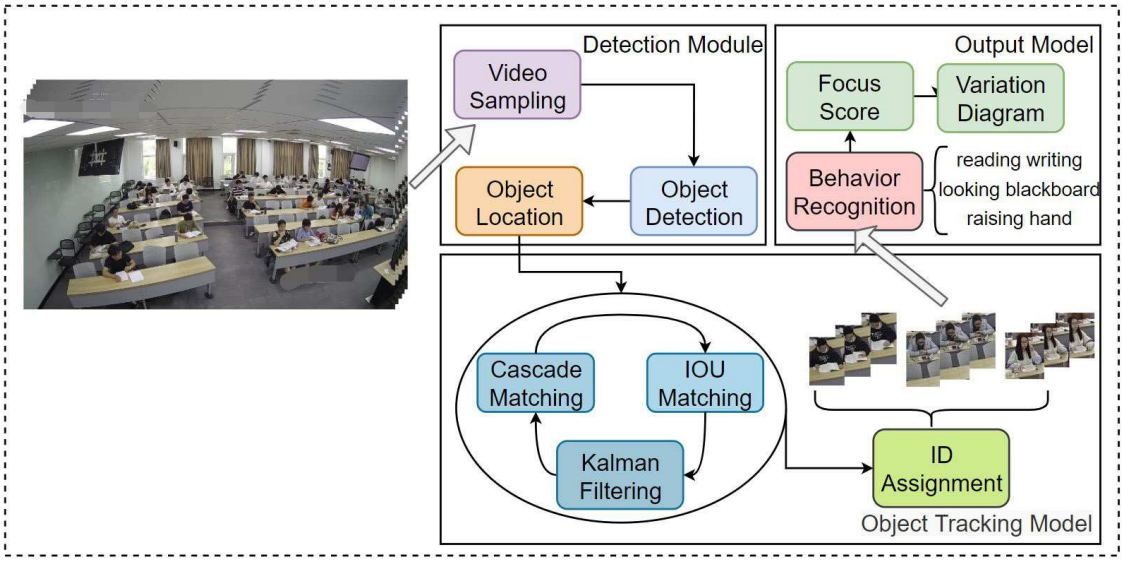


FIGURE 1. Overall framework

3.1. **Target detection module.** The classroom teaching dataset is fed into the detection module. First, the video is segmented into M parts using one minute as the segmentation interval; then each video segment is sampled using a sparse time sampling strategy. It is divided into T segments equally, and then a frame is randomly selected from each segment to form an input sequence with T frames; finally, the target detection is performed step by step in time order.

Target detection is the basis of subsequent tracking. Through target detection, the position of the student target in each frame of the video can be determined. And the accuracy of the target detection algorithm also affects the tracking accuracy. In this paper, YOLOv5 is used as the target detection network. It is a lightweight detection network improved based on YOLOv3. The YOLOv5 network structure consists of four parts: input side, backbone network, neck module and prediction module. As shown in Figure 2, the input uses Mosaic for data enhancement and performs adaptive anchor frame calculation and adaptive image scaling. The backbone network is designed with Focus and CSP1 structures, and Focus is used for slicing operation, and then features

are extracted by CBL, CSP1, and SPP structures. The neck module uses FPN and PAN structures, and features are fused more efficiently by a modified CSP2 structure. The basic structures of CBL, CSP1, and CSP2 are shown in Figure 2. The network uses CSP1 and CSP2 structures to improve the problem of reusing gradient information in the Res module's transformation process, thus reducing the computational effort. The prediction module used a combination of generalized intersection over union (GIOU) loss function and weighted non-maximum suppression (NMS) to make the network achieve better convergence.

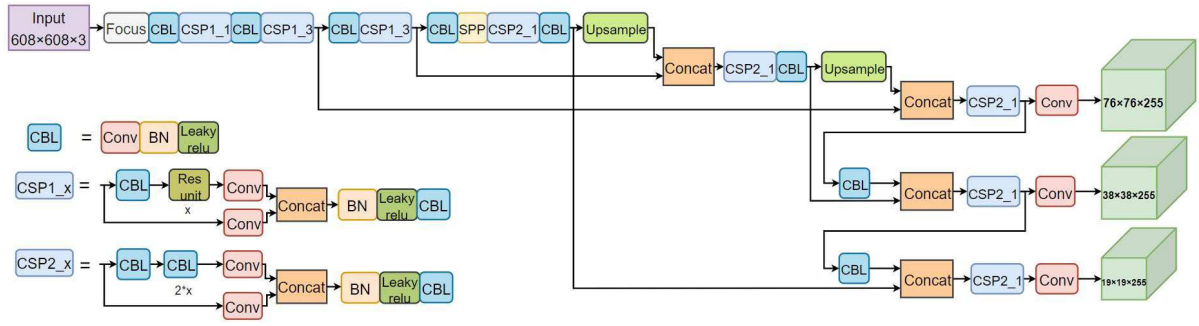


FIGURE 2. YOLOv5 network

3.2. **Tracking distribution module.** This paper combines the DeepSort algorithm to implement multi-target tracking in the classroom. The algorithm flow of the tracking assignment module is shown in Figure 3. Firstly, we obtain the target frame information of students located by the target detection module. Then use Kalman filter to predict the next frame of student target frame information, while compare the target frame information predicted by the detection module, and take the target frame with high confidence of both as the prediction result. Secondly, for cascade matching, the cost matrix of the Kalman filter's predicted result and target detection results are calculated using the appearance model (re-identification algorithm) and the motion model (martingale distance). The intersection over union (IOU) matching strategy is performed by calculating the IOU frame information between the prediction and detection frames. Finally, the filter parameters are updated according to the successfully matched prediction frames, and the successfully matched student targets are stored in the corresponding buffers based on the ID information. By looping the above operations, thus associating the student targets in the input sequence, the classroom student target tracking is realized and the ID sequence of the student single target is obtained.

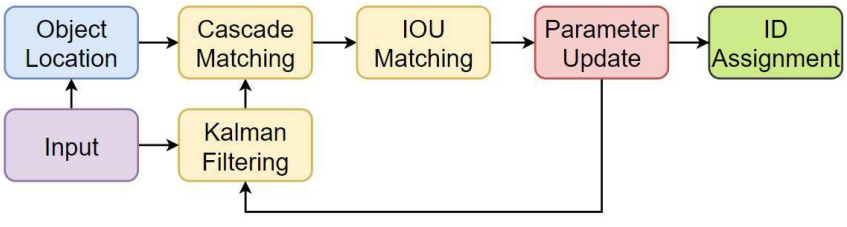


FIGURE 3. Tracking assignment model

3.3. **Output model.**

3.3.1. *Parts area division.* Human behavior can be seen as a combination of body part movements and requires precise positioning of the scenario when classifying specific parts. In this paper, the main study scenario is in the teaching classroom, so we focus on the upper body component actions. For example, reading a book usually consists of the head region and the hand-related region, with the head act "bowing the head" and the hand act "pressing the book". Raising the hand usually consists of the head area and the hand related area, the head act is "raising the head" and the hand act is "raising the hand". First, the coordinates of the key points of the human body are extracted from the image automatically by the human posture estimation method [23]. Then the position frame of each local part is calculated according to the image size and the key point coordinates. For example, in the head region, the minimum bounding box is calculated based on the head keypoints, and then the bounding box is expanded by 20% and adjusted according to the image boundaries to obtain the head position box. In a nutshell, the region box is generated from the key points, and then the final position box is generated by expanding and adjusting. Six local components are defined as shown in Figure 4: head, left wrist, right wrist, left shoulder, right shoulder, wrist and the middle region of the elbow.

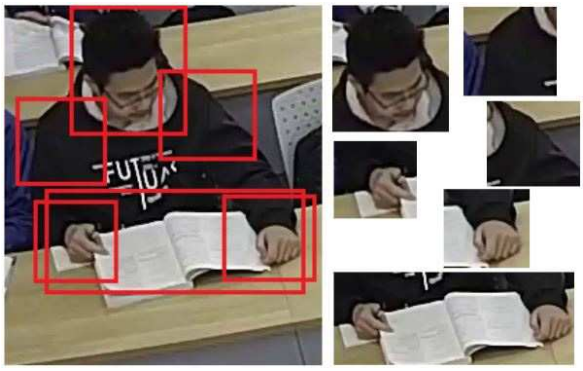


FIGURE 4. Part area division

3.3.2. *Behavior recognition.* In this section, a behavior recognition network based on the component attention mechanism is proposed for learning behavior recognition in classroom scenarios. The overall network structure is shown in Figure 5, where the model takes a single target student image X as input and the final predicted behavior category Y of the network as output. The overall model consists of two parts to extract global and local feature maps, respectively. For the global feature map extraction part, the network uses ResNet as the backbone network, Block1-Block5 denote the feature extraction layers of the network, and F_j, $j \in [1, 5]$ denotes the output feature map of each feature extraction layer. Finally, the feature extraction layer Block5 maps the input features into features and converts the global feature map into a global behavioral feature vector by operations such as pooling. For the component feature map extraction branch, the input image is assumed to be $X \in R^{H \times W \times 3}$. First, the network first uses the pose estimation module to obtain the key point coordinates of the student's body $P = \{(x_0, y_0), \ldots, (x_{17}, y_{17})\}$. Then the region partition module divides the global feature map into 6 local feature maps according to the defined part region partition method, and it transforms the global feature map F_3 into the local feature $F_{3,i}$ according to the key point coordinates and the local position box, where $i \in [1, 6]$. Each local feature map is then processed by a component attention network (PNet), which encodes the local features into the local feature vector

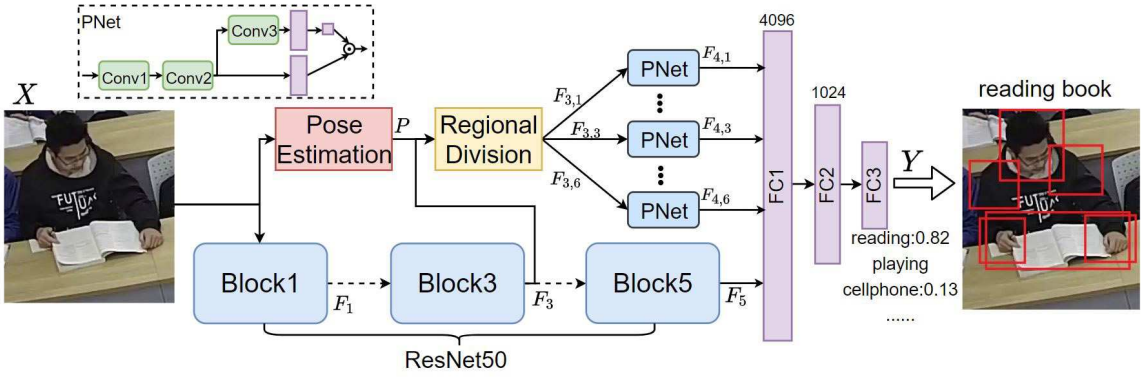


FIGURE 5. Behavior recognition network

$F_{4,i}$, where $i \in [1, 6]$. PNet consists of two branches, the feature extraction branch maps the input feature maps into local feature vectors through Conv1, Conv2 and pooling layers, and the attention branch calculates the weights of local features through Conv1, Conv2, Conv3, pooling layers and sigmoid activation functions. The computed weight values range from 0 to 1, with 1 indicating the most important and 0 indicating the least important. The weight values and the feature vector are multiplied to obtain the final local feature vector. Finally, the network stitches the local feature vector $F_{4,i}$ and the global feature vector F_5 to predict the student's behavior through multiple fully connected layers and output layers.

3.3.3. *Concentrate analysis.* In this paper, seven behaviors with high classroom student engagement were included in the behavior identification network, including reading or writing, looking at the blackboard, playing with cellphone, looking around, standing up, raising hands, and lying down. Each category (Clss) score is defined as shown in Table 1. The final student's score was between -2 and 2, normalized to the student's behavior score at a given time, as shown in Equation (1).

$$S_k(t) = W_k \times P_k(\text{Clss}) \tag{1}$$

where $S_k(t)$ denotes the behavior score of student k at moment t, W_k denotes the category weight, and $P_k(\text{Clss})$ denotes the predicted value of the behavior category of student k by the behavior recognition network.

TABLE 1. Behavior score definitions

Behavior	Standing up	Raising hands	Reading or writing	Looking at the blackboard	Looking around	Playing with cellphone	Lying down
Scores	2	2	1	1	-1	-2	-2

The overall classroom score at moment t is the set of all student behavior categories on the identification picture fused over the spatial domain. Therefore, the overall student score at this moment is obtained by summing up and averaging the behavior scores of all students. Let the number of students identified at the moment t be n_t, and the formula for the classroom score at moment t is shown below.

$$Frame(t) = \frac{\sum_{k=1}^{n_t} S_k(t)}{n_t} \tag{2}$$

The overall classroom behavior score is obtained by further integrating the overall classroom score in the time domain at moment t. Thus, the scores of all moments are

accumulated and averaged to obtain the overall effectiveness score of a class, and the formula is shown in (3), where MT denotes the classroom time after sampling.

$$S(\text{total}) = \frac{\sum_{t=1}^{MT} Frame(t)}{MT} \tag{3}$$

4. Experiment.

4.1. Dataset. This algorithm uses the COCO2017 dataset and the classroom dataset BOCD constructed by ourselves. Classroom scenes usually only see the upper body of students, while the annotation of the COCO dataset contains the whole body of the human body and is not applicable to classroom scenes. As shown in Figure 6, the key point labels of the COCO dataset are converted to human upper body target annotation boxes. In this paper, we use the data provided by the school classroom monitoring system as the initial data source to construct the BOCD, which comes from different classrooms and subjects, each containing 8 to 66 students. In this paper, the raw video data is first converted to image data at one frame per second and data cleaning is performed to remove some of the low quality and incorrect samples. Then the cleaned data is annotated. For the target detection part, the LabeImg image annotation tool developed based on python and QT GUI is used to save the tags as XML files. For the behavior classification data, a python based annotation script was developed to perform label saving in the form of txt files. The final obtained target detection part contains 5,000 samples and the behavior detection part has 3,173 samples. As shown in Table 2, the seven behaviors include reading or writing, looking at the blackboard, playing with cellphone, looking around, standing up, raising hands, and lying down. The number of students standing up, raising their hands, and lying down in the classroom scene was smaller than the number of students in the other four behaviors.

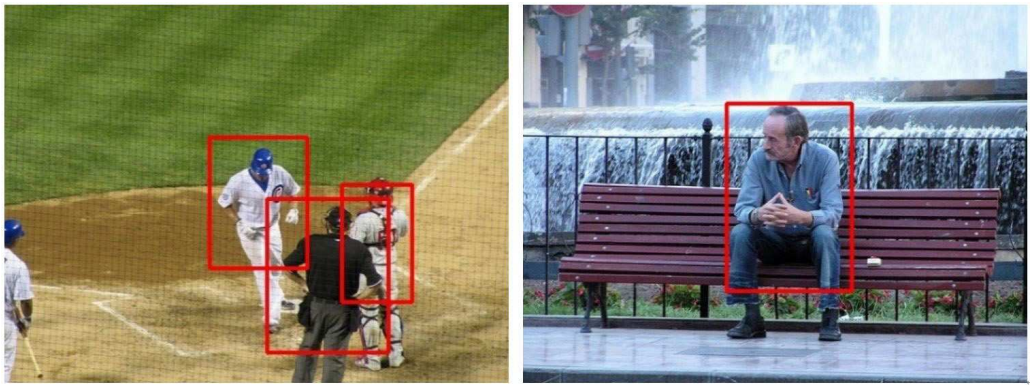


FIGURE 6. Processed COCO2017 dataset

TABLE 2. Number of behavior samples

Behavior	Reading or writing	Looking at the blackboard	Playing with cellphone	Looking around	Standing up	Raising hands	Lying down
Numbers	702	558	785	438	227	245	218

4.2. **Experimental results.**

4.2.1. *Object detection experiments and analysis.* Depending on the width and depth of the target detection network, the YOLOv5 algorithm is divided into four forms: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. In this paper, according to the actual classroom scenario, the detection accuracy and detection speed of different versions of the YOLOv5 network are weighed, and finally the YOLOv5s with the smallest network width and depth is chosen as the backbone of the target detection network. In the training process, the detection network is first pre-trained using the processed COCO2017 dataset, and then formally trained using the target detection part of the constructed dataset BOCD. The training parameters of the network are shown below, with an input size of 640×640, a data enhancement and adaptive image scaling method, an optimization algorithm Adam, an initial learning rate of 0.001, and a training batch of 299 [11].

As shown in Figure 7, the COCO dataset was used for training first and then the BOCD dataset was used for migration. The mAP0.5 value of the network is 98.87% and the mAP0.5:0.95 value is 88.27%. In contrast, the mAP0.5 value of the model without migration was 98.86% and the mAP0.5:0.95 value was 86.08%. Compared to pre-training, the mAP0.5 values remain the same without migration, and there is a 2 percentage point decrease in the mAP0.5:0.95 values.

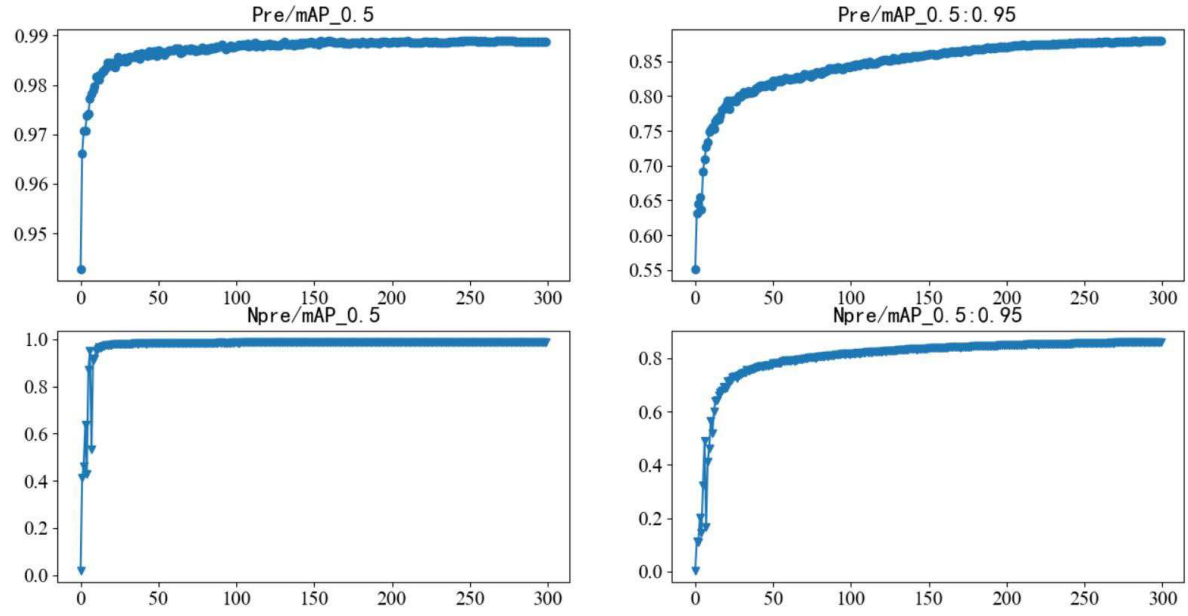
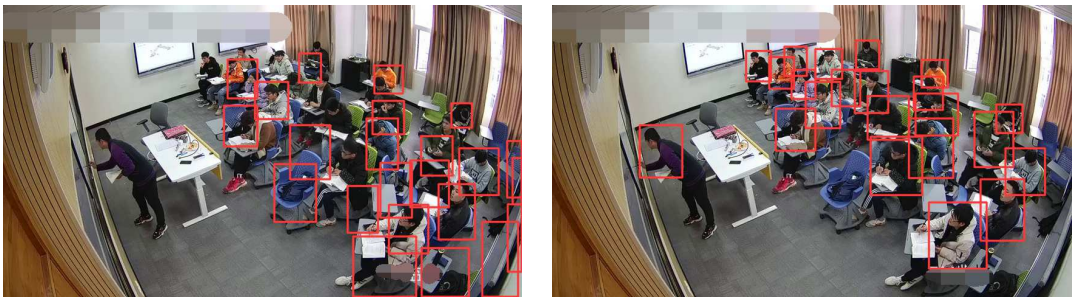


FIGURE 7. YOLOv5 algorithm performance

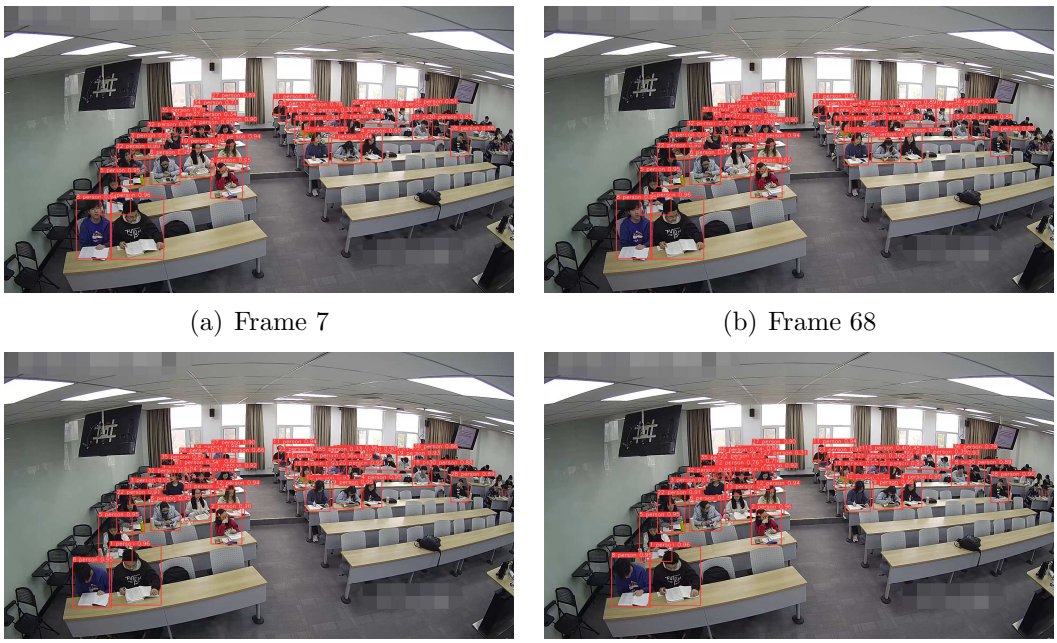
In this paper, the effectiveness of the object detection algorithm is tested in real classroom A. The camera view in classroom A is to the upper right, there are problems with blocking and small targets between students, and there are 21 student targets in the classroom. The target threshold is set to 0.3 for detection, and the detection effect is shown in Figure 8, where 8(a) indicates the detection result of the YOLOv5s model without pre-training, and 8(b) indicates the detection result of the YOLOv5s model with pre-training. The YOLOv5s without pre-training correctly detects 9 correct targets and there are invalid frames. The pre-trained YOLOv5s detects all 21 targets and the detection accuracy of the target boxes is higher than when there is no pre-training.



(a) (b)

FIGURE 8. Detection results

4.2.2. *Object tracking results and analysis.* In order to verify the effectiveness of the algorithm used in the paper for target tracking, classroom B was selected for testing the algorithm. The cameras in classroom B were divided into two left and right, the left camera video data of classroom B was tested in this paper, so the students in the right part of the picture were ignored. Figure 9 shows the effect of target tracking in frames 7, 68, 237, and 392. The probability of the target loss rate of students is low, and the target tracking is accurate. Among them, the student targets in the first four rows of the classroom are tracked well without ID number swapping, and the student ID numbers in the last three rows remain stable without and with a small amount of occlusion, and a small amount of ID number swapping occurs when the occlusion is serious. Therefore, this algorithm can achieve student target tracking in classroom teaching videos and get students single target ID sequences.



(a) Frame 7 (b) Frame 68

(c) Frame 237 (d) Frame 392

FIGURE 9. Tracking results

In the component attention-based student classroom concentration detection algorithm, this paper uses YOLOv5s combined with the DeepSort method for student target tracking. To verify the performance of student target tracking in classroom video scenarios, Table

TABLE 3. Comparison with other state-of-the-art methods

	Classroom1		Classrom2	
	FPS	Numbers	FPS	Numbers
Fair-MOT	13	9	11	30
Ours	30	9	22	35

3 shows the comparison between our method and other state-of-the-art methods in two different classrooms. Observing the experimental results in the table, we can find that our target tracking algorithm achieves better results in terms of tracking speed and accuracy, while Fair-MOT has a good detection effect but a slower detection speed. In summary, the algorithm of YOLOv5s+DeepSort is used to achieve the tracking of student targets in classroom teaching videos.

4.2.3. *Behavior recognition results and analysis.* To verify the effectiveness of the behavior recognition network, the network was tested using the behavior portion of the BOCD dataset. As shown in Figure 10, the confusion matrix demonstrates the model recognition performance, where the rows of the matrix represent the network prediction classes and the columns represent the true labels. The prediction results for each class are displayed in the matrix as numerical percentages. Observing the experimental results, it can be found that the accuracy of all seven categories of behavior is above 85%, with the highest accuracy of 95% for lying down, and the lowest accuracy of 85% and 86% for the two categories of standing up and looking at the blackboard, respectively. The model tended to confuse standing up with looking at the blackboard and reading and writing with playing with cellphone.

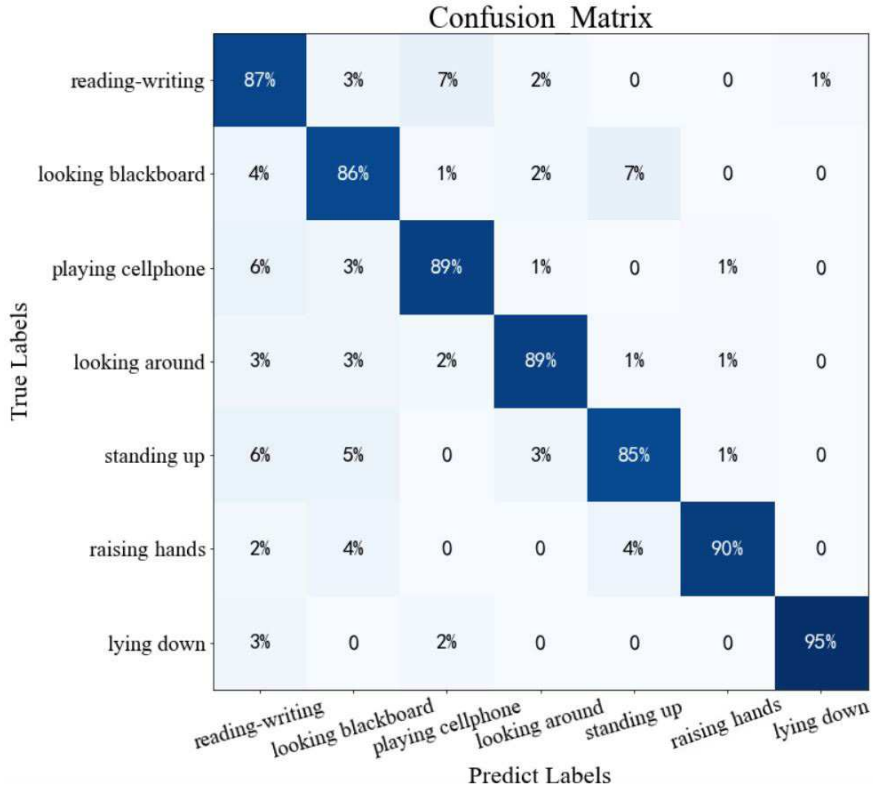
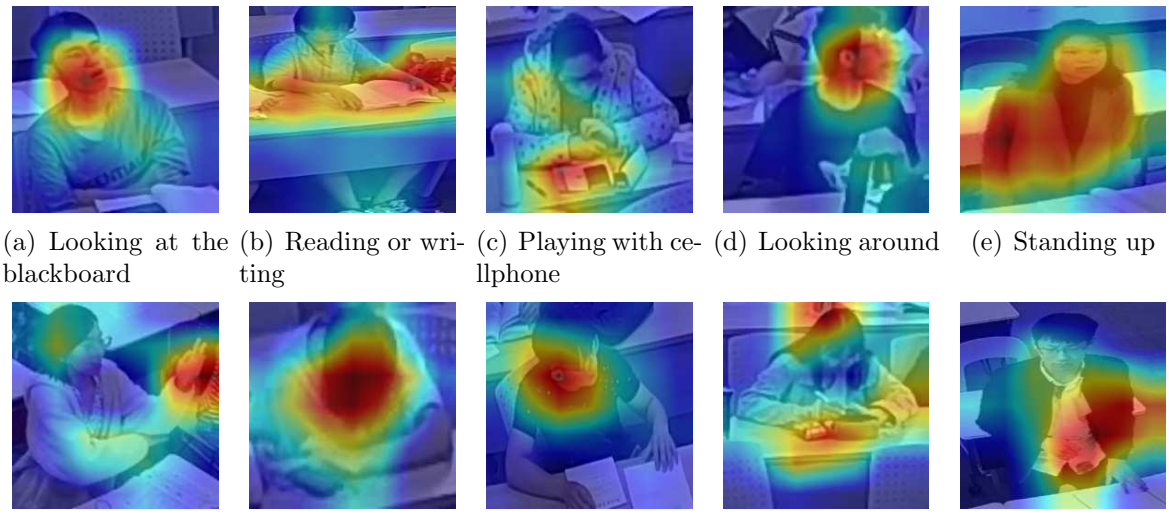


FIGURE 10. Behavioral classification confusion matrix

The attention area map for each behavioral category is shown in Figure 11. The head is an important region in the figure when looking at the blackboard and looking left or right, because looking at the blackboard and looking left or right is mainly a change of the head, which tends to look upward when looking at the blackboard and to the left or right when looking left or right. The behavior of looking at the blackboard in Figure 11 Error 1 was misclassified as left-right looking because of the camera viewpoint bias. The attention area for reading and writing and looking at cellphone is between the elbows, with the main focus on whether there is a book on the desk, a pen in hand, or a cellphone in hand. However, there are few pixels between the elbows in real classroom data, and it is sometimes difficult to distinguish them, as in Figure 11 Error 2 reading and writing behaviors were misclassified as looking at a cellphone. The attention area of standing and lying down focuses on the upper body, and the attention area of raising hands is on the hands. For example, in Figure 11 Error 3 the act of standing up was misclassified as lying down. Student behaviors in the classroom can present different positions and postures due to camera view and occlusion, and the attention area of the behavior can be biased.

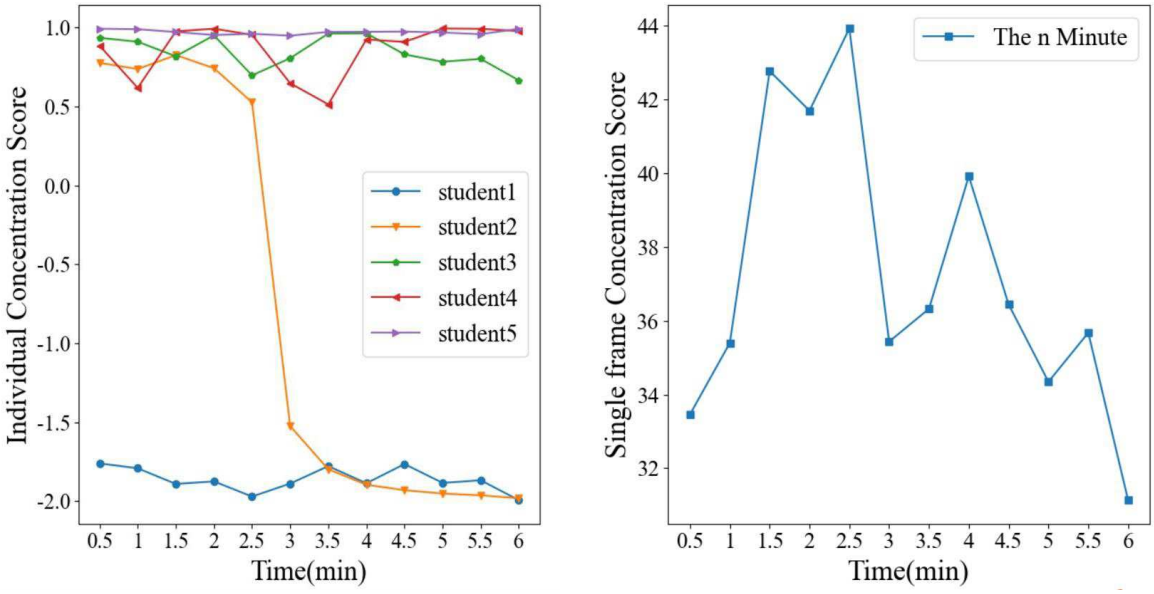


(a) Looking at the blackboard (b) Reading or writing (c) Playing with cellphone (d) Looking around (e) Standing up

(f) Raising hands (g) Lying down (h) Error 1 (i) Error 2 (j) Error 3

FIGURE 11. Behavioral attention area map

4.2.4. *Experimental analysis based on classroom videos.* Concentration experiments were conducted with classroom B's left camera video data. Six minutes of the test dataset were extracted from the video, and the target detection algorithm detected 35 people to test the algorithm's effectiveness in a traditional classroom. The video was input to the algorithm to obtain a line graph of student classroom listening concentration scores over time. As shown in Figure 12(a), the line graph of single-person concentration scores, Student 1 maintained low engagement throughout the test time; Student 2 had high engagement during the first two and a half minutes, significantly decreased in focus from two and a half minutes to three minutes, and remained low in the second three minutes; Students 3 through 5 maintained a high level of engagement throughout the test time; Student 5's attention remained stable during the test time, while Student 3 and Student 4's attention levels fluctuated significantly. As shown in Figure 12(b), a line graph of the overall classroom concentration in a single frame of this part of the algorithm, the overall classroom concentration trended higher and lower over the six minutes of the test dataset.



(a) Single concentration (b) Overall concentration

FIGURE 12. Diagram of students' classroom concentration

5. **Conclusions.** This paper proposes a classroom concentration evaluation algorithm based on component attention for detecting student concentration in classroom video scenes. Firstly, the input video is sampled and detected to obtain the student's location information; then, the classroom multi-target video is converted into a single target ID sequence of students using the target tracking ID assignment method; finally, the single sequence of students is input into the behavior recognition network to obtain the single person behavior, the learning concentration score is obtained by the concentration evaluation method, and the overall classroom concentration evaluation is obtained by the weighted fusion method. In order to meet the needs of the algorithm, we created a target detection and behavior classification dataset. Also, the COCO dataset was used to pre-train the target detection part to compensate for the problem of insufficient samples in the dataset. The experimental results show that the target detection and tracking algorithm used in this paper has a low miss detection rate and ID exchange rate. The detection accuracy of the component attention-based behavior recognition algorithm is higher, and the concentration evaluation method can demonstrate the overall state of the students and the classroom. In future research, we plan to improve the problem of poor target detection and recognition effect and ID interchange caused by occlusion between students. We plan to expand the constructed data set to increase the number of behavior types and sample size of target detection set. In addition, we plan to improve the focus evaluation system.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (Grant Numbers 62177012, 62001133, and 61967005). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] M. Pan, J. Wang and Z. Luo, Modelling study on learning affects for classroom teaching/learning auto-evaluation, *Science*, vol.6, no.3, pp.81-86, 2018.

[2] Q. Guo, Detection of head raising rate of students in classroom based on head posture recognition, *Traitement du Signal*, vol.37, no.5, pp.823-830, 2020.
[3] W. Huang, N. Li, Z. Qiu et al., An automatic recognition method for students' classroom behaviors based on image processing, *Traitement du Signal*, vol.37, no.3, pp.503-509, 2020.
[4] A. Pise, H. Vadapalli and I. Sanders, Facial emotion recognition using temporal relational network: An application to E-learning, *Multimedia Tools and Applications*, no.2, pp.1-21, 2020.
[5] S. K. Gupta, T. S. Ashwin and R. Guddeti, Students' affective content analysis in smart classroom environment using deep learning techniques, *Multimedia Tools and Applications*, vol.78, no.18, pp.25321-25348, 2019.
[6] T. S. Ashwin and R. M. R. Guddeti, Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues, *IEEE Access*, vol.7, pp.150693-150709, 2019.
[7] M. Tan, R. Pang and Q. V. Le, EfficientDet: Scalable and efficient object detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10781-10790, 2020.
[8] Q. Zhao, T. Sheng, Y. Wang et al., M2Det: A single-shot object detector based on multi-level feature pyramid network, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.33, no.1, pp.9259-9266, 2019.
[9] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, *arXiv Preprint*, arXiv: 2004.10934, 2020.
[10] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, *arXiv Preprint*, arXiv: 1804.02767, 2018.
[11] *YOLOv5*, https://github.com/ultralytics/yolov5, Accessed on May 30, 2020.
[12] Z. Ge, S. Liu, F. Wang et al., YOLOx: Exceeding YOLO series in 2021, *arXiv Preprint*, arXiv: 2107.08430, 2021.
[13] A. Bewley, Z. Ge, L. Ott et al., Simple online and realtime tracking, *2016 IEEE International Conference on Image Processing (ICIP)*, pp.3464-3468, 2016.
[14] N. Wojke and A. Bewley, Deep cosine metric learning for person re-identification, *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.748-756, 2018.
[15] Z. Wang, L. Zheng, Y. Liu et al., Towards real-time multi-object tracking, *European Conference on Computer Vision*, pp.107-122, 2020.
[16] Y. Zhang, C. Wang, X. Wang et al., FairMOT: On the fairness of detection and re-identification in multiple object tracking, *International Journal of Computer Vision*, vol.129, no.11, pp.3069-3087, 2021.
[17] Y. Li, K. Li and X. Wang, Recognizing actions in images by fusing multiple body structure cues, *Pattern Recognition*, vol.104, 107341, 2020.
[18] Y. Zheng, X. Zheng, X. Lu and S. Wu, Spatial attention based visual semantic learning for action recognition in still images, *Neurocomputing*, vol.413, pp.383-396, 2020.
[19] S. Yan, J. S. Smith, W. Lu and B. Zhang, Multibranch attention networks for action recognition in still images, *IEEE Trans. Cognitive and Developmental Systems*, vol.10, pp.1116-1125, 2018.
[20] J. Carreira and A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp.4724-4733, DOI: 10.1109/CVPR.2017.502, 2017.
[21] C. Feichtenhofer, H. Fan, J. Malik et al., Slowfast networks for video recognition, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.6202-6211, 2019.
[22] C. Feichtenhofer, X3D: Expanding architectures for efficient video recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.203-213, 2020.
[23] K. Sun, B. Xiao, D. Liu et al., Deep high-resolution representation learning for human pose estimation, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5693-5703, 2019.
[24] W. Yang, G.-L. Zhou, Z.-W. Gu, X.-D. Jiang and Z.-M. Lu, Safety helmet wearing detection based on an improved YOLOv3 scheme, *International Journal of Innovative Computing, Information and Control*, vol.18, no.3, pp.973-988, https://doi.org/10.24507/ijicic.18.03.973, 2022.
[25] T. L. Dang, G. T. Nguyen and T. Cao, Object tracking using improved deep_sort_yolov3 architecture, *ICIC Express Letters*, vol.14, no.10, pp.961-969, https://doi.org/10.24507/icicel.14.10.961, 2020.

Author Biography

Jianwen Mo received the B.S. degree from Department of Electronic Engineering, North China Electric Power University, in 1994, the M.S. degree from School of Mathematics and Computer Science, Guangxi Normal University, in 2002, and the Ph.D. degree from School of Electronic Engineering, Xidian University, in 2011.

He is currently a full-time associate professor at Guilin University of Electronic Technology, China. He presided over and mainly participated in a number of projects, including projects of the National Natural Science Fund, the Guangxi Natural Science Fund, and so on. His research interests include intelligent information processing, computer vision and image processing, machine learning, and artificial intelligence.

Rui Zhu obtained a B.S. degree in Communication Engineering from Dalian Nationalities University, China in 2018; he obtained an M.S. degree in Electronic and Communication Engineering from Guilin University of Electronic Technology, China in 2022.

His research interests include deep learning and intelligent image processing.

Hua Yuan received the B.S. degree in Computers and Applications and the M.S. degree in Electronic Information Engineering from the Guilin University of Electronic Technology, in 1999 and 2012, respectively.

He is currently a full-time lecturer at Guilin University of Electronic Technology, China. In recent years, he has presided over three department-level projects. He has authored or coauthored many articles in academic journals as the first author and received many national invention patents. His current research interests include intelligent information processing, computer vision and image processing, machine learning, and artificial intelligence.

Zhaoyu Shou received the B.S. degree, M.S. degree and the Ph.D. degree in Computer Science and Technology from Guilin University of Electronic Technology in 1999, 2004 and 2019, respectively.

He is currently a full-time professor at Guilin University of Electronic Technology, China. He presided over and mainly participated in more than 20 projects, including projects of National Natural Science Fund, National Development and Reform Commission, Guangxi Natural Science Fund, Guangxi Scientific Research and Technological Development Project, and so on. Prof. Shou's research interest covers digital image fusion, outlier detection, computer information management and education big data. Currently more attention is being paid to big data processing in education.