applied sciences

MDPI

Article

Research on Recognition and Analysis of Teacher–Student Behavior Based on a Blended Synchronous Classroom

Taojie Xu, Wei Deng *, Si Zhang, Yantao Wei and Qingtang Liu

School of Educational Information Technology, Central China Normal University, Wuhan 430079, China
* Correspondence: sdengwei@ccnu.edu.cn

Abstract: Due to the impact of the COVID-19 pandemic, many students are unable to attend face-to-face courses, Therefore, in this case, distance education should be promoted to replace face-to-face education. However, because of the imbalance of education in different regions, such as the imbalance of education resources between rural and urban areas, the quality of distance education may not be guaranteed. Therefore, in China and some regions, there have been efforts made to carry out blended synchronous classroom attempts. In hybrid synchronous classroom situations, teachers' workloads have increased, and it is difficult to fully understand students' learning efficiency and class participation. We use deep learning to identify the behaviors of teachers and students in a blended synchronous classroom-based situation, aiming to automate the analysis of classroom videos, which can help teachers in classroom reflection and summary in a blended synchronous classroom or face-to-face classroom. In the behavior recognition of students and teachers, we combine the head, hand, and body posture information of teachers and students and add the feature pyramid (FPN) and convolutional block attention module (CBAM) for comparative experiments. Finally, S–T (student–teacher) analysis and engagement analysis were carried out on the identification results.

Keywords: blended synchronous classroom; behavior recognition; learning analysis; deep learning

check for updates

Citation: Xu, T.; Deng, W.; Zhang, S.; Wei, Y.; Liu, Q. Research on Recognition and Analysis of Teacher–Student Behavior Based on a Blended Synchronous Classroom. *Appl. Sci.* **2023**, *13*, 3432. https://doi.org/10.3390/app13063432

Academic Editors: Hüseyin Kusetogullari and Chih-Hsiung Tu

Received: 13 January 2023
Revised: 2 March 2023
Accepted: 5 March 2023
Published: 8 March 2023

1. Introduction

At present, due to the outbreak of COVID-19 all over the world, some students are restricted by geographical location and cannot attend face-to-face courses. Although most universities provide online learning for affected students, online learning is equivalent to traditional face-to-face learning, but there are still many drawbacks, such as low learning efficiency caused by poor participation and persistence [1]. At the same time, some areas with relatively poor educational resources are quite different from those with relatively rich educational resources in the follow-up methods of online learning. For example, teachers in areas with relatively poor educational resources have insufficient experience in online learning methods, and it is difficult to carry out excellent educational activities, which leads to serious educational inequality during the epidemic situation. The use of a blended synchronous classroom can avoid these problems.

In the synchronized classroom, it is difficult for teachers to know whether learners are really as effective as teachers expect, and learners may not pay attention to teachers' instruction in the synchronized classroom [2]. Teachers should not only be satisfied with teaching students but also need to evaluate the level of students' participation in learning scenes because learning participation is a prerequisite for good learning results, and it is positively related to student's academic achievement and higher-level ability development [3]. Moreover, teachers have a large workload in the blended synchronous classroom, so it is difficult to fully understand students' learning efficiency and engagement in the course. It is obviously not possible for teachers to watch the recording and broadcast of a class every time and to analyze and reflect on it. In recent years, the latest achievements of artificial intelligence have been gradually applied to all aspects of education, such as intelligent education, data analysis, and prediction. An intelligent education system provides teachers and learners with timely and personalized guidance and feedback [4]. In fact, the intelligent recognition of classroom behavior based on machine learning has achieved initial results. There are two challenges in recognition of multi-person classroom teaching behaviors: one is to transform the discrimination and classification basis of artificial classroom behaviors into machine learning conditions, that is, how to make machines learn as many artificial behavior classification methods as possible to improve the recognition accuracy of machine learning; second, there are serious interference factors in real scenes, including the overlapping and changing of target behaviors, diversity of classroom scenes, etc.

Our research is mainly divided into two parts:

- In the scenario of hybrid synchronous classroom, we try to realize the automatic classification of teacher and student behaviors, the behavior of teachers and students will be classified under the improved S–T behavior classification code combined with the way of pose–action–behavior hierarchical recognition;
- By analyzing the behavior we know, teachers can obtain the results of teaching evaluation. In this paper, S–T behavior analysis and classroom engagement measurement are carried out on the results of behavior classification.

2. Background

2.1. Blended Synchronized Classroom

With the improvement of synchronous communication tools, the boundary between traditional face-to-face education mode and online education mode becomes blurred, while new education mode and mixed education mode begin to appear [5].

Blended synchronous classroom (BSC) is defined as students participating in face-to-face courses through remote media synchronization technology, such as video conferences, online conferences, or virtual worlds [6]. Blended learning can provide more educational opportunities for students and, in many ways, provide a more inclusive and fair learning experience for those who are geographically isolated or unable to be in class [7].

Yan Huang's research results [8] put forward that, when teachers promote mixed and synchronous curriculum, the main problem teachers face is cognitive overload caused by communication and distraction. The survey shows that students in the distance class still feel excluded from the main class because they are physically separated from the main class, especially when the distance class encounters technical difficulties and is not immediately solved. Therefore, it is very important to find an automatic way to identify and analyze the students' behaviors in the blended synchronous classroom.

2.2. S–T Behavior Analysis

In 2011, Yuan Jun and Ou Huanghai proposed the S–T (student–teacher) analysis method [9]. In the S–T analysis method, teaching behavior can only be divided into two types: teacher's (T) behavior and student's (S) behavior. T behavior is a teacher's visual and auditory information transmission behavior, while S behavior is all behaviors except T behavior. The main features of these analyses are the division, recognition, classification, and coding of teaching behaviors (including language behaviors). However, most of the research process is carried out manually, and there are some common problems such as complicated labeling operation, easy cause misoperation, and low efficiency of manual labeling [10].

Because the traditional classification of S–T behaviors has some problems, such as the ambiguity and semantic ambiguity of S–T behaviors, we re-classify and define S–T behaviors to eliminate the ambiguity and ambiguity between physical behaviors in class. Then, auxiliary discrimination rules are formulated to improve the fault tolerance and accuracy of automatic behavior recognition.The target detection model and behavior classification model are trained, and classroom behavior recognition and analysis experiments are carried out.

2.3. Student Behavior Recognition

Human behavior recognition has attracted considerable research attention in the field of computer vision, especially in the classroom environment. Human behavior recognition is a challenging task [11]. Under the influence of various factors, such as different illumination, complex background, multi-perspectives, and large intra-class differences, human behavior recognition algorithms are mainly divided into two types: (1) based on traditional machine learning and (2) based on deep learning. Each method has its advantages and disadvantages [12]. The key to an action recognition algorithm based on traditional machine learning is feature extraction. It usually takes a lot of effort to design features that meet requirements and simply implement them.

The main factors that affect video presentation include the complexity of the background, the change of human behavior speed, and the change of various perspectives. A convolutional neural network (CNN), with its strong scalability and end-to-end learning and training framework, has made breakthroughs in many fields, such as semantic segmentation, behavior recognition, and action classification. Action recognition based on the CNN framework can be divided into two types: one is a single-stream CNN framework, which uses only one stream (time or space) at a time. It is a dual-stream CNN framework, which uses both time stream and space stream. For the single-stream CNN framework, from the perspective of spatial CNN, a single video frame is input for behavior recognition, but it lacks the consistency of front and back actions. In terms of time CNN structure, human behavior information is usually obtained from dense optical flow. This kind of intensive time sampling leads to the high computational cost of video, and only a fixed number of video frames can be extracted from the video every once in a while. For the dual-stream CNN framework, the spatio-temporal information is integrated, and the recognition result is better than that of the single-stream CNN framework, such as three-dimensional convolution dual-stream neural network and track pool fusion dual-stream CNN network. Ji et al. [12] first proposed the concept of three-dimensional convolution, and used the three-dimensional convolution kernel to extract temporal and spatial features for behavior recognition [13]. A dual-stream convolutional neural network for behavior recognition is proposed, which is divided into a spatial stream convolutional network and a time stream convolutional network. The spatial convolution network takes a frame of an RGB image as input, which represents the apparent static information at a certain moment in the video. The time stream convolution network takes stacked optical stream images as input. The time flow convolution network takes several consecutive optical flow images as input to represent the motion information of objects. Finally, the classification results of the two networks are fused to obtain the final accuracy. This model breaks the intensive trajectory extraction algorithm in the field of improved behavior recognition. Tran et al. [14] proposed a new three-dimensional convolutional neural network (C3D). Three-dimensional continuous video frames are stacked in the network as the network input, a 3D convolution kernel is used to convolve the stacked cubes, and its time dimension is larger than that of a 2D convolution kernel. In this way, the motion information can be obtained from successive frames. The biggest advantage of this algorithm is that its recognition speed is greatly improved compared with the two-stream algorithm.

In 2020, aiming at the complex problem of student behavior recognition in video, Jisi et al. [15] proposed a new feature fusion network for educating students' behavior recognition. The new feature fusion network includes two main stages: feature extraction and classification. They combine spatial affine transformation networks and convolutional neural networks to extract more detailed features. Then, the weighted sum method is used to fuse the temporal and spatial features, and the softmax classifier is improved for classification and recognition. Zhang et al. [16] proposed an improved YOLOv3 target detection algorithm. By inserting the attention mechanism CBAM module into the original YOLOv3 shortcut structure, the effective features of students' classroom behavior can be quickly and effectively learned. Experimental results show that the improved YOLO-CBAM algorithm can improve the detection ability of small targets. When GIoU and Focal loss were included in the model, the mAP value could reach 0.95 and the F1 value could reach 0.879. Wu et al. [17] collected and constructed a dataset of teacher behavior using online open classroom teaching videos as data sources. Using this dataset, they explore action recognition methods based on RGB video and skeleton information and perform information fusion between them to improve recognition accuracy. Their experimental results show that the fusion of RGB information and skeleton information can improve recognition accuracy, and the early fusion effect is better than the later fusion effect.

In 2021, Lin et al. [18] proposed a student behavior identity based on skeletal posture estimation and human detection. They use OpenPose framework to collect skeleton data, and put forward an error correction scheme based on posture estimation and human detection technology to reduce false connections in skeleton data. The preprocessed skeleton data are then used to eliminate several joints that have weak influence on behavior classification. Secondly, feature vectors representing human posture are generated by feature extraction. The adopted features include standardized joint position, joint distance, and bone angle. Finally, the students' behaviors are classified to identify their behaviors. A deep neural network is constructed to classify actions. Chonggao et al. [19] proposed to combine the traditional clustering algorithm with the random forest algorithm, improve the traditional algorithm, and identify students' classroom behavior in real time by combining the human skeleton model. Wu et al. [20] proposed a moving target detection algorithm for student behavior recognition in class. Based on the region of interest (ROI) and face tracking, the authors proposed two algorithms to recognize the standing behavior of students in class. Moreover, a recognition algorithm was developed for hand raising in class based on skin color detection. Through experiments, the proposed algorithms were proved to be as effective in the recognition of student classroom behaviors.

In 2022, Mo et al. [21] proposed that an object detector extracts a single region from the keyframes as input to the network, and then the Multi-Task Heatmap Network (MTHN) module extracts an intermediate heatmap of multi-scale feature associations. Through the mapping relationship, the pose estimation and object detection tasks are constructed to obtain the keypoint and object position information. Finally, the key point behavior vector and measurement vector were used to model the behavior, and the classroom behavior detection algorithm based on the fully connected network was designed. Zhou et al. [22] proposed a new type of StuArt, a new automatic system designed to make personalized classroom observations that enable teachers to focus on the learning status of each student. StuArt can identify five representative student behaviors (hand raising, standing, sleeping, yawning, and smiling) that are highly correlated with engagement and track their changing trends throughout the course. To protect the privacy of students, all the changing trends are indexed by the seat number without any personal identifying information.

3. Materials and Methods

3.1. Encoding Method

3.1.1. Classification and Coding of Classroom S–T Behaviors

The existing S–T behavior classification method can comprehensively investigate verbal behavior and nonverbal behavior in the classroom; However, in the research of artificial intelligence behavior recognition, the method of fusing multimodal information is still in the discussion stage, and most classroom behavior research can only be based on single modal information, such as nonverbal behavior from images [23].

S–T analysis divides classroom behaviors into two categories: student behaviors (S) and teacher behaviors (T). By counting the number and sequence of S–T behaviors in the classroom, we can calculate the occupancy rate (Rt) of teacher behaviors and the conversion rate (Ch) of teacher–student behaviors. By drawing S–T curves, we can divide classroom types (lecture type, practice type, conversational type, and mixed type) to help teachers intuitively understand their own teaching situation and make timely adjustments. The traditional S–T classification table is shown in Table 1.

Table 1. Traditional S–T behavior classification table.

S behavior	1. Speaking; 2. Thinking and calculation; 3. Taking notes; 4. Doing experiments and finishing homework; 5. Silence; 6. Others
T behavior	1. Explanation; 2. Demonstration; 3. Writing (on the blackboard); 4. Media display; 5. Asking questions; 6. Evaluation and feedback

3.1.2. Physical Posture Characteristics of Classroom S–T Behavior

Design of Students' Behavior Characteristics in Class. Large-class teaching is widely used in schools in China. A large number of students in class leads to serious occlusion among students and various forms of action. However, the low resolution of the camera also makes the imaging of individual behavior blurred, especially for the students in the back row. The existing S–T behavior classification method is difficult to effectively distinguish various intra-class behaviors and inter-class behaviors on images. For example, in S behavior, students' thinking and silence behaviors are difficult to distinguish between calculation and finishing homework. In T behavior, teachers' questioning, explanation, and evaluation are difficult to distinguish.

How to effectively define various kinds of teaching behaviors is the first problem to be solved in the intelligent recognition task. Based on Table 1 and the actual classroom behavior, this paper proposes a new S–T behavior classification Table 2, which is suitable for image recognition. In this table, S behaviors are divided into six categories, and T behaviors are divided into three categories. Without the help of more information, the five behaviors in the first category of T behaviors are difficult to distinguish in the image, so they are classified into one category. The classification method of the table makes it difficult to confuse behaviors within the class, which will help the machine to recognize individual behaviors. Figure 1 shows seven kinds of behavior labeling methods based on local posture.

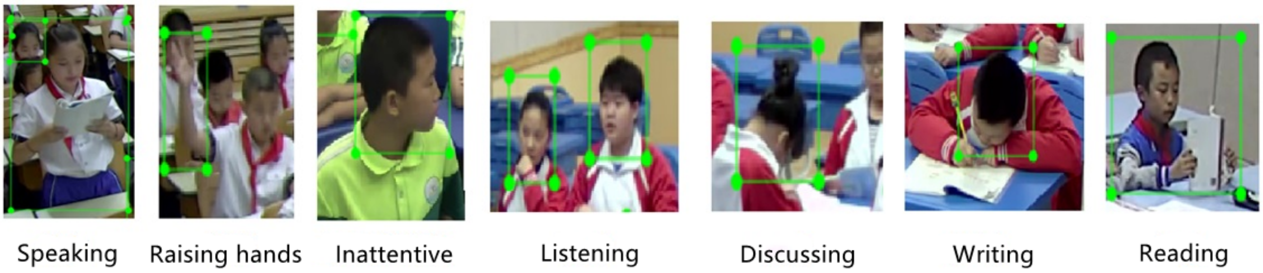


Figure 1. Seven kinds of behavior labeling methods based on local posture.

Table 2. Classification table of S–T behaviors in this paper.

S behavior	1. Raising hands; 2. Speaking; 3. Discussion; 4. Reading; 5. Writing; 6. Listening
T behavior	1. Asking questions, explanation, evaluation, demonstration, and media display; 2. Writing (on the blackboard); 3. Guide

Table 3 shows the local posture descriptions of the six main behaviors of students and whether they can be individually identified by image channels. It can be seen that we divide students' head postures into three categories: raise head, lower head, and turn head. Hand posture is divided into two categories: flat and raise. Body postures can be divided into three categories: stand, sit (without books), and sit (covered with books). Compared with teachers, the fuzziness among the six kinds of behaviors of students is lower, so the description of gestures is not very specific and only serves to distinguish these six kinds of behaviors. For different behaviors, it is often not necessary to use all local postures. Bold descriptions in the table are the main posture information of the corresponding behaviors. While simplifying the recognition rules brings convenience to recognition, it can also be seen that this description is not enough to deal with all situations. For example, “writing” should be judged more by the gesture of holding the pen than by the gesture of bowing the head. When discussing, students’ gestures are different. The former description is more inclined to “check all,” while the latter is “check accurately.” For this situation, the role of “situation” in the three-level model proposed in this paper will be reflected. When we take the class as a whole, the state of other students can correct some behaviors with no obvious characteristics. For example, although we can only recognize the “discussion” with the gesture of “turning head”, we can see from the total number of occurrences that all the students in the whole class are in a state of the discussion.

Table 3. Coding table of students’ behavior of human body local posture based on RGB images.

Behavior	**Local Attitude Description**			**Identified**
	Head	**Hand**	**Body**	
Listening	Raise Turn	Flat	Sit	Yes
Speaking	Raise	Null	**Stand**	Yes
Raising hands	Raise	Raise	Sit	Yes
Discussion	**Turn**	Null	Sit	Yes
Writing	**Lower**	Flat	Sit	Yes
Reading	Raise	Flat	**Sit (covered with books)**	Yes

Design of Teachers’ Behavior Characteristics in Class. (I) RGB image information.

As teachers’ behavior is closely related to verbal information, when image information is used alone for discrimination, there will be many ambiguous judgments. For example, we can not describe a teacher’s “Explanation behavior” with actions. To improve this problem, this paper adopts a hierarchical recognition method of posture–action behavior. We recognize teachers’ actions by their local postures instead of directly recognizing their behaviors because gestures and actions have a unique correspondence, while actions and behaviors have a “many-to-many” relationship. The recognition of teachers’ behaviors can be achieved by combining action recognition and situational reasoning. Figure 2 is the teacher’s behavior posture image.

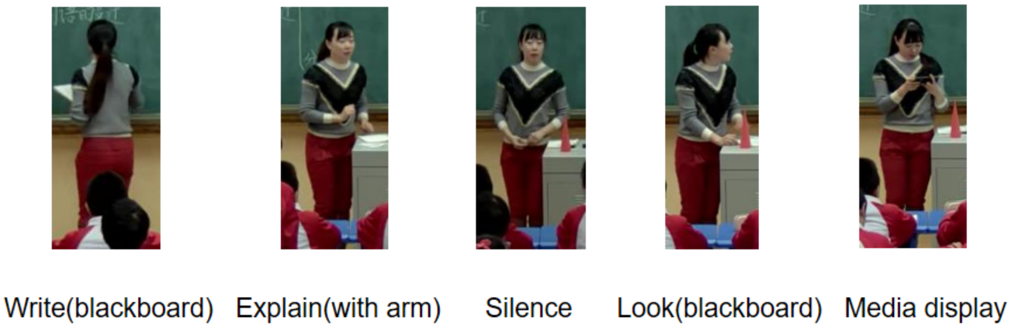


Figure 2. Teachers’ actions and gestures.

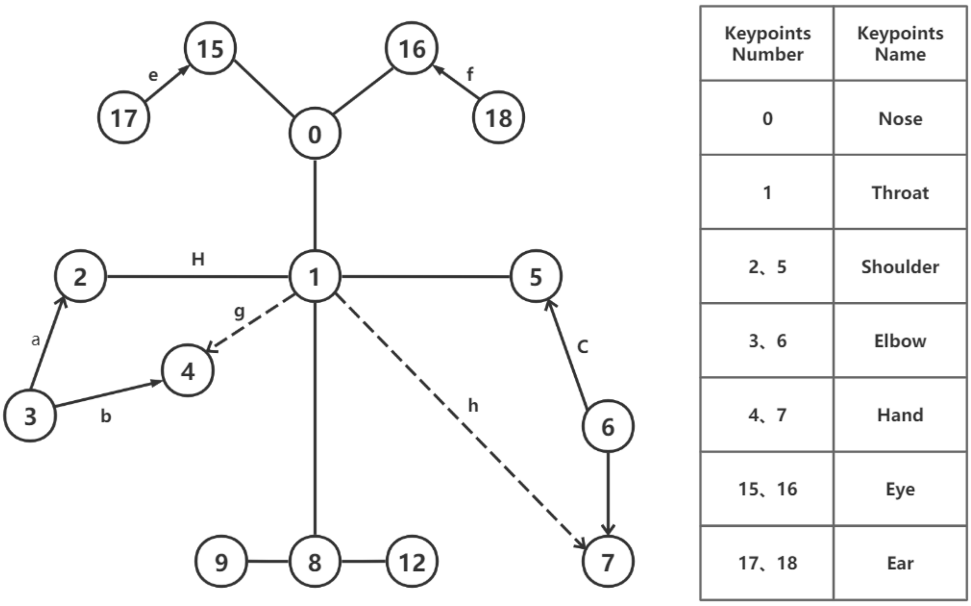
The following is the coding table of teachers’ actions. As can be seen from Table 4, it describes teachers’ posture in more detail and divides the head (left and right directions) and body into three categories. Positive represents facing the students (P), back represents facing the blackboard (B), and side represents the state between them (S); Head (up and down direction) describes the pitching state of the head, which can be divided into three categories (up (U), down (D), and flat (F)). The hands are divided into two categories (yes, no); yes means that the palm is over the waist, or the arm has a large bending; and no means the opposite. Compared with students' behavior, teachers' actions need the joint expression of various local gestures, and it is difficult to rely on a single gesture for expression. At the same time, because there is no feature design for the teacher's mouth shape when judging the behavior of the explanation, we only rely on the behavior of explanation (with the arm), and the behavior of explanation without gestures can not be identified. This is the drawback of the lack of information channels.

Table 4. Teacher's action coding table of human body local posture based on RGB images.

Behavior	Local Attitude Description				Identified
	Head (UD)	Head (LR)	Hand	Body	
Explanation (with the arm)	P-S-B	U-F	**Yes**	P-S-B	No
Writing (on the blackboard)	**B**	F-F-F	**Yes**	**B**	Yes
Silence	P-S-B	F-F	**No**	P-S-B	Yes
Looking (at the blackboard)	**S-B**	**U-F**	**No**	P-S-B	Yes
Media display	P	D	No	P	Yes

(II) Bone information

Human skeleton information is obtained by posture estimation technology, and this paper uses Openpose [24] as a tool. Human skeleton information refers to the coordinate information of human joints, which is concise, so it is often used in human behavior recognition. Similarly, because there is less information when it is missing, it will have a greater impact on the recognition effect. Therefore, considering the serious occlusion between students, the small target in the back row, and the serious lack of skeletal information, this paper only designs the skeletal features of teachers' behavior. A human body skeleton diagram is shown in Figure 3.



Keypoints Number	Keypoints Name
0	Nose
1	Throat
2、5	Shoulder
3、6	Elbow
4、7	Hand
15、16	Eye
17、18	Ear

Figure 3. Skeleton diagram of the human body.

The teacher's local posture and action are the two targets we recognize, so the bone features are designed to correspond to the local posture features of RGB images, as shown in Table 5.

Table 5. Coding table of teacher's body posture based on bone information.

Local Part of Human Body	Feature Description
Head	1. Rotation degree of lower head 2. Rotation degree of left and right shaking heads
Hand	3. Hand height 4. Arm bending degree
Body	5. Shoulder degree

3.2. Model Architecture

3.2.1. Method of Classroom Behavior Recognition

Student Behavior Recognition Method. This paper puts forward two recognition schemes for students' behavior.

- Using the student data obtained by the first labeling method (labeled as posture category and position), Faster R-CNN [25] is directly trained to recognize and locate the local posture of the students, and then the behavior is determined by combining the posture;
- Using Faster R-CNN as a human target detector, and additionally using VGG16 [26] as a behavior classification model. Use the VOC 2007 dataset [27] to train a Faster R-CNN target detection network, so that it can identify the human body area (no behavior recognition is involved), and use the student data obtained by the second labeling method (labeled as behavior category) to train the VGG16 classification model.

Framework Design of Teacher's Behavior Recognition. This paper presents a recognition method based on human body local posture, which involves the fusion of multiple information and the use of a feature pyramid (FPN) [28] and convolutional block attention module (CBAM) [29]. In order to verify the effects of different combinations, this paper designs nine different schemes, which are divided into two groups. The first group only involves the recognition of movements, and the second group includes the simultaneous recognition of four postures and movements (here, head 1 refers to the head posture in the left and right directions, while head 2 refers to the head posture in the up and down directions):

Scheme 1: The VGG16 model is used for transfer training to identify teachers' actions. It is a single-target output network using RGB image information, which can be used as a benchmark;

Scheme 2: Using the coordinate axis of bone information change and normalized 15 coordinate points as feature input, and using the random forest classifier, the purpose is to observe the function of single bone information;

Scheme 3: Action recognition using hand-designed skeletal features with the aim of comparing with the original coordinate information;

Scheme 4: VGG16 is used as the backbone network, and after the pyramid is built with the output characteristic maps of each layer, the P2 layer is used as the final output. Refer to Figure 4 for details.

Scheme 5: It is the same as the pyramid construction scheme of Scheme 4, but uses P2-P5 multi-layer fusion features to identify actions, in order to form a contrast with Scheme 4 and explore the functions of different layers of features, and its structure is shown in Figure 4 (VGG16 has not drawn all the structures to simplify the drawing, and the extracted features of VGG16 are C2–C5);

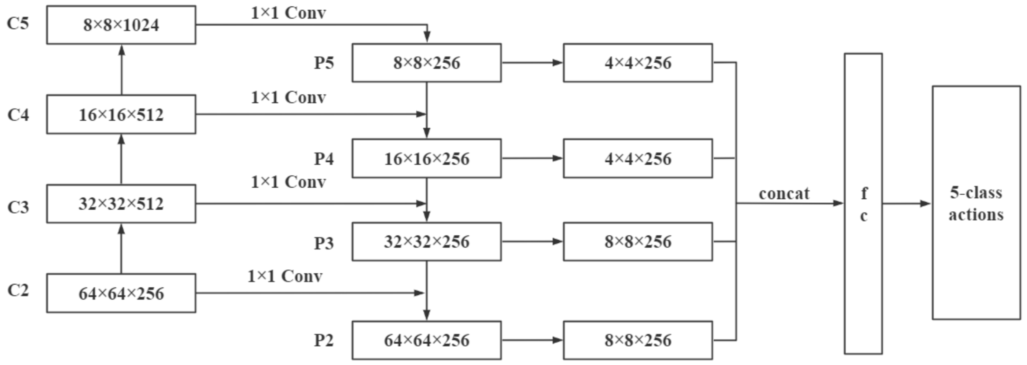


Figure 4. Scheme 5 network structure.

Scheme 6: Directly change the output of VGG16 to multi-channel as a benchmark for multi-objective tasks. Considering the influence of different local poses on actions, multi-output networks can realize feature sharing between local poses and actions;

Scheme 7: The feature maps of different scales using FPN technology are used as the information for target recognition of different scales. The larger the target, the higher the semantic level, while the smaller the target, the lower the semantic level. FPN technology makes the low-level features merge with the high-level information. As shown in Figure 5, the overall movement is relatively large in scale, so it is recognized at the upper level, while the head posture is small and recognized at the lower level. In addition, the network uses the horizontal CBAM module, which strengthens the feature information for different targets;

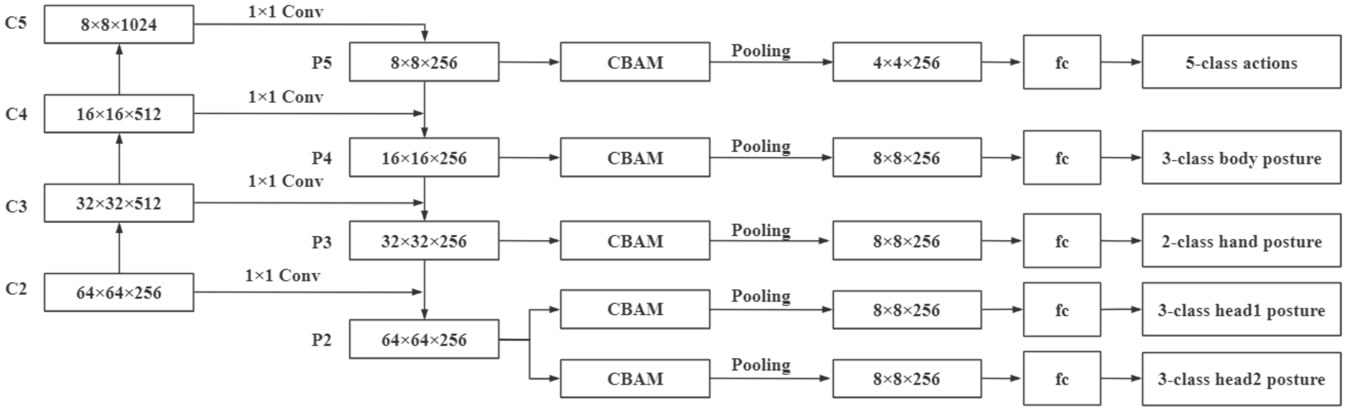


Figure 5. Scheme 7 network structure.

Scheme 8: The overall structure of the network is shown in Figure 6, which realizes bottom-up feature extraction and top-down feature fusion. Four pipelines are designed, and CBAM is added between every two levels to enhance feature expression, which can fuse all important information from high level to low level for each local gesture and action, and each pipeline is relatively independent. In the network's output, the features of the action pipeline will fuse the features of the other four local gestures to provide as effective information as possible for the action recognition. This design embodies the idea that the local gestures will affect the action representation in this study;

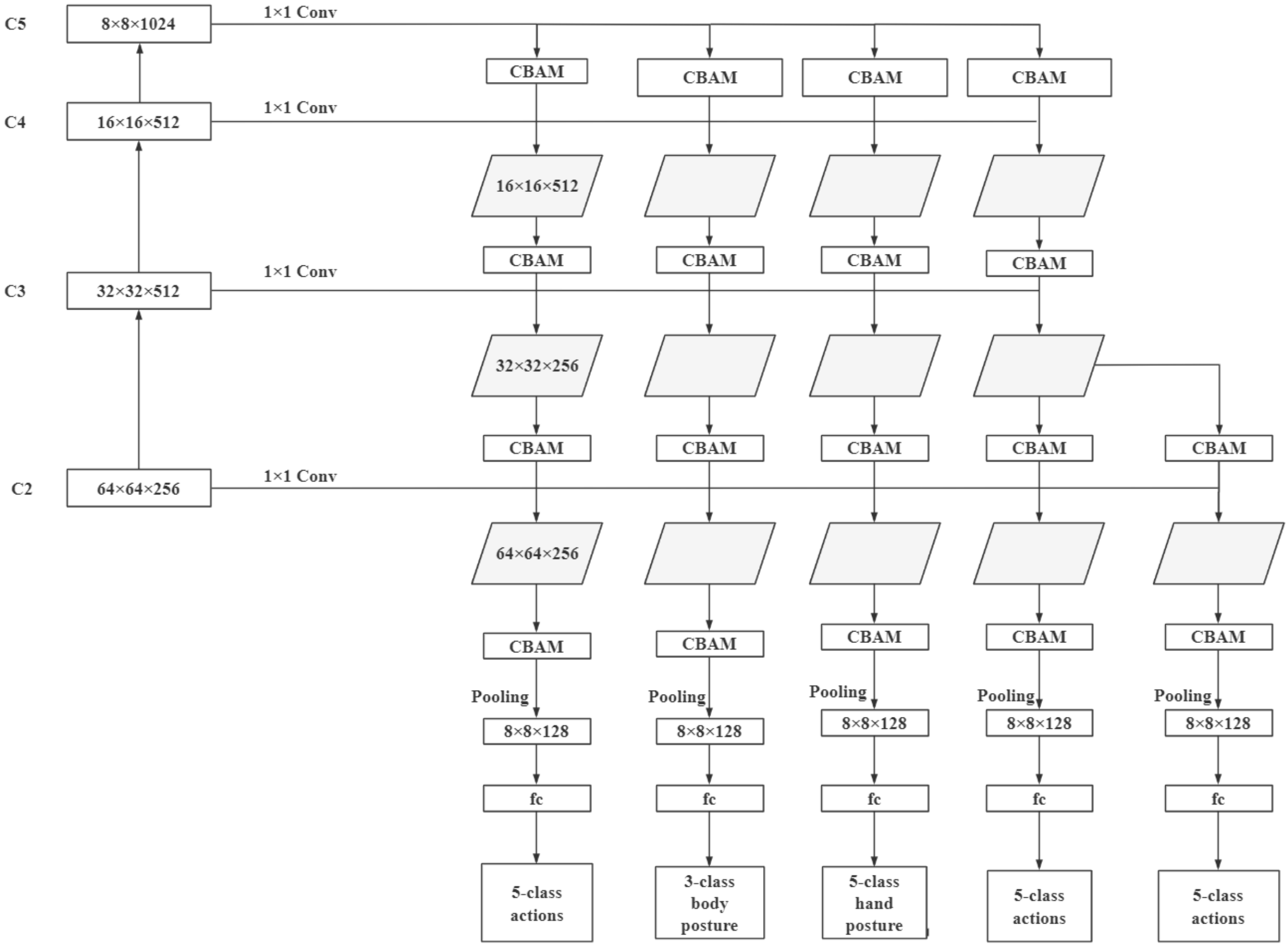


Figure 6. FPN-CBAM four-channel model.

Scheme 9: The model of Scheme 8 and the model of Scheme 2 are integrated for decision-making, with a weight of 1:1. Used to explore the effect of multi-information fusion.

3.2.2. Classroom S–T Behavior Recognition Process

The dataset of this paper comes from the actual blended synchronous classroom videos, including 100 local class recording classes in primary schools, junior high schools, and senior high schools. Due to the limitation of data sources, the perspective of classroom video in this paper is that the camera is facing the students. The camera uploads a frame of classroom images to the server every 30 s, so the video data in this paper is actually an image sequence of 80 frames per class, and the image resolution is 1920*1080 full HD. The dataset used in this paper is highly consistent with the blended synchronous classroom and face-to-face classroom scenarios.

This experiment is implemented on a Windows 10 operating system device, 64 GB memory, Nvidia Geforce GTX 3080 graphics card and Intel(R) Xeon(R) CPU E5-2620 v4, based on the Python language and Tensorflow deep learning framework. Figure 7 is the flow chart of the S–T behavior recognition experiment in this paper. This experiment process is mainly divided into three steps. The first and second steps are the training of an object detection model and classification model, and the third step is the image detection and judgment by using detection and classification model and S–T behavior discrimination rules.

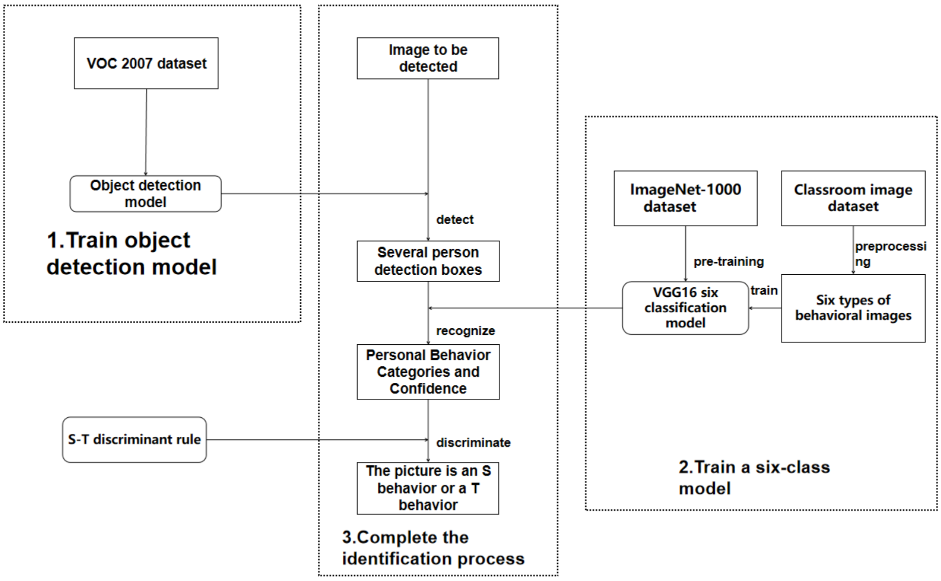

VOC 2007 dataset
Object detection model
1.Train object detection model
Image to be detected
detect
Several person detection boxes
recognize
Personal Behavior Categories and Confidence
discriminate
The picture is an S behavior or a T behavior
3.Complete the identification process
S-T discriminant rule
ImageNet-1000 dataset
Classroom image dataset
pre-training
preprocessing
VGG16 six classification model
train
Six types of behavioral images
2.Train a six-class model

Figure 7. Experimental process.

In the first step, we use the VOC 2007 dataset to train a Faster R-CNN model. Since the VOC 2007 dataset contains 20 categories, we only need the category person. Therefore, properly adjusting the algorithm only detects the person category, which can reduce the detection time.

In the second step, one frame of teaching video of 100 classes is taken at intervals of 30 s, and the positions and behaviors of students in the images are marked. The improperly marked and ambiguous pictures are eliminated, and the data are expanded (rotation, Gaussian blur, brightness adjustment, etc.) to obtain six kinds of behaviors, each with 3000 pictures. The VGG16 model pre-trained with ImageNet-1000 data set can be used for migration training. We divide 18,000 images into a training set and verification set according to the ratio of 8:2, and use 100 images in each category as the test set. The training set is used to train the model, and the test set is used to test the generalization of the model. The learning rate is 0.0005, with 25,000 rounds of six-category model training and 15,000 rounds of two-category model training. Train a VGG16 six-classification model, the purpose of which is to preliminarily judge what kind of behavior students have. Five binary classification models (listening and speaking are relative to the other five categories) are used for further discrimination when the two behaviors are seriously confused. The training process is the same as that of the six classification models.

In the third step, an image to be detected is input, and several person detection frames are output through the Faster R-CNN detection model. The images in these detection frames are input into a VGG16 classification model, and the behaviors and scores in the target detection frames can be obtained. Then, the S–T behavior discrimination rule is used to judge the category of the S–T behavior of the image to be detected.

4. Results

4.1. Recognition Results

4.1.1. Comparison of Two Schemes for Student Posture Recognition

Through the experiments of each test set, the recognition rate of Scheme 1 is 91.25%, and that of Scheme 2 is 82.5%. From the results, it can be seen that the recognition method based on a gesture in Scheme 1 is more accurate than that in Scheme 2. However, because the reasoning from gesture to behavior is combined with the classroom scene (assisted by the rule of artificial experience summary), strictly speaking, the results of Scheme 2 are not all due to the classification model. It can also be seen from here that, in addition to improving the performance of model recognition, the general rules based on specific scenes can effectively assist machines in solving recognition problems.

Table 6 is the mixed matrix of the recognition results of Scheme 1. We can observe the recognition of each behavior. It can be found that the accuracy rate of other behaviors except "writing" is high, especially the behaviors of "speaking", "listening", "reading", and "discussion". This shows that, as long as these behaviors are detected, the probability of misjudgment is low. In fact, 19.9% of the predicted "writing" behaviors are "listening" behaviors, which indicates that it is not enough to recognize this behavior only by the head angle. Because of the camera angle, the head angle fluctuates, and the boundary between bowing and looking up is not clear, so it is easy to make prediction errors when the bow angle is not large. According to the comprehensive F1 value, listening and discussing have the best recognition effect, followed by speaking and writing, while the recognition rates of raising hands and reading aloud need to be improved. The recognition of "raising hands" and "reading" depends on the posture of hands and body, and being easily blocked is the main reason for the poor recognition effect, while other behaviors depend more on the posture of head, which is much better.

Table 6. Mixed matrix of student behavior recognition.

		True Value								
		Raising Hands	**Speaking**	**Listening**	**Writing**	**Reading**	**Discussion**	**Accuracy Rate**	**Recall Rate**	**F1 Value**
PV	Raising hands	**0.828**	0	0.109	0	0	0.063	0.828	**0.453**	0.586
	Speaking	0	**0.931**	0	0	0	0.069	0.931	0.73	0.818
	Listening	0.038	0.006	**0.934**	0.01	0.012	0	0.934	0.954	0.944
	Writing	0.027	0.003	0.199	**0.756**	0.014	0	0.756	0.94	0.838
	Reading	0	0	0	0	**1**	0	1	**0.475**	0.644
	Discussion	0.005	0	0	0	0	**0.995**	0.995	0.973	0.984

4.1.2. Analysis of Teacher's Action Recognition

Table 7 shows the recognition results of nine schemes. From the comparison of Schemes 1 and 2, it can be concluded that RGB information has more noise than bone information, so a feature extraction method is needed to extract effective information from it. However, it is obvious that the feature extracted by VGG16 is not as effective as the bone information. It can be seen from Schemes 2 and 3 that the bone information (BCI) itself has very important discriminant information. After the artificial design, the accuracy rate of nine bone features reached 0.8, which shows that these feature designs are effective, but some key information is still missing, so the effect is not as good as the original bone coordinate information. Therefore, it is a direction to dig more discriminating bone features. The comparison of Schemes 1 and 4 directly shows the excellent performance of FPN technology. Compared with Schemes 4 and 5, it is found that the recognition rate of multi-layer features is not as good as that of one-layer features. It may be that multi-layer features bring more information but also noise.

From the comparison between Scheme 6 and Scheme 1, it can be seen that the attitude information does have an impact on the performance of the action. In Scheme 6, action recognition shares the characteristics of attitude recognition, and the performance is slightly improved. From the comparison between Scheme 7 and Scheme 6, it is found that, after adopting FPN and CBAM technology, the recognition of motion is improved, but the recognition of attitude is not. This shows that the CBAM module has a certain function, but the effect of attitude recognition is not significantly improved due to its single-layer prediction. In addition, there seems to be a relationship between different channels. When one kind of attitude recognition rate increases, the other kind will decrease. The effect of Scheme 8 is the best among all schemes, which shows the effectiveness of the model design. From the comparison of Schemes 8 and 9, it is found that decision fusion (action only) has no better effect. This is because there is little difference in the recognition effect between the two schemes, and the complementary information has no obvious improvement effect on the wrongly classified samples. More fusion methods are worth exploring.

Table 7. Recognition results of all schemes.

Schemes No.	Scheme Details	Recognition Accuracy				
		Action	Head1	Head2	Gestures	Body
1	VGG16	0.77	\	\	\	\
2	Bone coordinate information+Random forest	0.83	\	\	\	\
3	Manually designed bone features+Random forest	0.8	\	\	\	\
4	VGG16+FPN+P2 Forecast	0.81	\	\	\	\
5	VGG16+FPN+P2-P5 fusion prediction	0.79	\	\	\	\
6	VGG16+Multi-output	0.78	0.86	0.78	0.91	0.87
7	VGG16+FPN+CBAM+P2-P5 Separate prediction	0.81	0.87	**0.8**	0.9	0.85
8	VGG16+ FPN+Multilayer CBAM	**0.84**	**0.91**	0.79	**0.91**	**0.91**
9	VGG16+FPN+Multilayer CBAM+Bone Model Decision Fusion	0.84	0.91	0.79	0.91	0.91

In addition, we find that the recognition effect of the head (up and down direction) is obviously lower than that of the other three local poses, which indicates that the features of this local pose are not as discriminating as those of the other three, and it is more difficult for the model to capture the effective feature information of this pose. On the whole, the recognition results of all schemes still can not reach the ideal results. Too few training samples and similar information among action categories may be the main reasons for the poor recognition results.

4.2. Classroom S–T Behavior Recognition and Analysis

4.2.1. Evaluation Indicators

The performance of the recognition system can be measured by several indicators, such as accuracy, recall, F value, etc. In this paper, the accuracy rate is used, and the calculation method is as shown in the Formula (1). In this formula, True Positive (TP): predicts positive classes as positive classes; True Negative (TN): predicts negative classes to negative classes; False Positive (FP): Type I error in predicting the number of negative classes as positive classes; and False Negative (FN): Predict positive classes to negative classes → Type II error:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

For example, if there are 30 S behaviors and 40 T behaviors in a class, 20 S behaviors are judged correctly, and 30 T behaviors are judged correctly. Then, the accuracy rate is (20 + 30)/(30 + 40) = 71.4%.

4.2.2. Training and Test Results

We trained one six-classification model and five two-classification models, and the accuracy of the models is shown in Table 8.

Table 8. Accuracy of the classification model.

Classification Model	Six Categories	Listening Discussion	Listening Reading	Listening Writing	Listening Raising Hand	Listening Speaking
Accuracy Rate	82.50%	93.90%	75.60%	80.40%	88.40%	88.90%

Combined with Figure 8, in the S–T recognition task of a real scene, there is interference from desks, chairs, and other students around individuals, so there is a big gap between the accuracy rate of six categories and the accuracy rate of 97.0% of single behavior recognition. Among the two classification models, the recognition rate of listening–discussion is the highest because students turn sideways and turn around during the discussion, which has a large division with the posture of sitting up while listening. Listening–reading has the lowest recognition rate, which may be due to the fact that the head angle is different from that of listening, and the two are easily confused. In addition, the recognition rate of listening–raising hands and listening–speaking is high, but the three behaviors are easy to confuse in practice. This is because the behaviors in the test data are relatively standard, and the situation of sitting posture misconduct often occurs in the actual classroom.

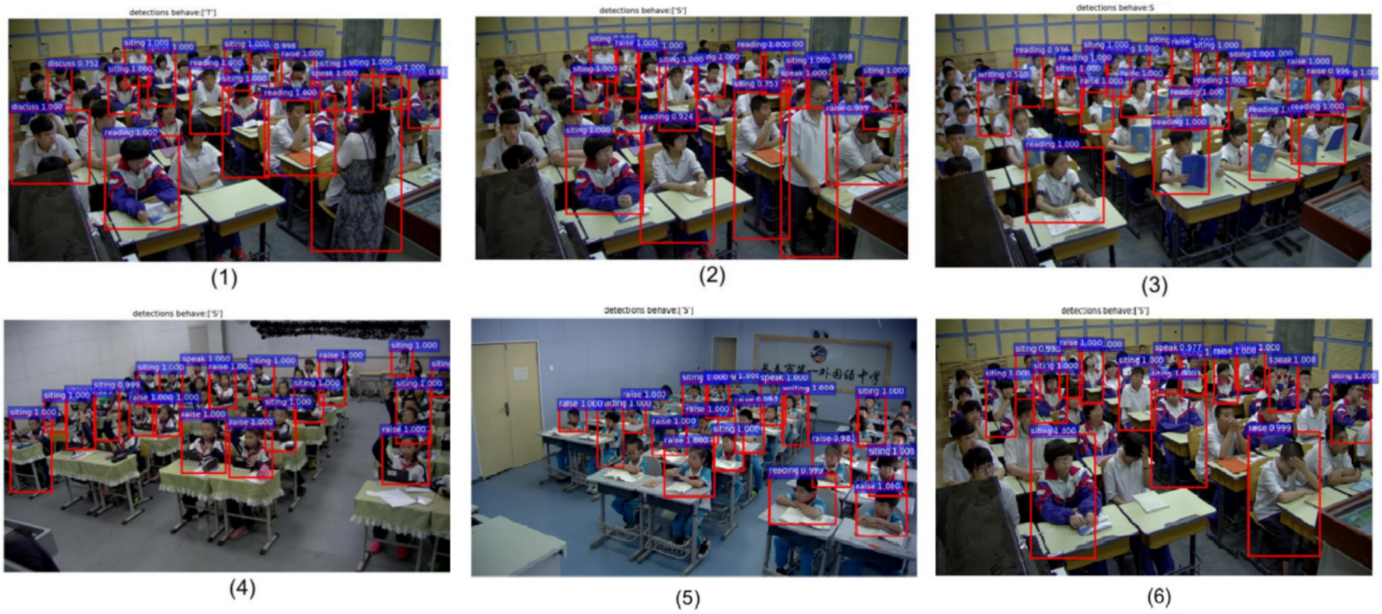


Figure 8. Detection and recognition in real scenes ((**1**–**6**) show the recognition results of different images).

We use the VGG16 classification model to recognize the behavior of the detected person class and use S–T behavior discrimination rules to judge S–T behavior.

Combining Figure 8 and Table 9, we find that, although there are some misjudgments in the correctly judged pictures, the reasons for the judgment are consistent with the actual situation, as shown in Figure 8(1–3). There are also cases where the judgment results are correct, but the reason is inconsistent with the actual situation, as shown in Figure 8(4). In addition, from the pictures with wrong judgment, we found that listening is easy to be mistaken for raising hands. In Figure 8(5), most middle school students are prone on the table, and in Figure 8(6), middle school students' casual sitting posture and random placement of their hands are the reasons for the wrong recognition. At this time, despite the existence of judgment rules, the final result still appears wrong. This shows that our S–T behavior judgment rule does reflect the actual situation to some extent and improves the recognition accuracy, but it can not completely correct the recognition errors.

We randomly selected 10 classes for testing, and the average accuracy rate of 10 classes reached 77.2%. In addition, combined with S–T analysis method, we divide classroom types into lecture type, practice type, conversational type, and mixed type so as to explore the influence of machine recognition results on S–T analysis. As shown in Table 10, we find that the correct rate of class type judgment is 80%, of which 125 and 149 are wrong in class prediction because of the low accuracy of S–T behavior recognition. We also selected a class with correct and wrong predictions to draw S–T curves, as shown in Figure 9. The accuracy of S–T behavior recognition in Class 16 is higher, and the S–T curve is closer to the real curve.

However, the students in Class 149 are more densely distributed, the camera resolution is lower, and the camera position is lower, which makes the overlap between students more serious. These reasons make the recognition rate in Class 149 much lower than that in Class 116.

Table 9. Comparison between manual and machine judgment results of classroom S–T behavior.

Picture	Judgment Reason	Forecast	Artificial Judgment Reason	Actual
1	Listening < 60% and speaking = 0	T behavior	Listening behavior dominates.	T behavior
2	Listening < 60% and speaking > 0	S behavior	Have the behavior of speaking.	S behavior
3	Reading and writing > 25%	S behavior	Have the behavior of reading.	S behavior
4	Listening > 60% and speaking > 0	S behavior	Have the behavior of raising hands.	S behavior
5	Listening < 30%.	S behavior	Listening behavior dominates.	T behavior
6	Listening > 60% and raising hands> 0	S behavior	Listening behavior dominates.	T behavior

Table 10. Actual test table of the algorithm.

Class No.	Accuracy Rate	Actual Classroom Type	Forecast Classroom Type
32	75.80%	Mixed type	Mixed type
35	71.10%	Conversational type	Conversational type
48	76.60%	Conversational type	Conversational type
102	79.10%	Mixed type	Mixed type
109	80.50%	Conversational type	Conversational type
116	85.70%	Mixed type	Mixed type
125	66.70%	**Mixed type**	**Conversational type**
141	80.00%	Practice type	Practice type
148	85%	Practice type	Practice type
149	71.70%	**Mixed type**	**Conversational type**

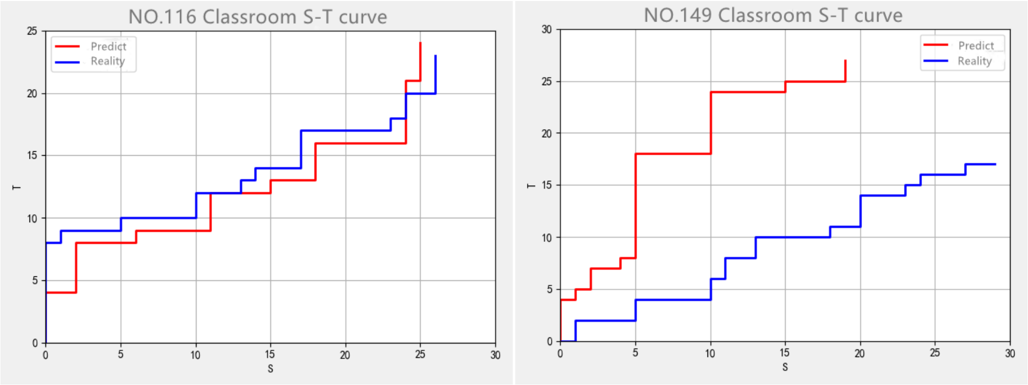


Figure 9. S–T curves of two classes.

Combined with the actual situation of the classroom with wrong prediction, the reasons for the wrong classroom behavior recognition are summarized as follows:

- The irregular posture of students affects the behavior recognition, such as the confusion between holding chin and raising hands, and the confusion between whispering and discussion;
- For the time being, the intelligent recognition method in this paper can not effectively distinguish the categories with similar characteristics, such as the difference of the

head angles between writing and listening, and the characteristics of students' pen holding in the writing behavior of images are too small;

- The image quality is not high, and the behavior is blocked. We obviously find that the recognition accuracy of the students sitting in the front row is higher than that in the back row. On the one hand, the students in the back row are blocked seriously; on the other hand, the students in the front row occupy more pixels in the image, so the display effect is better.

To sum up, the method in this paper has a better recognition effect in the classroom with fewer students, standardized students' posture, and better camera imaging quality.

4.3. Classroom Comparative Analysis Based on Behavior Recognition

Combined with the automatic classroom behavior recognition system, we performed behavior recognition and statistics on the teaching videos of five classes. It can be seen that there are significant differences in the distribution of students' behaviors such as raising their hands, reading and writing, and speaking in different teachers' classes (Figure 10). There are significant differences in the distribution of teachers' own behaviors, such as patrolling and writing on the blackboard (Figure 11).

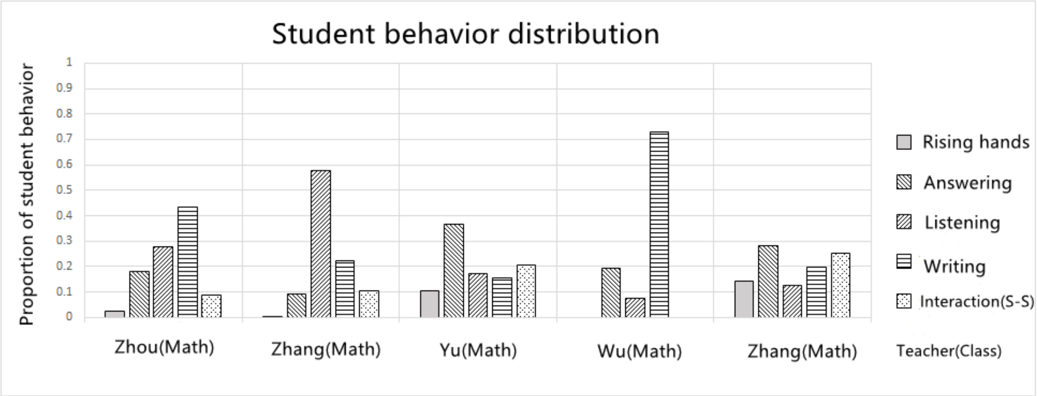


Figure 10. Comparison of students' behavior distribution in class.

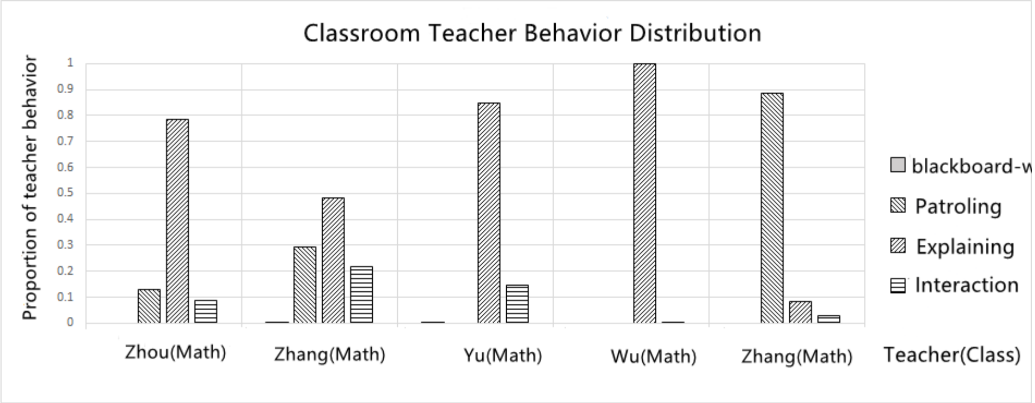


Figure 11. Comparison of teachers' behavior distribution in class.

Through the statistics of S–T behaviors in five classes, draw the Rt-Ch chart of five classes (Figure 12) and quickly compare the differences in teaching modes in different classes.

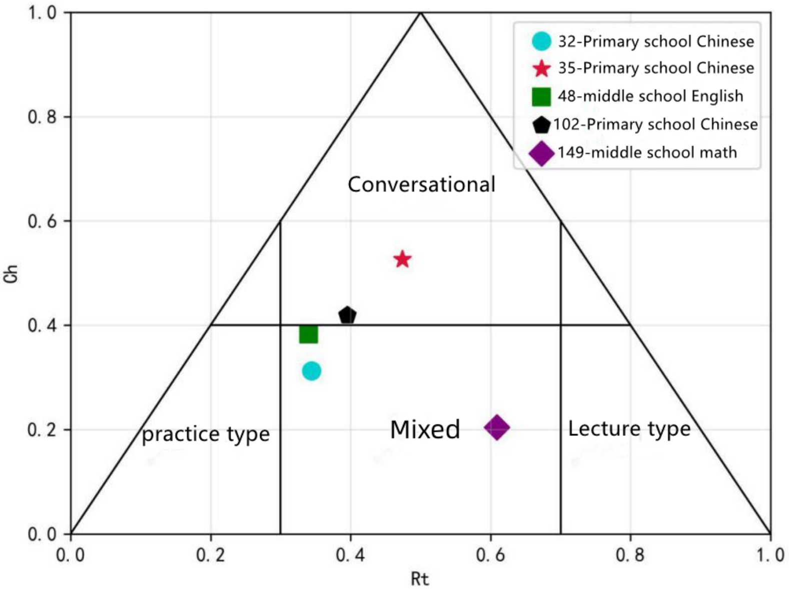


Figure 12. Comparison of classroom teaching modes.

4.4. Statistics and Visualization of Teaching Input in Classroom Teaching Videos

Classroom learners' explicit behavior includes interaction with teachers, peers, and the teaching environment, which reflects learners' internal cognition and emotion. In order to enhance the interpretability of students' behavior engagement, this paper uses students' head, hands, and body postures to identify students' behavior and predicts their behavior engagement. Based on these three nonverbal cues, the evaluation system, labeling method, and analysis method of behavioral engagement are established.

4.4.1. Evaluation Method of Engagement Degree

We recruited 22 students majoring in educational technology to score the engagement of seven main behaviors. If the scores of several raters for a certain kind of behavior are consistent, the engagement reflected by this behavior can be measured by a specific range of values. Firstly, the behavioral engagement E is divided into four grades, $E \in \{0--3\}$. Single pictures of 7 kinds of behaviors were selected, with 5 pictures of each kind, and 5 pictures showed the same kind of behaviors differently, with a total of 35 pictures. After the disorder, 22 students scored according to the images. After screening the data through SPSS analysis, the α value of 21 raters is 0.974. For each category, the α value of raters is all above 0.9, which indicates that the scores of various behaviors have high reliability. Finally, the average engagement level of each kind of behavior is taken as the final engagement value, and 0–1 is defined as low engagement, 1–2 as medium engagement, and 2–3 as high engagement. For the convenience of drawing and scoring, the scores are normalized as shown in Table 11.

Due to the space, this paper focuses on the exploration of S–T behavior recognition and only makes simple statistics on the automatic recognition data of classroom behavior. Data mining and detailed interpretation based on automatic recognition data of classroom behavior will be the next step of this research.

Table 11. Main behaviors and investment.

Behavior Categories	Raising Hands	Writing	Reading	Speaking	Discussion	Listening	Inattention Behavior
Engagement Score	2.8	2.6	2.5	2.4	2.3	1.4	0.7
Normalized Score	1	0.9	0.86	0.81	0.76	0.3	0

In the image, head posture, hand posture, and body posture, as the main body of behavior, is the most important information basis for engagement rating. Although there is a high consistency among the raters for each behavior, there are still subtle differences within the behavior. For example, when the discussion takes place, the raters give different grades to the students who turn around, sit in, and take a positive posture, which shows that the granularity of behavior and posture division needs to be refined and more visual information such as eyes and expressions are also subconsciously concerned by the raters.

4.4.2. Personal Real-Time Investment and Personal Phased Investment

Figure 13 shows the real-time change curve of five students' engagement in the whole class. We find that the engagement of No. 43, No. 29, and No. 36 is higher because their classroom performance is more active, and they raise their hands and speak many times, so their individual engagement in stages is higher, while No. 5 and No. 41 not only raise their hands and speak less, but also do not engage many times, so their individual engagement in stages is lower.

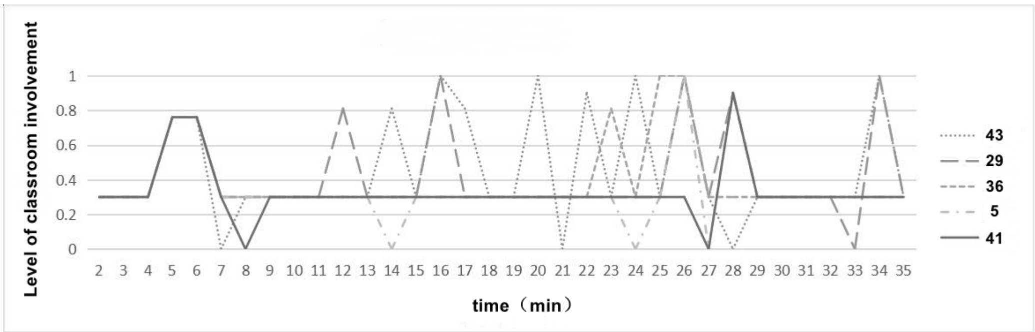


Figure 13. Behavior engagement curve of five students in class 121.

4.5. Class Real-Time Engagement

Figure 14 shows the real-time change curve of class engagement in Class No. 30, Class No. 121, and Class No. 51, with phased engagement of 0.41, 0.36, and 0.34. The appearance of an engagement peak (above 0.7) often represents the development of group activities, such as discussion, reading aloud, and answering (raising hands). The appearance of an engagement center (0.4–0.7) represents some students' participation in activities. The reason may be that teachers ask questions, but not all students are involved in answering. The appearance of a low engagement peak (0.3–0.4) represents a small number of students participating in activities, which may be due to the fact that only a small number of students raise their hands and speak. Compared with the three curves, the engagement of three classes fluctuates periodically from low level to high level (0.3 is the benchmark), which may be related to the design of teachers' teaching activities. All three teachers pay attention to the introduction of the curriculum and catch the attention of students at the beginning; Classroom No. 21 teachers pay more attention to the interaction with students all the time, and it is difficult to combine questions. There are many centers and peaks of engagement. Class No. 30 has a high peak of engagement and a lack of center, which shows that teachers pay attention to all activities, but it is difficult to ask questions. In the second half of class 51, there was no peak of engagement, and the engagement was lower than the 0.3 benchmark many times. The teacher needs to grasp the classroom rhythm and take care of more students.

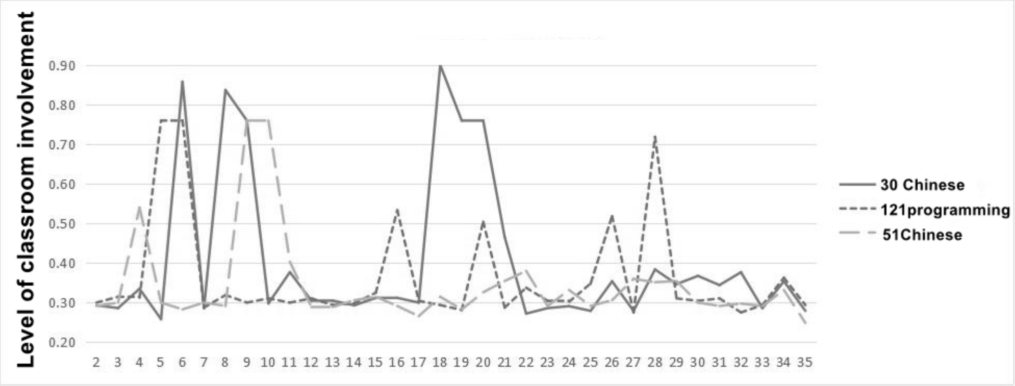
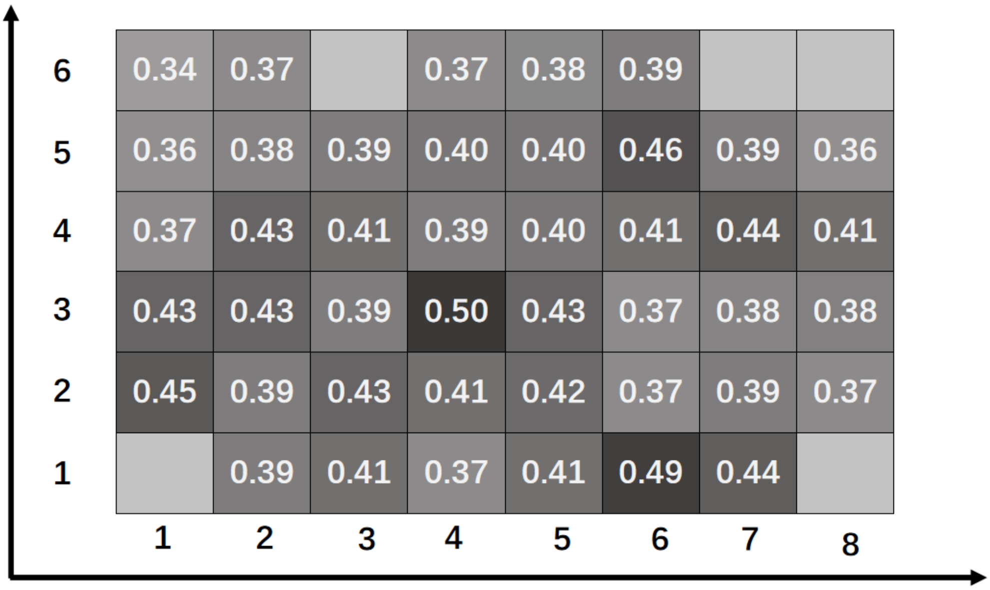


Figure 14. Changes of class engagement.

4.6. Spatial Distribution of Students' Phased Engagement

Figure 15 shows the spatial distribution of students' phased engagement in Class 30. From the heat map, it can be seen that the engagement of students in the front row is higher than that in the back row, and the engagement of students in the periphery is obviously lower than that of students in the center because the periphery students are more likely to be ignored by teachers, and they are not easily noticed when they are not paying attention. This situation may lead to the students' more relaxed mentality.



6	0.34	0.37		0.37	0.38	0.39		
5	0.36	0.38	0.39	0.40	0.40	0.46	0.39	0.36
4	0.37	0.43	0.41	0.39	0.40	0.41	0.44	0.41
3	0.43	0.43	0.39	0.50	0.43	0.37	0.38	0.38
2	0.45	0.39	0.43	0.41	0.42	0.37	0.39	0.37
1		0.39	0.41	0.37	0.41	0.49	0.44	
	1	2	3	4	5	6	7	8

Figure 15. Spatial distribution of students' phased engagement in Class 30.

5. Discussion

Aiming at the education in the synchronous classroom environment and combining with the current popular deep learning methods, this paper adopts the methods of object detection and image recognition to achieve the purpose of intelligent classroom behavior recognition and automatically analyzes the recognition results. The experimental results show that the deep learning method can quickly identify classroom pictures with an accuracy of 77.2%, and has great potential for improvement. This paper provides a reference case for the combination of deep learning and classroom teaching analysis. However, from the practice of this paper, in order to realize the application of automatic recognition and analysis of classroom teaching behavior, there are still the following problems to be solved:

(1) Video data quality is an important factor affecting the performance of classroom teaching behavior recognition. The first is video data recording standards and the es-

tablishment of a basic behavior database. In the process of education informatization, various smart classrooms have been established in different places, but the video recording resolution and camera distribution of these classrooms are different. The density and occlusion of students will also seriously affect the recognition performance. According to the research in this paper, to support the recognition function of deep learning, it is better to have a resolution above 4k so that a higher recognition rate can be achieved when the number of students in class is large. The only data in this paper are videos from the perspective of students, but to achieve complete teaching analysis, it is better to have complete videos from the perspective of students and teachers. The second is the integrity of the data collection. Most classroom teaching behaviors are verbal behaviors, while only nonverbal body behaviors are used in this paper. Voice and speech information, which play an important role in teaching, are not used, and information loss will seriously affect the behavior recognition rate.

(2) Artificial intelligence recognition technology needs to be further enriched and improved. Technically speaking, video-based body behavior recognition has been able to achieve a certain effect, but classroom teaching is a dynamic process, and its behavior before and after has a causal sequence connection. Therefore, in the following classroom teaching recognition research, contextual information and timing information should be fully used so as to further improve the recognition effect under the condition of simplifying the rules. In addition, the performance of classroom behavior recognition based on deep learning still has great room for improvement.

(3) Classroom teaching behavior classification and coding rules are the basis for automatic recognition and analysis of classroom teaching behavior. The teaching behavior classification table adopted in this paper is adapted from the classical teaching behavior classification table. However, for many classroom teaching behavior analysis tasks, the behavior and action categories of this classification table are insufficient. In the field of artificial intelligence, public behavior standards and behavior classification tables should be established as far as possible. Different from standardized data collection in the field of artificial intelligence, classroom teaching behavior analysis has various objectives and forms. Therefore, it is necessary to study and establish a targeted classification table of teaching behavior according to different objectives of teaching behavior analysis. In addition, on the basis of the classification of teaching behaviors, it is also necessary to study the mapping relationship between body movements and teaching behaviors in classroom teaching so as to establish a theoretical basis for automatic recognition and discrimination of classroom behaviors.

(4) The application of automatic analysis of classroom teaching behavior needs to be explored. The automatic recognition and analysis of classroom teaching behavior based on artificial intelligence is a new thing with the development of technology in recent years. This paper only makes a simple statistical comparison of teaching behavior. For classroom teaching research, the automatic analysis of classroom teaching behavior has a broad prospect, such as being able to be used to extract the classroom teaching behavior norms or according to the characteristics of classroom teaching behavior to predict the achievement or to evaluate the teaching behavior. However, for more front-line teachers and principals, the application demand for automatic recognition and analysis of classroom teaching behavior still needs to be further investigated and studied.

6. Conclusions

This paper explores the intelligent recognition of teacher–student behavior in blended classroom teaching scenarios. Based on the above problems, we propose a new machine learning S–T classification method and auxiliary discriminant rules. Faster R-CNN object detection network and VGG16 deep convolutional neural network are combined to study S–T behavior recognition of classroom video images. This study can be summarized into four parts:

(1) Since the traditional S–T behavior classification method has the problems of intra-class ambiguity and semantic ambiguity of S and T behaviors, we re-classify the S–T behaviors to eliminate the ambiguity and fuzziness between classroom physical behaviors;

(2) Formulate auxiliary discrimination rules to improve the error tolerance rate and accuracy of automatic action recognition;

(3) Train object detection models and action classification models for classroom action recognition. Its student posture recognition recognition rate reached 91.25%, and its teacher recognition network designed four pipelines, adding CBAM between every two levels to strengthen feature expression, which could fuse all important information from high-level to low-level for each local posture and action, and each pipeline was relatively independent, and its recognition rate reached 87%;

(4) A method to calculate classroom engagement based on students' head posture, hand posture, and body posture was proposed. The results show that it has a good target detection rate, recognition accuracy, and target tracking performance. The experiment shows that the proposed automatic recognition technology can calculate the S–T type of class and the real-time and stage engagement of individual and class.

Author Contributions: Conceptualization, T.X., W.D. and Q.L.; methodology, T.X., W.D. and S.Z.; software, T.X., W.D. and Y.W.; visualization, S.Z. and Y.W.; writing—original draft preparation, T.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by The Humanities and Social Science Research Planning Fund of the Ministry of Education (No. 20YJA880009), the key project of the Basic Education Research Association of Normal Universities "Internet + Education and Teaching Reform Practice Research" (No. CCNUTEIII 2021-09), and the AI and Faculty Empowerment Pilot Project CCNUAI&FE2022-03-24.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, J.; Yu, H.; Chen, N. Using blended synchronous classroom approach to promote learning performance in rural area. *Comput. Educ.* **2019**, *141*, 103619. [CrossRef]
2. Chirinda, B.; Ndlovu, M.; Spangenberg, E. Teaching mathematics during the COVID-19 lockdown in a context of historical disadvantage. *Educ. Sci.* **2021**, *11*, 177. [CrossRef]
3. Pascarella, E.T.; Seifert, T.A.; Blaich, C. How effective are the NSSE benchmarks in predicting important educational outcomes? *Chang. Mag. High. Learn.* **2010**, *42*, 16–22. [CrossRef]
4. Chen, L.; Chen, P.; Lin, Z. Artificial intelligence in education: A review. *IEEE Access* **2020**, *8*, 75264–75278. [CrossRef]
5. Bell, J.; Sawaya, S.; Cain, W. Synchromodal classes: Designing for shared learning experiences between face-to-face and online students. *Int. J. Des. Learn.* **2014**, *5*, 1. [CrossRef]
6. Bower, M.; Dalgarno, B.; Kennedy, G.E.; Lee, M.J.; Kenney, J. Design and implementation factors in blended synchronous learning environments: Outcomes from a cross-case analysis. *Comput. Educ.* **2015**, *86*, 1–17. [CrossRef]
7. Norberg, A. Blended learning and new education logistics in Northern Sweden. *Educ. Publ.* **2012**, *1*, 327–330.
8. Huang, Y.; Zhao, C.; Shu, F.; Huang, J. Investigating and analyzing teaching effect of blended synchronous classroom. In Proceedings of the 2017 International Conference of Educational Innovation through Technology (EITT), Osaka, Japan, 7–9 December 2017; pp. 134–135.
9. Jun, Y.; Huanghai, O. ST analysis of classroom teaching. In Proceedings of the 2011 6th International Conference on Computer Science & Education (ICCSE), Singapore, 3–5 August 2011; pp. 136–140.
10. Xu, L.; He, X.; Zhang, J.; Li, Y. Automatic classification of discourse in Chinese classroom based on multi-feature fusion. In Proceedings of the 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), Beijing, China, 28–31 August 2019; pp. 1–5.
11. Wang, X.; Yin, S.; Sun, K.; Li, H.; Liu, J.; Karim, S. GKFC-CNN: Modified Gaussian kernel fuzzy C-means and convolutional neural network for apple segmentation and recognition. *J. Appl. Sci. Eng.* **2020**, *23*, 555–561.
12. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef]

13. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
14. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
15. Jisi, A.; Yin, S. A new feature fusion network for student behavior recognition in education. *J. Appl. Sci. Eng.* **2021**, *24*, 133–140.
16. Zhang, Y.; Wu, Z.; Chen, X.; Dai, L.; Li, Z.; Zong, X.; Liu, T. Classroom behavior recognition based on improved yolov3. In Proceedings of the 2020 International Conference on Artificial Intelligence and Education (ICAIE), Tianjin, China, 26–28 June 2020; pp. 93–97.
17. Wu, D.; Chen, J.; Deng, W.; Wei, Y.; Luo, H.; Wei, Y. The recognition of teacher behavior based on multimodal information fusion. *Math. Probl. Eng.* **2020**, *2020*, 1–8. [CrossRef]
18. Lin, F.C.; Ngo, H.H.; Dow, C.R.; Lam, K.H.; Le, H.L. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors* **2021**, *21*, 5314. [CrossRef] [PubMed]
19. Chonggao, P. Simulation of student classroom behavior recognition based on cluster analysis and random forest algorithm. *J. Intell. Fuzzy Syst.* **2021**, *40*, 2421–2431. [CrossRef]
20. Wu, B.; Wang, C.; Huang, W.; Huang, D.; Peng, H. Recognition of Student Classroom Behaviors Based on Moving Target Detection. *Trait. Du Signal* **2021**, *38*, 215–220. [CrossRef]
21. Mo, J.; Zhu, R.; Yuan, H.; Shou, Z.; Chen, L. Student behavior recognition based on multitask learning. *Multimed. Tools Appl.* **2022**. [CrossRef]
22. Zhou, H.; Jiang, F.; Si, J.; Xiong, L.; Lu, H. StuArt: Individualized Classroom Observation of Students with Automatic Behavior Recognition and Tracking. *arXiv* **2022**, arXiv:2211.03127.
23. Chango, W.; Lara, J.A.; Cerezo, R.; Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1458. [CrossRef]
24. Cao, Z.; Martinez, G.H.; Simon, T.; Wei, S.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
25. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
28. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–17 June 2019; pp. 7036–7045.
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.