applied sciences

MDPI

Article

Research on Students' Action Behavior Recognition Method Based on Classroom Time-Series Images

Zhaoyu Shou [1], Mingbang Yan [1], Hui Wen [1,*], Jinghua Liu [1], Jianwen Mo [1] and Huibing Zhang [2]

[1] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; guilinshou@guet.edu.cn (Z.S.); 20022201053@mails.guet.edu.cn (M.Y.); liujinghua1763@gmail.com (J.L.); mo_jianwen@126.com (J.M.)

[2] School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; zhanghuibing@guet.edu.cn

* Correspondence: huiwen@guet.edu.cn

Abstract: Students' action behavior performance is an important part of classroom teaching evaluation. To detect the action behavior of students in classroom teaching videos, and based on the detection results, the action behavior sequence of individual students in the teaching time of knowledge points is obtained and analyzed. This paper proposes a method for recognizing students' action behaviors based on classroom time-series images. First, we propose an improved Asynchronous Interaction Aggregation (AIA) network for student action behavior detection. By adding a Multi-scale Temporal Attention (MsTA) module and a Multi-scale Channel Spatial Attention (MsCSA) module to the fast pathway and slow pathway, respectively, the accuracy of student action behavior recognition is improved in SlowFast, which is the backbone network of the improved AIA network,. Second, the Equalized Focal Loss function is introduced to improve the category imbalance that exists in the student action behavior dataset. Experimental results on the student action behavior dataset show that the method proposed in this paper can detect different action behaviors of students in the classroom and has better detection performance compared to the original AIA network. Finally, based on the results of action behavior recognition, the seat number is used as the index to obtain the action behavior sequence of individual students during the teaching time of knowledge points and the performance of students in this period is analyzed.

Keywords: action behavior recognition; asynchronous interaction aggregation network; attention mechanism; equalized focal loss

Citation: Shou, Z.; Yan, M.; Wen, H.; Liu, J.; Mo, J.; Zhang, H. Research on Students' Action Behavior Recognition Method Based on Classroom Time-Series Images. *Appl. Sci.* **2023**, *13*, 10426. https://doi.org/10.3390/app131810426

Academic Editor: Emanuel Guariglia

Received: 3 August 2023
Revised: 14 September 2023
Accepted: 15 September 2023
Published: 18 September 2023

1. Introduction

Classroom teaching activities have always been the focus of research in the field of education, and students are the main body of the activities. Recognizing and analyzing students' action behavior plays an important role in teaching evaluation [1]. Student action behavior analysis can help teachers understand the learning process of students and is an important means of measuring learning effectiveness. Traditional methods of analyzing student action behaviors rely on manual assessment, manual recording, and manual coding to collect and interpret behavioral performance data, which have the disadvantages of strong coding subjectivity, small sample size, and high time-consumption [2]. With the development of education informatization, smart classrooms are widely used; so, cameras and other devices in the classroom can be used to obtain massive data such as videos and images in the teaching process. For the acquired massive data, selecting the appropriate sampling scheme and decomposition algorithm can effectively compress the data while retaining the integrity of the data features [3]. Moreover, extracting effective features from the data based on machine learning methods can be used to intelligently recognize students' action behaviors. In recent years, some researchers have used traditional algorithms such as wavelet transform and discrete path transform to extract features from data [4,5], Although

the extracted features are more representative and generalized, the traditional methods still have problems such as manual intervention and poor robustness. In addition, some researchers have also applied fractal theory to the field of image processing [6].

Different from traditional machine learning methods, deep learning automatically learns feature representations of data such as videos and images, which is robust and can be used as an effective tool for detecting students' action behaviors [7]. Tang et al. [8] proposed a student posture detection method based on an improved Faster RCNN object detection model to recognize the behaviors of standing, sitting, and sleeping. Liu et al. [9] proposed an action detection framework based on YOLOv5, which is used to detect the classroom behavior of students in the monitoring system under different backgrounds. Although the above studies achieved good results in student action behavior detection, they did not consider the temporal features of the behavior and the interaction with the spatial contextual environment, and there are limitations in understanding the action behavior based on a single-frame image.

Different from student action behavior recognition based on static images, video behavior recognition can effectively utilize time-series information. Xie et al. [10] proposed a college students' classroom behavior recognition algorithm based on spatiotemporal representation learning, which is used to recognize students' abnormal behaviors such as sleeping and playing mobile phones. However, this type of research is only applicable to action behavior recognition in single-person scenarios, which cannot mark the spatial location of each student and cannot be applied to action behavior detection in multi-person classroom scenarios.

The purpose of spatiotemporal behavior detection is to locate each person in the video and recognize their action behaviors. Many existing studies have improved the performance of action behavior detection by modeling the interactions between the target person and the contextual environment. Students' action behaviors usually interact with other students, objects, etc., around them, and modeling these interactions can be used to recognize the students' action behaviors more effectively. In addition, the background of the classroom scene is complex, the students are crowded, the pixel area occupied by the front row and the back row students in the video is different, and the visual speed of each student's action behavior is also different; so, the extraction of video features needs to be optimized for these problems. Based on the above analysis, this paper proposes a student action behavior recognition method based on classroom time-series image data. Firstly, the improved Asynchronous Interaction Aggregation (AIA) network [11] is used to detect the action behavior of students in classroom videos. Then, based on the recognition results, the action behavior sequence of individual students during the teaching time of knowledge points is obtained and analyzed. In addition, to validate the effectiveness of the proposed action behavior detection algorithm, a student action behavior dataset is constructed and Equalized Focal Loss (EFL) [12] is introduced to improve the category imbalance problem existing in the dataset. The main contributions of this paper are as follows:

a. We propose an action behavior detection model based on an improved AIA network [11]. The Multi-scale Temporal Attention (MsTA) module and Multi-scale Channel Spatial Attention (MsCSA) module are added to the video backbone network SlowFast, which improves the accuracy of students' action recognition.
b. The EFL function [12] was introduced to dynamically adjust the categorization loss weights of different categories to improve the category imbalance problem existing in the dataset.
c. Experiments are conducted on a self-made student action behavior dataset. The experimental results show that the algorithm proposed in this paper improves the mean average precision (mAP) value of action behavior detection. Based on the results of action behavior recognition, the student's seat number is used as the index to analyze the sequence of students' action behavior during the teaching time of knowledge points.

2. Related Work

2.1. Video Behavior Recognition

The research on video behavior recognition is broadly divided into behavior recognition, temporal behavior detection, and spatiotemporal behavior detection. Behavior recognition is to classify the input video; temporal behavior detection needs to detect the start time and end time of the behavior and recognize the behavior of the human in that time; spatiotemporal behavior detection needs to locate the spatial position of the human in the video and recognize the duration and category of the behavior. In this paper, multiple students in classroom teaching videos are used as research objects to detect their action behaviors; therefore, the spatiotemporal behavior detection method is used to detect students' action behaviors in this paper.

Video is different from images, detecting human behavior requires extracting both appearance features and motion features of the frame sequence. Based on these features, researchers have proposed many effective networks to detect human behavior in video. Feichtenhofer et al. [13] proposed a dual pathway network SlowFast, where the Slow pathway and Fast pathway are responsible for the extraction of the appearance features and motion features, respectively. Literature [14] proposed an efficient X3D network that reduces the amount of computation and, at the same time, gives better results in video behavior recognition. Such studies are based on a two-stage approach to recognize human behavior: one stage is used to generate the human bounding box and the other stage is used for behavior recognition. Recently, a researcher proposed a one-stage network to detect human behavior in video [15], which decouples detection and behavior recognition into two branches: one branch is responsible for detecting humans in video and the other is responsible for recognizing the behavior.

With the application of transformer in the image processing field, Fan et al. [16] proposed a Multi-scale Transformer, which combines the multiscale feature hierarchies with the transformer to extract the video features at different levels so that the model can better understand the video content. In addition, behavior recognition methods based on interaction relationship modeling are also widely used. Tang et al. [11] proposed an AIA Network that models human–human interaction, human–object interaction, and temporal interaction. Pan et al. [17] proposed the Actor-Context-Actor Relation (ACAR) Network, which explicitly models higher-order relationships between humans based on their interactions with the background information. Zheng et al. [18] proposed a Multi-Relational Support Network (MRSN), which first models actor–context and actor–actor relationships separately and then models the two types of interactions at the relational level. Faure et al. [19] proposed a multi-modal Holistic Interaction Transformation (HIT) Network, which contains two branches: the RGB branch and the pose branch. The two branches extract appearance features and motion features, respectively; finally, the features extracted from the two branches are fused and input to the classification layer to detect human behavior.

2.2. Behavior Recognition in Classroom Scenarios

In recent years, there have been many scholars applying machine learning to students' action behavior recognition, which mainly focuses on recognizing students' action behaviors through data modalities such as human keypoints, still images, and videos. Therefore, this paper will introduce related research from these aspects.

In the study based on human keypoints, Zhang et al. [20] proposed a method to recognize students' postures in the classroom, which used an improved HRNet network to extract students' keypoints and then used a support vector machine to classify students' classroom behaviors. Zhou et al. [21] extracted the key information of human skeleton from student behavioral images and recognized students' classroom behaviors based on a deep convolutional neural network (CNN-10). Pang et al. [22] improved the traditional algorithm by combining the traditional cluster analysis algorithm and random forest algorithm with the human skeleton model to recognize students' classroom behaviors in

real-time. Ding et al. [23] used OpenPose to extract human keypoints and used a graph convolutional neural network to classify the features and recognize the abnormal behaviors of students.

In their study based on static images, Wu et al. [24] proposed a motion object detection algorithm for student behavior recognition in the classroom, recognizing students' standing behaviors based on the region of interest (ROI) and face tracking, and students' hand-raising behaviors based on skin color detection. Banerjee et al. [25] proposed an improved SSD object detection model to recognize the behavior of students and teachers in the teaching laboratory. Liu et al. [26] improved the two-stage object detection network and proposed a new ROI extractor SAA module and a new detection head RST module for student behavior detection in classroom scenes. Huang et al. [27] constructed a deep neural network to extract facial keypoints, recognize students' head postures and expressions, and classify classroom behaviors by combining the head gestures and expressions. Zheng et al. [28] improved the CBL module of the YOLOv5 model to detect students' classroom behavior from multiple perspectives to evaluate students' classroom attention.

In the study based on videos, Liu et al. [29] proposed a 3D multi-scale residual dense network based on heterogeneous view perception for recognizing students' classroom behaviors. Jisi et al. [30] combined a spatial affine transform network with a convolutional neural network to extract the spatiotemporal features of the video and fused the spatiotemporal features to classify students' behaviors by using a weighted sum method.

Although the above studies have achieved good results in the field of student action behavior detection, there are still some shortcomings: (1) human keypoint-based studies have difficulties in extracting human keypoints in student-intensive classroom scenarios and it is difficult to accurately extract the keypoints of each student; (2) static-image-based studies ignore the temporal features of the action behaviors and do not incorporate time-series context information; (3) video-based studies incorporate the time-series context information of action behaviors but ignore the interaction between students and the context of the environment; these methods only recognize individual action behaviors and do not apply to real multi-person classroom scenarios.

To address the above issues, this paper uses a spatiotemporal action detection model based on interaction relationship modeling to locate students in classroom videos while recognizing their action behaviors and optimizes the model for the difficulties existing in the classroom scenarios; based on the results of action behavior recognition, the student's seat number is used as the index to obtain the action behavior sequence of individual students during the teaching time of knowledge points to help teachers understand the students' learning process and take personalized intervention measures to improve the teaching effect.

3. Proposed Method

The overall system structure diagram of the student action behavior recognition model proposed in this paper is shown in Figure 1. Firstly, the spatiotemporal features of classroom video are extracted based on the backbone network in the improved AIA network. Secondly, students and objects such as books and mobile phones are detected by using the object detection model applicable to the classroom scenario; then, the spatiotemporal features of the video and the spatial coordinate information of the students are fused to detect the student's action behaviors. Finally, based on the action behavior recognition results during the teaching time of the knowledge points, the student's seat number is used as the index to correlate the student's action behavior sequence during the period.

3.1. Detection of Students' Action Behavior

The AIA network [11] has achieved good results in spatiotemporal action behavior detection; however, the network uses a fixed convolutional kernel to learn video features, which cannot capture the spatiotemporal information of multi-scale feature maps to enrich the feature space, and the detection accuracy of the small sample categories is lower in

a dataset with category imbalance. Therefore, in a real classroom environment, the AIA network does not have high accuracy in detecting action behaviors with small sample sizes and insignificant features. To improve the detection accuracy, this paper proposes an improved AIA network: firstly, it incorporates the MsCSA module into the slow pathway of the backbone network SlowFast [13], and incorporates the MsTA module into the fast pathway of SlowFast, which helps the network to extract the multiscale spatiotemporal features; second, the EFL function [12] is introduced to improve the detection performance of action-behavior categories with small sample sizes. The improved AIA network structure is shown in Figure 2 and consists of three parts: The first is part a, which uses an independent object detection network to detect students and objects in video keyframes; next is part b where, in this paper, the MsCSA module is added to the Res2, Res3, and Res4 stages of the slow pathway of the SlowFast network [13] and the MsTA module is added to the Res3, Res4, and Res5 stages of the fast pathway of the SlowFast network to help the network extract multi-scale spatial features and multi-scale temporal features. Finally, the AIA module in part c consists of the Asynchronous Memory Update (AMU) Algorithm and the Interaction Aggregation (IA) structure. The ROI Align algorithm [31] is used to extract the feature P_t of the student bounding box and the feature O_t of the object bounding box; store and read the memory feature M_t online using the AMU algorithm; input P_t, O_t, and M_t into the IAstructure to model multiple types of interaction relationships; and then classify the fused features to detect the student's action behavior.

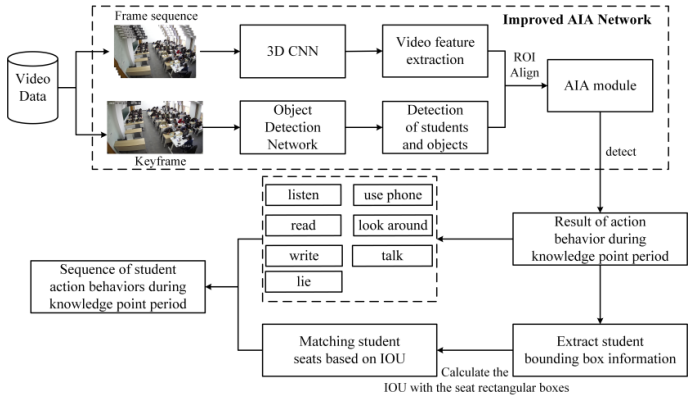

Figure 1. Overall system structure diagram of student action behavior recognition model.

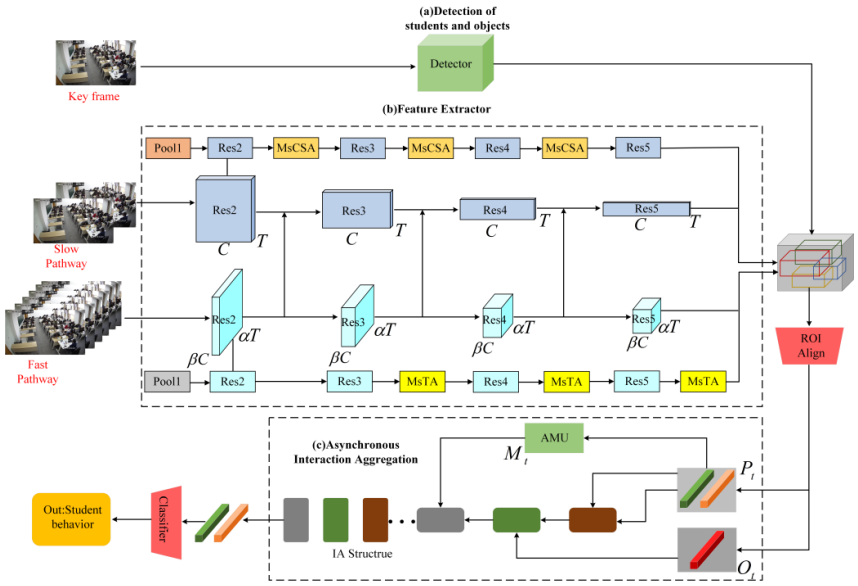

Figure 2. Improved AIA network structure.

3.1.1. Video Backbone Network

The video backbone network SlowFast is the b part of Figure 2. Literature [13] categorizes SlowFast into slow and fast pathways based on the difference in the sampling step of the frames, where the number of convolutional blocks in the residual layer is the same and only the number of convolutional kernels and output channels are different. The sampling step for slow pathway frames is $\tau = 16$, 1 frame every 16 frames, and the number of output channels is C to learn spatial semantic information with fewer frames and a larger number of channels; the sampling step for fast pathway frames is $\tau/\alpha(\alpha = 8)$, 1 frame every 2 frames, and the number of output channels is $\beta C\left(\beta = \frac{1}{8}\right)$ to learn the motion information with a larger number of frames and fewer number of channels. Lateral connections are established with the slow pathway in the Res2, Res3, and Res4 stages of the fast pathway, and the features of the fast pathway are fused with the features of the slow pathway after reconstruction.

Different from the traditional SlowFast network, this paper adds an MsCSA module after the Res2, Res3, and Res4 stages of the slow pathway and an MsTA module after the Res3, Res4, and Res5 stages of the fast pathway to extract multi-scale spatial features and multi-scale temporal features of the classroom teaching video and, thus, to improve the detection of the students' action behavior accuracy. In this paper, these two modules will be introduced in detail in Sections 3.1.4 and 3.1.5.

3.1.2. Feature Extraction

The video clip v_t is taken as input and the video features $f_t \in R^{C\times T\times H\times W}$ are extracted using the backbone network described in Section 3.1.1, where C, T, H, and W are channel, time, height, and width, respectively. The video features f_t are average pooled along the time dimension to obtain the feature map $I_t \in R^{C\times H\times W}$. Meanwhile, the Faster RCNN [32] is used to locate students in the keyframes (i.e., the center frames) of the video clip v_t, as well as objects, such as books and mobile phones, and obtains N_t student bounding boxes and K_t object bounding boxes. Based on the detected bounding boxes, the ROI Align algorithm [31] is used to extract the student features $P_t = \left[p_t^1, \cdots p_t^i, \cdots, p_t^{N_t}\right]$ and object features $O_t = \left[o_t^1, \cdots o_t^i, \cdots, o_t^{K_t}\right]$ of the video clip v_t along the spatial dimension from the feature map I_t. In addition, to model the long-term temporal context between different clips, the student features of multiple clips are deposited into the feature pool using the AMU algorithm and the student features of the neighboring clip are read from the feature pool each time of training; then, they are combined with those of the current clip to form a memory feature $M_t = [P_{t-L}, \cdots, P_t, \cdots, P_{t+L}]$, where L is the size of the time window. Thus, the memory feature M_t contains long-term semantics, which can provide useful temporal semantics to help recognize temporally relevant action behaviors such as reading and writing.

3.1.3. Modeling and Aggregation of Interactions

In addition to spatial and temporal features, student self-interactions, student–student interactions, student interactions with objects such as books and mobile phones, and long-term temporal interactions of the same student are crucial for understanding the student's action behaviors. Given different student features P_t, object features O_t, and memory features M_t, the IA structure can fuse these features to model and aggregate the above interaction types for more accurate action behavior detection.

The IA structure consists of multiple interaction blocks improved based on transformer blocks [33], each modeling a single type of interaction through an attention mechanism. Interaction blocks are divided into three types in total: P-Block, O-Block, and M-Block. P-Block is used to model human–human interaction (including self-interaction) in the same clip, and its query and key/value are both student features or enhanced student features; O-Block is used to model human–object interaction, and its key/value input is the object features O_t. Figure 3a shows an illustration of an O-Block; M-Block is used to model the

same human at different frames of time interaction, and its key/value input is the memory features M_t.

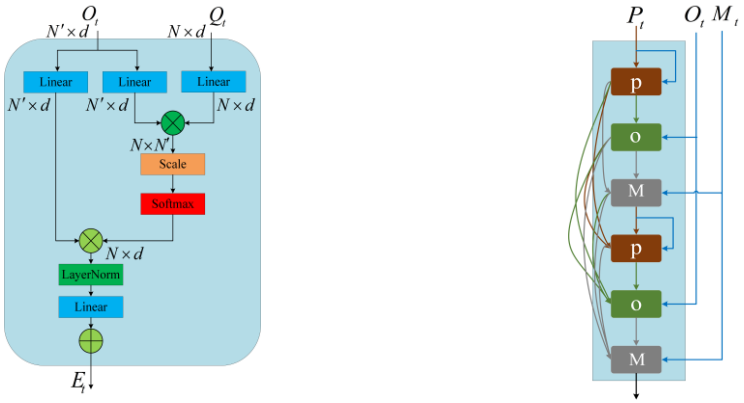

Figure 3. Interaction block and IA structure: (**a**) O-Block; (**b**) Dense Serial IA structure.

In a Dense Serial IA structure, each interaction block accepts the output of all previous interaction blocks, and different types of interactions interact with each other and are aggregated using learnable weights. The query for the ith block can be represented as

$$Q_{t,i} = \sum_{j \in C} W_j \odot E_{t,j} \tag{1}$$

where $\odot$ denotes the element-wise multiplication, C is the set of indices of previous blocks, W_j is a learnable d-dimensional vector normalized with a Softmax function among C, and $E_{t,j}$ is the enhanced output features from the jth block. Dense Serial IA is illustrated in Figure 3b.

The fused features from the Dense Serial IA structure are fed into the classifier to detect the student's action behavior.

3.1.4. MsTA Module

In the field of video action recognition, existing research uses fixed convolution kernels or operations to learn the temporal features of videos. The AIA network used in this paper also uses fixed $3 \times 1 \times 1$ convolution to learn the temporal features of students' action behaviors in the classroom scene. However, different action behaviors last for different durations; so, it is crucial to capture the multi-scale temporal features. This paper proposes a Multi-scale Temporal Convolution Unit (MsTCU) with different temporal convolution kernels, which utilizes different temporal sliding windows to extract the multi-scale temporal information.

The Double Attention (DA) block of A^2-Net [34] is applied to the fast pathway of the backbone network of the action detection model in this paper and cascaded with the abovementioned MsTCU to form an MsTA module, which extracts the local temporal features and global temporal features to model the long-distance interdependencies. The MsTCU and the DA block are described below.

MsTCU. As shown in Figure 4, the feature map $X \in R^{C \times T \times H \times W}$ first passes through the three branches of MsTCU, which are equipped with different temporal convolution kernels, respectively, and perform temporal convolution processing on the feature map X to obtain multi-scale temporal features as follows:

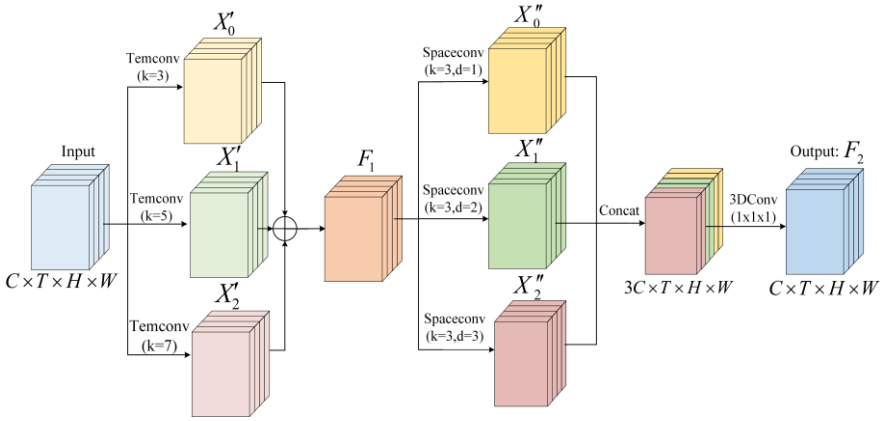

Figure 4. Multi-scale Temporal Convolution Unit.

(1) $X \in R^{C \times T \times H \times W}$ passes through three branches with temporal convolution kernel sizes of 3, 5, and 7, respectively, and the dimensionality of the output feature maps is constant in size, denoted as $\left[X'_0, X'_1, X'_2\right]$ Afterward, the feature maps of the three branches are summed up element by element:

$$F_1 = X'_0 + X'_1 + X'_2 \tag{2}$$

(2) To maintain spatiotemporal consistency, F_1 extracts the spatial features of the video through three dilated convolutional branches with convolutional kernel size 3 and dilation rates 1, 2, and 3, respectively. The output feature maps are represented using $\left[X''_0, X''_1, X''_2\right]$. Afterward, the feature maps of the three branches are concatenated together and the channel is reduced in dimension by $1 \times 1 \times 1$ convolution:

$$F_2 = f^{1\times1\times1}\left(Cat\left(\left[X''_0, X''_1, X''_2\right]\right)\right) \tag{3}$$

where $f^{1\times1\times1}$ denotes the $1 \times 1 \times 1$ convolution, $F_2 \in R^{C \times T \times H \times W}$.

DA block. As shown in Figure 5, let the input be $X \in R^{C \times T \times H \times W}$ and the local feature of each spatiotemporal location $i = 1, \cdots, THW$ be v_i. Then, define Equation (4):

$$z_i = F_{distr}\left(G_{gather}(X), v_i\right) \tag{4}$$

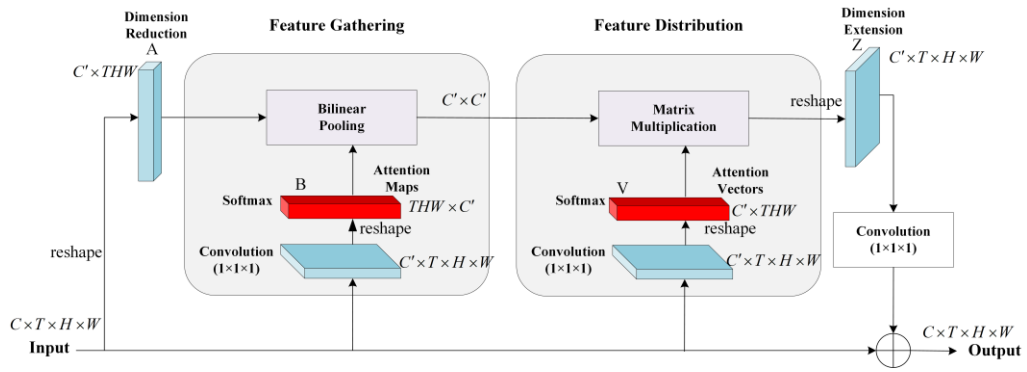

Figure 5. DA block.

There are two operations in the equation. First, the G_{gather} operation adaptively aggregates the features of the global context; then, the F_{distr} operation assigns the aggregated global context features back to each location i based on the local features v_i and outputs z_i. z_i can be obtained by steps (1) and (2):

(1) Feature Gathering. The DA block uses bilinear pooling to capture the second-order statistics of features to generate a global representation. Bilinear pooling is the summation of the outer product of all feature vectors (a_i, b_i) in the two input feature maps A and B:

$$G_{bilinear}(A, B) = AB^T = \sum_{\forall i} a_i {b_i}^T \tag{5}$$

where $A = [a_1, \cdots, a_{thw}] \in R^{m \times thw}$ and $B = [b_1, \cdots, b_{dhw}] \in R^{n \times thw}$. In CNNs, A and B can be the feature maps from the same layer, i.e., $A = B$, or from two different layers, i.e., $A = \phi(X; W_\phi)$ and $B = \theta(X; W_\theta)$, with parameters W_ϕ and W_θ.

By introducing the output variable $G = [g_1, \cdots, g_n] \in R^{m \times n}$ of the bilinear pooling and rewriting the second feature B as $B = \left[\overline{b}_1, \cdots, \overline{b}_n\right]$, where each $\overline{b}_i$ is a thw dimensional row vector, we can reformulate Equation (5) as

$$g_i = A\overline{b}^T = \sum_{\forall j} \overline{b}_{ij} a_j \tag{6}$$

Equation (5) shows that G can be viewed as a collection of visual elements in a sequence of frames, where each subset g_i is obtained by gathering local features weighted by $\overline{b}_i$; then, further applying a softmax onto B to ensure $\sum_j \overline{b}_{ij} = 1$, i.e., a valid attention weighting vector, gives the following second-order attention pooling process:

$$g_i = Asoftmax\left(\overline{b}_i\right)^T \tag{7}$$

(2) Feature Distribution. After the global features are collected from the frame sequence, the input X is first transformed by a convolutional layer to obtain the feature map V. The elements within V are normalized using the softmax function, i.e., $V = softmax(\rho(X; W_\rho))$, and $W\rho$ are parameters of the convolutional layer; then, the attention vectors v_i at each position of the feature map V are multiplied by the global features G and a subset of the global feature vectors can be adaptively assigned according to the weight magnitude of each vector v_i, as shown in Equation (8):

$$z_i = \sum_{\forall j} v_{ij} g_j = G_{gather}(X) v_i, where \sum_{\forall j} v_{ij} = 1 \tag{8}$$

(3) Substituting Equations (7) and (8) into Equation (4), the double attention operation can be expressed by Equation (9):

$$\begin{aligned} Z &= F_{distr}\left(G_{gather}(X), V\right) \\ &= G_{gather}(X) softmax(\rho(X; W\rho)) \\ &= \left[\phi(X; W_\phi) softmax(\theta(X; W_\theta))^T\right] softmax\left(\rho(X; W_\rho)\right) \end{aligned} \tag{9}$$

Based on the above process of double attention operation, the computational diagram of the DA block is briefly summarized as follows: the input feature map $X \in R^{C \times T \times H \times W}$ is passed through three different convolutional layers to obtain the feature maps A, B, and V. B and V need to be softmax normalized. A and B are subjected to bilinear pooling operation and then multiplied by V to obtain $Z \in R^{C' \times THW}$. Then, the $1 \times 1 \times 1$ convolution is used to reshape Z into $Z \in R^{C \times T \times H \times W}$.

Finally, residual connections are added to the MsTCU and DA block, respectively, to ensure the propagation of information across layers, as shown in Figure 6:

$$Y = Y' + X \tag{10}$$

where X is the input feature map and Y' is the output feature map of MsTCU (DA block).

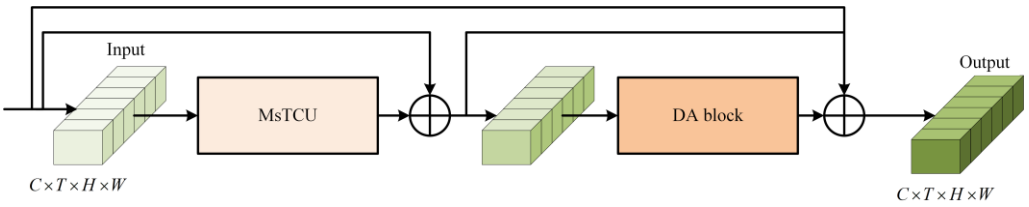

Figure 6. MsTA module.

3.1.5. MsCSA Module

The video data captured in this study have a large resolution; a large difference in pixel size occupied by front and back row students in space; and the need to extract small object features O_t, such as books and mobile phones around the students, along with other types of features, to model the interaction. Literature [35] shows that multi-scale receptive field is helpful for the network to notice the spatial position of different size objects. Therefore, a Multi-scale Spatial Attention (MsSA) block is added to the slow pathway of the video backbone network to obtain contextual information of a larger receptive field; at the same time, to allow the network to learn the weight response of the feature maps of different channels in the process, the MsSA block is cascaded with Gaussian Context Transformer (GCT) [36] to form an MsCSA module. The MsCSA module is shown in Figure 7.

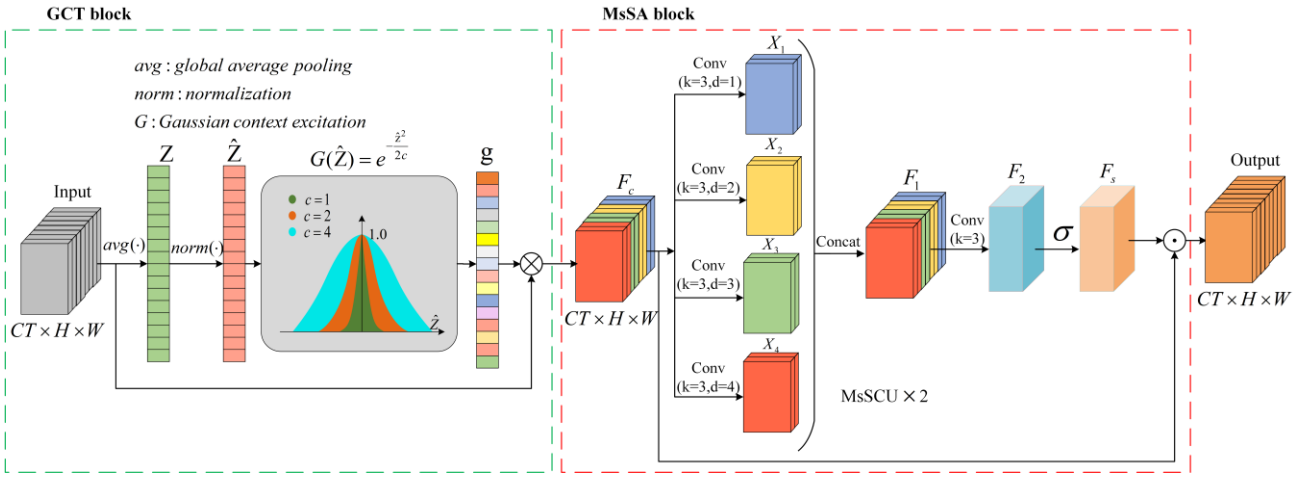

Figure 7. MsCSA module.

GCT block. The GCT block first performs the global average pooling operation on the input feature map X in the spatial dimension; then, it stabilizes the distribution of the global features by normalizing the channel vectors; finally, it obtains the attention map by using the Gaussian function to perform the excitation operation on the normalized global features. The specific process is as follows:

(1) Global Context Aggregation (GCA). Let the input $X \in R^{C\times T\times H\times W}$ reshape X into $R^{CT\times H\times W}$; the GCA can be expressed as

$$z = avg(X) = \left\{ z_k = \frac{1}{H\times W}\sum_{i=1}^{W}\sum_{j=1}^{H} X_k(i,j); k \in \{1,\cdots,CT\} \right\} \tag{11}$$

where CT is the number of channels, and H and W are the height and width of the feature map, respectively.

(2) Normalize. The GCT block computes the attentional activation value of the channel using the function $f(\cdot)$. Define $\hat{z}$ as the input to the function $f(\cdot)$, and $\hat{z}$ is represented by Equation (12):

$$\hat{z} = \frac{1}{\sigma}(z-\mu) \tag{12}$$

where $\mu = \frac{1}{C}\sum_{k=1}^{C} z_k$ denotes the mean of the global context z, and $z - \mu$ is the mean shift. $\sigma = \sqrt{\frac{1}{C}\sum_{k=1}^{C}(z_k - \mu)^2 + \varepsilon}$ denotes the standard deviation of the global context z, ε is a very small constant, and σ is introduced so that $\hat{z}$ remains in distribution with mean 0 and variance 1 for different input samples, thus ensuring that the output of $f(\cdot)$ is stable. Equation (11) is consistent with the result of normalizing z and, thus, can be expressed as $\hat{z} = norm(z)$.

(3) Gaussian Context Excitation (GCE). The GCT block substitutes the Gaussian function into $f(\hat{z})$ and defines the GCE operation as

$$g = f(\hat{z}) = e^{-\frac{\hat{z}^2}{2c^2}} \tag{13}$$

where c is a constant or learnable parameter and g is the attentional activation value.

The above steps are combined to form the GCT block, which is represented by Equation (14):

$$F_c = e^{-\frac{norm(avg(X))^2}{2c^2}} X \tag{14}$$

where $F_c \in R^{CT\times H\times W}$ is the feature map output by the GCT block.

MsSA block. After the GCT block, $F_c \in R^{CT\times H\times W}$ is input to the Multi-scale Spatial Convolution Unit (MsSCU) in the MsSA block to extract the spatial information at different scales. The process is shown below:

(1) F_c undergoes dilation convolution with a convolution kernel size of 3×3 and dilation rates of 1, 2, 3, and 4, respectively; then, F_c is sliced into four parts, denoted using $[X_1, X_2, X_3, X_4]$, with the number of channels in each part being $C' = \frac{CT}{r}$, and then concatenated in the channel dimension

$$F_1 = Cat([X_1, X_2, X_3, X_4]) \tag{15}$$

(2) $F_1 \in R^{\frac{4CT}{r}\times H\times W}$ undergoes a 3×3 convolution to further fuse the multi-scale features and recover the original number of channels CT to obtain the feature map $F_2 \in R^{CT\times H\times W}$:

$$F_2 = f^{3\times 3}(F_1) \tag{16}$$

where $f^{3\times 3}$ denotes the 3×3 convolution operation in the spatial dimension; for ease of representation, the above process is denoted as $F_2 = MsSCU(F_c)$. After two MsSCU, the spatial attention feature map is then obtained by the sigmoid function

$$F_s = \sigma(MsSCU(MsSCU(F_c))) \tag{17}$$

where $F_s \in R^{CT\times H\times W}$, σ is the sigmoid function, and F_c and F_s are multiplied element by element to obtain the feature map Y':

$$Y' = F_c \odot F_s \tag{18}$$

Reshape Y' to $Y' \in R^{C\times T\times H\times W}$ to keep the size of the input constant.

Similar to the MsTA module, residual connections are added to the MsCSA module to ensure the propagation of information across layers.

3.1.6. Equalized Focal Loss Function

The student action behavior dataset used in this paper suffers from a category imbalance problem, whereby most of the students show action behaviors such as listening and reading during the class period while a few students have action behaviors such as using mobile phones and lying on the table. The loss function used in the original AIA network is Focal Loss [37], but Li et al. [12] proved that Focal Loss does not deal with the foreground category imbalance problem; so, in this paper, we use the EFL proposed by

Li et al. [12] as the loss function of the action classifier in the AIA network to improve the category imbalance problem, specifically shown by Equation (19):

$$EFL(p_t) = -\sum_{j=1}^{C} \alpha_t \left(\frac{\gamma_b + \gamma_v^j}{\gamma_b} \right) (1 - p_t)^{\gamma_b + \gamma_v^j} \log(p_t) \tag{19}$$

where α_t is used to regulate the weights of positive and negative samples during training and p_t is the prediction score. $\left(\frac{\gamma_b + \gamma_v^j}{\gamma_b} \right)$ is the weight factor associated with the category, $\gamma_b + \gamma_v^j$ is the focusing factor for the jth category, γ_b is a constant, and γ_v^j can be expressed as

$$\gamma_v^j = s\left(1 - g^j\right) \tag{20}$$

where the hyper-parameter s is the scaling factor that determines the upper limit of γ_v^j in EFL, and the parameter g^j indicates the accumulated gradient ratio of positive samples to negative samples of the jth category during the training process, with the range of g^j set to [0,1].

Since g^j indicates the accumulated gradient ratio of positive samples and negative samples of the jth category during the training process, a larger value of g^j indicates that the category is trained to be balanced and a smaller value of g^j indicates that the category is trained to be unbalanced; so, the focusing factor $\gamma_b + \gamma_v^j$ and the prediction score p_t in the EFL, composed of $(1 - p_t)^{\gamma_b + \gamma_v^j}$ as the category-related weighting factor in the loss, can dynamically regulate the weight of the loss based on the cumulative gradient ratio of positive and negative samples in each category during the training process and the prediction score p_t to handle the problem of category imbalance.

3.2. Seat-Association-Based Analysis of Students' Action Behavior Sequence

The application scenario of this paper is in the classroom. The classroom scenario is densely populated with students, mutual occlusion, and the pixel area occupied by the face region in the video frame is small; thus, it is difficult for the face recognition technology to associate the student object across the video frames. During the class period, the position of students is fixed and the position shift of individual students in the video frame sequence is small; therefore, this paper associates the same student object through the student's seat across the video frames, tracks the changes of the same student's action behaviors in a continuous period, and obtains the action behavior sequence of individual students in the teaching time of knowledge points:

(1) Students' use of electronic devices in the classroom cannot be simply classified as playing with a mobile phone; so, this paper first obtains the recognition result set G_A of the action behavior detection model within the teaching time of knowledge point k. The number of students who use mobile phones in the set G_A is counted; if the ratio of the number of students is greater than 0.6, it is considered that the teacher published an in-class test and classifies the behavior as reading.

(2) In this paper, we first use the LabelImg tool to label a rectangular box for each student's seat in the classroom and save its coordinate information; then, we extract the student's bounding box information from G_A. We use the method of calculating the Complete Intersection over Union (CIoU) in the literature [38] to calculate the CIoU value of the seat rectangular box and the student's bounding box to measure the overlap between the two boxes, match the student's seat based on the maximum overlap, and then obtain the sequence of individual student's action behaviors by using the seat number as the index.

4. Experimental Results and Analysis

4.1. Datasets

Dataset 1: The AVA dataset [39] is a dataset oriented to the spatiotemporal action detection task. This dataset takes 1 frame per second as a keyframe to be labeled. There are 80 atomic action classes, including three major classes: posture actions, actions of human–human interaction, and actions of human–object interaction.

Dataset 2: UCF101-24 is a dataset oriented to the spatiotemporal action detection task. This dataset is labeled frame-by-frame and contains a total of 24 classes of actions.

Dataset 3: At present, there is no public student action behavior dataset; so, this paper constructs a dataset of students' action behaviors based on real classroom videos and annotated concerning the AVA dataset [39], which contains seven types of action behaviors, including listening, looking around, lying on the table, reading, writing, using mobile phone, and talking, as shown in Figure 8. The construction process of the dataset is as follows:

Figure 8. Example of student action behavior in the classroom.

Step 1: Collect real classroom videos from a university; screen and edit the videos; and select a total of 25 videos, each with a length of 5 min.

Step 2: Firstly, video frames are extracted according to the frame rate of 30 frames per second; then, every 30 frames are extracted as keyframes, which are used to label the students' positions and action behaviors.

Step 3: Use the VGG Image Annotation tool to label the position and action behavior categories of students in the keyframes, save the labeling information in CSV file format, and then process it into labeling files in AVA format.

In addition, to model the interaction between students and the environment in the classroom video, it is necessary to correctly find the objects that students interact with. Therefore, this paper trains a target detection model based on YOLOv5 that is suitable for classroom scenes and then uses the model to detect objects such as cell phones and books in key frames, extracts coordinate information and category indexes of cell phones and books based on the detection results, and generates labeled files in the format of the COCO dataset to be used as auxiliary training data.

4.2. Experimental Results on the Public Dataset

In the experiments on the AVA dataset, the model was trained using the pretraining parameters of the Kinetics-700 dataset for weight initialization. The input to the network is sixty-four frames, $\alpha = 8$,$\tau = 16$, i.e., thirty-two frames for the fast pathway and four frames for the slow pathway. To reduce the number of parameters, the short edges of the video frames are cropped to 256, two GPUs are used for training, and the SGD optimization algorithm is used with a batch size of four and an initial learning rate of 0.00025, which is adjusted in the first 2k iterations using linear warm-up. In this paper, we refer to the

literature [31] and evaluate the performance of the model using frame-level mAP with an IoU threshold of 0.5.

In this paper, the MsTA module and MsCSA module are fused to the AIA network; the classification loss function is replaced with the EFL function, which is compared with the SlowFast [13], X3D [14], ACAR [17], AIA [11], MRSN [18], DOAD [15], and HIT networks [19]; and the experimental results are shown in Table 1.

Table 1. Comparison with other models in the AVA dataset.

Model	Pretrain	mAP%
SlowFast, R101-NL	Kinetics-600	29.0
X3D	Kinetics-600	27.4
AIA	Kinetics-700	31.2
ACAR	Kinetics-700	31.9
MRSN	Kinetics-700	33.5
DOAD	Kinetics-700	28.5
HIT	Kinetics-700	32.6
Ours	Kinetics-700	32.0

ACAR and AIA models are results reproduced in the environment provided by the authors and the rest are reported results from the paper. In the experiments on the CUF101-24 dataset, the input to the network is 32 frames, $\alpha = 4$, $\tau = 4$; trained using a single GPU with a batch size of eight; and the other settings are the same as the experiments on the AVA dataset. The MsTA module and MsCSA module are fused to the AIA network and compared with the ACT [40], STEP [41], AIA [11], ACAR [17], MRSN [18], DOAD [15], and HIT networks [41]. The experimental results are shown in Table 2.

Table 2. Comparison with other models in the UCF101-24 dataset.

Model	Input	mAP%
ACT	V	69.5
STEP	V + F	75.0
AIA	V	81.7
ACAR	V	84.3
MESN	V	80.3
DOAD	V	74.8
HIT	V	84.8
Ours	V	82.2

V and F in Table 2 denote visual frames and optical flow, respectively.

Tables 1 and 2 show the results of comparing this paper's improved AIA network with other models on the AVA dataset and the UCF101-24 dataset, respectively: (1) The experimental results of the two datasets show that the improved AIA network in this paper outperforms most of the models and has good generalizability. (2) On the AVA dataset, the mAP values are lower than those of the MRSN model and the HIT model; however, the MRSN model requires pretraining of a base network to extract the features of the video clips as memory features and the HIT model requires an additional human pose estimation network to extract the keypoints of the human body. (3) On the UCF101-24 dataset, the mAP values are lower than those of the ACAR model and the HIT model but the ACAR model models actor–context interactions, which may generate background noise.

4.3. Student Action Behavior Detection Performance

4.3.1. Experimental Results and Analysis of the Student Action Behavior Dataset

In the experiments on the student action behavior dataset, in this paper, the video frames are cropped to 640 × 640, the batch size is set to four, the initial learning rate is 0.00025, and the input of the network is 64 frames. Figure 9 shows the curves of the accuracy

and loss values of the proposed method during the experimental process. The results show that the improved AIA network in this paper converges stably during the training process and achieves 92% accuracy.

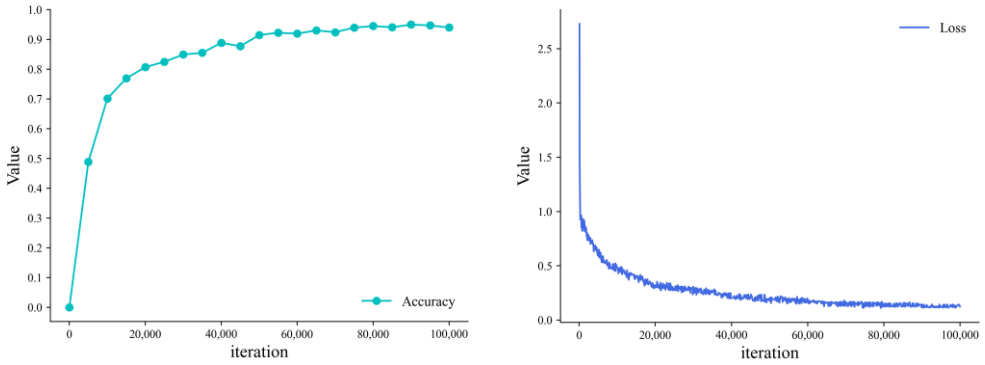

Figure 9. Accuracy and loss curves of the improved AIA network: (**a**) accuracy curve; (**b**) loss value curve.

To verify the effectiveness of the different modules added to the improved AIA network and the improved loss function, Table 3 lists the experimental results with the addition of the MsTA module and the MsCSA module, as well as the comparative results of replacing the loss function with the EFL function [12] and the EQLv2 function [42].

Table 3. Comparison of adding different modules and replacing loss functions.

Model	Precision%	Recall%	mAP%
AIA	78.4	77.6	76.5
AIA + MsTA	81.5	78.3	79.4
AIA + MsCSA	79.2	78.5	77.8
AIA + EQLv2	80.3	78.2	78.8
AIA + EFL	82.7	79.1	80.2
AIA + MsTA + MsCSA + EQLv2	82.2	78.8	80.1
AIA + MsTA + MsCSA + EFL (Ours)	83.9	80.4	81.3

As can be seen from Table 3, after integrating the MsTA module and the MsCSA module into the AIA network, the mAP values of student action detection improved. Students' different action behaviors during class have different durations and visual speeds, and the MsTA module can extract action behavior features with different time scales. For example, when students are listening carefully, they usually look up and face the teacher, and there is not much posture change between the current frame and the history frame; however, when students are looking around, the module needs to combine the temporal features of multi-frames to recognize the student's "looking around" behavior. When students are reading, writing, or using mobile phones, they will turn the book, take a pen, or touch the book or mobile phone and the MsCSA module expands the receptive field in the space and enhances the response of the mobile phone, book, and other objects in the feature map, thus enhancing the interaction features of the student and the object so that the network better recognizes the interaction behaviors such as reading, writing, or using mobile phones.

In addition, due to the problem of category imbalance in the dataset, there are fewer samples of behaviors such as lying on the table, using a mobile phone, and talking, and more samples of behaviors such as listening and reading. Both the EQLv2 function and the EFL function are used to improve the problem by adjusting the loss weights of the different categories. The experimental results in Table 3 show that the EFL function performs better in the study of this paper; so, the EFL is chosen as the improved AIA network's loss function.

Figure 10 shows the AP values of seven student action behaviors for the method proposed in this paper and the AIA network. The original AIA network has a good recognition effect for action behaviors with a large number of samples, such as listening, reading, and writing; however, the recognition accuracy of action behaviors with a small number of samples, such as lying on the table, using mobile phones, and talking, is low. The improved AIA network significantly improves the detection accuracy of the four categories of action behaviors such as looking around, lying on the table, using mobile phones, and talking, indicating that the method proposed in this paper can not only extract spatiotemporal features that are beneficial to the recognition of action behaviors but also deal with the problem of imbalance of the foreground category that exists in dataset 2, which has a greater improvement in the recognition accuracy of the categories of action behaviors with fewer samples. Although there are few sample instances of lying on the table, the features are obvious, and the detection accuracy reaches 80.7%. The detection accuracy of "talk" behavior is improved by 5.7% compared with that of the AIA network but is still significantly lower than the other action behaviors, mainly because the application scenario of this paper is densely populated with students and the head region of the students occupies a smaller area in the image, with less obvious features.

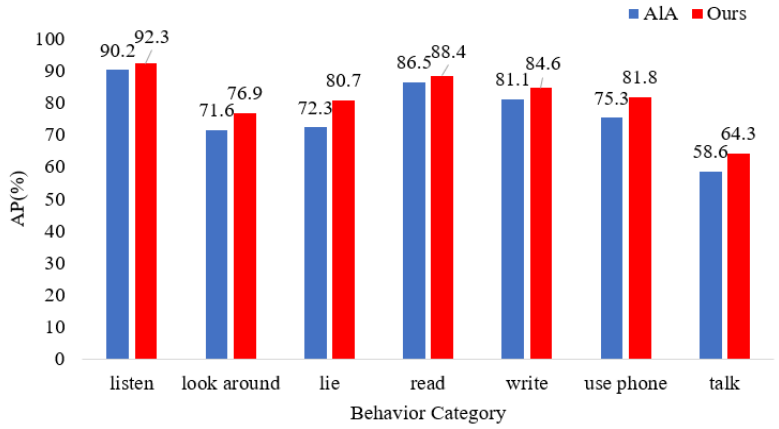

Figure 10. Comparison of average precision of seven action behavior categories.

Figure 11 shows an example of student action behavior detection at a certain moment in the classroom of the method proposed in this paper, which proves the effectiveness of the method.

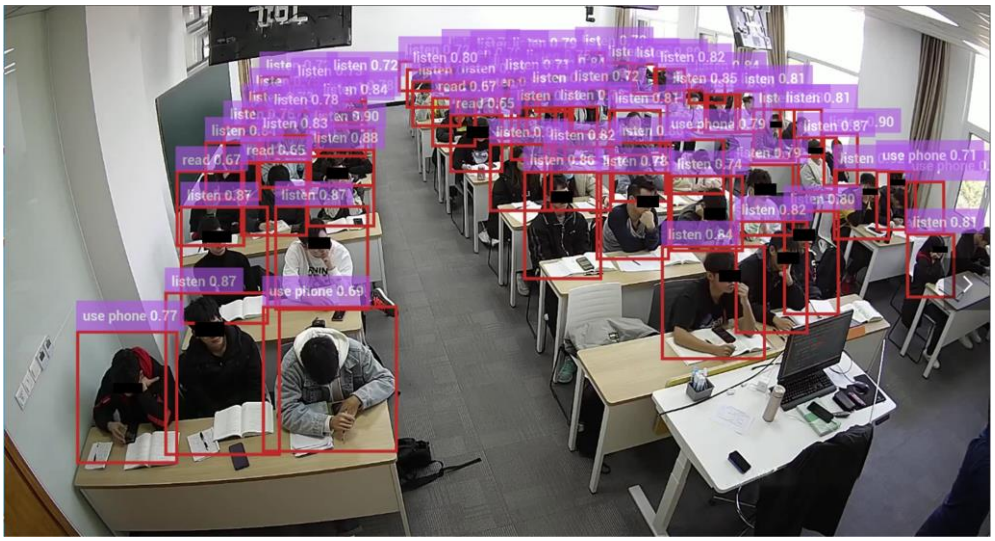

Figure 11. Example of improved AIA network detection.

In Figure 11, the improved AIA network model can detect students' action behaviors such as listening, reading, and using mobile phones during class. Since the number of students in the classroom is large and the boxes are dense, Figure 12 shows an example of the detection of action behaviors in a local area.

(**a**) (**b**)

Figure 12. Example of local region action behavior detection: (**a**) "read" behavior detection example; (**b**) "use phone" behavior detection example.

4.3.2. Comparative Experiments and Analysis

To objectively evaluate the performance of the network proposed in this paper for recognizing students' action behaviors in classroom scenarios, the improved AIA network is compared with the SlowFast [13], MviT [16], ACAR [17], and AIA [11] networks under the same experimental configuration conditions and dataset. The experimental results are shown in Table 4.

Table 4. Comparison with other models in the student action behavior dataset.

Model	Pretrain	mAP%
SlowFast	Kinetics-600	74.2
AIA	Kinetics-700	76.5
MViT	Kinetics-600	75.2
ACAR	Kinetics-700	76.1
Ours	Kinetics-700	81.3

As can be seen from Table 4, the mAP values of the network proposed in this paper are higher than those of other networks, which are 7.1%, 6.1%, 5.2%, and 4.8% higher than those of SlowFast, MviT, ACAR, and AIA networks, respectively, indicating that the improved model has better accuracy in the detection of spatiotemporal-oriented students' action behaviors. The mAP curves during the experiments of different models are shown in Figure 13.

4.4. Analysis of Students' Action Behavior Sequences

The action behavior of the students in the classroom is detected by the action behavior recognition method described in Section 3.1. Then, each student's seat is matched using the method described in Section 3.2. The seat matching result at a certain point in time is shown in Figure 14. The seat is denoted as (x, y), where x stands for rows of seats and y stands for columns of seats.

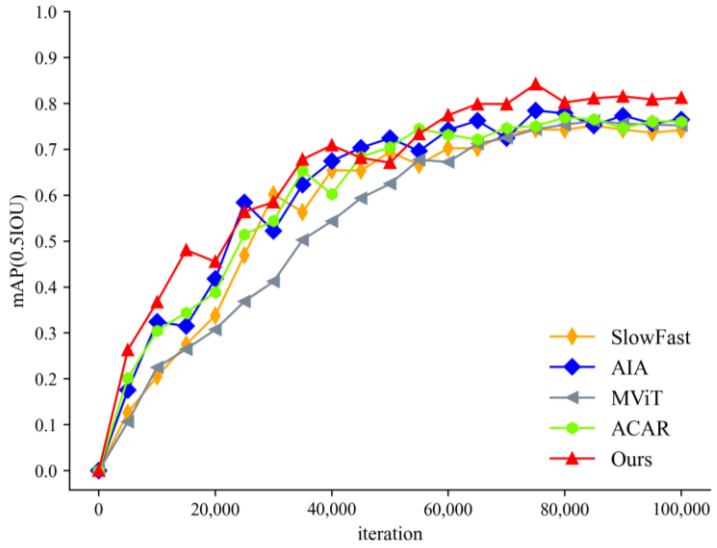

Figure 13. mAP curves for different models.

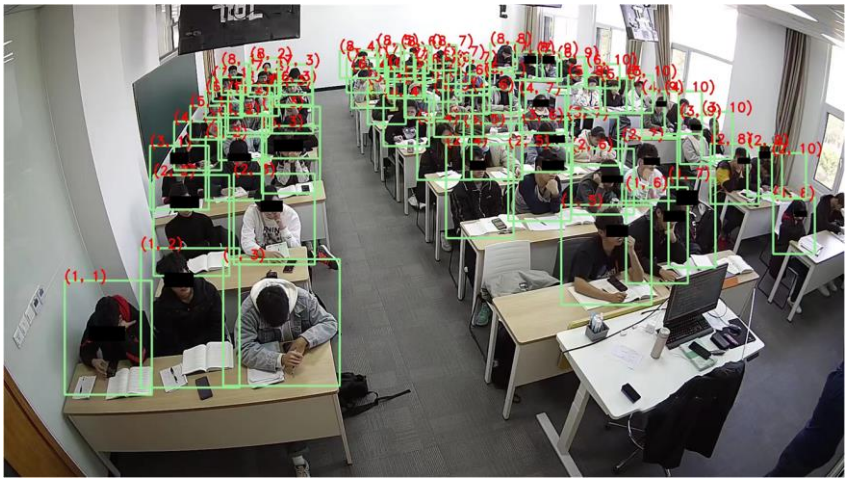

Figure 14. Example of student seat matching.

With the seat number as the index, the changes in students' action behavior are tracked, and the sequence of students' action behavior in the teaching time of knowledge points k is obtained, as shown in Table 5. Listening, writing, reading, looking around, using mobile phones, lying on the table, and talking are represented by 3, 2, 1, -1, -2, -3, and -4, respectively.

Table 5. The sequence of students' action behavior in the teaching time of knowledge point k.

Seat	10 s	20 s	30 s	40 s	50 s	60 s	70 s	80 s	90 s	100 s	110 s
(1,1)	−2	−2	3	3	3	3	3	3	3	3	3
(1,2)	1	−2	−2	3	1	3	3	3	3	3	3
(1,3)	1	3	−4	3	3	3	3	3	3	3	3
(1,5)	−4	−4	1	1	3	3	−4	−4	3	3	3
...	...	...	...	...	...	...	...	...	...	...	...
(5,9)	3	3	3	3	3	3	3	3	3	3	1
(5,10)	−1	3	−2	−2	−2	−2	−2	−2	−2	−2	−2

From Table 5, it can be seen that analyzing the sequence of students' action behavior in the teaching time of knowledge point k can understand the behavior performance of

students in class and help teachers to find students with negative behaviors, for example, if the student whose seat number is $(5, 10)$ continues to use a mobile phone during the period and is in a state of wandering, the teacher should remind the student after class so that the student can listen to the lectures attentively during the class period and review what the teacher has said promptly after the class.

5. Conclusions

Recognizing and analyzing students' action behaviors is of great significance in the research of teaching feedback and improving students' learning effectiveness. In this paper, we propose a student action detection model based on an improved AIA network. To improve the detection accuracy of the model, we add an MsTA module to the fast pathway of the video backbone network, and an MsCSA module to the slow pathway, to efficiently extract the multi-scale temporal and spatial information. The EFL function is introduced to improve the category imbalance problem that exists in the action behavior dataset. The experimental results show that the improved AIA network in this paper can detect different action behaviors of students and has higher detection accuracy compared with the original network. In addition, correlating students' action behavior sequences with seat numbers as indexes can find the students who have negative action behaviors during the class period, which helps teachers to understand students' learning efficiency during the class period. In the future, data such as students' facial expressions will be further combined to jointly analyze students' learning emotions. Meanwhile, more data are collected to create a rich and diverse dataset. In addition, this paper plans to deploy the algorithm to embedded devices for use in smart classrooms to help teachers understand students' learning and improve their learning outcomes.

Author Contributions: Conceptualization, Z.S. and M.Y.; methodology, M.Y.; software, M.Y.; validation, H.W., J.M. and H.Z.; formal analysis, H.W.; investigation, J.L.; resources, Z.S.; data curation, M.Y.; writing—original draft preparation, M.Y.; writing—review and editing, Z.S.; visualization, M.Y.; supervision, Z.S.; project administration, Z.S.; funding acquisition, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by The National Natural Science Foundation of China (62177012, 61967005, 62267003).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The student action behavior dataset is not available due to privacy restrictions. Other data are from https://www.crcv.ucf.edu/data/UCF101.php, accessed on 14 September 2023 (UCF101-24) and from https://research.google.com/ava/index.html, accessed on 14 September 2023 (AVA Actions Dataset).

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

References

1. Fu, R.; Wu, T.; Luo, Z.; Duan, F.; Qiao, X.; Guo, P.; IEEE. Learning Behavior Analysis in Classroom Based on Deep Learning. In Proceedings of the 10th International Conference on Intelligent Control and Information Processing, Marrakesh, Morocco, 14–19 December 2019; pp. 206–212.
2. Zhang, Y.; Wang, G. Research on Application of Intelligent Analysis in Monitoring of Classroom Teaching. In Proceedings of the 2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication, Shanghai, China, 23–25 April 2021; pp. 253–257.
3. Zheng, X.W.; Tang, Y.Y.; Zhou, J.T. A Framework of Adaptive Multiscale Wavelet Decomposition for Signals on Undirected Graphs. *IEEE Trans. Signal Process.* **2019**, *67*, 1696–1711. [CrossRef]
4. Yang, L.; Su, H.L.; Zhong, C.; Meng, Z.Q.; Luo, H.W.; Li, X.C.; Tang, Y.Y.; Lu, Y. Hyperspectral image classification using wavelet transform-based smooth ordering. *Int. J. Wavelets Multiresolution Inf. Process.* **2019**, *17*, 1950050. [CrossRef]
5. Guido, R.C.; Pedroso, F.; Contreras, R.C.; Rodrigues, L.C.; Guariglia, E.; Neto, J.S. Introducing the Discrete Path Transform (DPT) and its applications in signal analysis, artefact removal, and spoken word recognition. *Digit. Signal. Process.* **2021**, *117*, 103158. [CrossRef]
6. Guariglia, E. Primality, Fractality, and Image Analysis. *Entropy* **2019**, *21*, 304. [CrossRef]

7. Bui Ngoc, A.; Ngo Tung, S.; Phan Truong, L.; Le Phuong, C.; Nguyen Huu, T.; Nguyen Cong, D.; Nguyen Huu, T.; Aftab, M.U.; Tran Van, D. A Computer-Vision Based Application for Student Behavior Monitoring in Classroom. *Appl. Sci.* **2019**, *9*, 4729. [CrossRef]
8. Tang, L.; Gao, C.; Chen, X.; Zhao, Y. Pose detection in complex classroom environment based on improved Faster R-CNN. *IET Imag. Process.* **2019**, *13*, 451–457. [CrossRef]
9. Liu, S.; Zhang, J.; Su, W. An improved method of identifying learner's behaviors based on deep learning. *J. Supercomput.* **2022**, *78*, 12861–12872. [CrossRef]
10. Xie, Y.; Zhang, S.; Liu, Y. Abnormal Behavior Recognition in Classroom Pose Estimation of College Students Based on Spatiotemporal Representation Learning. *Trait. Du Signal* **2021**, *38*, 89–95. [CrossRef]
11. Tang, J.; Xia, J.; Mu, X.; Pang, B.; Lu, C. Asynchronous Interaction Aggregation for Action Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 71–87.
12. Li, B.; Yao, Y.; Tan, J.; Zhang, G.; Yu, F.; Lu, J.; Luo, Y. Equalized Focal Loss for Dense Long-Tailed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6980–6989.
13. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6201–6210.
14. Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 200–210.
15. Chang, S.; Wang, P.; Wang, F.; Feng, J.; Shou, M.Z. DOAD: Decoupled One Stage Action Detection Network. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, BC, Canada, 17–24 June 2023; pp. 3123–3232.
16. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6804–6815.
17. Pan, J.; Chen, S.; Shou, M.Z.; Liu, Y.; Shao, J.; Li, H. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 464–474.
18. Zheng, Y.D.; Chen, G.; Yuan, M.; Lu, T. MRSN: Multi-Relation Support Network for Video Action Detection. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo, Brisbane, Australia, 10–14 July 2023; pp. 1026–1031.
19. Faure, G.J.; Chen, M.H.; Lai, S.H. Holistic Interaction Transformer Network for Action Detection. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3329–3339.
20. Zhang, Y.; Zhu, T.; Ning, H.; Liu, Z. Classroom student posture recognition based on an improved high-resolution network. *Eurasip J. Wirel. Commun. Netw.* **2021**, *2021*, 140. [CrossRef]
21. Zhou, J.; Ran, F.; Li, G.; Peng, J.; Li, K.; Wang, Z. Classroom Learning Status Assessment Based on Deep Learning. *Math. Probl. Eng.* **2022**, *2022*, 7049458. [CrossRef]
22. Pang, C. Simulation of student classroom behavior recognition based on cluster analysis and random forest algorithm. *J. Intell. Fuzzy Syst.* **2021**, *40*, 2421–2431. [CrossRef]
23. Ding, Y.; Bao, K.; Zhang, J. An Intelligent System for Detecting Abnormal Behavior in Students Based on the Human Skeleton and Deep Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 3819409. [CrossRef]
24. Wu, B.; Wang, C.; Huang, W.; Huang, D.; Peng, H. Recognition of Student Classroom Behaviors Based on Moving Target Detection. *Traitement Du Signal* **2021**, *38*, 215–220. [CrossRef]
25. Banerjee, S.; Ashwin, T.S.; Guddeti, R.M.R. Multimodal behavior analysis in computer-enabled laboratories using nonverbal cues. *Signal Image Video Process.* **2020**, *14*, 1617–1624. [CrossRef]
26. Liu, M.; Meng, F.; Wu, Q.; Xu, L.; Liao, Q. Behaviour detection in crowded classroom scenes via enhancing features robust to scale and perspective variations. *IET Imag. Process.* **2021**, *15*, 3466–3475. [CrossRef]
27. Huang, W.; Li, N.; Qiu, Z.; Jiang, N.; Wu, B.; Liu, B. An Automatic Recognition Method for Students' Classroom Behaviors Based on Image Processing. *Traitement Du Signal* **2020**, *37*, 503–509. [CrossRef]
28. Zheng, Z.; Liang, G.; Luo, H.; Yin, H. Attention assessment based on multi-view classroom behaviour recognition. *IET Comput. Vis.* **2022**. [CrossRef]
29. Liu, H.; Liu, Y.; Zhang, R.; Wu, X. Student Behavior Recognition From Heterogeneous View Perception in Class Based on 3-D Multiscale Residual Dense Network for the Analysis of Case Teaching. *Front. Neurorobotics* **2021**, *15*, 675827. [CrossRef]
30. Jisi, A.; Yin, S. A new feature fusion network for student behavior recognition in education. *J. Appl. Sci. Eng.* **2021**, *24*, 133–140.
31. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 16th IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
34. Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J. A^2-Nets: Double Attention Networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018.

35. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z.; IEEE. Scale-Aware Trident Networks for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6053–6062.
36. Ruan, D.; Wang, D.; Zheng, Y.; Zheng, N.; Zheng, M. Gaussian Context Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15124–15133.
37. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
38. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D.; Assoc Advancement Artificial, I. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the 34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
39. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6047–6056.
40. Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; Schmid, C. Action Tubelet Detector for Spatio-Temporal Action Localization. In Proceedings of the 16th IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4415–4423.
41. Yang, X.; Yang, X.; Liu, M.-Y.; Xiao, F.; Davis, L.; Kautz, J.; Soc, I.C. STEP: Spatio-Temporal Progressive Learning for Video Action Detection. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 264–272.
42. Tan, J.R.; Lu, X.; Zhang, G.; Yin, C.Q.; Li, Q.Q. Equalization Loss v2: A New Gradient Balance Approach for Long-tailed Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 1685–1694.