




Development and Application of a Deep Learning-Based Image Processing System for Classroom Behavior Analysis

Qiuju Wang¹, Qinde Jiang^{2*}, Fengguo Liu³

¹ School of Humanities and Education, Liaodong University, Dandong 118001, China

² Academic Affairs Office, Hechi University, Hechi 546300, China

³ Party Committee Organization Department, Liaodong University, Dandong 118001, China

Corresponding Author Email: 16020@hcnu.edu.cn

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420419>

ABSTRACT

Received: 7 February 2025

Revised: 8 June 2025

Accepted: 19 June 2025

Available online: 14 August 2025

Keywords:

deep learning, classroom behavior analysis, image processing, temporal 2D convolution, behavior recognition

With the ongoing advancement of educational informatization, leveraging advanced technological methods to improve classroom teaching quality has become a significant focus of educational research. The application of deep learning-based image processing technology in the education field has gradually attracted attention. By automatically analyzing classroom videos, student behaviors can be objectively recorded and evaluated, helping teachers better understand teaching effectiveness and make timely adjustments to teaching strategies. Although some current studies have attempted to apply deep learning to classroom behavior analysis, challenges such as a heavy reliance on manual feature extraction and insufficient correlation between sequential data remain. To address these issues, a deep learning-based image processing system for classroom behavior analysis was proposed. The main research contributions include a) the development of a temporal 2D convolution model for classroom behavior analysis to extract temporal information from image data; b) the design of a method to expand the receptive field of temporal 2D convolution, enhancing the ability to perceive behaviors at different time scales; c) the construction of a classroom behavior recognition network to improve the accuracy and robustness of behavior recognition. This research aims to provide an efficient and accurate solution for classroom behavior analysis and promote the development of educational informatization.

1. INTRODUCTION

With the continuous advancement of educational informatization, how to effectively leverage advanced technologies to enhance classroom teaching quality has become an important research topic [1-4]. In recent years, deep learning-based image processing technologies have achieved significant accomplishments across various fields, and their application in education has gradually gained attention [5-9]. By automating the analysis of classroom videos, student behaviors can be objectively recorded and assessed, enabling teachers to better understand teaching effectiveness and adjust teaching strategies in a timely manner. Consequently, the development of a deep learning-based image processing system for classroom behavior analysis holds considerable practical significance.

Classroom behavior analysis not only reflects student engagement and learning status but also provides data support for personalized teaching, thereby enabling precision in teaching [10-13]. Through the quantitative analysis of student behaviors during class, issues in the teaching process can be identified, which can subsequently inform the optimization of teaching design and improve the overall quality of education. Furthermore, classroom behavior analysis systems can be

employed in teaching research, providing empirical data that contribute to the advancement of educational theory [14-16]. Therefore, research into deep learning-based image processing systems for classroom behavior analysis is not only of theoretical significance but also holds wide practical value.

Although some studies have attempted to apply deep learning to classroom behavior analysis, there remain several limitations [17-22]. For example, traditional behavior analysis methods largely rely on manual feature extraction, which fails to fully exploit the information contained in image data, leading to insufficient accuracy and robustness in the analysis results [23-25]. Within the theoretical framework of deep learning, Convolutional Neural Networks (CNNs) are the foundational architecture for visual feature extraction. Their hierarchical feature learning mechanism provides a crucial basis for the analysis of image sequences [26]. In the domain of temporal modeling, temporal models such as Recurrent Neural Networks (RNNs) and their variants, notably Long Short-Term Memory (LSTM) [27], have been widely employed to capture temporal dependencies via the transmission of hidden states. However, these models are often constrained by issues such as vanishing gradients and limited computational efficiency. Additionally, existing CNNs often overlook the temporal correlations when processing sequential

data, which negatively affects the recognition performance of classroom behaviors. In recent years, artificial intelligence technologies-particularly image processing techniques driven by deep learning-have demonstrated growing potential within the field of education and have garnered increasing scholarly attention. CNNs have been employed to identify students' postures in classroom environments, marking preliminary efforts toward the automation of behavioral analysis. Additional studies have focused respectively on video-based assessments of student attentiveness and analyses of classroom engagement, thereby underscoring the value of deep learning in interpreting classroom dynamics. Moreover, sequential models have been applied to capture patterns of teacher-student interaction. These investigations have collectively indicated that deep learning-based automated analysis of classroom behavior from image and video data offers promising pathways for enhancing the objectivity of teaching assessments, optimizing instructional strategies, and promoting personalized learning. Nevertheless, current methodologies still face significant challenges in fully extracting complex temporal information embedded in classroom behavior image sequences, effectively perceiving behavior variations across different time scales, and constructing robust recognition models capable of maintaining high generalizability. Therefore, there is an urgent need to develop a more effective deep learning model that can better capture and analyze classroom behaviors.

In response to the aforementioned issues, a deep learning-based image processing system for classroom behavior analysis was proposed in this study. The primary research content includes three key components: a) the development of a temporal 2D convolution model for classroom behavior analysis to fully exploit the temporal information embedded in image data; b) the design of a method to expand the receptive field of the temporal 2D convolution, enhancing the model's ability to perceive behaviors at different time scales; c) the construction of a classroom behavior recognition network to improve the accuracy and robustness of behavior recognition. Through these efforts, this research aims to provide an efficient and accurate solution for classroom behavior analysis and contribute to the advancement of educational informatization.

2. CONSTRUCTION OF THE TEMPORAL 2D CONVOLUTION FOR CLASSROOM BEHAVIOR ANALYSIS

In the classroom teaching environment, student behaviors often contain rich temporal information. For example, actions, such as raising a hand, answering questions, and reading, exhibit specific temporal features. The detection and analysis of these behaviors play a critical role in enhancing teaching quality and classroom management. Conventional classroom behavior analysis methods have primarily relied on classical algorithms in computer vision, such as background subtraction, optical flow, and Histogram of Oriented Gradients (HOG) combined with Support Vector Machine (SVM) classifiers. These approaches have heavily depended on handcrafted feature design and extraction by researchers. However, this manual feature engineering paradigm presents several critical limitations:

a) Limited expressive capacity and high subjectivity: Handcrafted features often fail to comprehensively and

effectively capture the rich visual and spatiotemporal information embedded within the complex and dynamic behavior patterns present in classroom environments. These methods are particularly inadequate when addressing subtle motions or occlusions. The design and selection of features are highly reliant on expert knowledge, which introduces a significant degree of subjectivity and potential bias.

b) Weak temporal modeling capability: Classroom behavior is inherently a temporally dependent dynamic process. Traditional approaches have generally been limited to single-frame or short-sequence analysis, lacking the ability to model long-term behavioral evolution. As a result, they are often unable to distinguish between behaviors with similar initial or terminal states but different intermediate trajectories.

c) Insufficient robustness and generalizability: Manually crafted features tend to be sensitive to variations in lighting, viewpoint, occlusion, and background interference. Consequently, the performance of such models deteriorates across different classroom settings, camera angles, and student populations, undermining the stability and reliability of behavioral analysis outcomes.

These limitations have hindered the effective utilization of the vast amount of information embedded in classroom image sequences, often resulting in behavior recognition outcomes that fall short of the accuracy and robustness required for practical deployment. In contrast, deep learning technologies-particularly end-to-end deep neural networks-have demonstrated superior capability in automated feature learning. Such networks can directly learn robust and discriminative feature representations from raw image data. Therefore, the present study centers on the development of a deep learning-based solution, with a primary objective of overcoming the bottlenecks associated with manual feature extraction by designing advanced network architectures capable of automatically learning and fully exploiting the spatiotemporal information contained in classroom imagery.

To effectively analyze the temporal information in classroom behavior video images and improve the model's temporal perception capability without significantly increasing computational load, two improvements to the CNN were proposed in this study. First, a temporal 2D convolution model was constructed, utilizing the advantages of 2D convolution in spatial feature extraction while incorporating convolution operations in the temporal dimension to enhance the model's ability to perceive temporal information. Second, a method for expanding the receptive field of the temporal 2D convolution was designed, which appropriately enlarges the convolution kernel's receptive field to ensure that it covers the duration of most classroom behaviors. This method not only effectively captures long-duration behavioral features but also improves the model's temporal analysis capability without a significant increase in computational complexity.

The primary innovations and advantages of the proposed T2D-Conv module over existing approaches are as follows: a) Efficient temporal modeling: In contrast to computationally intensive 3D convolutional networks, T2D-Conv adopts a compositional strategy of "spatial-first, then-temporal," which significantly reduces both parameter count and computational complexity. This design enables more efficient training and deployment, particularly in classroom behavior analysis scenarios that require the processing of extended temporal sequences. b) Explicit temporal awareness: Unlike conventional 2D convolutional networks that are limited to processing single frames or simple frame stacks, T2D-Conv

incorporates 1D temporal convolutions following spatial feature extraction. This enables explicit modeling of inter-frame temporal dependencies and allows the dynamic characteristics of behaviors to be effectively captured. Consequently, the limitations of traditional 2D methods in temporal information utilization are addressed. c) Modularity and flexibility: The T2D-Conv module is designed as a modular component that can be flexibly integrated into existing 2D CNN backbones. Temporal modeling can be achieved simply by appending a 1D convolutional layer along the temporal dimension to the output feature maps of the backbone network. This design facilitates transfer learning and modular network construction. Therefore, the T2D-Conv module serves as the foundational component for constructing an efficient and temporally aware classroom behavior analysis model. It provides a robust temporal feature representation for the subsequent components proposed in later sections, including the dilated temporal convolution and the behavior recognition network.

In the context of classroom behavior analysis, conventional 2D convolution methods, although highly effective in spatial feature extraction, struggle to capture the temporal information inherent in behaviors. To address this limitation, an innovative temporal 2D convolution method was proposed. The core idea of this method is to effectively compress the information from multiple images into a single image, thereby allowing 2D convolution to perceive temporal relationships even when processing a single image. However, directly stacking multiple images may distort the original spatial information, making it challenging to accurately classify classroom behaviors. To resolve this issue, an adaptive filtering gate (AFG) mechanism was introduced, which selectively retains and combines information from images at different time points, thereby preserving spatial features while enhancing the model's ability to perceive temporal information. The core principle of this mechanism lies in learning a dynamic, input-dependent weight mask capable of adaptively emphasizing feature channels that are critical to the current task while suppressing those that are irrelevant or potentially disruptive. AFG adaptively assigns an importance weight to each feature channel based on the content of the input features. A weight value approaching 1 indicates that the corresponding channel conveys features essential to the current input and should be retained or even amplified. Conversely, a weight approaching 0 suggests that the channel likely contains noise or redundant information and should therefore be attenuated. Given the complexity and variability of classroom environments, extracted temporal features inevitably include noise or irrelevant information. Through dynamic weighting, AFG effectively filters out such interferences, thereby enabling the model to focus on feature cues that are genuinely pertinent to behavior recognition. This contributes to enhanced recognition accuracy and robustness under complex scenarios. Fundamentally, the mechanism operates as a form of feature calibration; rather than introducing new features, it reassigns significance to existing feature channels to improve the quality of feature representation.

To construct a temporal 2D convolution model for classroom behavior analysis, it is first necessary to understand how to perceive 3D temporal information at the 2D image level. Based on the understanding of conventional convolution operations, the temporal convolution operation can be decomposed into shifting and convolution operations. In 1D convolution, convolution weights $Q=\{q_1, q_2, q_3\}$ are convolved

with the input sequence $A=\{A_1, A_2, \dots, A_u\}$, where the sequence is shifted through displacement, and weighted sums are computed at each position to produce the output. When extending this concept to 2D image sequences, it becomes necessary to apply displacement operations on multiple consecutive image frames and combine them using convolutional kernel weights, thus enabling the perception of temporal information. Convolution can then be defined as $Y=Conv1D(W, X)B=CONV1D(Q, A)$, as expressed in the following formula:

$$B_u = q_1 A_{u-1} + q_2 A_u + q_3 A_{u+1} \quad (1)$$

The displacement operation formula is as follows:

$$\begin{aligned} A_u^{-1} &= A_{u-1} \\ A_u^0 &= A_u \\ A_u^{+1} &= A_{u+1} \end{aligned} \quad (2)$$

The convolution operation after displacement is expressed as:

$$B = q_1 A_u^{-1} + q_2 A_u^0 + q_3 A_u^{+1} \quad (3)$$

Specifically, in the construction of temporal 2D convolution, for each image frame, displacement operations along the time dimension were performed, with each image at different time points being weighted accordingly. It is assumed that each frame of the input video image sequence is a 2D matrix. By applying the displacement operation, each frame of the image was shifted along the time dimension, and corresponding convolution kernel weights were introduced to perform a weighted sum on each image frame. As a result, the outcome of the displacement and weighted summation not only contained the spatial information of the current frame but also integrated the temporal information of the preceding and succeeding frames. Let the 2D convolution kernel q be of size (l, v) , and let the resolution of the input image a be (u, k) . The conventional definition of 2D convolution is given by the following equation:

$$CONV(q, a) = \sum_l \sum_v a(u+l, k+v) q(l, v) \quad (4)$$

Let the batch size be represented by V , the number of channels by Z , the time dimension by S , and the resolution of a single image by u and k . The input vector for the temporal image sequence of classroom behavior can thus be represented as $[V, Z, S, u, k]$. For different time displacements of the channels, $Z_{X1}=\{Z_1, \dots, Z_{Z/3}\}$ and $Z_{X2}=\{Z_{Z/3+1}, \dots, Z_Z\}$, the displacement expression for the time dimension channels is given by:

$$\begin{aligned} A_s^{Z1} &= a_{s-1}^{Z1} \\ A_s^{Z2} &= a_s^{Z2} \end{aligned} \quad (5)$$

To address the limitations of fixed channel displacement in temporal feature extraction, a filtering gate mechanism was introduced within the temporal 2D convolution structure to enhance the effectiveness of classroom behavior analysis. Specifically, the feature map processed by each network layer represents the result of each image frame after 2D convolution,

with these feature maps arranged in chronological order. By adding a filtering gate at each layer, the network was enabled to dynamically select which channels should be moved and aggregated along the time dimension, allowing the model to flexibly adapt to variations in the duration of different

behaviors. The output dimension of the filtering gate was consistent with the number of channels, and its weights were trained through backpropagation, thereby adjusting the extraction of temporal features in real time for each channel.

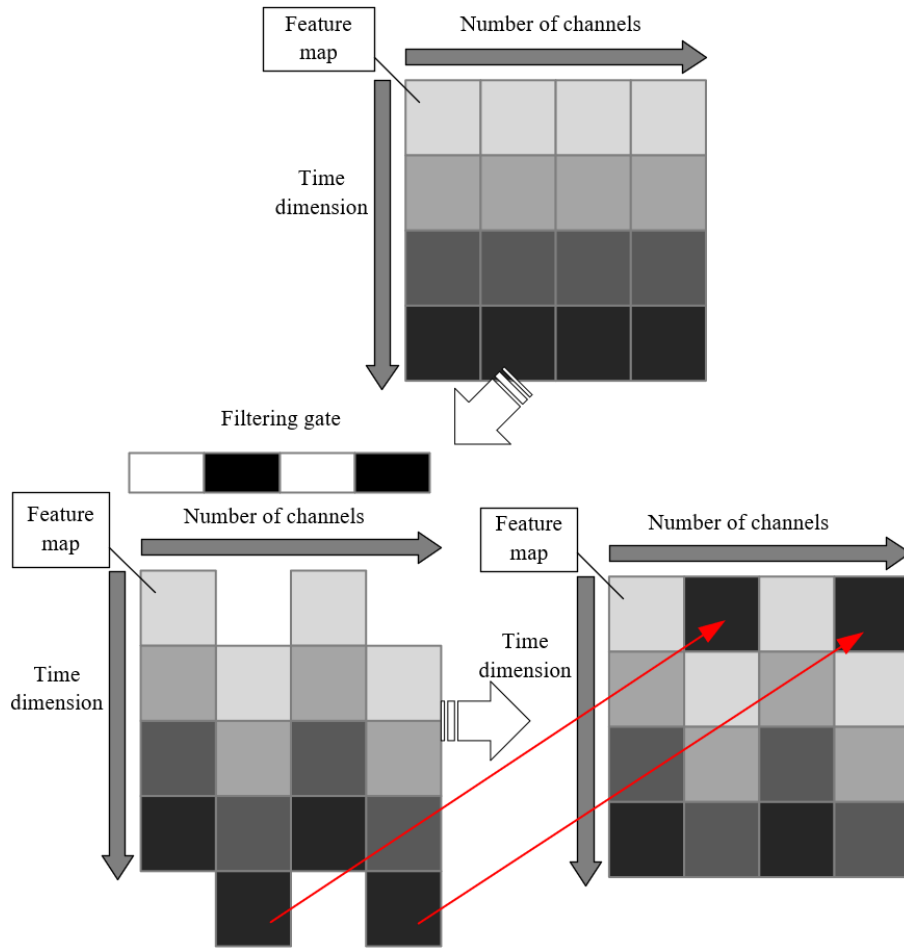


Figure 1. Example of the role of the filtering gate

The role of the filtering gate is to process its output values through a sigmoid activation function, ensuring that the values are distributed within the range $[0, 1]$. Figure 1 illustrates an example of the filtering gate's function. A threshold of 50% was used as the boundary, with channels whose output exceeds 0.5 undergoing temporal shifting, while those below 0.5 remain unchanged. This filtering gate mechanism allows the network to effectively extract temporal feature information without significantly increasing computational complexity. Let the input video image sequence consist of a series of frames. After 2D convolution, the feature maps obtained can determine, based on the output of the filtering gate, whether temporal shifting and accumulation should occur. Let the filtering gate before the activation function be represented by c . The computational formula for this layer is as follows:

$$\delta(c) = \frac{1}{1 + e^{(c)}} \quad (6)$$

As for the filtering gate output, the channel sequence that needs to be shifted is represented by $T = \{a | \delta(a)\}$, and the sequence that remain unchanged is represented by T' . The resulting channel displacement after applying the filtering gate is then given by:

$$\begin{aligned} A_s^T &= a_{s-1}^T \\ A_s^{T'} &= a_s^{T'} \end{aligned} \quad (7)$$

Further, the A convolution computation after temporal channel displacement is expressed as:

$$CONV(q, A) = \sum_l \sum_v A(u+l, k+v) q(l, v) \quad (8)$$

3. CALCULATION OF THE EXPANDED TEMPORAL 2D CONVOLUTION PERCEPTION RANGE FOR CLASSROOM BEHAVIOR ANALYSIS

When processing classroom behavior video images, the temporal perception range of the feature maps generated by the temporal 2D convolution must be considered. By stacking a sufficient number of temporal 2D convolutional layers, a broader temporal perception range and richer semantic features can, in theory, be obtained. However, the depth of the neural network cannot be increased indefinitely, as both computational cost and training difficulty rise significantly with the increase in layers. To effectively explore and optimize this process, a feedforward neural network was discussed in

this study using the nonlinear unit as the unit. This approach aids in a clearer analysis and understanding of the role and contribution of the temporal 2D convolution at specific layers. Let the neural network $D(a;r)$ consist of M stacked nonlinear units. Assuming that the computation of the neural network is denoted as G and the activation function is represented by h , the net input c and output x of the m -th layer of the neural network can be expressed by the following equation:

$$\begin{aligned} c^{(m)} &= G\left(x^{(m-1)}\right) \\ x^{(m)} &= h\left(x^{(m)}\right) \end{aligned} \quad (9)$$

To harness the advantages of deep networks, it is necessary to address issues such as overfitting, gradient vanishing, and gradient explosion effectively. The classroom behavior analysis model utilized in this study expands the perception range through multiple layers of the temporal 2D convolution, capturing the temporal dynamics within the video data to identify and analyze students' behaviors. To prevent overfitting, several regularization techniques were employed. These include L2 regularization, dropout, and data augmentation methods. L2 regularization involves adding a weight penalty term to the loss function to suppress excessively large network parameters, thereby reducing the risk of overfitting. Dropout works by randomly deactivating a portion of neurons during training, enhancing the model's robustness and preventing dependence on specific nodes. Additionally, data augmentation generates more diverse training samples through operations such as rotation, translation, and scaling, further enhancing the model's generalization capability. The combined use of these methods helps the deep neural network to better learn features during training, rather than memorizing the training data.

To address the issues of gradient vanishing and explosion, several advanced techniques and optimization strategies were introduced in this study. For example, batch normalization was employed to standardize the input of each batch of data, ensuring that the input distribution of each layer remains stable. This technique accelerates training convergence and reduces the risk of gradient vanishing. Additionally, appropriate weight initialization methods were adopted to ensure that the initial weights are within a suitable range, thereby preventing the gradual disappearance or explosion of gradients during both forward and backward propagation. Furthermore, the selection of suitable activation functions, such as ReLU and its variants, can effectively mitigate the gradient vanishing problem and improve training efficiency. Specifically, the parameter update at the M -th layer requires the computation of the gradient of the loss Z with respect to the layer. This gradient depends on the error term of the layer, denoted as $\sigma = \partial\gamma/\partial c^{(m)}$. According to the chain rule, $\sigma^{(m)}$ is related to the error term $\sigma^{(m+1)}$ of the subsequent layer:

$$\sigma^{(m)} = \frac{\partial c^{(m+1)}}{\partial c^{(m)}} \cdot \sigma^{(m+1)} \quad (10)$$

Let $\varepsilon^{(m)} = \partial c^{(m+1)}/\partial c^{(m)}$, then it leads to the following expression:

$$\sigma^{(m)} = \varepsilon^{(m)} \sigma^{(m+1)} \quad (11)$$

Assuming that the output and input dimensions of the neural network are consistent, the function G that needs to be fitted

can be divided into two parts as follows:

$$c^{(m_2)} = G\left(x^{(m-1)}\right) = x^{(m-1)} + D\left(x^{(m-1)}\right) \quad (12)$$

For computational convenience, if no activation functions are applied, the following expression can be obtained:

$$\begin{aligned} x^{(m_2)} &= x^{(m_2-1)} + D\left(x^{(m_2-1)}\right) \\ &= x^{(m_2-2)} + D\left(x^{(m_2-2)}\right) + D\left(x^{(m_2-1)}\right) \end{aligned} \quad (13)$$

By recursively expanding this expression, the final result can be derived as:

$$x^{(m_2)} = x^{(m_1)} + \sum_{u=m_1}^{m_2-1} D\left(x^{(u)}\right) \quad (14)$$

The gradient of the final loss γ with respect to a lower-layer output can be expanded as:

$$\begin{aligned} \frac{\partial \gamma}{\partial x^{(m_1)}} &= \frac{\partial \gamma}{\partial x^{(m_2)}} \frac{\partial x^{(m_2)}}{\partial x^{(m_1)}} \\ &= \frac{\partial \gamma}{\partial x^{(m_2)}} \left(1 + \frac{\partial x^{(m_2)}}{\partial x^{(m_1)}} \sum_{u=m_1}^{m_2-1} D\left(x^{(u)}\right) \right) \end{aligned} \quad (15)$$

The choice of optimization algorithm and hyperparameter tuning are also crucial for the training of deep neural networks. The Adam optimizer was employed in this study, which adjusts the update step size of each layer's parameters through an adaptive learning rate. The incorporation of momentum terms helps to reduce oscillations, further stabilizing the training process.

As previously described, the standard T2D-Conv module captures sequential frame information through 1D convolutional kernels. However, classroom behaviors exhibit pronounced multi-scale temporal characteristics: certain actions occur over brief durations, whereas others span significantly longer periods. To enhance the model's ability to perceive long-duration behavioral patterns-while avoiding the computational burden and optimization challenges associated with excessively deep convolutional stacks-a dilated convolution mechanism was introduced into the temporal 1D convolutional component, forming the D-T2D-Conv module.

The primary advantage of dilated convolution lies in its ability to exponentially expand the receptive field. By simply increasing the dilation rate, D-T2D-Conv significantly extends the model's temporal receptive field without increasing the number of convolutional parameters or the network depth. This enhancement markedly improves the model's capability to capture long-range dependencies, thereby enabling the effective recognition of classroom behaviors that span dozens or even hundreds of frames. In contrast, achieving an equivalent receptive field with standard convolution would necessitate deep layer stacking, substantially increasing model complexity and training difficulty.

In the system design, multiple groups of D-T2D-Conv modules with different dilation rates were applied across various network hierarchies, forming a multi-scale temporal feature extraction pyramid. Lower layers employed smaller dilation rates to capture fine-grained, short-term actions, while upper layers utilized larger dilation rates to model coarse-

grained, long-duration behavioral patterns. This architectural strategy constitutes a core technique for enhancing the model's perceptual capacity across diverse temporal scales in classroom behavior recognition.

4. CONSTRUCTION OF THE CLASSROOM BEHAVIOR RECOGNITION NETWORK

The goal of this study is to develop an efficient and real-time neural network architecture for accurately recognizing and analyzing student behavior in the classroom. To achieve this, the network architecture must not only possess strong feature extraction capabilities but also maintain computational efficiency, avoiding performance degradation due to excessive parameters and layers that may affect real-time processing. To address these challenges, a core structure based on the temporal 2D convolution and dilated perception was proposed in this study, which cleverly balances performance with computational efficiency.

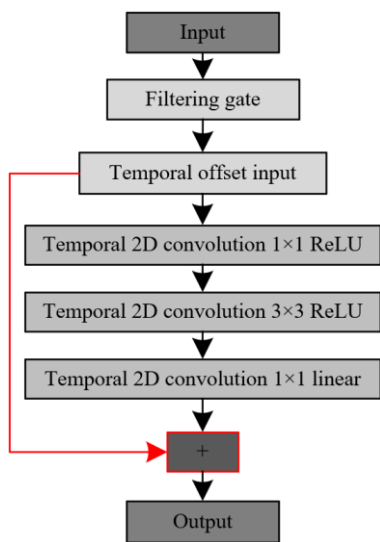


Figure 2. Temporal offset dilation structure

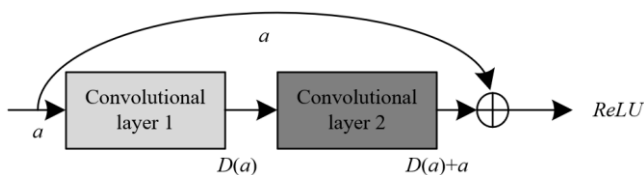


Figure 3. Identity mapping structure

The network adopts a temporal offset dilation core structure that integrates the temporal 2D convolution and dilated perception techniques. Specifically, the input to the network first passes through a filtering gate channel to filter out key temporal information. The filtered channel data is then processed with a temporal offset, transforming it into input with temporal information. This processing method effectively captures temporal features in video data, catering to the requirements of classroom behavior analysis. To prevent the gradient vanishing problem caused by the multi-layer stacking of the temporal offset dilation structure shown in Figure 2, an identity mapping structure, as shown in Figure 3, was introduced after the temporal offset input. The addition of the identity mapping ensures that gradients are successfully

transmitted, maintaining training stability in the network. In each layer, to maintain the equivalence of the input and output matrix dimensions, a 1×1 temporal 2D convolution was used for dimensionality reduction, followed by a 3×3 temporal 2D convolution for feature extraction. A 1×1 temporal 2D convolution was then employed for dimensionality expansion, preparing for the subsequent identity mapping addition.

The core of the network architecture lies in efficiently extracting and processing temporal features to achieve accurate recognition of various classroom behaviors. Initially, a preprocessing module was employed, where three-layer 3×3 convolution operations progressively compress the input high-resolution video frames ($224 \times 224 \times 3$) into lower-resolution feature maps ($56 \times 56 \times 24$), while simultaneously increasing the number of feature channels. This operation significantly reduces the computational load while enhancing the expressive capacity of the feature maps, thereby laying a solid foundation for subsequent temporal feature extraction. At this stage, by reducing the resolution and increasing the number of channels, the computational load of the input was reduced from approximately 150,000 to around 70,000, which is roughly half of the original amount, making the network more efficient when processing large volumes of video frames.

The core structure of the network, the temporal offset dilation structure, gradually expands the temporal receptive field by stacking 12 layers. Each layer of the temporal offset dilation structure ensures stable gradient propagation through identity mapping and utilizes the 1×1 convolution with a stride of 2 to perform dimensionality expansion. This approach allows deeper features to be extracted while halving the feature map and doubling the dimensionality. Each layer applies temporal offset processing to the input feature map, enabling the network to effectively capture and integrate behavioral information spanning up to 6 seconds. Finally, through average pooling, the feature map can be converted into a highly compact 1×1 feature vector, which can then be connected to a fully connected layer to output the final behavior category. This architecture not only ensures the full utilization of temporal information but also maintains computational efficiency and feature richness through judicious dimensionality reduction and expansion, providing efficient and precise technical support for classroom behavior recognition. Assuming that the probability of the current action belonging to the u -th action category is represented by T_u , the numerical output of the u -th action category is denoted as C_u , and the category is denoted as Z . The output category formula is as follows:

$$T_u = \frac{e^{C_u}}{\sum_{z=1}^Z e^{C_z}} \quad (16)$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, the classroom behavior recognition network architecture describes a multi-level deep learning model specifically designed for the analysis and recognition of behaviors in classroom environments. Each input image to the network contains multiple temporal frames. A series of different convolutional layers, combined with the temporal offset dilation structure, was used to progressively extract features from low-level to high-level. In the initial stages,

multiple standard convolutional layers were used for spatial feature extraction, with the output channel count increasing, while convolution operations with a stride of 1 were employed to maintain spatial resolution while ensuring feature dimension consistency. As the network deepened, the temporal offset dilation convolution structure was introduced to enhance the ability to perceive temporal information. Particularly, through the use of different dilation rates and repetition counts, the network was able to model and extract temporal features across various time scales, effectively capturing the dynamic changes in classroom behaviors.

Table 1. Classroom behavior recognition network architecture

Input	Network Type	Number of Output Channels	Repetition Count of the Layer	Stride
11*215*215*3	Standard 3×3 convolution	31	1	2
11*125*125*31	Standard 3×3 convolution	15	1	1
11*125*125*31	Standard 3×3 convolution	23	1	2
11*55*55*23	Temporal offset dilation structure	23	3	1
11*55*55*23	Standard 1×1 convolution	31	1	2
11*27*27*31	Temporal offset dilation structure	31	2	1
11*27*27*31	Standard 1×1 convolution	62	1	2
11*12*12*62	Temporal offset dilation structure	62	3	1
11*12*12*62	Standard 1×1 convolution	95	1	1
11*12*12*94	Temporal offset dilation structure	95	3	1
11*12*12*94	Standard 1×1 convolution	158	1	2
11*7*7*158	Temporal offset dilation structure	158	1	1
11*7*7*335	Standard 1×1 convolution	1125	1	1
11*7*7*1128	Average pooling	-	1	-
11*1*1*1128	Fully connected	4	1	-

In the experimental section, the first task was the prediction and detection of classroom behavior boundaries, with the goal of distinguishing between intervals containing classroom behaviors and those without. Figure 4 illustrates the conceptual approach for boundary prediction, which effectively segments long videos into multiple segments. Each segment was marked with the action intervals of both students and teachers, allowing the model to focus on analyzing only those parts of the video that contain classroom behavior. This segmentation improves data processing efficiency and model accuracy. This step not only reduces interference from irrelevant data but also provides more precise input data for the subsequent behavior recognition network. The segments selected through this method serve as the input for the

classroom behavior recognition network, enabling further detailed action recognition.



Figure 4. Prediction of classroom behavior boundaries to identify the student-teacher action intervals

Table 2. Results of classroom behavior recognition testing

Dataset	Correct Action Accuracy	Incorrect Action Accuracy	False Positive Rate	False Negative Rate
Training dataset	95%	91%	0%	0%
Testing dataset	87%	85%	22%	12%

Based on the classroom behavior recognition test results presented in Table 2, it can be observed that the system's performance differs between the training and testing datasets. On the training dataset, the model achieved high accuracy, with a correct action accuracy of 95%, an incorrect action accuracy of 91%, and both false positive and false negative rates of 0%, indicating that the model was able to effectively identify classroom behaviors during training without any misclassifications or omissions. In contrast, the results for the testing dataset were slightly lower, with a correct action accuracy of 87%, an incorrect action accuracy of 85%, a false positive rate of 22%, and a false negative rate of 12%. This suggests that, while the model's behavior recognition ability remains highly accurate in practical applications, the increased diversity and complexity of the testing data led to a higher occurrence of false positives and false negatives.

Table 3. Performance comparison of different models

Network Model	Backbone Network	Number of Sample Frames	Weight	mAP
Classroom behavior recognition network	Temporal expansion network	11	8.8 M	44.6
Two-stream networks	VGG	8	46.2 M	38.9
SlowFast networks	ResNet	31	34.5 M	42.3
Temporal segment networks	ResNet	8	32.6 M	37.8
Inflated 3D ConvNet	Inception-v1	8	11.2 M	18.6

As shown in Table 3, the proposed classroom behavior recognition network demonstrates excellent performance in terms of mean Average Precision (mAP), achieving a value of 44.6, which outperforms several other mainstream network models. The model employs a temporal expansion network as

the backbone and was trained with a sample size of 11 frames, with a weight parameter of 8.8 M. In comparison, the two-stream networks, which use the Visual Geometry Group (VGG) backbone, have a higher weight (46.2 M) but achieve an mAP of 38.9, which is slightly lower than the proposed network. SlowFast networks, based on the ResNet backbone and with 31 sample frames, achieve an mAP of 42.3, indicating good performance in spatiotemporal feature extraction, although it does not surpass the proposed network. Additionally, temporal segment networks and inflated 3D ConvNet, which use the ResNet and Inception-v1 backbones, respectively, show weaker performance with mAP values of 37.8 and 18.6. The comparison results clearly show that the classroom behavior recognition network proposed in this study performs superiorly across several aspects, particularly in accurately identifying classroom behaviors. While other models, such as two-stream networks and SlowFast networks, also perform well in processing temporal and spatial features, their more complex network structures and higher parameter weights limit their computational efficiency and generalization ability. The temporal expansion network designed in this study, with fewer parameters (8.8 M) and a reasonable frame count (11 frames), achieves a balance between accuracy, computational efficiency, and robustness.

Table 4. Comparison of processing speeds of different models

Network Model	FLOPs	Weight	Processing Time
<i>Inflated 3D ConvNet</i>	315 G	34.6 M	156.2 ms
<i>Two-Stream Networks</i>	63 G	46.2 M	31.2 ms
<i>SlowFast Networks</i>	32 G	28.5 M	24.3 ms
<i>Classroom behavior recognition network</i>	16 G	8.8 M	12.4 ms

As shown in Table 4, the classroom behavior recognition network proposed in this study demonstrates a significant advantage in processing speed. The model has a floating-point operations (FLOPs) count of 16 G, with a weight of 8.8 M and a processing time of only 12.4 ms, indicating high computational efficiency. In contrast, although the inflated 3D ConvNet and two-stream networks perform well in terms of

recognition accuracy, their processing times are 156.2 ms and 31.2 ms, respectively. Furthermore, the former has a FLOPs count of 315 G, and the latter 63 G, suggesting that these models incur higher computational costs. Even the SlowFast Networks, with a FLOPs count of 32 G, has a processing time of 24.3 ms, which still does not match the performance of the proposed network.

To provide a more intuitive and in-depth understanding of system performance beyond quantitative metrics, a detailed qualitative analysis was also conducted. A comprehensive review of numerous video samples was undertaken to summarize representative cases of successful recognition and to examine challenging scenarios that led to misclassification, thereby revealing both the strengths and current limitations of the system. In the successfully recognized cases, the system demonstrated robust capability in capturing a variety of typical classroom behaviors. For instance, in the recognition of the “raising hand to ask a question” behavior, the model was able to accurately track the full sequence of motion—from the initial lift of the arm, through the maintained raised-hand posture, to the eventual lowering of the hand—while consistently producing high-confidence outputs. This outcome substantiates the effectiveness of the T2D-Conv model in modeling the dynamic evolution of continuous actions. In the detection of the “taking notes” behavior, the recognition results remained stable even when students momentarily looked up at the blackboard, thereby interrupting the writing motion. This stability was attributed to the moderate window-overlap strategy and the model’s temporal contextual memory, which allowed temporal segments to be associated across time, effectively preventing misclassification due to short-term interruptions. For the “attentive listening” state, high-confidence results were continuously produced when students maintained a standard seated posture facing the lectern. Even in the presence of minor background disturbances, the AFG mechanism was presumed to have effectively suppressed noise activation in non-critical regions, maintaining focus on the target student. In addition, for long-duration “inattentive” behaviors, the use of high-level D-T2D-Conv modules with large dilation rates enabled the successful capture of these prolonged deviation patterns spanning dozens of frames.

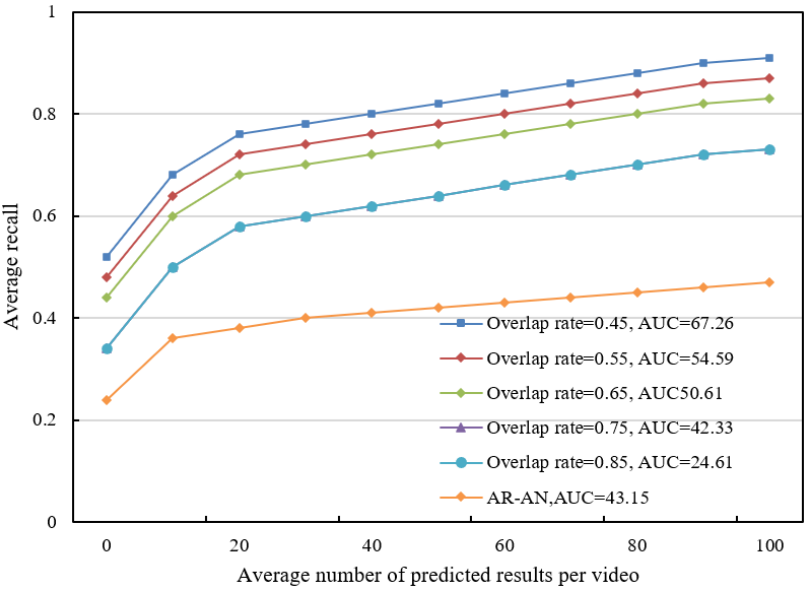


Figure 5. Testing results on the actual classroom video dataset

Based on the data presented in Figure 5, which shows the variation in average recall with respect to the average number of predicted results per video, differences in performance at varying overlap rates can be observed. When the overlap rate is 0.45, the average recall increases gradually from 0.52 to 0.91, with an Area Under the Curve (AUC) of 67.26. As the number of predicted results increases, the average recall steadily rises, indicating that the model performs well in capturing actual behavioral segments at lower overlap rates. When the overlap rate is 0.55, the initial average recall is 0.48, which eventually reaches 0.87, with an AUC of 54.59. As the overlap rate increases to 0.65, the initial recall decreases to 0.44, and it eventually reaches 0.83, with an AUC of 50.61. Further increasing the overlap rate to 0.75 results in an initial recall of 0.34, which rises to 0.73, with an AUC of only 42.33. At an overlap rate of 0.85, the initial recall remains 0.34, but the final recall reaches only 0.73, with an AUC of 24.61. In comparison, the AR-AN method shows a gradual increase in average recall from 0.24 to 0.47, with an AUC of 43.15, overall underperforming. From the experimental results, it can be observed that the proposed deep learning-based classroom behavior analysis system demonstrates a superior recall at lower overlap rates. The average recall steadily improves as the number of predicted results increases, suggesting that the model can accurately capture time segments that overlap with actual behavior segments. However, as the overlap rate increases, the recall performance of the system significantly decreases, and the AUC value also drops sharply, reflecting the model's difficulty in maintaining high accuracy at higher overlap rates. Particularly at an overlap rate of 0.85, despite the increase in the number of predicted results, the final recall shows only limited improvement, with the AUC dropping to 24.61, indicating a significant decline in the model's recognition performance in this scenario.

To assess the system's sensitivity to behavior boundary delineation, performance was evaluated under different sliding window overlap ratios during input video segmentation for prediction generation. The experimental results demonstrated that as the overlap ratio increased from 0% to 50%, the overall recognition accuracy initially improved and then plateaued. Notably, optimal or near-optimal performance was typically achieved when the overlap ratio ranged between 25% and 33%. The performance variation under different overlap settings can be attributed to the following factors: a) Continuity of behaviors and boundary ambiguity: Classroom behaviors generally evolve continuously, with indistinct onset and offset boundaries. When the window overlap is too low, a severe discontinuity is introduced between adjacent segments. As a result, behavioral sequences may be split at the window boundaries, causing the model to observe only partial behavior segments. This fragmentation leads to incomplete feature representations and increases the likelihood of misclassification or missed detections. b) Temporal modeling and contextual information: The proposed D-T2D-Conv module relies on adequate contextual frames to effectively model temporal dynamics. A moderate overlap ratio ensures that a substantial number of frames are shared across adjacent windows. Consequently, even when a behavior spans across window boundaries, a relatively complete contextual representation is preserved within multiple overlapping windows. More comprehensive information from preceding and succeeding frames is utilized when predicting behaviors near the center of each sliding window, thereby enhancing the accuracy of assessing the complete behavioral process and

mitigating the adverse effects of boundary segmentation. c) Computational redundancy and noise amplification: While high overlap ratios theoretically minimize boundary segmentation issues, they also introduce significant redundancy due to repeated frames across adjacent windows. This redundancy leads to increased computational cost from repeated feature extraction and inference across highly similar windows, thereby reducing processing efficiency. Moreover, the high degree of content similarity among overlapping windows may amplify transient or local noise or mispredictions during result fusion, rather than providing diverse and complementary information. In some cases, this effect may even slightly degrade overall performance or result in diminishing returns.

Considering both recognition accuracy and computational efficiency, an overlap ratio of 25% to 33% is recommended. This configuration strikes an effective balance between minimizing behavior boundary fragmentation, ensuring sufficient contextual representation, and controlling computational redundancy. This parameter serves as a critical factor in maximizing the performance of the proposed multi-scale temporal modeling network. Furthermore, the findings reinforce the importance of effectively leveraging temporal context information to enhance the robustness of behavior recognition.

Table 5. Comparison of testing results on the actual classroom video dataset with other mainstream models

Network Model	mIoU	mAP@0.5IoU	mAP@0.7IoU	FPS
<i>Two-stream networks</i>	62.26	81.23	42.31	75
<i>SlowFast networks</i>	72.15	91.25	61.25	42
<i>Temporal segment networks</i>	76.36	87.26	64.58	31
<i>Inflated 3D ConvNet</i>	51.23	57.54	9.36	41
<i>Classroom behavior recognition network</i>	83.21	96.36	81.26	75

Based on the comparison of the test results from the actual classroom video dataset with other mainstream models provided in Table 5, the proposed classroom behavior recognition network demonstrates superior performance across multiple evaluation metrics. The model achieves a mean intersection over union (mIoU) of 83.21, significantly outperforming other leading network models, such as the two-stream networks (62.26) and SlowFast networks (72.15). On the mAP@0.5IoU metric, the proposed network also leads all comparison models with a score of 96.36, especially surpassing the inflated 3D ConvNet (57.54). The model continues to maintain a leading position in mAP@0.7IoU with a score of 81.26, well above SlowFast networks (61.25) and temporal segment networks (64.58). Additionally, in terms of frames per second (FPS), the processing speed of the proposed classroom behavior recognition network is comparable to that of the two-stream networks, both achieving 75 FPS, indicating high real-time performance suitable for real-time classroom behavior analysis applications. The comparison results clearly indicate that the proposed network excels in both accuracy and efficiency. Particularly in the mIoU, mAP@0.5IoU, and mAP@0.7IoU metrics, the model significantly outperforms other mainstream models, demonstrating a strong ability to accurately recognize classroom behaviors and more precisely segment and identify behavior-related time intervals. Furthermore, the network's real-time performance is excellent, with a processing speed of 75 FPS, matching that of the two-

stream networks, ensuring efficient real-time feedback capability. In contrast, the inflated 3D ConvNet performs poorly across multiple metrics, particularly in mAP@0.7IoU, where it scores only 9.36, highlighting its deficiencies in both accuracy and robustness. Overall, the proposed network not only offers significant advantages in recognition accuracy but also excels in computational efficiency, showing greater potential for practical applications, especially in meeting the high efficiency and real-time requirements of actual classroom behavior analysis. Figure 6 provides an example of classroom behavior tracking and judgment.



Figure 6. Example of classroom behavior tracking and judgment

To evaluate the temporal and spatial computational efficiency, a comparative experiment was conducted using 3D convolution as the baseline method on the ClassroomAction dataset. As presented in Table 6, the T2D-Conv module achieved a 68.3% reduction in FLOPs, a 52.7% decrease in the number of model parameters, and a 41.2% improvement in per-frame inference time relative to the standard 3D convolution. These gains can be attributed to the structural optimization offered by the T2D-Conv, wherein the spatial and temporal convolution operations are decoupled, thereby preserving the temporal feature modeling capacity while significantly enhancing computational efficiency. Such lightweight architectural design is particularly advantageous for real-time analysis in classroom scenarios, as it reduces the computational burden on edge devices and facilitates practical system deployment.

Table 6. Comparative evaluation of temporal and spatial computational complexity

Method	FLOPs (G)	Parameters (M)	Inference Time (ms/frame)
3D-Conv	12.6	28.4	32.7
T2D-Conv	3.9	13.4	19.2
Performance improvement	↓68.3%	↓52.7%	↓41.2%
Method	FLOPs (G)	Parameters (M)	Inference time (ms/frame)
3D-Conv	12.6	28.4	32.7

To further validate the effectiveness of the AFG mechanism, a pair of comparative experiments was designed, consisting of Model A with the mechanism enabled and Model B without it. On the standard test set of the ClassroomAction dataset, Model A achieved an average recognition accuracy of 89.7%, representing a 4.5 percentage point improvement over Model B, which yielded an accuracy of 85.2%, as detailed in Table 7. Under robustness evaluation conditions with 10% Gaussian noise added to the input images, the performance of Model A decreased to 82.3%, while Model B dropped more sharply to 75.1%. These results demonstrate the AFG mechanism's capacity to suppress noise interference effectively. By

dynamically adjusting the weight distribution across spatiotemporal features, the mechanism enables more precise focus on salient behavioral cues while mitigating the impact of background noise and irrelevant information. Consequently, the mechanism enhances the practical applicability of the model in real-world classroom environments.

Table 7. Comparative results for the AFG mechanism

Experimental Condition	Model A (with the Mechanism)	Model B (without the Mechanism)
Standard test set	89.7%	85.2%
10% Gaussian noise	82.3%	75.1%

6. CONCLUSION

The deep learning-based classroom behavior analysis image processing system proposed in this study demonstrated significant advantages in the task of classroom behavior recognition. By constructing a temporal 2D convolution model designed for classroom behavior analysis, the temporal information within image data was thoroughly explored. Additionally, a method to enhance the receptive field of the dilated temporal 2D convolution was devised, further strengthening the model's ability to perceive behaviors at different time scales. Moreover, the proposed classroom behavior recognition network not only achieved breakthroughs in accuracy but also outperformed current mainstream models, such as two-stream networks and SlowFast networks, across multiple evaluation metrics. Particularly, it exhibited outstanding performance in metrics such as mIoU, mAP@0.5IoU, and mAP@0.7IoU, thereby validating the method's efficiency and robustness in classroom behavior recognition. More importantly, the model also demonstrated strong real-time performance, with a processing speed of 75 FPS, which meets the real-time feedback requirements in actual classroom environments.

However, despite the excellent results achieved in multiple aspects in this study, some limitations remain. First, the model's performance is constrained by the diversity and complexity of the dataset. The current experiments were conducted only on a specific classroom video dataset, and future work should focus on validating the model's generalization capability on a wider range of scenarios and datasets. Second, while a balance between accuracy and computational efficiency was achieved, further optimization of the model may require deeper algorithmic refinements to reduce computational resource consumption and improve real-time performance. Despite the advances achieved by the proposed deep learning-based classroom behavior analysis system, several challenges and promising directions remain to be further explored: a) Enhancing robustness under dynamic and complex classroom environments and improving recognition of intricate behaviors: A performance decline has been observed in scenarios involving extreme lighting variations, severe occlusions, unconventional camera angles, or densely interactive student behaviors. Future efforts should be directed toward the development of more robust spatiotemporal feature representation learning methods and interference-resistant mechanisms. b) Integrating multimodal data to enrich analytical depth and dimensionality: Reliance on visual data alone has proven insufficient for capturing the full complexity of classroom dynamics. The incorporation of

multimodal information-such as audio signals, textual content, and even physiological indicators-is regarded as a necessary progression for achieving deeper behavioral understanding. c) Model lightweighting and optimization for real-time performance: Real-time processing capabilities still require further enhancement, particularly under high-resolution and multi-stream video conditions. Additional optimization is needed to support real-time feedback in large-scale or routine classroom deployments.

In the area of multimodal data integration, a hybrid strategy combining feature- and decision-level fusion is planned to be adopted. Feature-level fusion is first performed using a cross-modal attention mechanism to integrate video frames, audio signals, and textual data derived from classroom environments. Subsequently, decision-level fusion is conducted through ensemble learning techniques to integrate the classification outcomes of each modality. Regarding model lightweighting, a joint optimization scheme combining channel pruning and knowledge distillation is proposed. Initially, redundant convolutional channels are pruned based on a Taylor expansion-based channel importance evaluation algorithm. Thereafter, knowledge distillation is employed, whereby a pre-trained teacher model guides a lightweight student model. This approach allows recognition accuracy to be preserved while further improving operational efficiency on mobile devices. Continued research along these technical trajectories is anticipated to contribute to the development of a more robust, intelligent, and practical classroom behavior analysis system, thereby offering a stronger technological foundation for the advancement of intelligent education.

ACKNOWLEDGMENT

This paper was funded by 2024 Annual Doctoral Research Startup Fund Project of Liaodong University (Second Batch): Construction of an Excellent Professional Growth System for Primary School Teachers in Local Undergraduate Colleges Based on Digital Technology (Grant No.: 2024BS025); and 2025 Guangxi Humanities and Social Sciences Development Research Center "Scientific Research Project · Study on the Construction of University Faculty in the New Era" (Project No. JSDWY2025007), titled "Teaching Paradigm Shifts and Capacity Rebuilding Amid Educational Digital Transformation."

REFERENCES

- [1] DeJaeghere, J., Duong, B.H., Dao, V. (2023). Quality of teaching and learning: The role of metacognitive teaching strategies in higher-performing classrooms in Vietnam. *Educational Research for Policy and Practice*, 22(2): 239-258. <https://doi.org/10.1007/s10671-023-09330-x>
- [2] Saqlain, M. (2023). Evaluating the Readability of English Instructional Materials in Pakistani Universities: A Deep Learning and Statistical Approach. *Education Science and Management*, 1(2): 101-110. <https://doi.org/10.56578/esm010204>
- [3] Zhang, L.J., Wu, J.Z., Wei, J.X., Yu, X.Y., Yu, J., Yuan, B. (2023). Enhanced laboratory safety education through interactive applications of machine learning-boosted image processing technologies. *Traitement du Signal*, 40(6): 2623-2633. <https://doi.org/10.18280/ts.400624>
- [4] Yan, J., Wang, N., Wei, Y.M., Han, M.L. (2023). Personalized learning pathway generation for online education through image recognition. *Traitement du Signal*, 40(6): 2799-2808. <https://doi.org/10.18280/ts.400640>
- [5] Lakhani, P., Gray, D.L., Pett, C.R., Nagy, P., Shih, G. (2018). Hello world deep learning in medical imaging. *Journal of Digital Imaging*, 31: 283-289. <https://doi.org/10.1007/s10278-018-0079-6>
- [6] Pain, C.D., Egan, G.F., Chen, Z. (2022). Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement. *European Journal of Nuclear Medicine and Molecular Imaging*, 49(9): 3098-3118. <https://doi.org/10.1007/s00259-022-05746-4>
- [7] Schilling, M.P., Schmelzer, S., Klinger, L., Reischl, M. (2022). KaIDA: A modular tool for assisting image annotation in deep learning. *Journal of Integrative Bioinformatics*, 19(4): 20220018. <https://doi.org/10.1515/jib-2022-0018>
- [8] Mishra, R.K., Urolagin, S., Jothi, J.A.A., Gaur, P. (2022). Deep hybrid learning for facial expression binary classifications and predictions. *Image and Vision Computing*, 128: 104573. <https://doi.org/10.1016/j.imavis.2022.104573>
- [9] Zhang, X. (2021). Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4819-4838.
- [10] Gold, B., Foerster, S., Holodyski, M. (2013). Evaluation of a video-based training to foster the professional vision of classroom management in elementary classrooms. *Zeitschrift Fur Padagogische Psychologie*, 27(3): 141-155.
- [11] Wang, S., Cheng, L., Liu, D., Qin, J., Hu, G. (2022). Classroom video image emotion analysis method for online teaching quality evaluation. *Traitement Du Signal*, 39(5): 1767-1774. <https://doi.org/10.18280/ts.390535>
- [12] Prilop, C.N., Weber, K.E., Kleinknecht, M. (2021). The role of expert feedback in the development of pre-service teachers' professional vision of classroom management in an online blended learning environment. *Teaching and Teacher Education*, 99: 103276. <https://doi.org/10.1016/j.tate.2020.103276>
- [13] Scherr, R.E. (2009). Video analysis for insight and coding: Examples from tutorials in introductory physics. *Physical Review Special Topics-Physics Education Research*, 5(2): 020106. <https://doi.org/10.1103/PhysRevSTPER.5.020106>
- [14] Dalland, C.P., Klette, K., Svenkerud, S. (2020). Video studies and the challenge of selecting time scales. *International Journal of Research & Method in Education*, 43(1): 53-66. <https://doi.org/10.1080/1743727X.2018.1563062>
- [15] Koc, M. (2011). Let's make a movie: Investigating pre-service teachers' reflections on using video-recorded role playing cases in Turkey. *Teaching and Teacher Education*, 27(1): 95-106. <https://doi.org/10.1016/j.tate.2010.07.006>
- [16] Gold, B., Hellermann, C., Holodyski, M. (2017). Effects of video-based trainings for promoting self-efficacy in elementary classroom management. *Zeitschrift für Erziehungswissenschaft*, 20: 115-136.

- <https://doi.org/10.1007/s11618-017-0727-5>
- [17] Zhou, Y., Wang, J., Zhang, J. (2024). A multimodal image recognition system for student behavior analysis in smart classrooms in universities. *Traitement du Signal*, 41(6): 3285-3293. <https://doi.org/10.18280/ts.410644>
- [18] Lin, J., Li, J., Chen, J. (2022). An analysis of English classroom behavior by intelligent image recognition in IoT. *International Journal of System Assurance Engineering and Management*, 13(Suppl 3): 1063-1071. <https://doi.org/10.1007/s13198-021-01327-0>
- [19] Aspiranti, K.B., Bebech, A., Ruffo, B., Skinner, C.H. (2019). Classroom management in self-contained classrooms for children with autism: Extending research on the color wheel system. *Behavior Analysis in Practice*, 12: 143-153. <https://doi.org/10.1007/s40617-018-0264-6>
- [20] Fan, Z., Liu, J. (2022). Correlation analysis between teachers' teaching psychological behavior and classroom development based on data analysis. *Frontiers in Psychology*, 13: 905029. <https://doi.org/10.3389/fpsyg.2022.905029>
- [21] Li, L., Chen, C.P., Wang, L., Liang, K., Bao, W. (2023). Exploring artificial intelligence in smart education: Real-time classroom behavior analysis with embedded devices. *Sustainability*, 15(10): 7940. <https://doi.org/10.3390/su15107940>
- [22] Kale, U. (2008). Levels of interaction and proximity: Content analysis of video-based classroom cases. *The Internet and Higher Education*, 11(2): 119-128. <https://doi.org/10.1016/j.iheduc.2008.06.004>
- [23] Trout, K.P., Adkins, M., Bekker, J., Harlacher, A., Ramirez, F., Swingler, A., Wagner, C. (2021). Using iPads for video analysis physics labs in times of social isolation. *The Physics Teacher*, 59(5): 370-372. <https://doi.org/10.1119/10.0004893>
- [24] Hofman, J. (2023). Classroom management and teacher emotions in secondary mathematics teaching: A qualitative video-based single case study. *Education Inquiry*, 14(3): 389-405. <https://doi.org/10.1080/20004508.2022.2028441>
- [25] DeCuir-Gunby, J.T., Marshall, P.L., McCulloch, A.W. (2012). Using mixed methods to analyze video data: A mathematics teacher professional development example. *Journal of Mixed Methods Research*, 6(3): 199-216. <https://doi.org/10.1177/1558689811421174>
- [26] El-Shafai, W., Mahmoud, A.A., Ali, A.M., El-Rabaie, E.S.M., Taha, T.E., El-Fishawy, A.S., El-Samie, F.E.A. (2024). Efficient classification of different medical image multimodalities based on simple CNN architecture and augmentation algorithms. *Journal of Optics*, 53(2): 775-787. <https://doi.org/10.1007/s12596-022-01089-3>
- [27] Gaafar, A.S., Dahr, J.M., Hamoud, A.K. (2022). Comparative analysis of performance of deep learning classification approach based on LSTM-RNN for textual and image datasets. *Informatica*, 46(5): 21-28. <https://doi.org/10.31449/inf.v46i5.3872>