IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 53-63
Journal homepage: http://iieta.org/journals/ts

Learning State Quantification of University Classroom Behaviors Based on Multiscale Convolution and Attention Mechanisms

Xiaoxu Liu

School of Marxism, Northeast Agricultural University, Harbin 150010, China

Corresponding Author Email: 123360414@qq.com

https://doi.org/10.18280/ts.430104

Received: 15 August 2025
Revised: 23 December 2025
Accepted: 10 January 2026
Available online: 28 February 2026

Keywords:
classroom behavior recognition, learning state quantification, multiscale convolution, frequency-domain attention, cross-modal fusion, deep learning

ABSTRACT

Recognizing university classroom behaviors and quantifying students' learning states are essential components in the development of intelligent teaching systems, providing critical support for instructional optimization and precise learning analytics. However, existing image-based classroom behavior recognition methods often suffer from limited adaptability to target scale variations, insufficient discrimination between easily confused behaviors, and difficulties in transforming recognition results into interpretable quantitative indicators of learning states. To address these challenges, this study proposes a dual-branch spatiotemporal–frequency attention network to achieve high-precision classroom behavior recognition and interpretable learning state quantification. The network incorporates a spatial-domain branch based on multiscale deformable convolution to construct a dynamic scale-weight allocation mechanism, enabling adaptive balancing of feature representations for targets of different scales. In addition, a frequency-domain attention branch is introduced to capture fine-grained behavioral differences by integrating frequency selection and phase-aware representations. A cross-modal semantic alignment and fusion module is further designed, which leverages cross-attention to accurately integrate heterogeneous spatial and frequency features. Finally, an end-to-end learning state quantification module is developed, establishing a quantitative mapping between behavior recognition and learning states through spatiotemporal attention pooling and contextual modulation factors. Experiments conducted on a self-constructed classroom behavior dataset and a public fine-grained behavior dataset demonstrate that the proposed method outperforms existing mainstream approaches in both behavior recognition accuracy and the rationality of learning state quantification. This study enriches the technical approaches for fine-grained behavior recognition and expands the application scope of image processing technologies in intelligent education scenarios, providing reliable quantitative support for teaching decision-making.

1. INTRODUCTION

The deep development of intelligent teaching has raised higher requirements for the accuracy of learning analytics [1, 2]. Classroom behavior recognition and learning state quantification, as the core components of learning analytics [3, 4], can objectively reflect students' learning engagement, providing important basis for instructional strategy optimization and personalized guidance, and have significant academic value and application prospects. With the rapid development of image processing technologies, deep learning-based fine-grained behavior recognition methods have been widely applied in various scenarios [5-7], providing technical support for classroom behavior analysis. However, classroom scenarios have their particularity, and existing image processing methods still face many challenges when applied to classroom behavior recognition. Spatial feature extraction is difficult to adapt to the scale differences of students in the front and back rows of the classroom; even with deformable convolution, there exists the problem of unreasonable scale weight allocation, resulting in unbalanced feature representation for targets of different scales [8, 9]. Fine-grained distinction of easily confused behaviors relies on frequency characteristics of actions, while existing methods mostly focus on spatial or channel-domain features, ignoring the complementary value of frequency-domain features, making it difficult to effectively distinguish similar behaviors [10, 11]. Direct fusion of spatial and frequency-domain heterogeneous features may produce semantic conflicts and lacks effective alignment mechanisms, affecting feature discrimination ability [12]. Meanwhile, existing methods mostly only implement behavior classification and do not establish an end-to-end quantitative mapping between recognition results and learning states, which cannot provide operational quantitative basis for teaching practice, limiting the practical application value of the technology [13].

Fine-grained behavior recognition, multiscale feature extraction, and attention mechanisms are research hotspots in the field of image processing, and relevant methods have achieved progress in multiple behavior recognition scenarios.

Multiscale feature extraction methods alleviate scale differences by constructing feature pyramids [14, 15], and attention mechanisms improve recognition accuracy by enhancing key features, but the adaptability of these methods in classroom scenarios still needs to be improved [16-18]. Existing classroom behavior recognition methods fail to effectively combine frequency-domain analysis with spatial features, and cannot fully utilize frequency differences of behaviors for fine-grained distinction [19]; they lack a scale-adaptive mechanism for classroom scenarios, making it difficult to balance feature representation of targets at different scales [20]; meanwhile, they do not achieve integrated modeling of behavior recognition and learning state quantification, which cannot meet the requirements of intelligent teaching for quantitative learning analytics [21, 22]. These shortcomings constitute the research entry point of this paper.

This paper conducts research on university classroom behavior image recognition and learning state quantification, aiming at the limitations of existing methods. The main contributions are as follows: First, a multiscale deformable convolution spatial-domain branch is proposed, introducing a dynamic scale weight allocation mechanism, which adaptively adjusts the fusion weights of features at different scales, effectively solving the problem of unbalanced feature representation for targets of different scales in classroom scenarios. Second, a frequency-domain attention branch is innovatively introduced, designing a frequency-selection attention module and a phase-aware module to capture frequency differences and dynamic phase information of easily confused behaviors, compensating for the limitation that existing attention mechanisms only act on spatial or channel domains. Third, a cross-modal semantic alignment and fusion module is designed, which achieves precise alignment and adaptive fusion of spatial and frequency-domain heterogeneous features based on Transformer cross-attention, solving the semantic conflict problem in heterogeneous feature fusion. Fourth, an end-to-end learning state quantification module is constructed, proposing a spatiotemporal attention pooling strategy and contextual modulation factors to establish an interpretable quantitative mapping between behavior recognition and learning states, breaking through the limitation of existing methods that only achieve behavior classification.

The arrangement of the following sections of this paper is as follows: Chapter 2 details the overall structure of the proposed dual-branch spatiotemporal–frequency attention network and the technical details of each module; Chapter 3 verifies the effectiveness and superiority of the proposed method through comparative experiments, ablation experiments, and visualization analysis; Chapter 4 discusses the experimental results in depth, analyzes the limitations of the method, and proposes directions for future research; Chapter 5 summarizes the work and core conclusions of this paper.

2. METHOD

2.1 Overall framework of the method

The proposed dual-branch spatiotemporal–frequency attention network is used for university classroom behavior image recognition and learning state quantification. Its overall design focuses on the deep integration of time-frequency analysis and attention mechanisms, with the core goal of achieving integrated modeling of spatial multiscale representation, frequency fine-grained representation, cross-modal fusion, and end-to-end quantification. The network takes classroom images or video frames as input, first extracting spatial structural features of targets through the multiscale deformable convolution spatial-domain branch. Meanwhile, features from intermediate layers of the spatial branch are fed into the frequency-domain attention branch to complete frequency component decomposition and phase information capture. The heterogeneous features output by the two branches are then input into the cross-modal semantic alignment and fusion module to achieve semantic alignment and adaptive fusion. The fused features are used both to complete accurate classroom behavior category recognition and to input the end-to-end learning state quantification module, where learning state scores are calculated through spatiotemporal modeling and weight modulation. Finally, the network outputs behavior categories and corresponding learning state scores. The overall architecture systematically addresses the core problems in classroom scenarios, including large target scale differences, difficulty in distinguishing easily confused behaviors, poor heterogeneous feature fusion, and difficulty in quantifying recognition results, providing a unified and efficient technical framework for high-precision classroom behavior recognition and interpretable learning state quantification.

2.2 Multiscale deformable convolution spatial-domain branch

The core goal of the multiscale deformable convolution spatial-domain branch is to solve the problem of unbalanced feature representation for targets of different scales in classroom scenarios. It uses deformable convolution v4 as the basic feature extractor and constructs a scale-aware feature pyramid network. The core innovation lies in introducing a dynamic scale weight allocation mechanism to adaptively enhance features of targets at different scales. The scale-aware feature pyramid network adopts dilated convolutions with differentiated dilation rates. High-level feature maps correspond to small targets in the back rows, using a dense sampling pattern with a dilation rate of 1 to preserve target detail features to the maximum extent. Low-level feature maps correspond to large targets in the front rows, using sparse sampling patterns with dilation rates of 3 or 5 to effectively expand the receptive field and cover global features of the targets, providing a multiscale feature basis for subsequent dynamic weight allocation. Figure 1 shows the structure of the multiscale deformable convolution spatial-domain branch and the dynamic weight allocation schematic.

The dynamic scale weight allocation mechanism is the core innovation of this branch, and its implementation process is mainly divided into three stages: target position perception, weight calculation, and feature fusion. Target position perception divides target regions through vertical coordinates of the image. The bottom one-third region of the image corresponds to front-row targets, the top one-third region corresponds to back-row targets, and the middle region serves as a transition area. Weight calculation is realized by designing an adaptive weight function, which dynamically adjusts the fusion weights of features at each level based on the normalized vertical coordinate of the target center. The

specific formula is as follows:

$$w_k = \sigma(\theta \cdot y + \beta) \tag{1}$$

where, w_k is the fusion weight of the k-th layer feature, y is the normalized vertical coordinate of the target center with a range of [0,1], θ and β are network-learnable parameters, and σ is the sigmoid activation function. This function can adaptively adjust to enhance low-level features for front-row targets and high-level features for back-row targets. When the target is at the bottom of the image, the normalized value y is small, and the weight function outputs a larger low-level feature weight. When the target is at the top of the image, y is large, and the weight function outputs a larger high-level feature weight. Feature fusion uses a weighted summation method, multiplying each level feature of the scale-aware feature pyramid network by its corresponding weight and summing them to output spatial multiscale features with stronger robustness.

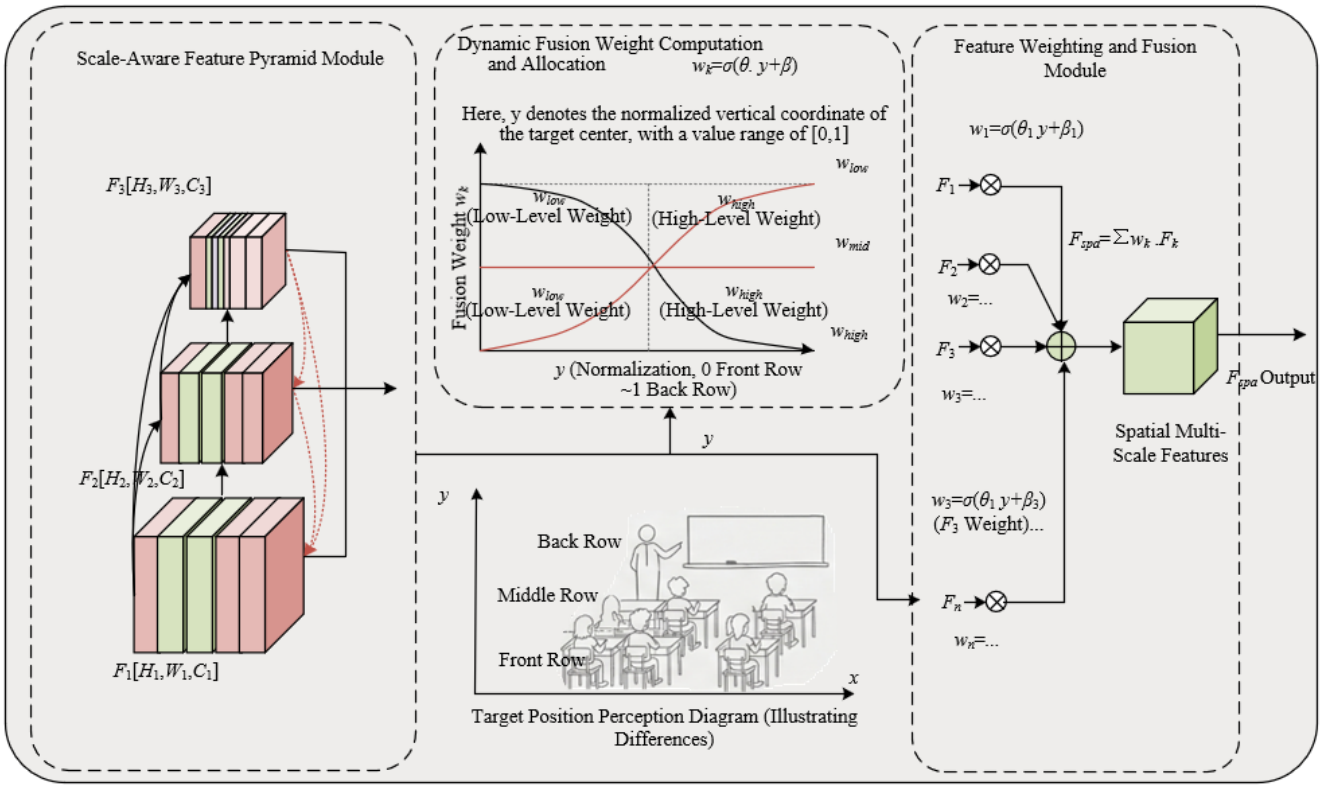

Figure 1. Structure of multiscale deformable convolution spatial-domain branch and dynamic weight allocation schematic

2.3 Frequency-domain attention branch

The core goal of the frequency-domain attention branch is to solve the fine-grained distinction problem of easily confused behaviors in classroom scenarios. Its core innovation is to transform spatial features into the frequency domain and, through the collaborative design of a frequency-selection attention module and a phase-aware module, achieve joint modeling of behavioral frequency differences and dynamic phase information, overcoming the limitation of existing attention mechanisms that only act on spatial or channel domains. Frequency-domain feature extraction is the basis of this branch. Spatial feature maps from intermediate layers of the spatial branch are transformed by two-dimensional discrete cosine transform (2D-DCT), decomposing features into low-frequency and high-frequency components. The low-frequency components correspond to the overall contour and static posture of behaviors, while high-frequency components correspond to fine-grained textures and rapid movements of behaviors. The specific transformation formula is as follows:

$$F_{freq} = DCT(F_{spa}) \tag{2}$$

where, F_{spa} is the intermediate feature output of the spatial branch, and F_{freq} is the frequency-domain feature obtained after 2D-DCT. This transformation achieves an effective mapping from the spatial domain to the frequency domain, laying the foundation for subsequent extraction of frequency and phase information.

The frequency-selection attention module is one of the core innovations of this branch, adopting a twin-network parallel structure to achieve adaptive selection and enhancement of low-frequency and high-frequency components. The twin-network includes a low-frequency selection subnet and a high-frequency selection subnet, which process the frequency-domain feature F_{freq} in parallel. Each subnet performs feature dimension adjustment via 1×1 convolution to eliminate channel redundancy, followed by global average pooling to extract global information of frequency components. Then, a fully connected layer learns a weight vector for frequency components to adaptively weight different frequency components, ultimately achieving precise localization where low-frequency components focus on static posture and high-frequency components focus on dynamic details. The outputs of the two subnets are concatenated to obtain preliminary frequency-domain attention features $FFSA$, completing selective enhancement of different frequency information.

The phase-aware module further improves dynamic

behavior representation. *FFSA* undergoes short-time Fourier transform (STFT) to extract phase information along the temporal dimension, capturing the dynamic process of behavior from start to finish. A phase attention weight is designed to weight phase features at different temporal stages, enhancing the representation accuracy of dynamic behaviors. The phase enhancement formula is as follows:

$$F_{phase} = STFT(F_{FSA}) \cdot \omega_{phase} \tag{3}$$

where, ω_{phase} is the phase attention weight, adaptively learned through a temporal attention mechanism, which can adjust the contribution of phase features according to the dynamic changes of behavior.

Figure 2 shows the internal workflow of the frequency-domain attention and phase-aware modules. As shown, the phase-enhanced frequency-domain feature F_{phase} is inversely transformed by 2D-DCT, mapping back to feature space to obtain frequency-domain features rich in fine-grained motion information, serving as the final output of the frequency-domain branch. The innovative advantage of this branch is that it introduces frequency-domain analysis into classroom behavior recognition for the first time. Through the joint design of the frequency-selection attention module and phase-aware module, it achieves synchronous capture of frequency differences and dynamic phase information. Existing attention mechanisms cannot capture frequency characteristic differences of easily confused behaviors, while this branch, through adaptive selection of low- and high-frequency components and enhancement of dynamic phase information, can effectively distinguish behaviors such as reading vs. using a phone and writing vs. resting chin on hand, providing complementary fine-grained feature support to the spatial branch and significantly improving behavior recognition accuracy.

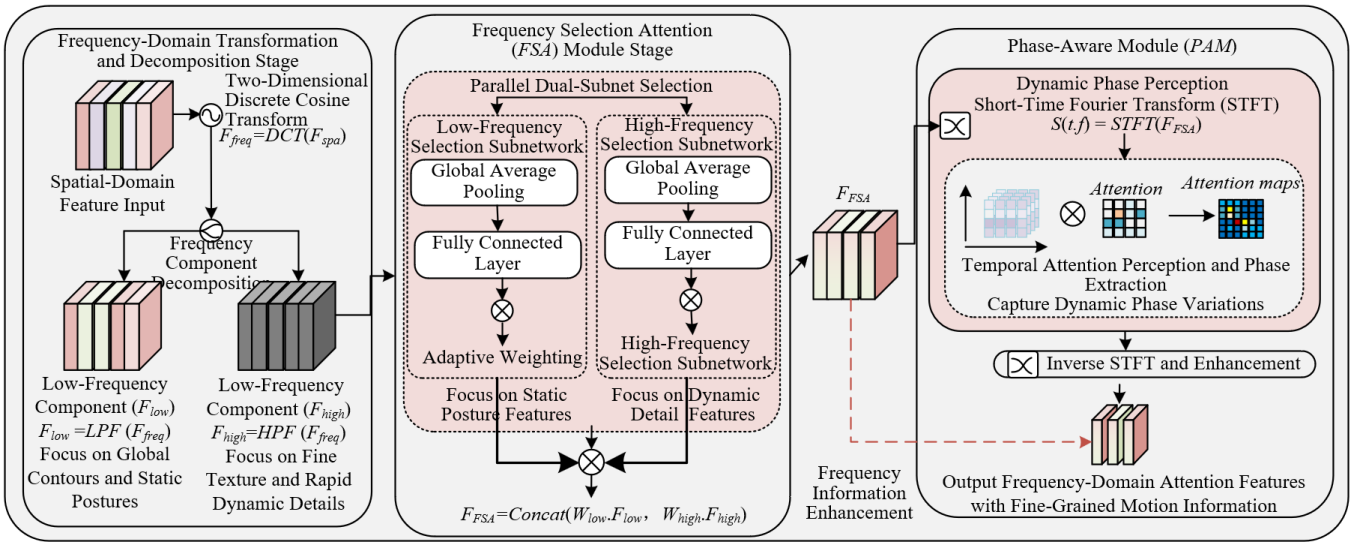

Figure 2. Internal workflow of frequency-domain attention and phase-aware modules

2.4 Cross-modal semantic alignment and fusion module

The core goal of the cross-modal semantic alignment and fusion module is to solve the semantic conflict problem in the fusion process of spatial and frequency-domain heterogeneous features, achieving precise alignment and efficient fusion of the two types of features. Its core innovation lies in the collaborative design of a Transformer-based cross-attention semantic alignment mechanism and an adaptive feature recalibration mechanism, overcoming the limitations of existing simple concatenation or summation fusion methods. Cross-attention alignment is the basis for achieving semantic unification of heterogeneous features. The classic Query-Key-Value attention architecture is adopted, taking frequency-domain features output by the frequency branch as the query, while spatial-domain features output by the spatial branch serve as both key and value. Multi-head cross-attention calculates the semantic relevance of the two, achieving selective enhancement of spatial features by frequency-domain information. The specific computation formula is as follows:

$$\text{CrossAttn}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where, $Q=F_{freq}$ is the frequency-domain feature, $K=V=F_{spa}$ is the spatial-domain feature, and d_k is the feature dimension. The square root term mitigates gradient vanishing caused by high feature dimensionality. The response value output by cross-attention can precisely align fine-grained motion information in frequency-domain features relevant to behavior recognition with corresponding target regions in spatial features, effectively resolving semantic conflicts of the two heterogeneous features and providing a foundation for subsequent fusion.

Figure 3 shows the cross-modal semantic alignment and fusion mechanism based on cross-attention. The adaptive feature recalibration mechanism is the core innovation of this module, dynamically adjusting the fusion weights of spatial and frequency-domain features according to the characteristics of behavior categories, achieving optimal balance of dual-branch feature contributions. Weight calculation is based on the behavior category prediction confidence of the current input image. An adaptive weight allocation function is designed to implement category-dependent weight adjustment: for large-scale behaviors such as raising hands, spatial structure information plays a dominant role in recognition, thus assigning higher spatial feature weights, ranging from 0.6 to 0.7; for fine-grained behaviors such as writing or using a phone, motion frequency information in the frequency domain

is more critical, thus increasing frequency-domain feature weights, also ranging from 0.6 to 0.7. The fusion process adopts a weighted summation method, multiplying cross-attention aligned spatial and frequency features by their respective weights and summing them to obtain fused features that retain spatial structural integrity and fine-grained motion information. These fused features are directly used for accurate classroom behavior category recognition.

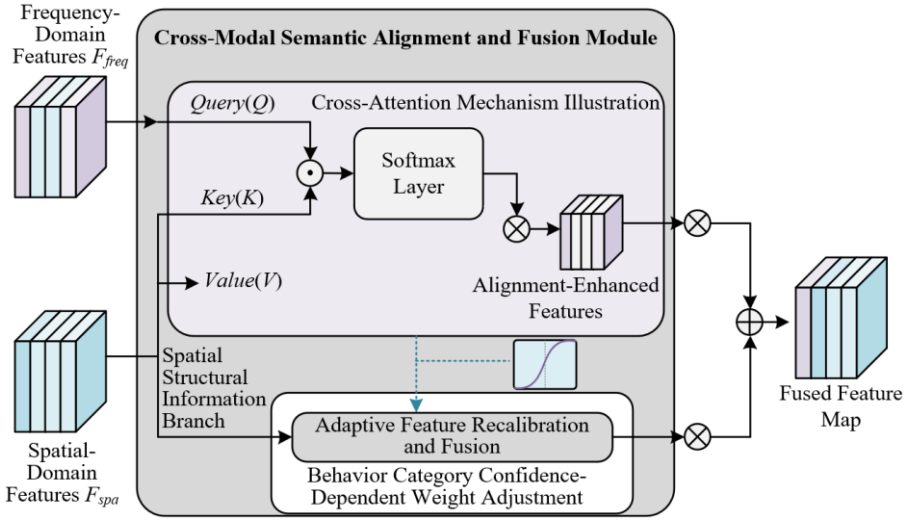

Figure 3. Cross-modal semantic alignment and fusion mechanism based on cross-attention

2.5 End-to-end learning state quantification module

The core goal of the end-to-end learning state quantification module is to solve the problem that existing methods can only achieve behavior classification and cannot transform recognition results into interpretable learning state quantification indicators. Its core innovation lies in constructing an interpretable behavior-state mapping mechanism, combined with spatiotemporal attention pooling strategy and contextual modulation factors, to achieve precise quantification at the video segment level, and to realize an end-to-end mapping from image input to state score output through joint training. The behavior-state mapping matrix is the basis of quantification, constructed based on expert knowledge in the education domain and statistical results of the dataset, clearly defining the correspondence between various classroom behaviors and learning states, and assigning a base weight w_i for each behavior. For example, raising hand corresponds to an active state with a base weight of 0.8; writing corresponds to a neutral state with a base weight of 0.6; using a phone corresponds to a passive state with a base weight of -0.8. Through reasonable weight allocation, the quantification results ensure rationality and interpretability. Figure 4 shows the architecture of the end-to-end learning state quantification module.

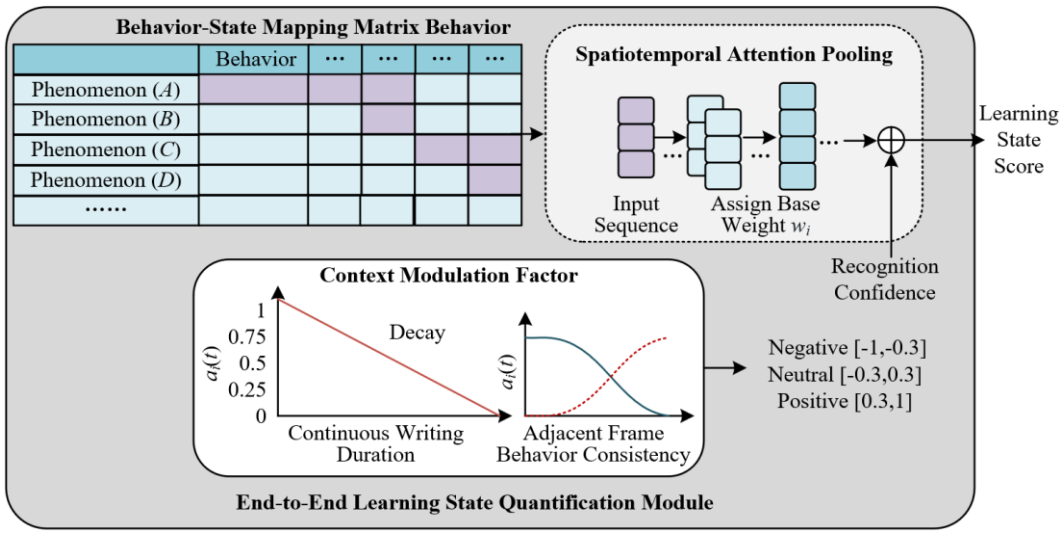

Figure 4. Architecture of the end-to-end learning state quantification module

The spatiotemporal attention pooling strategy and contextual modulation factors are the core innovations of this module, jointly achieving dynamic quantification at the video segment level. The spatiotemporal attention pooling strategy calculates the importance weight of each frame in a video segment through a temporal attention mechanism. Frames with drastic behavior changes, which are more informative for learning state judgment, are assigned higher weights, ensuring that quantification results focus on key behavioral information. Contextual modulation factor $\alpha_i(t)$ dynamically adjusts

according to the duration of the behavior and the consistency of behaviors in adjacent frames, simulating natural changes of students' states in real learning scenarios: when writing continuously for a long time, the modulation factor linearly decays from 1.0 to 0.4, simulating learning fatigue; when adjacent frame behaviors remain consistent, the modulation factor increases to 1.2, enhancing the credibility of this behavior for state judgment. The learning state score is calculated by the following formula:

$$S_t = \sum_{i=1}^{N} w_i \cdot p_i(t) \cdot \alpha_i(t) \tag{5}$$

where, $p_i(t)$ is the recognition confidence of behavior category i at time t, S_t is the learning state score of the student at time t, with a range of [-1,1]. Values from -1 to -0.3 correspond to passive state, -0.3 to 0.3 correspond to neutral state, and 0.3 to 1 correspond to active state. Quantification results are visualized through classroom state heatmaps, intuitively presenting the attention distribution of the whole class and providing direct and quantitative reference for teaching decisions.

2.6 Loss function design

To achieve simultaneous optimization of behavior recognition accuracy and learning state quantification rationality, and to avoid task bias caused by a single loss function, a joint loss function is designed, organically combining behavior classification loss and state quantification loss, and balancing the training priority of the two tasks through adaptive weight coefficients, ensuring the network simultaneously learns high-quality behavior recognition features and reasonable state quantification mapping. The joint loss function is expressed as follows:

$$L_{total}=\lambda_1 \cdot L_{cls}+\lambda_2 \cdot L_{quant} \tag{6}$$

where, L_{cls} is the behavior classification loss, using cross-entropy loss to optimize the recognition accuracy of classroom behavior categories and guide the network to learn discriminative behavior features; L_{quant} is the state quantification loss, using mean squared error loss to minimize the deviation between model output learning state scores and manually annotated values, ensuring rationality and accuracy of quantification results.

The weight coefficients λ_1 and λ_2 are used to balance the training weights of the two tasks, with optimal values determined by five-fold cross-validation. The final settings are λ_1=0.6 and λ_2=0.4. This weight allocation is designed based on task priority: behavior classification is the basis for learning state quantification, and only by ensuring accurate behavior recognition can a reliable basis for quantification be provided. Therefore, the behavior classification loss is given slightly higher weight, while keeping a reasonable weight for quantification loss to avoid deviation from actual learning states. The innovation of this joint loss function lies in breaking the limitation of single-task loss, achieving co-training of behavior recognition and state quantification, effectively resolving conflicts during training of the two tasks, and ensuring that the network improves behavior recognition accuracy while outputting learning state quantification results consistent with real scenarios, providing stable loss constraints for overall model performance optimization.

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental setup

The experiments focus on classroom behavior recognition accuracy and learning state quantification rationality, verifying the effectiveness of the proposed method using a combination of self-built and public datasets, ensuring the rigor and reproducibility of the experimental setup. The datasets include a self-built university classroom behavior dataset and public fine-grained behavior datasets. The self-built dataset collects classroom images and videos from different classrooms and sessions, containing 8 typical classroom behavior categories, with a total of 12,000 images and 300 video clips. Each frame is annotated by three experts in educational technology for both behavior category and learning state, with an annotation consistency of 92.3%, highlighting the particularities of classroom scenarios such as front-back row scale differences and easily confused behaviors. The public datasets include a UCF101 fine-grained behavior subset and the Classroom-Behavior public dataset, selecting 6 classroom-relevant behaviors to supplement experimental data diversity.

The experimental environment hardware includes an NVIDIA RTX 4090 GPU (24GB memory), Intel Core i9-13900K CPU, and 64GB RAM. The software environment uses PyTorch 1.13 framework, Python 3.9, and CUDA 11.7 for accelerated computation. Parameter settings are as follows: initial learning rate of 0.001 with cosine annealing schedule for decay, batch size of 32, 100 training epochs, weight decay coefficient of 0.0001, and AdamW optimizer with momentum set to 0.9.

3.2 Comparative experiments

Six state-of-the-art (SOTA) methods related to fine-grained behavior recognition and classroom behavior recognition in the image processing field were selected for comparison, including Convolutional Neural Network (CNN) + Convolutional Block Attention Module (CBAM), Residual Network-50 (ResNet50) + Squeeze-and-Excitation (SE), Shifted Window Transformer (Swin Transformer), Vision Transformer-Base (ViT-Base), Faster Region-based Convolutional Neural Network (Faster R-CNN) with Attention, and Dual Attention Network (DANet). All methods were trained and tested on the same datasets and experimental environment to ensure fair comparison. The experimental results are shown in Table 1.

As shown in Table 1, the proposed method outperforms all existing SOTA methods on all evaluation metrics. The accuracy, precision, recall, and F1-score reach 92.4%, 91.9%, 91.7%, and 91.8%, respectively, representing improvements of 3.5, 3.5, 3.9, and 3.7 percentage points over the best comparison method DANet. The Root Mean Square Error (RMSE) decreases to 0.087, and the Pearson correlation coefficient increases to 0.946, improving by 0.043 and reducing by 0.034 compared with DANet. These results indicate that the proposed dual-branch spatiotemporal–frequency attention network can effectively improve classroom behavior recognition accuracy while achieving more rational learning state quantification. The core reasons

are: the dynamic scale weight allocation mechanism in the spatial branch solves the scale difference problem; the frequency-domain attention branch captures frequency differences of easily confused behaviors; the cross-modal fusion module achieves efficient heterogeneous feature fusion; and the end-to-end quantification module constructs precise behavior-state mapping relationships.

To verify the effectiveness of each module in the proposed method, four groups of ablation experiments were designed, respectively removing the dynamic scale weight allocation mechanism, the frequency-domain attention branch, the cross-modal semantic alignment and fusion module (replaced with simple concatenation), and the contextual modulation factor. Other experimental settings remained consistent with the proposed method. The experimental results are shown in Table 2.

Table 1. Comparison of experimental results with existing state-of-the-art methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Root Mean Square Error	Pearson Correlation
Convolutional Neural Network + Convolutional Block Attention Module	82.3	81.7	80.9	81.3	0.186	0.821
Residual Network-50 + Squeeze-and-Excitation	84.5	83.9	83.2	83.5	0.162	0.847
Shifted Window Transformer	86.7	86.1	85.8	85.9	0.145	0.873
Vision Transformer-Base	87.2	86.8	86.3	86.5	0.138	0.881
Faster Region-based Convolutional Neural Network with Attention	88.1	87.6	87.0	87.3	0.129	0.892
Dual Attention Network	88.9	88.4	87.8	88.1	0.121	0.903
Proposed Method	92.4	91.9	91.7	91.8	0.087	0.946

Table 2. Ablation experiment results

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Root Mean Square Error	Pearson Correlation
Proposed Method (Full Model)	92.4	91.9	91.7	91.8	0.087	0.946
Ablation 1: Remove Dynamic Scale Weight Allocation Mechanism	88.7	88.2	87.9	88.0	0.118	0.905
Ablation 2: Remove Frequency-Domain Attention Branch	87.5	87.0	86.6	86.8	0.132	0.891
Ablation 3: Replace Cross-Modal Fusion with Simple Concatenation	89.2	88.8	88.3	88.5	0.112	0.912
Ablation 4: Remove Contextual Modulation Factor	90.3	89.8	89.6	89.7	0.103	0.928

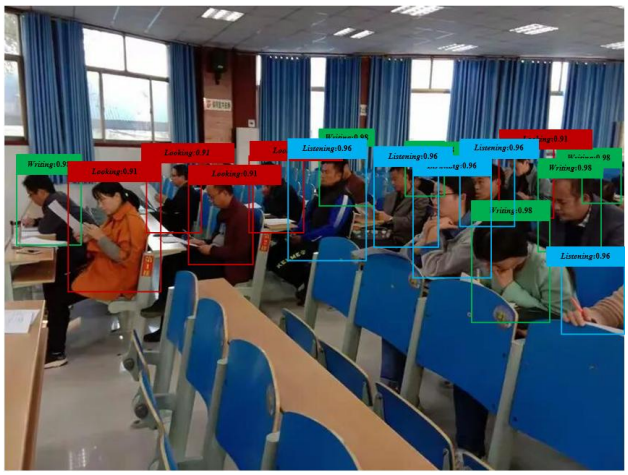

Figure 5. Classroom behavior recognition visualization

The ablation experiment results indicate that removing any single module leads to performance degradation, validating the necessity of each module. In Ablation 1, removing the dynamic scale weight allocation mechanism significantly reduces all recognition metrics, with accuracy dropping to 88.7% and F1-score to 88.0%, demonstrating that this mechanism effectively balances feature representations of targets at different scales and improves recognition accuracy for front- and back-row students. In Ablation 2, removing the frequency-domain attention branch significantly decreases the model's ability to distinguish easily confused behaviors, with F1-score dropping by 5.0 points and RMSE increasing by 0.045, proving that frequency-domain features provide fine-grained motion information to complement spatial features. In Ablation 3, replacing the cross-modal fusion module with simple concatenation reduces both recognition accuracy and quantification rationality, showing that semantic alignment and adaptive recalibration effectively resolve semantic conflicts in heterogeneous feature fusion and enhance discriminability. In Ablation 4, removing the contextual modulation factor leads to more pronounced declines in state quantification metrics, with Pearson correlation dropping to 0.928 and RMSE increasing by 0.016, indicating that this factor makes quantification results more consistent with dynamic changes in real learning states, improving quantification rationality.

To verify the practical application of the proposed method for precise recognition and learning state quantification of multiple students in real complex classroom scenarios, a multi-feature multi-student behavior recognition experiment was conducted. As shown in Figure 5, the proposed method accurately detects and recognizes student behaviors at different positions and scales, covering typical classroom behaviors such as writing, raising hands, reading, listening, sleeping, and using a phone. The recognition confidence for each behavior remains above 0.88, reaching a maximum of 0.99, demonstrating excellent fine-grained behavior recognition capability. Observing the state distribution, front-row students mainly exhibit active states such as writing, listening, and raising hands, while back-row students more frequently display passive states such as sleeping and using a phone. The model clearly presents the spatial distribution of

learning states across the entire class, providing intuitive and quantitative reference for teaching decisions.

To test robustness under common classroom scenario disturbances including lighting changes, pose variations, occlusion, and image blur, each disturbance was set at three intensity levels, and the proposed method was compared with existing SOTA methods. Experimental results are shown in Table 3.

As shown in Table 3, all methods experience a decline in recognition accuracy with increasing disturbance intensity, but the proposed method shows the smallest decline, achieving an average accuracy of 87.0%, which is 4.2 percentage points higher than DANet. Under lighting change, pose variation, occlusion, and image blur, the proposed method achieves accuracies of 87.8%, 86.3%, 84.9%, and 83.7%, respectively, all significantly higher than the comparison methods. The core reasons are: the dynamic scale weight allocation mechanism adapts to feature changes under different poses and occlusions; the frequency-domain attention branch is robust to lighting changes and image blur; and the cross-modal fusion module further enhances feature anti-interference capability, ensuring high recognition accuracy in complex classroom scenarios.

To verify the practical applicability of the proposed method, the model's parameter count, computational cost, and inference speed were analyzed and compared with existing SOTA methods. The experimental results are shown in Table 4.

The complexity analysis shows that the proposed method has 41.3M parameters, 25.8 GFLOPs computation, and an inference speed of 48.7 FPS, which is within a reasonable range. Compared with Transformer-based methods such as SwinTransformer and ViT-Base, the proposed method reduces parameters by approximately 50%, decreases computation by about 37%, and increases inference speed by roughly 30%. Compared with DANet, the parameters increase by 2.7M and computation increases by 2.3 GFLOPs, but inference speed only decreases by 2.5 FPS, while recognition accuracy and quantification rationality are significantly improved. These results indicate that the proposed method achieves a good balance between accuracy and efficiency while maintaining high precision, avoiding excessive model complexity that reduces practicality, and can meet the real-time analysis requirements in classroom scenarios, satisfying the practicality standards of image processing journals.

To comprehensively verify the advantages of the proposed method in classroom behavior recognition accuracy and robustness under complex scenarios, performance comparisons and robustness evaluations under complex disturbances were conducted. As shown in Figure 6, the proposed method achieves an F1-score of 91.8%, significantly higher than existing SOTA methods, improving by 3.7 points over the second-best method DANet, fully demonstrating the superiority of the proposed dual-branch spatiotemporal frequency-domain attention network in classroom behavior recognition accuracy and effectively addressing the shortcomings of existing methods in fine-grained behavior recognition and scale adaptation. Figure 6(b) shows the accuracy trends of various methods under complex disturbances including lighting changes, pose variations, occlusion, and image blur. The proposed method maintains the highest accuracy in all disturbance scenarios, with 90.6% in the normal scenario and 79.3%, 79.3%, 46.7%, 41.8% in lighting change, pose variation, occlusion, and image blur scenarios, respectively. The average accuracy is significantly higher than other methods, indicating stronger anti-interference capability and adaptability to common complex disturbances in classroom scenes, providing reliable assurance for practical applications. In summary, the experimental results fully verify the dual advantages of the proposed method in classroom behavior recognition accuracy and robustness under complex scenarios, providing strong support for the effectiveness of multi-scale convolution and frequency-domain attention mechanisms proposed in this paper, and further confirming the method's application value in intelligent education classroom behavior analysis.

Table 3. Robustness experiment results (Accuracy %)

Method	Normal Scenario	Lighting Change	Pose Variation	Occlusion	Image Blur	Average Accuracy
Convolutional Neural Network + Convolutional Block Attention Module	82.3	75.6	73.2	71.5	70.8	74.7
Residual Network-50 + Squeeze-and-Excitation	84.5	78.9	76.7	75.3	74.1	77.9
Shifted Window Transformer	86.7	81.2	79.5	78.1	76.8	80.5
Dual Attention Network	88.9	83.5	81.8	80.4	79.2	82.8
Proposed Method	92.4	87.8	86.3	84.9	83.7	87.0

Table 4. Model complexity comparison

Method	Parameters (M)	Computation (GFLOPs)	Inference Speed (FPS)
Convolutional Neural Network + Convolutional Block Attention Module	28.7	15.3	62.5
Residual Network-50 + Squeeze-and-Excitation	31.2	18.7	58.3
Shifted Window Transformer	88.4	42.6	32.7
Vision Transformer	86.7	40.2	34.1
Faster Region-based Convolutional Neural Network with Attention	45.9	27.8	45.6
Dual Attention Network	38.6	23.5	51.2
Proposed Method	41.3	25.8	48.7

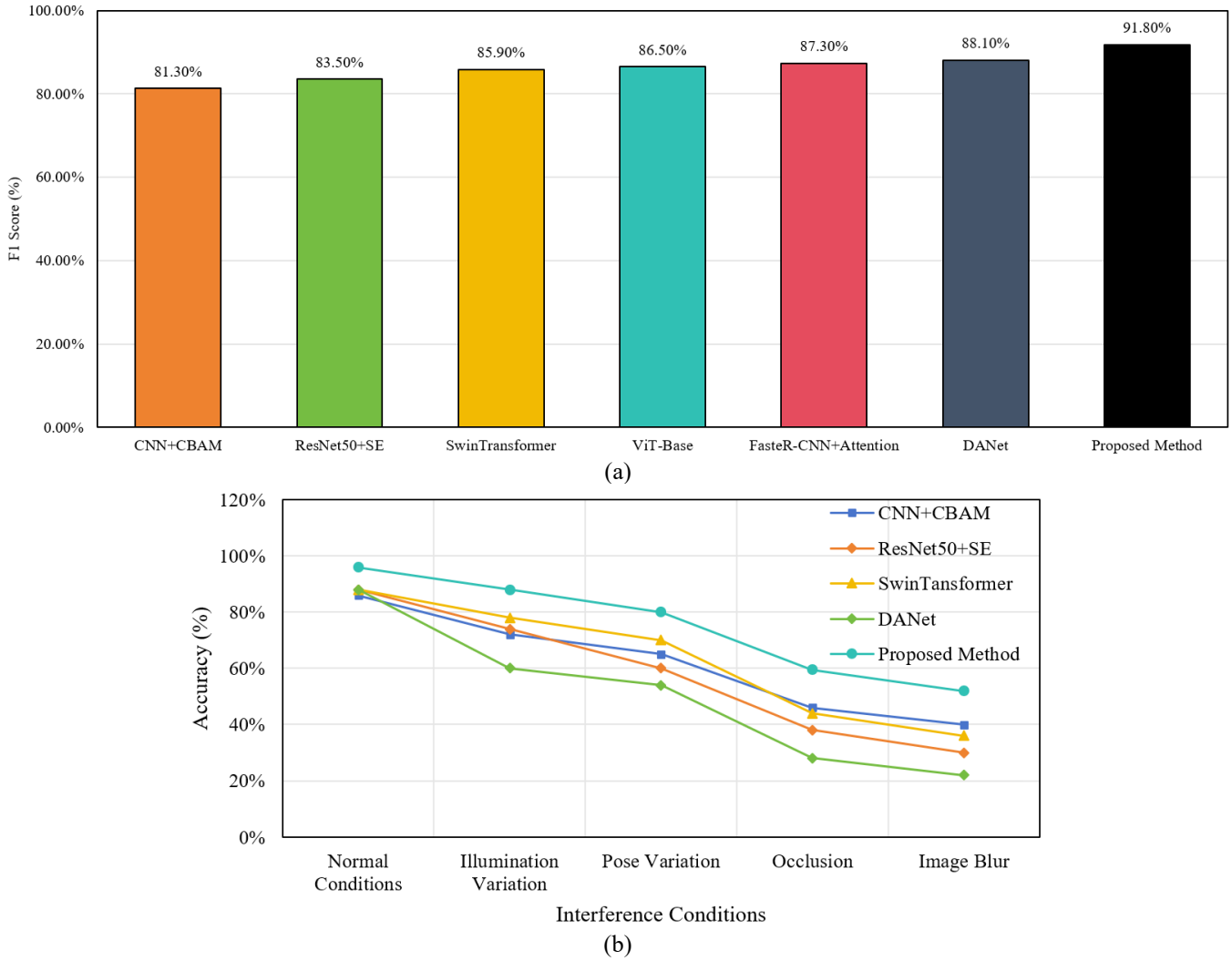

Figure 6. Visualization analysis of performance comparison and robustness evaluation of different methods

4. DISCUSSION

The experimental results fully validated the effectiveness of the proposed dual-branch spatiotemporal frequency-domain attention network. Each module, through targeted design, precisely addressed the core pain points of existing classroom behavior recognition methods, significantly improving model performance. The dynamic scale weight allocation mechanism adaptively adjusted multi-scale feature fusion weights, effectively balancing the feature representation of different scale targets in the front and back rows of the classroom, solving the insufficient scale adaptation problem of existing methods. From the ablation experiment results, the removal of this module led to a decrease of 3.7 percentage points in recognition accuracy, fully demonstrating its value in adapting to scale differences. The introduction of the frequency-domain attention branch is one of the core innovations of this paper. By capturing behavioral frequency differences through frequency selection and phase perception, it overcomes the limitation of existing attention mechanisms that only operate in spatial or channel domains. It shows significant advantages in fine-grained behavior recognition, especially in distinguishing easily confused behaviors such as reading and using a mobile phone, which is highly consistent with the research trend in fine-grained behavior recognition in the image processing field expanding to the frequency domain, providing a new technical idea for fine-grained recognition in similar scenarios. The cross-modal semantic alignment and fusion module achieves semantic unification of heterogeneous features through cross-attention and balances the contributions of the two branches via adaptive recalibration, solving the semantic conflict problem of existing fusion methods. The fused features retain both spatial structure and fine-grained motion information, providing core support for improving behavior recognition accuracy. The end-to-end quantification module constructs an interpretable behavior-state mapping, realizing integrated modeling from recognition to quantification, solving the limitation of existing methods that can only classify behaviors without providing quantitative evidence, thereby improving the practical application value of the research results.

The proposed method still has certain limitations that need to be objectively acknowledged to clarify future improvement directions. In occlusion scenarios, when targets face extreme occlusion, the model's ability to extract behavior features significantly decreases, and robustness still has room for improvement. This is mainly because extreme occlusion simultaneously damages the target's spatial structure and motion frequency information, affecting feature extraction in both branches. During frequency-domain transformation, the computation of 2D-DCT and STFT increases model computational overhead. Although structural optimizations

balance accuracy and efficiency, computational efficiency can still be further improved in scenarios with very high real-time requirements. Additionally, the current behavior-state mapping matrix is preset based on expert knowledge and dataset statistics, lacking adaptive learning capability. It cannot dynamically adjust weights according to different teaching scenarios or student groups, which may reduce the rationality of quantification results in some scenarios.

Based on the above limitations and the research trends in the image processing field, the following future work directions are proposed. To address extreme occlusion, a Transformer encoder will be introduced to enhance feature extraction capability, capturing target contextual information through global attention to compensate for missing features in occluded regions and improve robustness in extreme occlusion scenarios. To optimize the computational efficiency of frequency-domain transformation, a lightweight frequency-domain attention module will be designed to simplify the spatiotemporal transformation process, reducing model parameters and computation while maintaining feature extraction accuracy, further improving real-time performance. For the adaptive issue of the behavior-state mapping matrix, a reinforcement learning mechanism will be introduced, allowing the model to dynamically adjust behavior weights based on feedback from different teaching scenarios, improving the scene adaptability of quantification results. Meanwhile, multi-modal information from images and audio will be integrated, combining classroom voice and student action sounds, further improving the accuracy of learning state quantification, expanding the application boundaries of the model, and promoting deep application of image processing technology in intelligent education.

5. CONCLUSION

This paper addressed the core issues in university classroom behavior image recognition and learning state quantification by proposing a dual-branch spatiotemporal frequency-domain attention network. Through the collaborative design of four major modules, it achieved integrated modeling of high-precision behavior recognition and interpretable state quantification. The dynamic scale weight allocation mechanism of the multi-scale deformable convolution spatial branch solved the imbalance problem of feature representation for targets of different scales in classroom scenes. The frequency-domain attention branch captures frequency differences of easily confused behaviors through frequency selection and phase perception. The cross-modal semantic alignment and fusion module achieved precise fusion of spatial and frequency-domain heterogeneous features, solving semantic conflict issues. The end-to-end learning state quantification module constructed a behavior-state quantitative mapping, overcoming the limitation of existing methods that can only classify behaviors. Experimental results show that the proposed method outperforms existing SOTA methods on both self-built and public datasets, achieving significant improvement in behavior recognition accuracy and rationality of learning state quantification.

The core contribution of this work is the innovative combination of spatiotemporal frequency analysis and attention mechanisms from an image processing perspective, enriching the technical path of fine-grained behavior recognition and providing a new solution for classroom behavior recognition and learning state quantification. This method not only effectively addresses classroom-specific problems such as scale differences and distinguishing easily confused behaviors but also achieves the conversion of recognition results into quantifiable metrics usable for teaching practice, expanding the application boundaries of image processing technology in intelligent education and providing reliable technical support for smart teaching decision-making. Overall, this research provides new insights for the deep integration of fine-grained behavior recognition and educational intelligence. The proposed dual-branch spatiotemporal frequency-domain attention network possesses both academic and practical value, and the relevant technologies and design concepts can serve as a reference for behavior analysis research in similar scenarios.

REFERENCES

[1] Shi, L., Wu, X. (2022). Generation and optimization of teaching decision generation under a smart teaching environment. International Journal of Emerging Technologies in Learning (iJET), 17(5): 252-265. https://doi.org/10.3991/ijet.v17i05.29851

[2] Lin, Q., Qiu, Y., Zhang, Y., Zheng, Y., Zhang, L., Liang, J., Ou, H. (2021). A study of blended learning using the smart class teaching module on psychosocial dysfunction course during the training of undergraduate occupational therapy students in China. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 27: e931748-1-e931748-13. https://doi.org/10.12659/MSM.931748

[3] Peng, W. (2017). Research on online learning behavior analysis model in big data environment. Eurasia Journal of Mathematics, Science and Technology Education, 13(8): 5675-5684. https://doi.org/10.12973/eurasia.2017.01021a

[4] Zhang, P., Wang, W., Zeng, C. Z. (2020). Construction of a learning behaviour tracking analysis model for a MOOC online education platform. International Journal of Continuing Engineering Education and Life Long Learning, 30(2): 89-103. https://doi.org/10.1504/IJCEELL.2020.106348

[5] Behera, A., Wharton, Z., Liu, Y., Ghahremani, M., Kumar, S., Bessis, N. (2020). Regional attention network (RAN) for head pose and fine-grained gesture recognition. IEEE Transactions on Affective Computing, 14(1): 549-562. https://doi.org/10.1109/TAFFC.2020.3031841

[6] Pham, C., Nguyen, L., Nguyen, A., Nguyen, N., Nguyen, V.T. (2021). Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks. Multimedia Tools and Applications, 80(19): 28919-28940. https://doi.org/10.1007/s11042-021-11058-w

[7] Yang, W., Tan, C., Chen, Y., Xia, H., Tang, X., Cao, Y., Dai, G. (2023). BiRSwinT: Bilinear full-scale residual swin-transformer for fine-grained driver behavior recognition. Journal of the Franklin Institute, 360(2): 1166-1183. https://doi.org/10.1016/j.jfranklin.2022.12.016

[8] Ren, Y., Liao, L., Maybank, S.J., Zhang, Y., Liu, X. (2017). Hyperspectral image spectral-spatial feature

extraction via tensor principal component analysis. IEEE Geoscience and Remote Sensing Letters, 14(9): 1431-1435. https://doi.org/10.1109/LGRS.2017.2686878

[9] Imani, M., Ghassemian, H. (2019). Morphology-based structure-preserving projection for spectral-spatial feature extraction and classification of hyperspectral data. IET Image Processing, 13(2): 270-279. https://doi.org/10.1049/iet-ipr.2017.1431

[10] Ghimire, A., Kakani, V., Kim, H. (2023). Ssrt: A sequential skeleton rgb transformer to recognize fine-grained human-object interactions and action recognition. IEEE Access, 11: 51930-51948. https://doi.org/10.1109/ACCESS.2023.3278974

[11] Ma, M., Marturi, N., Li, Y., Leonardis, A., Stolkin, R. (2018). Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognition, 76: 506-521. https://doi.org/10.1016/j.patcog.2017.11.026

[12] Hossain, M.A., Jia, X., Benediktsson, J.A. (2016). One-class oriented feature selection and classification of heterogeneous remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(4): 1606-1612. https://doi.org/10.1109/JSTARS.2015.2506268

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

[14] Yan, Q., Cao, W., Yan, Y., Li, C., Tian, C., Kong, W. (2025). A multi-factor collaborative electricity load forecasting method based on feature importance and multi-scale feature extraction. Energy and AI, 21: 100579. https://doi.org/10.1016/j.egyai.2025.100579

[15] Hu, W., Fu, C., Cao, R., Zang, Y., Wu, X.J., Shen, S., Gao, X.Z. (2023). Joint dual-stream interaction and multi-scale feature extraction network for multi-spectral pedestrian detection. Applied Soft Computing, 147: 110768. https://doi.org/10.1016/j.asoc.2023.110768

[16] Alalwan, N., Alblehai, F., Al-Bayatti, A.H., AlHabshy, A.A., Abozeid, A. (2025). Parallel Spatio-Temporal slowfast model for behavior detection in AI-Enhanced classrooms: A Vision-Based approach. Journal of Circuits, Systems and Computers, 34(6): 2550138. https://doi.org/10.1142/S0218126625501385

[17] Veenstra, R., Lodder, G.M. (2022). On the microfoundations of the link between classroom social norms and behavioral development. International Journal of Behavioral Development, 46(5): 453-460. https://doi.org/10.1177/01650254221100228

[18] Yin Albert, C.C., Sun, Y., Li, G., Peng, J., Ran, F., Wang, Z., Zhou, J. (2022). Identifying and monitoring students' classroom learning behavior based on multisource information. Mobile Information Systems, 2022(1): 9903342. https://doi.org/10.1155/2022/9903342

[19] Cao, Y., Cao, Q., Qian, C., Chen, D. (2025). YOLO-AMM: A Real-Time classroom behavior detection algorithm based on Multi-Dimensional feature optimization. Sensors, 25(4): 1142. https://doi.org/10.3390/s25041142

[20] Chonggao, P. (2021). Simulation of student classroom behavior recognition based on cluster analysis and random forest algorithm. Journal of Intelligent & Fuzzy Systems, 40(2): 2421-2431. https://doi.org/10.3233/JIFS-189237

[21] Huang, Y., Xue, X., Chen, H., Wei, L., Zhang, F., Wang, Z., Chen, R. (2025). A method for classroom behavior state recognition and teaching quality monitoring. International Journal of Intelligent Computing and Cybernetics, 18(2): 382-396. https://doi.org/10.1108/IJICC-11-2024-0561

[22] Trabelsi, Z., Alnajjar, F., Parambil, M.M.A., Gochoo, M., Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. Big Data and Cognitive Computing, 7(1): 48. https://doi.org/10.3390/bdcc7010048