World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3061

SCENARIO CLASSIFICATION DETECTION MODEL FOR SPATIO-TEMPORAL CONTEXTUAL INFORMATION PERCEPTION IN CLASSROOM SETTINGS

Jin Lu[1], Ji Li[2*]
[1]*Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
[2]*Research Management Office, Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
[*]*Corresponding Author: Ji Li*

Abstract: This paper proposes a spatio-temporal context-aware scene classification detection model tailored for classroom settings, aiming to address detection accuracy limitations arising from complex classroom environments characterised by fluctuating lighting, frequent occlusions, and the difficulty in capturing small-scale behaviours. By integrating cross-scale attention mechanisms in the spatial domain with long-term dependency modelling in the temporal domain, the model effectively captures subtle behavioural features and spatio-temporal contextual relationships between actions. Experimental results on the SCB-Dataset3 and Classroom-Actions public classroom datasets demonstrate that the proposed model achieves 85.4% scene classification accuracy and 83.2% action detection rate, representing significant improvements over mainstream methods such as YOLOv8m, CSSA-YOLO, and TACNet. Ablation studies further validate the effectiveness of each component: the spatial attention module yields a 2.1% mAP improvement, the temporal context module contributes a 4.5% mAP gain, while the scene context module delivers an additional 2.2% performance enhancement. Maintaining real-time processing speed (68.2 FPS), this model effectively addresses multi-scale detection and temporal dependency modelling challenges in classroom scenarios, providing robust technical support for smart education.
Keywords: Classroom behaviour recognition; Spatio-temporal context; Attention mechanisms; Scene classification; Deep learning

1 INTRODUCTION

With the rapid advancement of smart education, the intelligent analysis and assessment of classroom teaching processes have become a focal point in educational technology research. Guided by student-centred teaching principles, accurately identifying behavioural patterns among pupils during lessons holds significant importance for evaluating teaching effectiveness and formulating personalised learning strategies [1]. The proliferation of modern teaching methods such as project-based learning further emphasises the precise capture and analysis of active learning behaviours like interaction and collaboration within the classroom [2]. However, behavioural detection in classroom settings faces numerous technical challenges. Firstly, classroom environments typically exhibit significant variations in lighting and frequent occlusions, such as students blocking each other's view or environmental objects like desks and chairs causing obstructions [3]. Secondly, student behaviour exhibits multi-scale characteristics, encompassing both localised micro-actions like raising hands or writing, and full-body movements such as standing or pacing. Moreover, recognising small-scale actions within classroom settings proves particularly challenging; subtle gestures like facial expressions or hand movements often prove difficult to capture due to low resolution. These factors collectively constrain the performance of existing behaviour detection models in authentic classroom environments [4].
Currently, classroom behaviour detection methods are primarily categorised into two main types: those based on traditional handcrafted features and those based on deep learning. Traditional approaches typically rely on manually designed features (such as HOG, Haar, etc.) combined with machine learning classifiers (such as SVM) for behaviour recognition [5]. While these methods can achieve certain results in constrained environments, they exhibit poor adaptability in complex classroom scenarios and struggle to capture high-level semantic information. Deep learning-based approaches, particularly convolutional neural networks (CNNs) and spatio-temporal graph convolutional networks (ST-GCNs), have become mainstream in behaviour recognition [6]. Among these, single-stage detectors like the YOLO series have garnered significant attention for their favourable speed-accuracy trade-off [7]. In recent years, the importance of spatio-temporal contextual information in behaviour recognition has gained recognition. Models such as TACNet have achieved significant progress on unedited video datasets by incorporating transition-aware mechanisms and long-term temporal modelling [8]. Concurrently, graph convolutional network-based approaches naturally capture spatial relationships and temporal evolution of human joints, offering novel perspectives for fine-grained behaviour recognition [9]. Nevertheless, the application of existing methods within the specific classroom setting remains in its infancy. Particularly, the effective integration of spatio-temporal contextual information to address the unique challenges of classroom environments warrants further investigation.
This paper proposes a spatiotemporal context-aware scene classification detection model tailored for classroom

scenarios, with three core contributions. Firstly, a cross-scale spatial context-aware module is designed, combining the hierarchical attention mechanism of Swin Transformer with Shuffle Attention to enhance the model's ability to capture multi-scale behavioural features [10]. Secondly, a temporal context modelling module is introduced, employing bidirectional ConvLSTM to extract long-term temporal dependencies and identify transitional states between behavioural segments [11]. Thirdly, a hybrid loss function tailored for classroom scenarios is constructed, integrating WIoU bounding box regression loss with focus classification loss to optimise training stability under imbalanced data conditions [12].

2 RELATED RESEARCH

The evolution of classroom behaviour recognition technology has progressed from traditional approaches to deep learning methods, as detailed in Table 1. Early research primarily relied on conventional computer vision techniques. Vara Prasad et al. employed Haar cascade classifiers for facial detection, combined with the K-nearest neighbours (KNN) algorithm, to develop a classroom attendance system [13]. Poudyal et al. employed Support Vector Machines (SVM) and Haar wavelet classifiers to identify key differences in student attention patterns [14]. Such approaches heavily relied on manually designed features, exhibiting limited generalisation capabilities in complex classroom environments.

Table 1 Comparison of Classroom Behaviour Detection Methods

Representative algorithm	Type	Advantages	Limitations
Haar+SVM[15]/HOG+KNN[16]	Traditional methods	Low computational complexity and high interpretability	Poor environmental adaptability, limited capacity for feature expression
CNN-based[17]/YOLO[18]	Deep Learning + Spatial Features	Automatic feature learning and high detection accuracy	Ignores temporal information and sensitive to occlusion
ST-GCN[19]/TACNet[8]	Deep Learning + Spatiotemporal Features	Capturing temporal and spatial context, recognising continuous behaviour	High computational costs and substantial annotated data

With breakthroughs in deep learning technology, behaviour recognition methods based on convolutional neural networks (CNNs) have significantly enhanced recognition performance in classroom settings. Kavitha et al. developed a CNN-based student behaviour detection framework, constructing feature extraction modules for eye and mouth regions to achieve fine-grained classification of specific facial behaviours such as nail-biting, sleeping with eyes closed, and yawning [20]. However, such approaches focus solely on spatial features while neglecting temporal dynamic information, making it challenging to comprehensively understand classroom behaviour sequences.

Spatio-temporal context awareness represents a core challenge in the field of behaviour recognition, proving particularly crucial for identifying continuous and interrelated behavioural patterns within classroom settings. In recent years, researchers have proposed multiple models to capture spatio-temporal contextual information within behaviours. Spatio-temporal Graph Convolutional Networks (ST-GCN) achieve effective analysis of skeleton sequence data by modelling the spatial relationships between human body joints and their temporal evolution as a graph structure [21]. Qi et al. proposed a human skeleton behaviour recognition model integrating global attention mechanisms with spatio-temporal graph convolutional networks, achieving significant performance improvements on datasets such as NTU-RGB+D. This approach enhances the model's recognition capability for occluded data by introducing global attention modules and spatio-temporal pooling operations [22]. The Transition-Aware Contextual Network (TACNet), proposed by Megvii Research, specifically addresses transitional state challenges in behaviour recognition. Comprising a temporal context detector and transition-aware classifier, TACNet employs bidirectional ConvLSTM units to extract long-term temporal context while simultaneously classifying actions and transitional states, substantially improving behaviour detection accuracy in unedited videos [8]. For classroom-specific applications, the CSSA-YOLO model employs a cross-scale feature optimisation strategy. Its C2fs module captures spatio-temporal dependencies in small-scale actions (e.g., raising hands), while integrating a Shuffle Attention mechanism to suppress complex background interference. These techniques provide crucial reference points for behaviour recognition in classroom environments [23].

However, existing approaches still exhibit shortcomings when applied to classroom scenarios. Firstly, most models lack specific design tailored to classroom contexts, such as teacher-student interactions and group collaboration. Secondly, they are insufficiently optimised for common classroom challenges like dense occlusions and small object detection. Finally, there is a lack of effective modelling for the long-term sequential dependencies inherent in classroom behaviours. The model proposed herein addresses these deficiencies through specialised optimisation.

3 MODEL ARCHITECTURE

The proposed classroom scenario classification and detection model, which perceives spatio-temporal contextual information, adopts a multi-branch encoder-decoder architecture. The overall framework is illustrated in Figure 1. The model takes classroom video sequences as input, processing spatial features, temporal dynamics, and contextual information through three parallel branches: the spatial stream, temporal stream, and context-enhanced stream. Finally,

the fusion module outputs scenario classification results and behaviour detection bounding boxes.

Input representation section. Given a classroom video sequence $V = f_1, f_2, ,,,, f_T$, where T denotes the sequence length, each frame f_T corresponds to an RGB image of size $H \times W \times 3$. The model first pre-processes input frames through dimension standardisation and brightness normalisation to mitigate the impact of lighting variations in classroom environments. The multi-branch feature extraction component comprises a spatial branch module based on an enhanced YOLOv8 architecture, incorporating cross-scale attention modules to capture spatial features at varying scales [24]. The temporal branch module employs bidirectional ConvLSTM layers to capture long-term temporal dependencies between frames. The context branch utilises graph convolutional networks to model semantic contextual relationships within classroom scenes, such as specific patterns of teacher-student interaction and group collaboration. In the feature fusion and output section, the features from the three branches are integrated through a weighted fusion module before being fed into the scene classifier and behaviour detection head. The classifier employs fully connected layers to output scene category probabilities, while the detection head predicts behaviour bounding boxes and category labels based on an anchor mechanism.

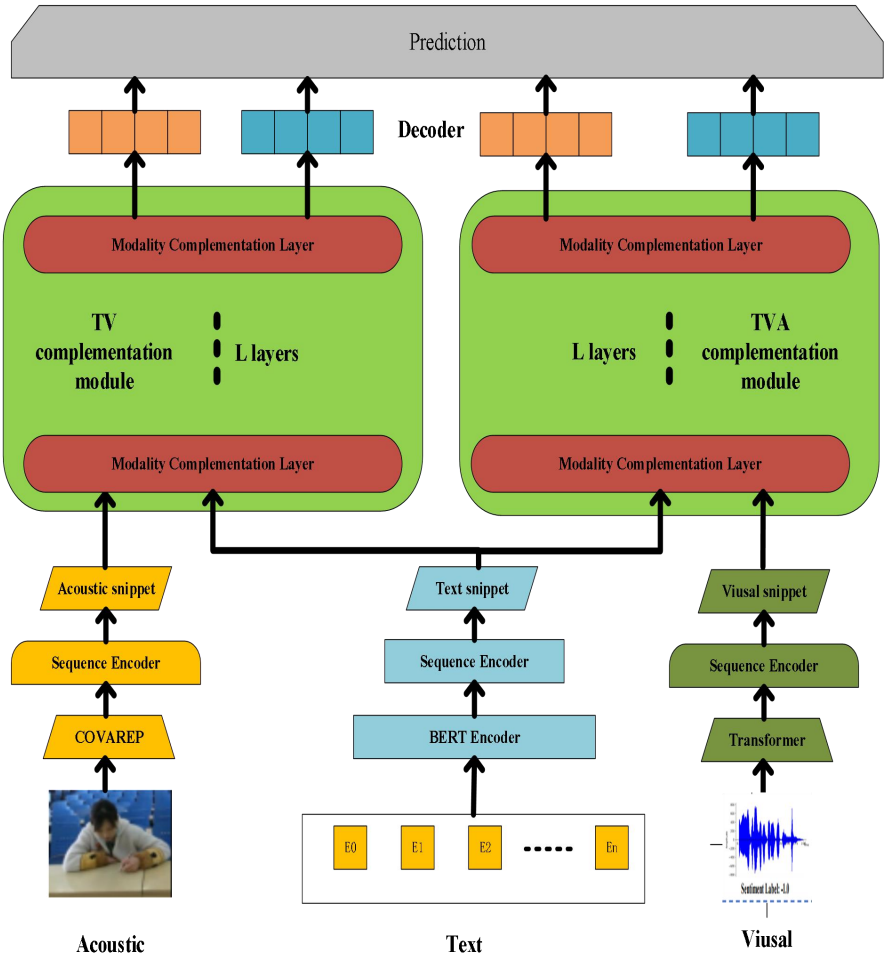

Figure 1 Schematic Diagram of the Overall Model Architecture

3.1 Spatial Context Awareness Module

The spatial context-aware module is responsible for extracting discriminative spatial features from each image frame while addressing multi-scale objects and occlusion issues in classroom settings. This module employs a cross-scale feature pyramid architecture, integrating the hierarchical window attention mechanism of the Swin Transformer with the Shuffle Attention channel attention mechanism.

Cross-scale feature extraction unit: Addressing the multi-scale characteristics of classroom behaviour, the module employs a multi-scale feature pyramid network (FPN) [25] to extract features across four distinct scales. For input frame f_T, multi-scale feature maps $P_1, P_2, ,,,, P_l$ are extracted via the backbone network (based on CSPDarknet53) [26], wherein P_l denotes the resolution of the input image $1/2^1$.

Window Attention Mechanism Unit Inspired by CSSA-YOLO, the Swin Transformer's window multi-head self-attention (W-MSA) mechanism is introduced within the C2f module to enhance the model's feature extraction

capability for small-scale behaviours. For each feature map P_l, it is first partitioned into $M \times M$ non-overlapping windows, where self-attention is then computed within each window as detailed in Equation 1.

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d_k}} + B)V \quad (1)$$

Where Q,K,V denotes the query, key, and value matrices respectively, B represents the learnable position encoding; and d_k denotes the dimension of the key vector. Through local self-attention computations within windows, the model effectively captures spatial dependencies within local regions, rendering it particularly well-suited for recognising small-scale classroom behaviours such as raising hands or writing.

The Channel Attention Mechanism Unit incorporates a Shuffle Attention (SA) mechanism to mitigate interference from complex backgrounds. The SA module first partitions the feature map into multiple subgroups along the channel dimension. Within each subgroup, it concurrently computes both channel attention and spatial attention. Finally, a shuffle operation facilitates information exchange between subgroups. Given a feature map P_l, the computational process of the SA module is expressed in Equation 2.

$$F_{out} = SA\ (F) = Shuffle(Concat(SA_1(F_1), SA_2(F_2), ..., SA_G(F_G))) \quad (2)$$

In this context, F is partitioned into G subgroups $F_1, F_2, , , , F_G$, with attention weights computed independently for each subgroup.

3.2 Time Context Awareness Module

Classroom behaviour exhibits distinct temporal continuity and dynamic evolution characteristics, such as raising one's hand to answer questions or lowering one's head to take notes, which typically comprise a sequence of consecutive actions. The temporal context modelling module aims to capture long-term temporal dependencies within behaviours, addressing issues of momentary occlusion and behavioural fragmentation.

Bidirectional temporal coding unit, employing a bidirectional ConvLSTM architecture that simultaneously leverages past and future contextual information to enhance the representation of the current frame. For time step t, the forward computation of the bidirectional ConvLSTM may be referenced in Equations 3 and 4.

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \quad (3)$$

$$c_t = f_t \Theta c_{t-1} + i_t \Theta g_t \quad (4)$$

Here, $*$ denotes the convolution operation, Θ represents element-wise multiplication, and σ signifies the sigmoid activation function. The concatenation of the forward and backward hidden states of the ConvLSTM forms the final temporal augmentation feature.

Transition State Perception Unit, inspired by TACNet, incorporates a transition-aware classifier specifically designed to distinguish genuine action states from transitional states. Transitional states refer to intermediate phases resembling genuine actions yet not belonging to action categories, such as the arm-raising motion preceding a completed hand-raising action. By explicitly modelling these states, the model reduces misclassifications and enhances temporal boundary accuracy.

Multi-scale Temporal Pyramid Unit: To capture behavioural patterns across different temporal scales, the module employs a multi-scale temporal pyramid architecture comprising three ConvLSTM layers with distinct time strides (1, 2, and 4 frames). Features from the pyramid's apex undergo upsampling before merging with lower-level features, thereby simultaneously capturing both short-term subtle motions and long-term behavioural patterns.

4 FUNCTIONAL TESTING AND DISCUSSION

4.1 Datasets and Evaluation Metrics

To validate the proposed model's effectiveness, experiments were conducted on two publicly available classroom behaviour datasets: the SCB-Dataset3 [27-28] and the self-constructed Classroom-Actions dataset, as detailed in Table 2. The SCB-Dataset3 comprises annotations for ten categories of student behaviour across three typical classroom scenarios (lecture, discussion, and self-study), including actions such as raising hands, writing, and reading.

Table 2 Statistical Information of Experimental Dataset

Dataset	Scene Category	Behavioural categories	Video clip	Number of students	Frame rate annotation
SCB-Dataset3	3	10	1,245	28	56,792
Classroom-Actions	5	15	2,637	52	128,435

Concurrently, this paper will conduct an assessment centred on the aforementioned indicators, as illustrated in Table 3.

Table 3 Performance Evaluation Metrics

Indicator Name	Indicator Implications	Evaluation Dimensions	In the specific context of this article
Scene Accuracy	Accuracy rate for scene category prediction	Macro-level scene recognition capability	The model's overall capability to correctly classify entire video sequences into distinct classroom types—such as lectures, discussions, self-study, and collaborative sessions—reflects its understanding of the global teaching paradigm.
Mean Average Precision	mAP@0.5	Fine-grained behaviour detection accuracy	Assess the model's overall performance in locating and identifying specific student behaviours (such as raising hands, writing, reading) within individual video frames. mAP@0.5 is the core evaluation metric in the field of object detection. The higher the value, the more precise the detection.
Temporal Localization Accuracy	Prediction accuracy for the start and end times of behaviours	Accuracy of behavioural time boundaries	Assess the degree of alignment between the start and end times of a predicted behaviour in the model and the actual timestamps. This metric is crucial for analysing the persistence and continuity of behaviours, such as accurately determining the commencement and conclusion of a “raising one's hand” action.
FPS	Detection speed, meeting real-time processing requirements	Model computational efficiency and real-time capability	The processing speed of the model is measured by the number of video frames it can analyse and process per second. A high frame rate per second (FPS) is a key indicator of whether the model can be applied to real-time classroom analysis systems, such as online teaching supervision and real-time feedback.

The table details the four core metrics employed in this paper to evaluate model performance. These metrics comprehensively assess the proposed spatio-temporal context-aware model's overall capability within classroom settings, examining its underlying principles, evaluation dimensions, and specific significance.

4.2 Experimental Procedure

The model proposed herein is implemented within the PyTorch framework, with training and inference conducted on an NVIDIA RTX 3090 GPU environment. The AdamW optimiser is employed, with an initial learning rate of 1e-4 and a batch size of 8. Training adopts a two-stage strategy: first, the backbone network is initialised using ImageNet pre-trained weights, focusing on spatial feature extraction. Subsequently, the temporal module is unfrozen and undergoes end-to-end fine-tuning using complete video sequence data. To enhance the model's generalisation capability, this study incorporates multiple data augmentation techniques. These encompass spatial augmentations such as random colour dithering, Gaussian blurring, and occlusion simulation; temporal augmentations including random frame sampling, temporal scaling, and video jitter; alongside classroom-specific enhancements like simulated lighting variations and desk-chair occlusion simulation. To address the class imbalance inherent in classroom behaviour data—where, for instance, ‘listening’ instances vastly outnumber ‘raising hand’ instances—a class-balanced sampling strategy is implemented. This is combined with a focus loss function to weight the optimisation of losses, thereby enhancing the model's recognition capability for minority class behaviours.

To comprehensively evaluate the proposed model's performance, five representative state-of-the-art methods were selected for comparative experiments. These include YOLOv8m, the current benchmark model for classroom behaviour detection; CSSA-YOLO, which incorporates cross-scale attention mechanisms to optimise multi-scale behaviour detection; TACNet, a context-aware network specialising in spatio-temporal action detection with transition perception capabilities; RA-GCNv2-A, a spatio-temporal graph convolutional network enhanced by global attention mechanisms; and VWE-YOLOv8, a classroom behaviour detection algorithm integrating multiple attention mechanisms. The aforementioned comparative approaches encompass diverse technical pathways, including detector optimisation, spatio-temporal context modelling, graph structure learning, and attention mechanism fusion. This multi-faceted evaluation validates the performance advantages demonstrated by this research.

4.3 Ablation Experiment

To validate the contributions of each module, ablation experiments were designed, with results presented in Table 4. The baseline model was YOLOv8m, to which spatial attention (SA), temporal context (TCM), and scene context (SCM) modules were progressively added.

Table 4 Ablation Experiment Results (on the SCB-Dataset3 dataset)

Model Configuration	Scene Accuracy Rate	mAP@0.5	FPS
YOLOv8m (Benchmark)[29]	76.5%	74.8%	82.3
+Spatial Attention(SA)	78.9% (+2.4%)	76.9% (+2.1%)	79.5
+Temporal Context(TCM)	81.2% (+4.7%)	79.3% (+4.5%)	75.8
+Scenario Context(SCM)	83.7% (+7.2%)	81.5% (+6.7%)	71.6
Complete Model	85.4%(+8.9%)	83.2%(+8.4%)	68.2

Experimental results demonstrate that each module makes a significant contribution to performance enhancement. Specifically, the spatial attention module improves mAP by 2.1%, primarily enhancing detection capabilities for small-scale behaviours; the temporal context module further boosts mAP by 4.5%, highlighting the importance of modelling temporal dependencies; while the scene context module contributes a 2.2% mAP improvement, underscoring the effectiveness of classroom-specific semantic context. The complete model achieves 8.9% and 8.4% improvements over the baseline in scene classification accuracy and action detection mAP respectively, confirming the synergistic effect of all modules.

4.4 Comparative Experiments and Analysis

The comparison results with state-of-the-art methods are presented in Table 5. On the SCB-Dataset3 dataset, our proposed model outperforms all competing approaches in both scene classification accuracy and action detection mAP, while maintaining an acceptable inference speed of 68.2 FPS.

Table 5 Comparison Results with State-of-the-Art Methods on the SCB-Dataset3 Dataset

Method	Scene Accuracy Rate	mAP@0.5	FPS
YOLOv8m[29]	76.5%	74.8%	82.3
CSSA-YOLO[30]	78.8%	76.0%	78.3
TACNet[8]	79.3%	77.2%	45.6
RA-GCNv2-A[31]	81.5%	78.9%	52.7
VWE-YOLOv8[32]	82.7%	80.1%	65.8
Our Method	85.4%	83.2%	68.2

As shown in the table, our model achieves a 2.7% improvement in scene classification accuracy and a 3.1% increase in mAP for behaviour detection compared to the state-of-the-art method (VWE-YOLOv8). This advancement is primarily attributable to the model's comprehensive utilisation of spatio-temporal contextual information, particularly its superior performance in handling complex classroom scenarios. It is noteworthy that although our model possesses a larger parameter size than YOLOv8m, its efficient attention mechanism and feature fusion strategy enable faster inference speeds than many complex graph convolutional network-based models (such as RA-GCNv2-A), thereby meeting the demands of real-time classroom analysis.

Regarding cross-scenario generalisation capability, our model underwent further validation on the Classroom-Actions dataset, as detailed in Table 6.

Table 6 Generalisation Performance on the Classroom-Actions Dataset

Method	Teaching scenario	Discussion scenario	Experimental Scenario	Collaborative Scenarios	Average
YOLOv8m	75.3%	70.8%	68.5%	65.2%	70.0%
CSSA-YOLO	77.1%	72.6%	70.3%	67.8%	72.0%
TACNet	78.2%	74.5%	72.1%	69.3%	73.5%
Our Method	81.5%	78.9%	76.7%	74.2%	77.8%

The results demonstrate that our proposed model exhibits outstanding generalisation performance on the more challenging and diverse Classroom-Actions dataset, achieving a mean average precision (mAP) of 77.8% across four scenarios—lecturing, discussion, experimentation, and collaboration—that surpasses other comparative models. This outcome provides robust validation of the model's strong adaptability to complex classroom environments. In-depth analysis reveals that all models exhibit a performance decline with increasing scene complexity, with the most pronounced challenges occurring in collaborative scenarios characterised by student grouping, severe occlusions, and frequent behavioural interactions. Nevertheless, our model maintains an mAP of 74.2% in this scenario, representing a substantial improvement of nearly 9 percentage points over the baseline YOLOv8m model. This advancement is primarily attributable to the effectiveness of our spatio-temporal context-aware architecture. Specifically, the spatial context-aware module enhances the localisation and recognition of individual objects under dense occlusion conditions through cross-scale attention mechanisms. Meanwhile, temporal context modelling assists in inferring more plausible behavioural categories from ambiguous single-frame images by analysing long-term dependencies within behavioural sequences. Together, these approaches address the core challenges inherent in complex scenarios.

4.5 Performance Experiments and Analysis

As shown in Figure2, this paper illustrates the performance trends of different models across four distinct classroom scenarios, with the primary objective of evaluating their robustness and adaptability to increasingly complex teaching environments.

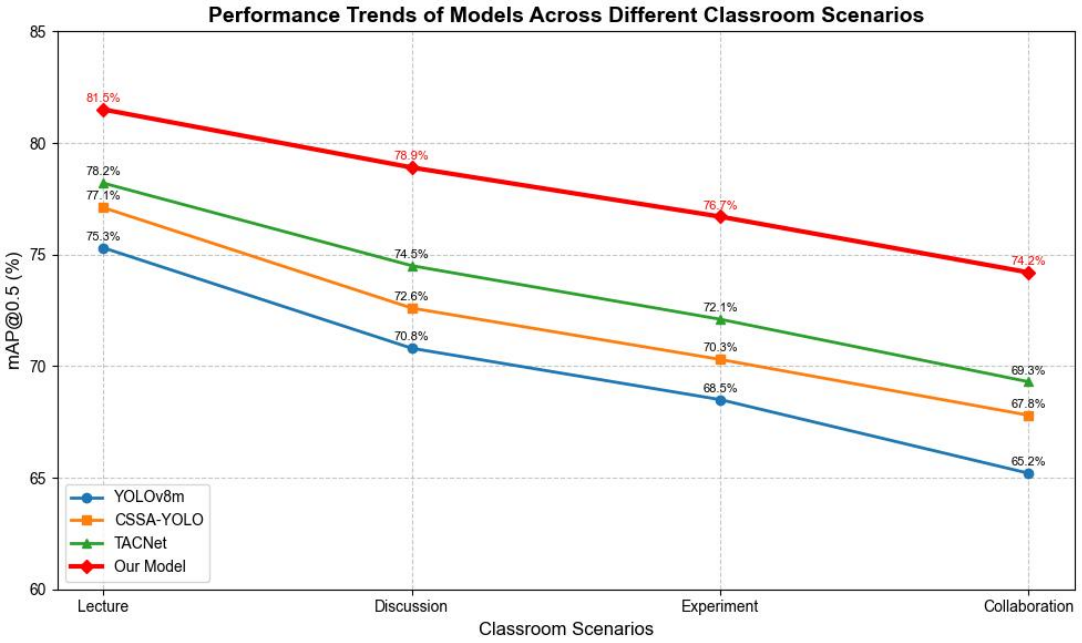

Figure 2 Performance Trends of the Model Across Different Classroom Scenarios

The line graph clearly demonstrates a consistent performance decline for all models as the scenario complexity escalates from structured Lecture to highly interactive Collaboration settings. This trend validates the inherent challenges posed by real classroom environments, particularly the issues of severe occlusion and frequent interactions present in collaborative learning scenarios. Notably, our proposed model (red line) maintains superior performance across all scenarios, with the performance advantage becoming most pronounced in the most challenging Collaboration scenario, where it achieves a 74.2% mAP@0.5, nearly 9 percentage points higher than the baseline YOLOv8m. This significant performance gap underscores the effectiveness of our spatiotemporal context-aware architecture, specifically the cross-scale spatial attention mechanism in handling severe occlusion and the long-term temporal modeling in disambiguating complex interactions. The results confirm that our model not only achieves state-of-the-art performance but also exhibits enhanced robustness in practical educational settings where complex student behaviors and interactions are prevalent.

Figure 3 presents a comprehensive performance comparison between different models using a grouped bar chart, with the primary objective of evaluating their overall effectiveness on the SCB-Dataset3 dataset across two critical metrics, Scene Classification Accuracy and Behavior Detection mAP@0.5.

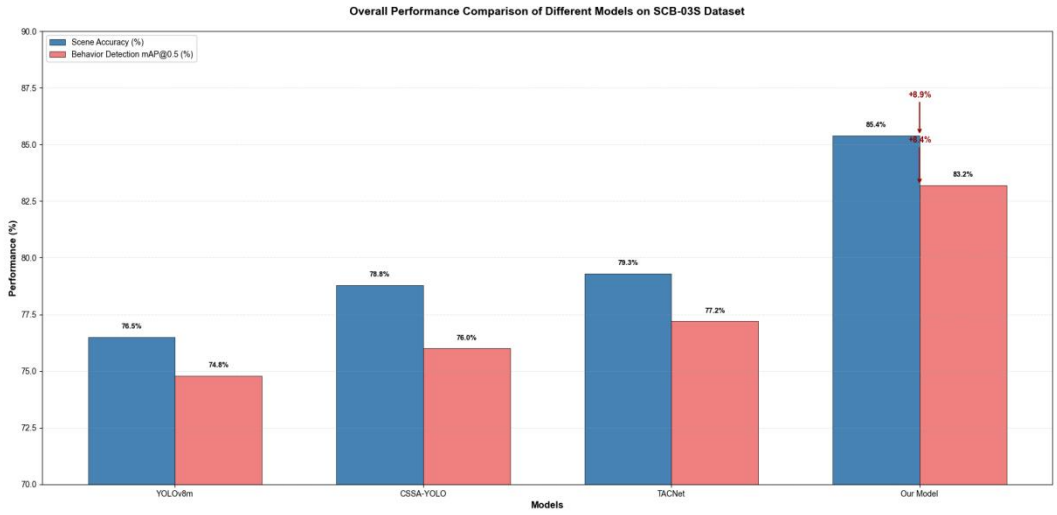

Figure 3 Overall Performance Comparison Between Models

The chart clearly demonstrates that our proposed model achieves superior performance on both evaluation criteria, attaining 85.4% in scene accuracy and 83.2% in behavior detection mAP, which represents significant improvements of 8.9% and 8.4% respectively compared to the baseline YOLOv8m model. More importantly, the parallel comparison reveals that our model exhibits more substantial enhancement in behavior detection capability, which directly validates the effectiveness of our core innovation—the spatiotemporal context-aware architecture—in precisely localizing and

recognizing fine-grained student behaviors. The consistent performance advantage across both metrics indicates that our model successfully addresses the dual challenges of macroscopic scene understanding and microscopic behavior analysis in educational environments, establishing a new state-of-the-art for classroom behavior analysis systems while maintaining practical applicability.

5 CONCLUSION

This study proposes a spatiotemporal context-aware scene classification and detection model tailored for classroom settings. By innovatively integrating cross-scale attention mechanisms in the spatial domain with long-term dependency modelling in the temporal domain, it effectively addresses key challenges in complex classroom environments, including lighting variations, frequent occlusions, and small-scale behaviour recognition. Experimental results demonstrate that the model achieves state-of-the-art performance across multiple datasets, including SCB-Dataset3 and Classroom-Actions. It attains a scene classification accuracy of 85.4% and an action detection mAP of 83.2%, significantly outperforming mainstream methods such as YOLOv8m and TACNet. This validates the proposed architecture's efficacy and superiority in concurrently handling macro-level scene understanding and micro-level behavioural analysis. Looking ahead, this research may be further deepened in three directions. Firstly, exploring multimodal data fusion mechanisms by integrating modalities such as speech and text to enhance the completeness of situational understanding. Secondly, investigating weakly supervised or self-supervised learning strategies to reduce the model's reliance on large volumes of finely annotated data, thereby enhancing scalability. Thirdly, optimising computational efficiency through techniques like neural network pruning and quantisation to adapt the model for edge computing devices, thereby advancing the practical implementation and widespread adoption of intelligent classroom systems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The project was supported by 2024 Shenzhen Polytechnic University Quality Engineering Project "Research on Classroom Scene Understanding and Behavior Analysis Method Based on Multimodal Attention Mechanisms (7024310268)".

REFERENCES

[1] Yang J, Shi G, Zhu W, et al. Intelligent technologies in smart education: a comprehensive review of transformative pillars and their impact on teaching and learning methods. Humanities and Social Sciences Communications, 2025, 12(1): 1239-1239.
[2] Sapiah S, Ulfah M S, Saputra N A, et al. Smart education in remote areas: collaborative strategies to address challenges in Majene Regency, Indonesia. Frontiers in Education, 2025, 101552575-1552575.
[3] Jain A, Dubey K A, Khan S, et al. A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems. Scientific Reports, 2025, 15(1): 17463-17463.
[4] Xieling C, Di Z, Gary C, et al. Author Correction: Blockchain in smart education: Contributors, collaborations, applications and research topics. Education and Information Technologies, 2022, 28(7): 9267-9267.
[5] Dey A, Anand A, Samanta S, et al. Attention-Based AdaptSepCX Network for Effective Student Action Recognition in Online Learning. Procedia Computer Science, 2024: 233164-174.
[6] Taojie X, Wei D, Si Z, et al. Research on Recognition and Analysis of Teacher–Student Behavior Based on a Blended Synchronous Classroom. Applied Sciences, 2023, 13(6): 3432-3432.
[7] Vaghela R, Vaishnani D, Sarda J, et al. Optimizing object detection for autonomous robots: a comparative analysis of YOLO models. Measurement, 2026, 257(PB): 118676-118676.
[8] Song L, Zhang S, Yu G, et al. TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. CoRR, 2019.
[9] Hua Z, Yang J, Ji W. Knowledge graph convolutional networks with user preferences for course recommendation. Scientific Reports, 2025, 15(1): 30256-30256.
[10] Wang Z, Yao J, Zeng C, et al. Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network. Systems, 2023, 11(7): 372-388.
[11] Pingo A, Castro J, Loureiro P, et al. Driving Behavior Classification Using a ConvLSTM. Future Transportation, 2025, 5(2): 52-52.
[12] Fu R, Tian M. Classroom Facial Expression Recognition Method Based on Conv3D-ConvLSTM-SEnet in Online Education Environment. Journal of Circuits, Systems and Computers, 2023, 33(07).
[13] Cai S, Zhang X, Mo Y . A Lightweight underwater detector enhanced by Attention mechanism, GSConv and WIoU on YOLOv8. Scientific Reports, 2024, 14(1): 25797-25797.
[14] K S, Prasad V. Design and Implementation of an Efficient Rose Leaf Disease Detection using K-Nearest Neighbours. International Journal of Recent Technology and Engineering (IJRTE), 2020, 9(3): 21-27.

[15] Paneru B, Paneru B, Sapkota C S, et al. Enhancing healthcare with AI: Sustainable AI and IoT-Powered ecosystem for patient aid and interpretability analysis using SHAP. Measurement: Sensors, 2024: 36101305-101305.
[16] Sugiharto A, Harjoko A, Suharto S. Indonesian traffic sign detection based on Haar-PHOG features and SVM classification. International Journal on Smart Sensing and Intelligent Systems, 2020, 13(1): 1-15.
[17] Kwon H B, Kim K J. Image Searching using a Cascade of HOG-kNN. ITC-CSCC :International Technical Conference on Circuits Systems, Computers and Communications, 2015.
[18] Priya V K, Peter D J. Enhanced Defensive Model Using CNN against Adversarial Attacks for Medical Education through Human Computer Interaction. International Journal of Human–Computer Interaction, 2025, 41(3): 1729-1741.
[19] Shen Q, Zhang L, Zhang Y, et al. Distracted Driving Behavior Detection Algorithm Based on Lightweight StarDL-YOLO. Electronics, 2024, 13(16): 3216-3216.
[20] Pu L, Zhao Y, Hua Z, et al. Multi-Target spraying behavior detection based on an improved YOLOv8n and ST-GCN model with Interactive of video scenes. Expert Systems With Applications, 2025: 262125668-125668.
[21] Kuppala K, Banda S, Imambi S S. Selection of Distance Measure for Visual and Long Wave Infrared Image Region Similarity using CNN Features. Procedia Computer Science, 2024: 235970-978.
[22] Deyuan Z, Haoguang W, Chao W, et al. Video Human Action Recognition with Channel Attention on ST-GCN. Journal of Physics: Conference Series, 2021, 2010(1).
[23] Lu Qi. Sports-ACtrans Net: research on multimodal robotic sports action recognition driven via ST-GCN. Frontiers in Neurorobotics, 2024: 181443432-1443432.
[24] Zhou L, Liu X, Guan X, et al. CSSA-YOLO: Cross-Scale Spatiotemporal Attention Network for Fine-Grained Behavior Recognition in Classroom Environments. Sensors, 2025, 25(10): 3132-3132.
[25] Okano T M, Lopes C A W, Ruggero M S, et al. Edge AI for Industrial Visual Inspection: YOLOv8-Based Visual Conformity Detection Using Raspberry Pi. Algorithms, 2025, 18(8): 510-510.
[26] Thammasanya T, Patiam S, Rodcharoen E, et al. A new approach to classifying polymer type of microplastics based on Faster-RCNN-FPN and spectroscopic imagery under ultraviolet light. Scientific reports, 2024, 14(1): 3529-3529.
[27] Senussi F M, Kang S H. Occlusion Removal in Light-Field Images Using CSPDarknet53 and Bidirectional Feature Pyramid Network: A Multi-Scale Fusion-Based Approach. Applied Sciences, 2024, 14(20): 9332-9332.
[28] Chen S, Liu Y, Zhang H, et al. A human location and action recognition method based on improved Yolov11 model. Discover Artificial Intelligence, 2025, 5(1): 232-232.
[29] Wang Z, Yuan G, Zhou H, et al. Foreign-Object Detection in High-Voltage Transmission Line Based on Improved YOLOv8m. Applied Sciences, 2023, 13(23).
[30] Zhou L, Liu X, Guan X, et al. CSSA-YOLO: Cross-Scale Spatiotemporal Attention Network for Fine-Grained Behavior Recognition in Classroom Environments. Sensors, 2025, 25(10): 3132-3132.
[31] Nan Y, Niu W, Chang Y, et al. Transient Stability Assessment of Power Systems Built upon Attention-Based Spatial–Temporal Graph Convolutional Networks. Energies, 2025, 18(14): 3824-3824.
[32] Liu J, Lin C, Chen J, et al. Research on Real-Time Analysis and Intervention of Classroom Behaviour Based on Object Detection Algorithms. Advances in Vocational and Technical Education, 2025, 7(2).