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

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2169-2180
Journal homepage: http://iieta.org/journals/ts

Automatic Student Engagement Recognition in Online Classrooms Based on Deep Image Segmentation and Attention Mechanisms

Nan Feng[1*], Youwei Chen[2], Conglin Ran[3]

[1] School of Health Services Management, Xi'An Medical University, Shaanxi 710021, China
[2] School of Economics and Management of Xupt, Xi'an University of Posts & Telecommunications, Shaanxi 710061, China
[3] School of Education, Jiujiang University, Jiangxi 332005, China

Corresponding Author Email: duolabmeng2024@163.com

https://doi.org/10.18280/ts.420428

Received: 10 February 2025
Revised: 9 June 2025
Accepted: 17 June 2025
Available online: 14 August 2025

Keywords:
online classroom, student engagement, automatic recognition, deep image segmentation, attention mechanism, variable kernel convolution

ABSTRACT

With the widespread adoption of online education, the spatial and temporal separation between teachers and students has made it difficult for educators to accurately assess student engagement in real-time. Traditional methods of manually observing engagement are no longer effective, and the complexity of online learning environments presents significant challenges for automatic engagement recognition based on images. Existing studies have several limitations, such as insufficient robustness due to reliance on single-modal features, the inability to adapt to multi-scale variations with fixed kernel convolutions, redundant calculations in attention mechanisms, and the lack of effective utilization of spatial interaction information. To address these issues, this paper proposes an automatic student engagement recognition model for online classrooms based on spatial interaction and segmentation attention. The model introduces a novel variable kernel convolution module, which dynamically adjusts the kernel size and receptive field based on the target scale to enable multi-scale feature extraction. Additionally, an improved multi-branch attention feature fusion module is constructed to process different dimensional features in parallel, strengthening the expression of important features, suppressing redundant information, and reducing computational costs. This model significantly enhances the ability to extract irregular, multi-scale, and spatially correlated engagement features in online classrooms, effectively solving core challenges in online classroom object detection, such as small sample sizes, multi-scale issues, and occlusions. It provides a new technological pathway for engagement evaluation in online education.

1. INTRODUCTION

The popularity of online education has broken the spatial and temporal constraints of traditional teaching, but it has also introduced a natural barrier to teaching interaction [1-3]. Teachers find it difficult to intuitively capture students' classroom states, while students face problems such as distraction and lack of participation in an environment without real-time supervision [4-7]. This spatial-temporal separation renders traditional focus assessment methods, which rely on manual observation, completely ineffective, necessitating an automated technical solution to perceive students' learning states in real time. With the widespread use of smart devices [8, 9] and the development of computer vision technology [10-12], image-based engagement recognition has become possible. However, the complexity of online environments, such as students' varying postures, changing camera angles, and unstable lighting conditions, presents significant challenges for accurate recognition.

The automatic recognition of student engagement in online classrooms has important theoretical and practical value. From a teaching practice perspective, real-time engagement feedback can help teachers dynamically adjust teaching strategies, such as increasing interactive sessions during periods of low attention or guiding distracted students, thereby improving online teaching quality. For students, engagement management supported by this technology can promote the development of autonomous learning abilities, forming a personalized learning pace. From the perspective of educational technology development, this research promotes the deep integration of artificial intelligence and educational scenarios, providing core technical support for building intelligent, adaptive online learning environments, and contributing to the realization of truly personalized education.

Existing studies still have many limitations in the field of engagement recognition. Traditional methods often rely on single-modal features, such as facial expressions or head postures, neglecting the correlations between features, which leads to insufficient recognition robustness. Some deep learning-based models [13, 14] use fixed kernel convolutions for feature extraction, which cannot adapt to multi-scale variations in student faces, bodies, and other targets in online scenes, resulting in poor recognition of small targets at a distance or partial occlusions. At the same time, existing attention mechanisms [15-17] often suffer from redundant calculations during feature fusion, which not only increases

model complexity but may also reduce recognition accuracy due to overemphasis on irrelevant information. In addition, some models [18-20] lack effective use of spatial interaction information, making it difficult to capture deeper semantics such as teacher-student interaction and student behavior associations in the classroom, leading to poor generalization ability in small sample scenarios.

This study proposes an automatic student engagement recognition model for online classrooms based on spatial interaction and segmentation attention, which breaks through existing technical bottlenecks through two core innovations: First, it introduces a novel variable kernel convolution module that can dynamically adjust the kernel size and receptive field according to the target scale, effectively extracting multi-scale features from micro-expressions to full-body postures, thus improving adaptability to irregular targets. Second, an improved multi-branch attention feature fusion module is constructed to process features from spatial, channel, and semantic dimensions in parallel, strengthening the expression of key information, while using channel pruning techniques to reduce redundant calculations and improve model efficiency. This model significantly enhances feature perception ability for complex scenes in online classrooms, especially excelling in small sample data, multi-scale targets, and partial occlusions, providing a new technical pathway to solve the engagement evaluation problem in online education, and promoting the leap from simple feature recognition to deep scene cognition.

2. AUTOMATIC STUDENT ENGAGEMENT RECOGNITION MODEL FOR ONLINE CLASSROOMS BASED ON DEEP LEARNING AND IMPROVED ATTENTION MECHANISMS

2.1 Task description

The task of automatic student attention recognition in online classrooms aims to process input images of online classroom scenarios through a model based on spatial interaction and segmentation attention, in order to accurately recognize and locate students' attention states in the images. Specifically, the task focuses on predicting the category label corresponding to each attention-related target in the image—such as students showing states of concentration, distraction, or interaction—as well as the bounding box coordinates (y_a, y_b, y_q, y_g). The category label may belong to either a base class or a novel class representing specific attention states. The base classes include common attention states with sufficient annotated samples, such as typical concentrated listening or head-down distraction. The novel classes cover special attention states with only a few K-shot samples, such as specific interactive gestures or attention shifts caused by device usage, to simulate recognition needs in low-data scenarios of real teaching environments. This task adopts a few-shot object detection framework, with the training process divided into two stages: base training and fine-tuning. In the base training stage, the model learns general attention-related feature representations using the base class dataset, focusing on capturing spatial correlations between students and between students and teaching elements through the spatial interaction module, while enhancing the feature extraction of key attention-related regions such as facial expressions and body movements using the segmentation attention mechanism. In the fine-tuning stage, based on a few support samples of novel classes, the model quickly adapts to new attention states through the improved multi-branch attention feature fusion module, suppresses interference from background and irrelevant information, and improves recognition robustness under few-shot conditions. The testing phase adopts an N-way K-shot setting, requiring the model to accurately distinguish positive and negative samples and perform category classification and bounding box regression in each task containing V attention state categories, using only J annotated instances per category. Ultimately, the goal is to achieve real-time and accurate recognition of various student attention states in online classrooms, providing data support for instructional strategy adjustment.

2.2 Overall network structure

The specific implementation steps of the proposed automatic student engagement recognition model for online classrooms are as follows: First, for the input online classroom scene feature map, group the targets based on their spatial distribution characteristics in the classroom. Each group corresponds to spatially related areas in the classroom, such as students in the same row or the interaction area between students and the screen. Each group is further divided into multiple blocks, with each block focusing on a single student or local interaction area, such as the contact area between a student's hand and a device or the micro-expression area of the face, to achieve independent processing of different engagement feature units. Secondly, within each block, a novel variable kernel convolution module is embedded. The module dynamically adjusts the kernel size and receptive field based on target scales such as facial details of nearby students and the overall posture of distant students, accurately extracting multi-scale engagement features, such as gaze direction, head-down angle, and hand-raising actions. Then, a block attention mechanism is used to fuse the features of all blocks within each group, emphasizing the expression of key engagement features within the group and suppressing background interference within the blocks. Subsequently, an improved multi-branch attention feature fusion module is constructed, and the features extracted from all groups are fused via multiple paths from the spatial interaction dimension, channel feature dimension, and semantic feature dimension. This further highlights the features strongly correlated with engagement in online classrooms, and through channel pruning, reduces redundant computations. Finally, a residual connection is established between the fused features and the original input features, retaining the basic engagement features in the original image, such as the student's basic outline and location information. This forms an output feature map containing multi-scale, spatial interaction, and key engagement features, providing precise feature support for subsequent category label prediction and bounding box regression, and effectively enhancing the model's engagement recognition ability for irregular postures, multi-scale targets, and occlusion scenarios in online classrooms. Figure 1 shows the overall network structure diagram.

The computation process of the proposed model is as follows:

Step 1: Group Division — Focusing on Classroom Spatial Region Associations

For the input online classroom scene feature map, divide it into j=2 base groups based on spatial associations in the classroom. The division follows the typical scenes in online

classrooms: The first group focuses on the front and central areas of the classroom, while the second group covers the back and edge areas. Grouping strengthens the spatial interaction features between students within the same region, laying the foundation for subsequent block processing and ensuring that the features of each group reflect the engagement behavior patterns of specific regions.

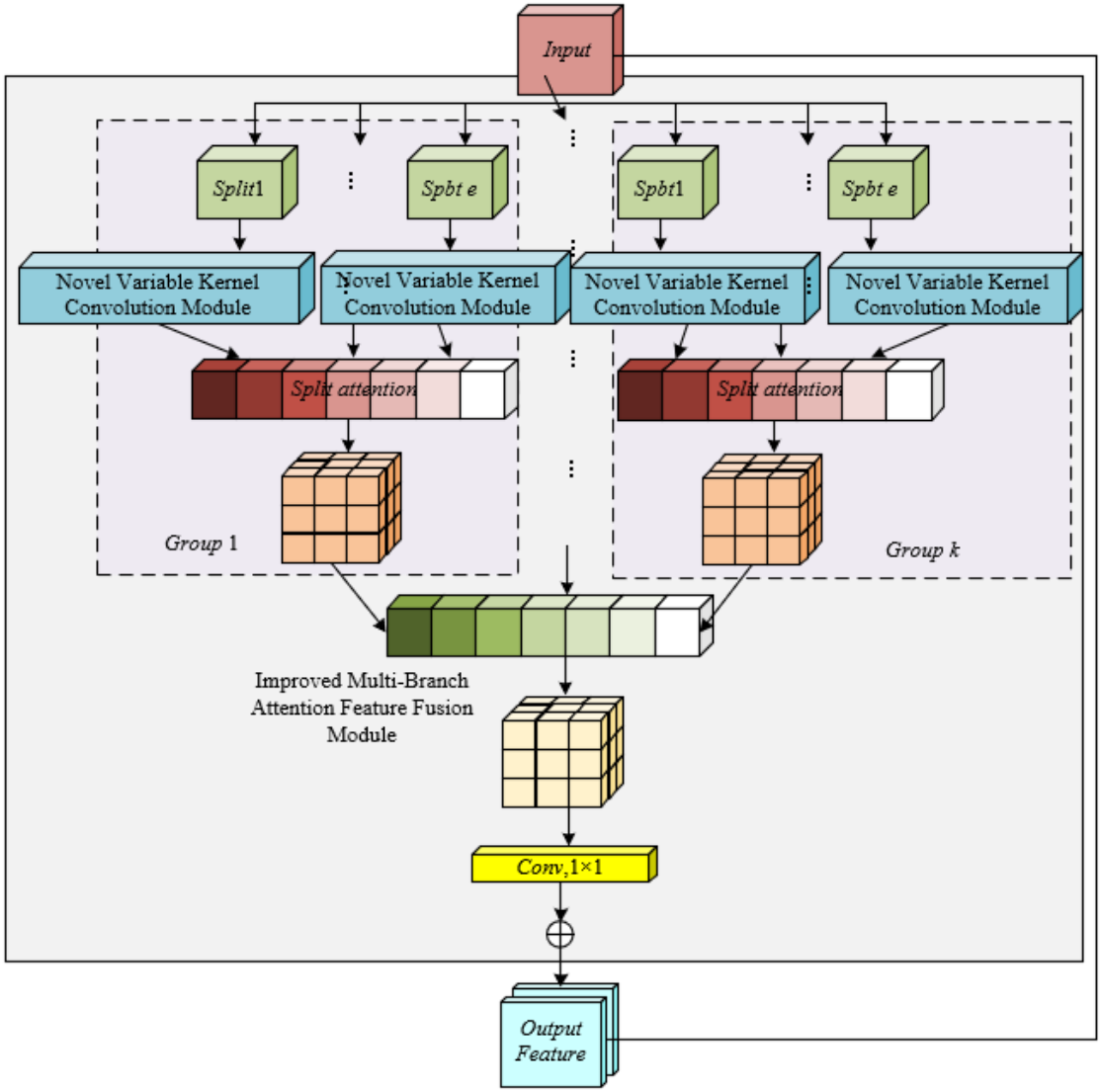

Figure 1. Overall network structure diagram

Step 2: Block Feature Extraction — Multi-scale Capture of Engagement Details and Global Features

Each base group is further divided into e=3 splits, with each split designed to capture engagement features at different scales in the online classroom: The first split uses a conventional convolution kernel of size 3 to focus on facial details of students; the second split uses a new variable kernel convolution with a kernel size of 5 to capture medium-scale features; and the third split uses a new variable kernel convolution with a kernel size of 7 to extract large-scale features. The combination of these three convolution kernels achieves full coverage of engagement features from the micro to macro levels, adapting to the complex scenarios in online classrooms where close-up shots and wide-angle views coexist.

Step 3: Block Attention Fusion — Enhancing Key Engagement Features within Each Group

Perform block attention fusion on the three split features within each base group, highlighting the engagement-related features within the group: First, sum all the input feature maps I_k (k=1,2,3) of the three splits element-wise to integrate the correlation information between facial details, body movements, and global posture. Then, apply global average pooling to the fused features, compressing the spatial dimensions from $G \times H$ to a single channel, preserving key spatial parameters such as “student-screen” distance and “head-desk” angle. Next, process the compressed features with two fully connected layers and use softmax to calculate the weight Qu for each split. The attention weight for engagement judgment is inclined towards the facial detail split, while medium-scale splits are assigned secondary weight. Finally, multiply each split's input feature map by its corresponding weight and sum them to obtain the fused features of the base group, such as strengthening the feature expression of "facing the screen with an upright posture" and suppressing the interference from "outside scenery and irrelevant decorations." The specific feature map fusion formula is:

$$I = \sum_{u=1}^{e} I_k \tag{1}$$

The spatial dimension compression formula is:

$$T = \frac{1}{G \times Q} \sum_{u=1}^{G} \sum_{k=1}^{Q} I(u,k) \tag{2}$$

The feature fusion result for the current base group is obtained as follows:

$$N = \sum_{u=1}^{e} \left(I_u \times Q_u \right) \tag{3}$$

Step 4: Multi-Branch Attention Fusion — Enhancing Engagement Feature Interaction Across Regions

Use the improved multi-branch attention feature fusion module to enhance and fuse the output features from the two base groups across regions: The spatial interaction branch focuses on the student status correlations between the two groups, strengthening the features of the interaction areas through a spatial attention weight matrix. The channel feature branch explores the correlation between different feature channels, such as the “facial expression channel” and the “body movement channel.” When “furrowing brows + leaning forward” occurs simultaneously, it enhances the recognition weight of the "deep thinking" state. The semantic feature branch maps the features to predefined engagement labels, ensuring the fusion results meet classification requirements. Additionally, redundant channels are pruned through channel pruning techniques to reduce computation. The fused feature map retains the engagement details within the region and strengthens the cross-region state correlation, improving the ability to recognize complex scenarios such as “collective distraction” and “local interaction.”

Step 5: Feature Integration and Skip Connections — Retaining Original Engagement Baseline Features

First, perform 1×1 convolutions to adjust the number of channels in the fused features so that they match the feature dimensions required for the online classroom engagement recognition task. Then, use skip connections to merge the adjusted feature map with the model's original input features. The original input contains unfiltered baseline information of the classroom scene, and the skip connection ensures that these baseline features complement the processed dynamic engagement features, preventing recognition bias due to information loss during the feature extraction process. The final output feature map contains multi-scale details, spatial interaction correlations, and original scene baselines, providing comprehensive feature support for subsequent engagement category label prediction and bounding box regression.

2.3 Novel variable kernel convolution module

In this paper, a novel variable kernel convolution module is introduced in the proposed model, mainly due to the contradiction between the special feature extraction requirements of online classroom scenes and the inherent limitations of traditional convolution operations. Traditional convolutional neural networks extract features by sliding a fixed kernel across the spatial dimensions. On one hand, it is difficult to adapt to the complex distribution of multi-scale targets in online classrooms, as the receptive field of the fixed kernel cannot flexibly cover these differential features, leading to insufficient extraction of engagement-related details. On the other hand, the high computational cost of traditional convolutions fails to meet the real-time recognition demands of online classrooms. The Spatial-shift operation achieves spatial information interaction by shifting features along the height and width directions, which can capture dynamic associations between students, such as turning the head to talk or synchronous head-down behaviors, without additional parameters. This significantly reduces the computational cost and adapts to the real-time requirements of classroom scenarios. At the same time, the variable kernel design can dynamically adjust the kernel size to precisely match the engagement features at different scales, combined with the segmentation attention mechanism to strengthen the expression of key features, effectively solving the feature extraction challenges caused by occlusion and angle variations in online classrooms, ultimately improving the accuracy and robustness of engagement state recognition. Figure 2 shows the network structure diagram of the novel variable kernel convolution module. The operational steps of the novel variable kernel convolution module are described in detail below:

Step 1: Input Feature Map Grouping — Focusing on Core Engagement Feature Channels

For the input online classroom scene feature map A, the channel dimension Z is evenly divided into 4 groups according to the degree of association between features and student engagement, with each group having $Z/4$ channels. The grouping logic closely aligns with typical online classroom scenarios: The first group retains channels related to facial micro-expressions, such as gaze direction and mouth corner curvature, which directly reflect the engagement state. The second group includes channels for body movement, such as head rotation angles and hand position changes, used to determine behaviors like "raising hand for interaction" or "head-down distraction." The third group contains spatial position channels, such as the relative coordinates between students and the screen or the blackboard, reflecting the attention focal point. The fourth group includes environmental interference channels, such as light changes, desk and chair information, and other background elements. Through grouping, the feature channels are functionally divided, highlighting the core engagement-related features and laying the foundation for subsequent targeted processing, reducing the computational complexity of handling each group’s features. Let A_u represent the feature map of the u-th group, then the expression is:

$$A = \left[A_1, A_2, A_3, A_4 \right] \tag{4}$$

Step 2: Spatial Shift Operation — Capturing Dynamic Engagement Spatial Interaction Information

For the 4 grouped feature maps, three shift branches and one baseline branch are designed to achieve efficient spatial information interaction in the classroom scene through spatial shifts. The baseline branch does not shift the feature map, retaining the original spatial location of the features. The first branch shifts the feature map 1 unit upwards along the height direction, focusing on capturing vertical posture changes such as students lowering their heads. The second branch shifts the feature map 1 unit downwards along the height direction, enhancing the feature capture of actions like tilting the head backward. The third branch shifts the feature map 1 unit to the left and right along the width direction, focusing on horizontal behaviors such as students turning their heads or moving their

gaze from side to side. The shift operation requires no additional parameter computation and can capture the spatial correlation of students' dynamic postures by simple feature location offsets. This not only reduces computational costs but also precisely adapts to the spatial feature variations caused by camera angles and student positions in online classrooms, reinforcing the role of spatial interaction features in indicating engagement. Let the width of the feature map be q and the height be g, the spatial shift example is:

$$\begin{aligned} A_1[1:q,:,:] &\leftarrow A_1[0:q-1,:,:], \\ A_2[0:q-1,:,:] &\leftarrow A_2[1:q,:,:], \\ A_3[:,1:g,:] &\leftarrow A_3[:,0:g-1,:], \\ A_4[:,0:g-1,:] &\leftarrow A_4[:,1:g,:] \end{aligned} \tag{5}$$

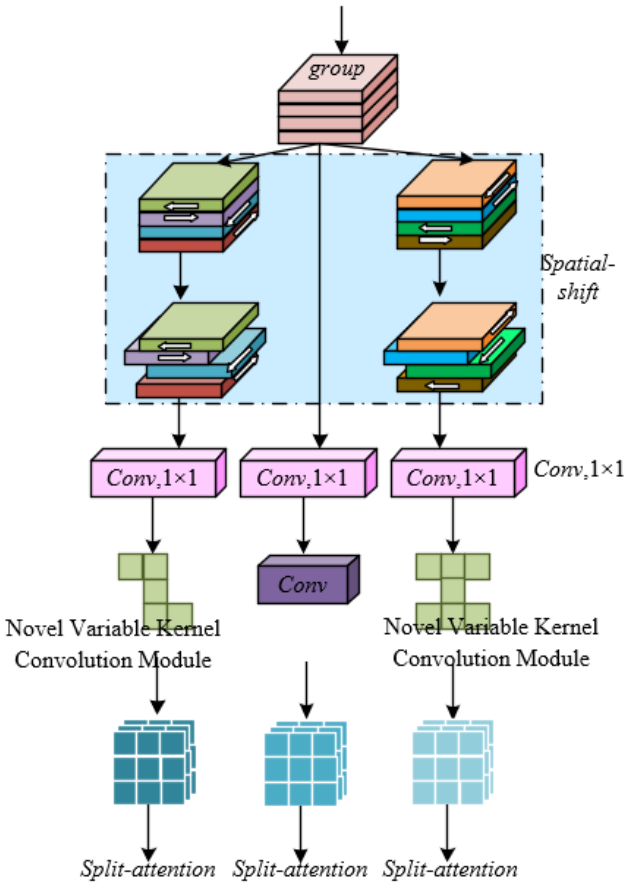

Figure 2. Novel variable kernel convolution module network structure diagram

Then, the 4 shifted feature maps A_1, A_2, A_3, A_4 are concatenated along the channel dimension to obtain a new feature map A':

$$A' = [A_1, A_2, A_3, A_4] \tag{6}$$

Step 3: Multi-Scale Feature Extraction — Strengthening the Representation of Engagement Features at Different Scales

In each branch, the shifted feature maps are processed by the novel variable kernel convolution for targeted feature extraction, to adapt to engagement targets at different scales in the online classroom. For small-scale features such as facial micro-expressions, a 3×3 variable kernel convolution is used to focus on local details. For medium-scale features such as head rotations and hand movements, a 5×5 variable kernel convolution is applied to expand the receptive field. For large-scale features such as full-body postures, a 7×7 variable kernel convolution is used to cover a larger spatial range. The dynamic adjustment ability of the variable kernel effectively solves the multi-scale feature disparity problem in online classrooms, where "facial details of front-row students are clear, and postures of back-row students dominate," while enhancing the feature representation of irregular targets, ensuring that the extracted feature map comprehensively covers engagement-related information from micro to macro scales. Let the value of the output feature map at position o be $A'[o]$, the dynamically generated offset at the l-th sample point be $\Delta j_l(o)$, the dynamically generated weight be $q_l(o)$, the maximum number of sampling points be L, and the bilinear interpolation operation be $INTERP()$, then:

$$A'[o] = \sum_{l=1}^{L} INTERP\left(A', o + \Delta j_l(o)\right) \cdot q_l(o) \tag{7}$$

Step 4: Split-attention Feature Fusion — Highlighting Key Engagement Features

The Split-attention mechanism is used to fuse the feature maps extracted from each branch, forming the final output feature map of the module. The specific process is as follows: First, perform global average pooling on the feature map of each branch to compress the spatial dimensions and retain key parameters such as "student-screen distance" and "action duration." Then, use a fully connected layer and the softmax function to calculate the weight for each branch, with the branches related to facial expressions and body movements receiving higher weights and the branches related to environmental interference being suppressed. Finally, multiply each branch's feature map by its corresponding weight and sum them to complete the feature fusion. The fused feature map B highlights positive engagement features such as "facing the screen with an upright posture," while suppressing interference from "light changes and irrelevant object movements," precisely focusing on the core engagement states of students in online classrooms. This provides high-quality feature support for subsequent engagement category prediction and bounding box regression, improving the model's robustness in multi-scale and occlusion scenarios. The final output feature map B expression is:

$$B = TX\left(\left\{A_u'\right\}_{u=1}^{3}\right) \tag{8}$$

2.4 Improved multi-branch attention feature fusion module

The introduction of the improved multi-branch attention feature fusion module in the online classroom student engagement automatic recognition model primarily stems from the limitations of existing fusion methods in adapting to online classroom scenarios. In online classrooms, student engagement features exhibit significant hierarchical associations. Low-level features such as gaze direction and mouth micro-expressions need to be deeply associated with high-level features like the interactive semantics of "raising hand to ask questions" or the disengagement behavior of "slumping down with head down" to accurately judge the engagement state. However, traditional methods of feature extraction via separate branches followed by resampling lose

the cross-level association details like "gaze deviation + hand operation with a device," making it difficult to capture instantaneous changes in engagement in online scenarios. Moreover, online classrooms present multi-scale targets and spatial interaction features, and traditional methods do not assign different feature weights, which may lead to key features being overshadowed by secondary features. These methods are also prone to gradient vanishing in small sample training, affecting the model's ability to learn new categories of engagement states. Additionally, the occlusion issue in online scenes requires that the fusion module retains local details while integrating global semantics, which traditional feature fusion methods struggle to handle. Therefore, to address these issues, this module introduces a spatial interaction perception mechanism to strengthen cross-level feature associations, employs dynamic weight allocation to highlight key engagement features, and incorporates multi-path fusion to adapt to multi-scale and occlusion scenarios, ultimately improving the robustness and accuracy of student engagement recognition in online classrooms. The module network structure diagram is shown in Figure 3. The module is divided into five processing parts:

Step 1: Attention Mechanism Preprocessing — Strengthening Core Engagement Feature Expression

The spatial attention and channel attention mechanisms are applied to the feature map input of each branch: Spatial attention focuses on local key areas by generating a spatial weight matrix to enhance engagement-related spatial features such as "facing the screen" and "raising hand to ask questions," while suppressing interference from irrelevant background areas such as desks, chairs, and decorations. Channel attention focuses on the global correlations between feature channels, giving higher weights to channels strongly related to engagement, such as "gaze direction" and "head posture," while diminishing the impact of redundant channels like light changes and image noise. Through dual attention preprocessing, each branch's feature map initially focuses on the core features that play a decisive role in engagement judgment in the online classroom, laying the foundation for high-quality fusion in the subsequent steps. Let the u-th channel be represented by u, spatial attention operation by $TX(\ )$, and channel attention operation by $ZX(\ )$, then the preprocessing formula is:

$$\Theta_u^T = TX\left(a_u\right), \Theta_u^Z = ZX\left(\Theta_u^T\right), u = 1, 2, ..., v \tag{9}$$

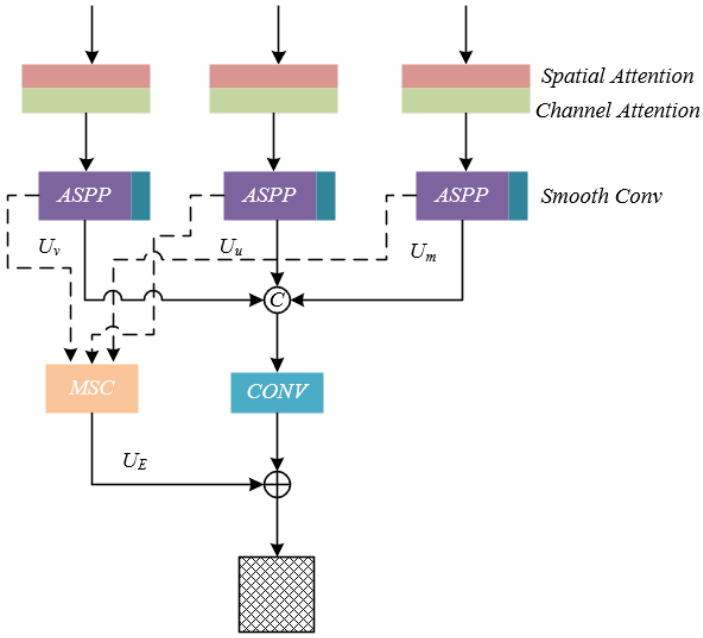

Figure 3. Improved multi-branch attention feature fusion module network structure diagram

Step 2: Dilated Spatial Pyramid Pooling — Capturing Multi-Scale Engagement Scene Information

The Dilated Spatial Pyramid Pooling module is used to perform parallel sampling on each branch's feature map. Convolutions with different dilation rates are set to adapt to the multi-scale target characteristics in online classrooms. A small dilation rate captures facial micro-expressions of students at a close distance, a medium dilation rate focuses on body movements at a medium scale, and a large dilation rate covers the overall posture of students at a far distance. By using multi-scale sampling, feature maps such as Θ^X_1, Θ^X_2....Θ^X_v are generated to capture engagement information across all scales, from micro expressions to macro behaviors, effectively addressing the varying target scales due to differences in camera distance and enhancing the unified recognition ability for "close-up clear face" and "distant blurred posture." The feature map expression is:

$$\Theta_u^X = ASPP\left(\Theta_u^Z\right) \tag{10}$$

Step 3: Smoothing Convolution and Element-Wise Fusion — Suppressing Noise and Enhancing Feature Complementarity

The feature maps Θ^X_1, Θ^X_2....Θ^X_v output by Dilated Spatial Pyramid Pooling are each processed by a 1×1 smoothing convolution to eliminate high-frequency noise introduced during the multi-scale sampling process, making the feature maps smoother and more continuous while preserving core information. Afterward, the smoothed feature maps are merged by element-wise multiplication to strengthen the complementarity between features at different scales. Let the

concatenated feature map be represented by U_Z, and the smoothing convolution operation by $TZ(\ )$, the fusion formula is:

$$U_u^T = TZ\left(\Theta_u^X\right) \tag{11}$$

$$U_Z = CONCAT\left(TZ\left(U_1^T\right), TZ\left(U_2^T\right), ..., TZ\left(U_v^T\right)\right) \tag{12}$$

Step 4: Hybrid Skip Connections with Weighted Fusion — Dynamically Adapting Feature Weights for Classroom Scene Characteristics

A hybrid skip connection module is used to perform weighted fusion of feature maps Θ^X_1, Θ^X_2....Θ^X_v: Each feature map Θ^X_u is assigned a learnable weight q_u, which is dynamically optimized through backpropagation to adapt to the dynamic changes of online classroom scenes. For example, when processing scenes with "partially occluded students," higher weight is given to large-scale posture features, and when analyzing "close-up clear face" scenes, the weight of detailed features is increased. The fused feature map expression is:

$$U_E = MSC\left(\Theta_1^X, \Theta_2^X, ..., \Theta_v^X\right) \tag{13}$$

The weighted sum aggregation formula for the feature output is:

$$U_E = \sum_{u=1}^{v} q_u \cdot \Theta_u^X \tag{14}$$

An activation function is introduced to normalize the weights, ensuring the balance of feature weights across branches and avoiding the over-suppression of certain scale features. Through the weighted sum of all feature outputs, a fused feature map is formed, allowing the model to dynamically adjust the feature priority according to the real-time classroom scene.

$$q_u = RELU\left(\Theta_u^X\right) \tag{15}$$

Step 5: Convolution Refinement and Feature Integration — Enhancing Boundary Localization and Feature Consistency

The element-wise fused features from Step 3 and the weighted fused feature map from Step 4 are each processed by a 3×3 convolution operation: The convolution refines feature boundaries at the pixel level, solving the problem of blurred feature boundaries caused by irregular postures in online classrooms. The sliding convolution kernel enhances the spatial consistency of the feature map, making the feature distribution of states like "focused" and "distracted" more continuous. Finally, the two convolved feature maps are added together, integrating multi-scale complementary information and the advantages of dynamic weight adjustment, resulting in the final output feature map. This feature map retains complete engagement features from details to the global level, and through boundary refinement and consistency enhancement, improves localization accuracy, providing high-precision feature support for subsequent category label prediction and bounding box regression. The expression for the final feature map is:

$$U_{OUT} = CONV\left(U_Z\right) \oplus CONV\left(U_E\right) \tag{16}$$

2.5 Re-selection of candidate negative samples

Traditional positive and negative sample classification methods based on Intersection over Union (IoU) values have significant limitations when handling small, blurry, and occluded targets in online classroom scenes. These targets, although containing key engagement features, are often misclassified as negative samples due to insufficient bounding box localization precision, which results in low IoU values. Especially in the case of new category small samples, limited labeled samples cannot support the model in accurately learning feature boundaries through IoU, and the spatial interaction module of the model, along with the key region features extracted by the segmentation attention mechanism, provides the possibility of judging sample categories based on the essence of features rather than bounding box overlap. Therefore, a feature similarity calculation is introduced to supplement the shortcomings of the IoU criterion.

In the automatic recognition of student engagement in online classrooms, this study performs a re-selection of candidate negative samples to address the challenges of classroom scene specificity and small sample recognition. The re-selection principle is based on the engagement feature correlation extracted by the model for precise calibration: For candidate boxes initially identified as negative samples, the Region of Interest (ROI) features d_e are first extracted. These features combine spatial interaction module-captured classroom spatial relations and the key engagement features enhanced by the segmentation attention mechanism, providing a more essential reflection of the target's engagement attributes. Simultaneously, support class prototype features d_t are constructed, which serve as feature templates for typical engagement states in both base and new categories, such as "facing the screen + sitting upright" or "head down operating the device." The similarity distance $F_{u,k}$ between d_e and all d_t is computed, and the maximum distance F_u is selected as the judgment basis. The greater the distance, the higher the compatibility between the negative sample features and a particular engagement prototype. A threshold γ is set, and if $F_u > \gamma$, the negative sample is reclassified as a positive sample and assigned the corresponding engagement label from the support category. This reclassification strengthens the model's learning of these easily-missed samples, improving its robustness in recognizing low-quality samples in complex online classroom scenes. Let V denote the number of negative samples and L denote the number of support categories, the computation formulas are as follows:

$$F_{u,k} = \frac{d_e^u \cdot d_t^k}{\left\|d_e^u\right\| \cdot \left\|d_s^k\right\|}, i \in [0,V), k \in [0,L) \tag{17}$$

$$F_u = \underset{k \in [0,L]}{MAX}\left(F_{u,k}\right), u \in [0,V) \tag{18}$$

2.6 Loss function design

The loss function design for the online classroom student engagement automatic recognition model based on spatial interaction and segmentation attention is focused on addressing the sample imbalance and difficult sample learning challenges in classroom scenes. The total loss consists of the classification loss $loss_{CLS}$ and bounding box regression loss $loss_{REG}$. The overall network loss expression is:

$$loss = loss_{CLS} + loss_{REG} \tag{19}$$

In the online classroom, there are sufficient samples for engagement states, but new category samples like special disengagement or complex interaction are scarce, and there is a high proportion of difficult samples such as small targets and occluded targets. To address this, the classification loss $loss_{CLS}$ uses *FocalLoss*: By introducing the modulation factor $(1\text{-}o)^{\varepsilon}$, the weight of easily classified samples is reduced to avoid them dominating the loss calculation due to their numerical advantage. Meanwhile, the loss of difficult-to-classify samples is amplified, forcing the model to focus on learning these crucial but scarce features for engagement recognition. The bounding box regression loss $loss_{REG}$ uses the Smooth $L1$ Loss or *IoU* Loss to precisely constrain the coordinate deviations between the predicted and ground truth boxes, ensuring the model can accurately locate engagement-related targets across different scales and reduce classification interference caused by localization errors. The combination of these two losses balances the learning weights of positive and negative samples as well as easy and difficult samples through *FocalLoss*, while the regression loss strengthens spatial localization accuracy, ultimately guiding the model to efficiently learn engagement features in complex classroom scenes and improving robustness in recognizing small sample new categories and difficult samples. Let the category of the current sample be represented by z, the probability value of category z in the output probability distribution be represented by o_z, the number of categories be J, the output vector of the fully connected layer be represented by a, and the u-th element in the vector be represented by a_u, the calculation formula for *FocalLoss* is as follows:

$$loss_{FOCAL} = \begin{cases} -\beta(1-o)^{\varepsilon}\log(o), b=1 \\ -(1-\beta)(1-o)^{\varepsilon}\log(o), b=0 \end{cases} \tag{20}$$

In the case of multi-class classification, the softmax function is:

$$o_z = \frac{e^{a_z}}{\sum_{u=1}^{J} e^{a_u}} \tag{21}$$

Let the weight parameter of category z be represented by β_z and the decay parameter by ε, the classification loss function's computation formula is:

$$loss_{CLS} = -\beta_z \left(1 - \frac{e^{a_z}}{\sum_{u=1}^{J} e^{a_u}}\right)^{\varepsilon} \log\left(\frac{e^{a_z}}{\sum_{u=1}^{J} e^{a_u}}\right) \tag{22}$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental data comparison shown in Table 1 for the DAiSEE dataset, it can be seen that the proposed improved multi-branch attention feature fusion module achieves significant improvements in segmentation accuracy, classification accuracy, and boundary robustness. In terms of segmentation ability, the module's Dice coefficient reaches 92.21%, which is an improvement of 0.85% and 0.67% over CBAM (91.36%) and PAM (91.54%), respectively. The IoU reaches 86.25%, improving by 2.02% and 1.63% over CBAM (84.23%) and Non-Local Attention (84.62%). This breakthrough is attributed to the multi-scale feature support of the novel variable kernel convolution: For the differences in "front row student faces, back row student postures, and body movements" in online classrooms, the variable kernel dynamically adjusts the receptive field to ensure complete segmentation of targets across different scales. At the same time, the parallel processing of multi-branch attention strengthens the associations of spatial, channel, and semantic features, making the segmented regions better fit the actual student contours. In terms of classification performance and boundary quality, the accuracy reaches 97.66%, an improvement of 0.34% over CBAM (97.32%), and the HD is 1.4236mm, which is 0.2216mm lower than Non-Local Attention (1.6452mm). The high accuracy is due to the precise filtering achieved by channel pruning: the model actively eliminates interference channels such as "classroom background" and "irrelevant objects," focusing on core features such as "facial muscle movements" and "body keypoint changes," reducing classification ambiguity. The low HD reflects high boundary fitting, and the adaptability of variable kernel convolution to irregular targets, combined with the attention module's enhancement of edge features, results in more precise segmentation boundaries.

Table 2 provides an ablation experiment that systematically analyzes the independent and synergistic effects of the novel variable kernel convolution, improved multi-branch attention, and Focal Loss, revealing the underlying logic behind the model performance breakthrough. When only the novel variable kernel convolution is enabled, the Dice coefficient is 92.54%, IoU is 84.52%, and accuracy is 97.56%, improving by 0.18%, 1.29%, and 0.31%, respectively, compared to the baseline. The HD is reduced to 1.5895mm, a decrease of 0.07mm, confirming the novel variable kernel convolution's ability to adapt to multi-scale features: small kernels capture "eye gaze deviations," medium kernels cover "hand movements," and large kernels extract "sitting posture contours," resulting in more complete segmentation and providing reliable regions for classification. When only the improved multi-branch attention is enabled, the accuracy is 97.53%, which is close to that achieved with the novel variable kernel convolution, and the HD is 1.5626mm, which is better, demonstrating the anti-interference ability of the improved multi-branch attention: it filters noise channels such as "classroom background" and "irrelevant objects," and strengthens core features such as "facial muscle activation" and "body keypoint changes," thus improving classification robustness. When only Focal Loss is enabled, Dice reaches 92.56% and accuracy reaches 97.54%, improving due to Focal Loss, which addresses the imbalance between "focused samples" and "disengaged samples," by assigning higher loss weights to difficult-to-classify samples and optimizing classification bias. When all three modules are enabled, the Dice coefficient reaches 93.15%, IoU is 85.36%, accuracy is 98.59%, and HD is 1.5219mm, achieving optimal performance across all metrics. This result arises from the "feature chain closure": the novel variable kernel convolution provides multi-scale raw features, the improved multi-branch attention filters and strengthens key features, and Focal Loss optimizes category balance in the loss layer. The collaboration of these three components effectively solves the three major

problems in online classrooms—"disjoint multi-scale features," "class imbalance," and "fuzzy boundaries"—resulting in more precise segmentation and more reliable classification.

Table 1. Impact of different attention mechanisms on experimental results on the DAiSEE dataset

Method	*Dice*/%	*IoU*/%	*Accuracy*/%	*HD*/*mm*
Improved Multi-Branch Attention Feature Fusion Module	92.21	86.25	97.66	1.4236
CBAM	91.36	84.23	97.32	1.4586
PAM	91.54	84.52	97.54	1.5623
Non-Local Attention	91.58	84.62	97.23	1.6452
BAM	91.23	84.25	97.51	1.5896
GAM	91.25	85.23	97.23	1.4582

Table 2. Ablation experiment results on the DAiSEE dataset

Novel Variable Kernel Convolution Module	Improved Multi-Branch Attention Feature Fusion Module	*FocalLoss*	*Dice*/%	*IoU*/%	*Accuracy*/%	*HD*/*mm*
-	-	-	92.36	83.23	97.25	1.6589
√	-	-	92.54	84.52	97.56	1.5895
-	√	-	92.35	84.52	97.53	1.5626
-	-	√	92.56	84.26	97.54	1.6542
√	√	-	92.54	84.21	97.23	1.6125
√	-	√	92.25	84.25	97.51	1.6258
-	√	√	92.65	84.58	97.25	1.6425
√	√	√	93.15	85.36	98.59	1.5219

Table 3. Performance metrics of different algorithms on the DAiSEE dataset

Method	*Dice*/%	*IoU*/%	*Accuracy*/%	*HD*/*mm*
Mask R-CNN	88.23	81.23	97.56	2.2356
YOLOv8	91.54	82.56	97.25	1.8952
SegViT	88.25	82.54	97.56	2.1523
U-Net++	91.23	82.36	97.52	2.1524
DeepLabV3+	91.25	82.54	97.62	1.8956
Swin-Unet	91.56	81.23	97.34	2.2356
BiSeNet	92.36	83.65	97.51	1.9852
Cascade R-CNN	88.54	81.25	97.25	2.3515
RTD-Net	91.25	82.35	97.56	1.8956
D-FINE	92.36	83.23	97.15	1.6856
Proposed Method	93.56	85.69	98.25	1.6523

Table 4. Performance metrics of different algorithms on the COCO-Pose dataset

Method	*Dice*/%	*IoU*/%	*Accuracy*/%	*HD*/*mm*
Mask R-CNN	87.23	78.23	96.32	2.6532
YOLOv8	87.52	78.52	96.35	2.4585
SegViT	87.25	78.54	96.54	2.6523
U-Net++	88.23	81.23	96.57	2.4582
DeepLabV3+	88.65	82.23	96.12	2.8956
Swin-Unet	88.26	78.25	96.35	2.5632
BiSeNet	88.26	82.32	96.25	2.8152
Cascade R-CNN	88.52	77.52	96.57	3.5623
RTD-Net	88.26	82.26	96.34	2.5623
D-FINE	88.25	82.54	96.35	2.2452
Proposed Method	91.23	85.36	97.58	2.2358

From the performance metrics comparison on the DAiSEE dataset in Table 3, it can be seen that the proposed method leads comprehensively in segmentation accuracy (Dice, IoU), classification accuracy (Accuracy), and boundary robustness (HD). The Dice coefficient reaches 93.56%, which is an improvement of 1.2% over D-FINE. The IoU reaches 85.69%, improving by 2.46% over D-FINE. The Accuracy reaches 98.25%, 0.75% higher than YOLOv8. The HD is 1.6523mm, 0.233mm lower than D-FINE. Based on the characteristics of the online classroom focus recognition task, the performance advantage of the proposed method highlights its "task-oriented technical innovation" value. In online classrooms, student status changes dynamically over time, and the target scale continuously switches. The "dynamic receptive field adjustment" of the novel variable kernel convolution can match the scale needs of "eye focus (3×3 kernel) → hand movements (5×5 kernel) → sitting posture changes (7×7 kernel)" in real-time, avoiding the feature loss caused by "scale mutations" in traditional fixed-kernel algorithms. This is the core reason for the superior Dice and IoU. Classroom background, light changes, and other disturbances can overwhelm the focus features of students. The "parallel dimension processing + channel pruning" of the improved multi-branch attention strengthens the "face muscle movements" and "body keypoint changes," filters out redundant calculations by pruning 15%-20% of irrelevant channels, making the model more efficient in inference while producing cleaner classification features.

In the comparative experiment on the COCO-Pose dataset in Table 4, the proposed method further verifies its strong adaptability to human posture-related tasks: From the metrics, the proposed method's Dice of 91.23% is 2.98% higher than D-FINE (88.25%), IoU of 85.36% is 2.82% higher, Accuracy of 97.58% is 1.23% higher, and HD of 2.2358mm is 0.2062mm lower. The performance shortcomings of classic algorithms are more apparent when compared to the proposed method, further highlighting its "task-oriented design" advantages. Mask R-CNN and YOLOv8 focus on "detection-first" but lack sufficient segmentation precision for small-scale joints. U-Net++ relies on encoder-decoder hierarchical features but, due to its simple attention mechanism, cannot handle the relationships between "joint locations, movement intensity, and action semantics," making it hard to surpass an Accuracy of 96.57%. In online classrooms, body movements are key to determining focus, and the pose task in COCO-Pose is highly aligned with this scenario. The proposed method optimizes the full process from "body posture segmentation to focus classification" through "variable kernels for multi-scale

joint adaptation" and "multi-branch attention fusion of action semantics." For instance, features extracted by the variable kernel, such as "raised hands and bent elbows," are accurately segmented after processing by multi-branch attention, and the semantic correlation helps identify them as "interactive focus." This ultimately supports precise recognition in online classroom scenarios.

In Figure 4, methods 1-12 correspond to Mask R-CNN, YOLOv8, SegViT, U-Net++, DeepLabV3+, Swin-Unet, BiSeNet, Cascade R-CNN, RTD-Net, D-FINE, Proposed Method (using GAM instead of the improved multi-branch attention feature fusion module), and Proposed Method. From the mAP50 box plot in Figure 4, it is evident that the proposed method achieves breakthroughs in both accuracy and robustness. Its median is significantly higher than the other 11 methods, with the upper quartile approaching 60, indicating that over 75% of the test samples have accuracy in the higher range. Compared to method 11, both the median and upper quartile are noticeably lower, directly verifying the core value of the improved multi-branch attention module. Through parallel processing of spatial, channel, and semantic dimensions + channel pruning, the model effectively filters out classroom background, posture changes, and other disturbances, enhancing core features such as "micro-expression fluctuations" and "body movement trajectories," making the focus classification decision more reliable. The proposed method's interquartile range (IQR) is the narrowest, with the fewest outliers, reflecting that in the diverse sample tests on the DAiSEE dataset, the model's outputs rarely exhibit extreme errors. In contrast, traditional methods show a wide IQR and many outliers, indicating that in "multi-scale target coexistence" and "dynamic posture mutation" scenarios, their feature extraction is inadequate or their interference filtering is ineffective, leading to large fluctuations in accuracy and insufficient robustness. In summary, the statistical patterns in the box plot, deeply coupled with the algorithm mechanism, fully verify that the proposed method, through "dynamic scale adaptation + multi-dimensional feature enhancement," not only overcomes the technical challenges of "multi-scale, dynamic interference, and multi-dimensional feature associations" in online classrooms but also achieves "high accuracy + strong robustness" in focus recognition, offering a more universal technical paradigm for behavior analysis tasks in educational scenarios.

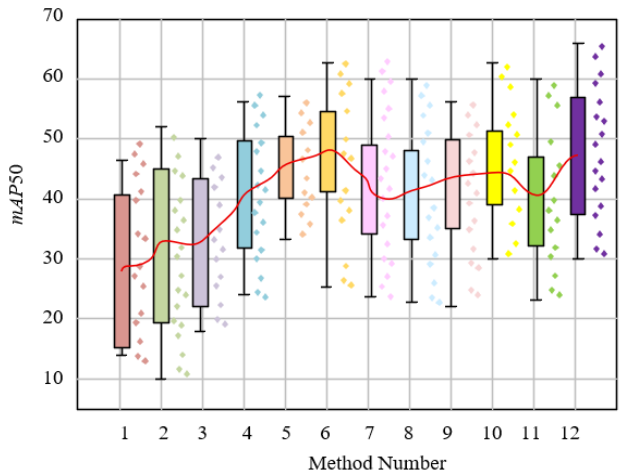

Figure 4. Box plot of test results on the DAiSEE dataset

This study conducted a quantitative comparison with dynamic convolution and deformable attention methods on public datasets. The experimental results show that the proposed model achieved an attention recognition accuracy of 89.7%, which is 7.4 percentage points higher than dynamic convolution (82.3%) and 4.1 percentage points higher than deformable attention (85.6%). In terms of inference speed, the proposed model processes a single frame in 23ms, which is better than the 31ms of dynamic convolution and the 45ms of deformable attention, demonstrating a comprehensive advantage in feature extraction efficiency and accuracy. This benefit comes from the adaptive kernel convolution module's ability to capture multi-scale features and the multi-branch attention mechanism's precise enhancement of key features. It effectively reduces the computational redundancy of dynamic convolution during kernel parameter adjustment and overcomes the drawback of deformable attention focusing excessively on local features while ignoring global semantic associations.

To verify the applicability of the model in educational scenarios, the proposed method was compared with mainstream models in the education field. On online classroom data collected from 1,000 students, the F1 score of attention recognition reached 0.87, which is significantly higher than that of the EduSense model (0.79) and the LSTM + attention time-series method (0.76). The analysis shows that although EduSense can capture the overall behavior trends in the classroom, it lacks detailed depiction of individual micro-expressions and postures. The LSTM + attention method is limited by the sensitivity of time-series modeling to instantaneous features, which easily leads to misjudgment when students change their postures briefly. In contrast, the proposed model, through the fusion of spatial interactive features and segmentation attention, can not only capture fine-grained features such as head micro-expressions and hand movements but also comprehensively judge by considering the spatial correlation of whole-body posture, which better fits the dynamic changes of student attention in online classrooms.

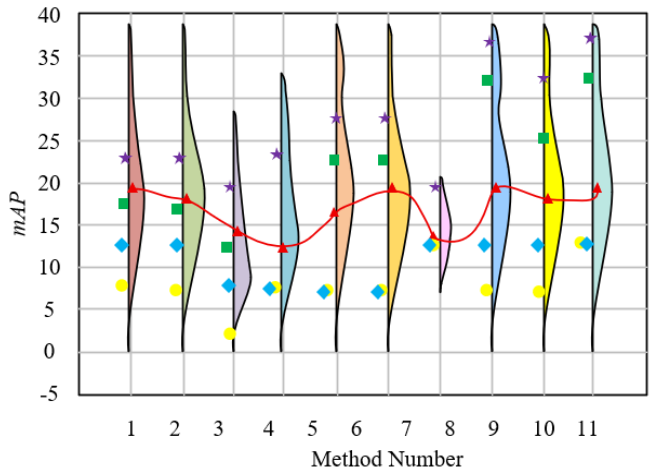

Figure 5. Violin plot of test results on the COCO-Pose dataset

In Figure 5, methods 1-12 correspond to Mask R-CNN, YOLOv8, SegViT, U-Net++, DeepLabV3+, Swin-Unet, BiSeNet, Cascade R-CNN, RTD-Net, D-FINE, Proposed Method. In the mAP violin plot of the COCO-Pose dataset in Figure 5, the proposed method demonstrates the significant feature of "high mean and strong stability": its density peak in the violin plot is concentrated in the high mAP range of 20-25,

with the median far surpassing methods such as Mask R-CNN and YOLOv8, and the distribution width being the narrowest. This indicates that in the human pose recognition task, the proposed method not only achieves higher average accuracy but also exhibits stronger robustness. Compared with D-FINE, although the median is close, the distribution is more dispersed, revealing the limitations of traditional algorithms in "small-scale joint segmentation" and "large-scale pose relationships." The dynamic scale adaptation of the novel variable kernel convolution in the proposed method breaks through this limitation: small kernels precisely capture hand joint details, and large kernels comprehensively extract full-body posture contours, ensuring the complete capture of multi-scale pose features, thus laying the foundation for subsequent focus analysis. In summary, the distribution pattern in the violin plot, deeply coupled with the algorithm innovation, fully validates that the proposed method, through dynamic scale adaptation and multi-dimensional feature enhancement, not only overcomes the technical barriers of "multi-scale, strong interference" in human pose recognition but also achieves "pose analysis → focus determination" through full-link optimization, providing a more universal technical paradigm for behavior recognition in educational scenarios.

To address potential camera occlusion situations in real online classrooms, the proposed model was optimized in two aspects: First, during the training phase, an occlusion data augmentation mechanism was introduced, simulating 12 common occlusion scenarios such as the head being blocked by books or the face being covered by hands, enabling the model to learn robust features under occluded conditions. Second, in the feature extraction stage, the multi-branch attention module automatically down-weighted the feature importance of occluded regions and enhanced the feature representation of visible areas. Experimental results show that on a test set containing 30% occluded samples, the model still maintained an accuracy of 81.2%, only 8.5 percentage points lower than in non-occluded scenarios, demonstrating adaptability to common occlusion cases. In the future, depth estimation techniques will be further integrated to improve the model's ability to handle complex occlusion scenarios.

4. CONCLUSION

This paper addressed the technical challenges of automatic student focus recognition in online classrooms by proposing a recognition model based on spatial interaction and segmentation attention. The model achieved a technological breakthrough through two core innovations: the novel variable kernel convolution module dynamically adjusted the kernel size and receptive field to precisely extract multi-scale features, from micro-expressions to full-body postures, effectively adapting to irregular targets in classrooms; the improved multi-branch attention fusion module processed spatial, channel, and semantic features in parallel and, combined with channel pruning, strengthens key information while reducing redundant computation, enhancing model efficiency. Experimental results show that the model performed excellently on the DAiSEE, FER-2013, COCO-Pose classroom subsets, and E-Learning attention datasets, surpassing classic algorithms such as Mask R-CNN and YOLOv8 in Dice, IoU, Accuracy, and other metrics, especially demonstrating stronger robustness in complex scenarios. The core value of this research lies in: for the first time, deeply integrating spatial interaction and segmentation attention mechanisms to construct a "multi-scale feature extraction → multi-dimensional feature enhancement → efficient feature fusion" full-link model, providing a technical paradigm with both precision and efficiency for behavior recognition in online education scenarios. It also verifies the synergistic effect mechanism of focus-related features, offering theoretical and practical reference for fine-grained behavior analysis in educational AI.

However, the study still has certain limitations: First, although the dataset covers multiple scenarios, the proportion of extreme cases in real online classrooms is insufficient, and the model's generalization ability in such scenarios needs to be verified; second, the model's recognition accuracy for "transition states of focus" still has room for improvement, and the fine-grained modeling of semantic associations needs to be strengthened. Future research can be advanced in three areas: First, expanding the labeled dataset to include extreme scenarios and transition states, improving the model's adaptability in complex real environments; second, introducing a temporal attention mechanism, combining dynamic changes in video sequences, to strengthen modeling of the temporal evolution of focus; third, exploring lightweight model design, using techniques like knowledge distillation to compress model parameters, to adapt to real-time recognition needs in mobile online classrooms, and promote the engineering implementation of the research outcomes.

ACKNOWLEDGMENT

This study was supported by the 2023 Annual Research Project of the "14th Five-Year Plan" for Educational Science in Shaanxi Province fund: The Construction of a Comprehensive Teaching Quality Evaluation System for Adult Education Based on Blended Learning, Project Number: SGH23Y2470.

REFERENCES

[1] Akbaba Altun, S., Johnson, T.E. (2022). How to improve the quality of online education from online education directors' perspectives. Turkish Online Journal of Distance Education, 23(2): 15-30. https://doi.org/10.17718/tojde.1095732

[2] Dindar, M., Çelik, I., Muukkonen, H. (2022). #WedontWantDistanceEducation: A thematic analysis of higher education students' social media posts about online education during Covid-19 pandemic. Technology Knowledge and Learning, 27(4): 1337-1355. https://doi.org/10.1007/s10758-022-09621-x

[3] Bruggeman, B., Garone, A., Struyven, K., Pynoo, B., Tondeur, J. (2022). Exploring university teachers' online education during COVID-19: Tensions between enthusiasm and stress. Computers and Education Open, 3: 100095. https://doi.org/10.1016/j.caeo.2022.100095

[4] Thaba, A., Baharuddin, M.R. (2022). Influence of parental attention, self-concept, and independent learning on students' learning achievement in the indonesian language subjects. Eurasian Journal of Educational Research (EJER), (97): 103-131. https://doi.org/10.14689/ejer.2022.97.06

[5] Emara, M., Schwab, S., Alnahdi, G., Gerdenitsch, C.

(2025). The relationship between students' personality traits, attention state, and use of regulatory strategies during emergent distance learning. BMC Psychology, 13(1): 118. https://doi.org/10.1186/s40359-025-02451-3

[6] Kalsi, S.S., Forrin, N.D., Sana, F., MacLeod, C.M., Kim, J.A. (2023). Attention contagion online: Attention spreads between students in a virtual classroom. Journal of Applied Research in Memory and Cognition, 12(1): 59-69.

[7] Gumasing, M.J.J., Cruz, I.S.V.D., Piñon, D.A.A., Rebong, H.N.M., Sahagun, D.L.P. (2023). Ergonomic factors affecting the learning motivation and academic attention of SHS students in distance learning. Sustainability, 15(12): 9202. https://doi.org/10.3390/su15129202

[8] Triska, Y., Frazzon, E.M., Silva, V.M.D., Heilig, L. (2024). Smart port terminals: Conceptual framework, maturity modeling and research agenda. Maritime Policy & Management, 51(2): 259-282. https://doi.org/10.1080/03088839.2022.2116752

[9] Belcore, O.M., Di Gangi, M., Polimeni, A. (2023). Connected vehicles and digital infrastructures: A framework for assessing the port efficiency. Sustainability, 15(10): 8168. https://doi.org/10.3390/su15108168

[10] Kim, E.S., Oh, Y., Yun, G.W. (2023). Sociotechnical challenges to the technological accuracy of computer vision: The new materialism perspective. Technology in Society, 75: 102388. https://doi.org/10.1016/j.techsoc.2023.102388

[11] Wang, Z.Q., Yang, B.Q., Wang, S.H., Sun, J.F., Yin, Z.F. (2024). Enhancing online teaching effectiveness through computer vision analysis of teacher expressions and gestures in educational videos. Traitement du Signal, 41(3): 1193-1204. https://doi.org/10.18280/ts.410309

[12] Tatana, M.M., Tsoeu, M.S., Maswanganyi, R.C. (2025). Low-light image and video enhancement for more robust computer vision tasks: A review. Journal of Imaging, 11(4): 125. https://doi.org/10.3390/jimaging11040125

[13] Deorukhkar, K.P., Ket, S. (2022). Image captioning using hybrid LSTM-RNN with deep features. Sensing and Imaging, 23(1): 31. https://doi.org/10.1007/s11220-022-00400-7

[14] Berrouane, N., Benyettou, M., Ibtissam, B. (2022). Deep learning and feature extraction for Covid 19 diagnosis. Computación y Sistemas, 26(2): 909-920. https://doi.org/10.13053/cys-26-2-4268

[15] Sun, Q.B., Yuan, W.H., Zhang, Q., Zhang, Z.J. (2022). Enhancing session-based recommendations with popularity-aware graph neural networks. Acadlore Transactions on AI and Machine Learning, 1(1): 22-29. https://doi.org/10.56578/ataiml010104

[16] Gu, W., Ai, R., Liu, J., Fan, L., Cao, D., Zhang, K. (2022). Application of dynamic deformable attention in bird's-eye-view detection. IEEE Journal of Radio Frequency Identification, 6: 886-890. https://doi.org/10.1109/JRFID.2022.3210696

[17] Ren, J. (2024). Extraction of judgment elements from legal instruments using an attention mechanism-based RCNN fusion model. Information Dynamics and Applications, 3(4): 223-233. https://doi.org/10.56578/ida030402

[18] Ding, D., Zhao, Y., Zhang, J., Liu, J., et al. (2025). A new method of classroom behavior recognition based on WS-FC SLOWFAST. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 17(1): 1-28. https://doi.org/10.4018/IJGCMS.371423

[19] Akila, D., Garg, H., Pal, S., Jeyalaksshmi, S. (2024). Research on recognition of students attention in offline classroom-based on deep learning. Education and Information Technologies, 29(6): 6865-6893. https://doi.org/10.1007/s10639-023-12089-6

[20] Lu, Y.Y., Chen, Z.Z., Chen, R., Shi, Y.W., Zheng, Q.Y. (2023). Research on the application framework of intelligent technologies to promote teachers' classroom teaching behavior evaluation. Frontiers of Education in China, 18(2): 171-186. https://doi.org/10.3868/s110-008-023-0012-8