SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 1.1, (2026), pp. 1724-1742
Open Access. Online & Print

DOI: 10.5281/zenodo.19053170

STUDENT CLASSROOM BEHAVIOR DETECTION METHOD USING TERNARY ATTENTION

Jiezhang Min[1], Haihong Li[2*], Renyuan Cui[3], Tiantian Li[4], Zhaowei Huang[5]

[1]Institute for History and Culture of Science and Technology, Guangxi Minzu University, Nanning 530006, China, 202592142210067@stu.gxmzu.edu.cn, https://orcid.org/0009-0005-9253-9787
[2]Faculty of Arts and Social Sciences, Universiti Malaya, 50603 Kuala Lumpur, Malaysia, s2191486@siswa.um.edu.my, https://orcid.org/0009-0008-0480-298X
[3]School of Public Administration, Hengshui University, Hengshui 053000, China, cuirenyuan@hsnc.edu.cn, https://orcid.org/0009-0008-5479-1895
[4]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China, litiantian0809@outlook.com, https://orcid.org/0009-0006-5006-1946
[5]Faculty of Arts and Social Sciences, Universiti Malaya, 50603 Kuala Lumpur, Malaysia, 23104374@siswa.um.edu.my, https://orcid.org/0009-0006-8502-888X

Received: 20/10/2025
Accepted: 22/11/2025

Corresponding Author: Haihong Li (s2191486@siswa.um.edu.my)

ABSTRACT

Target detection of students' classroom behavior recognition scenarios, followed by systematic analysis and evaluation, significantly contributes to enhancing instructional quality and fostering students' healthy development. Existing target detection for classroom behavior recognition scenarios suffers from large inconsistencies between the classification task and the localization task and the lack of a ternary attention mechanism, which makes it difficult for teachers to accurately assess the interest and attention levels of students, thus affecting the quality and performance of teaching. In order to solve these problems, the study have made three innovations, including quality focal loss, ternary attention mechanism module, data enhancement operation and related training. The target detection method proposed in this paper is compared with seven state-of-the-art target detection algorithms in a comparative experiment on the same dataset. Our results show that all models exhibit good performance, reaching 0.99476 for the map_0.5 metric, 0.96298 for the map_0.5:0.95 metric, and 0.99658 for the recall metric, which outperforms the other seven methods in all three metrics. In addition to this, the study validated the innovation points and verified the ablation experiment of the study, which can be better applied to the scenarios of intensive targeting of students' classroom behavioral actions, high variance of the targeting actions, and small amount of data.

KEYWORDS: Target Detection, Classroom Behavior Recognition, Ternary Attention, Ablation Experiment.

1. INTRODUCTION

Artificial intelligence technology is spearheading a paradigm shift in educational monitoring, with intelligent algorithm-based classroom behaviour recognition emerging as a core direction in educational informatisation [1]. Such systems possess dual capabilities: real-time adjustment of teaching strategies and quantitative assessment of students' cognitive/psychological health states [2]. Research indicates that teachers' capacity for quantitative analysis of student engagement characteristics constitutes a key indicator of teaching efficacy, particularly within blended learning contexts [3]. Deep neural networks (DNNs), serving as the core technology, process complex behavioral classifications through multi-layered feature abstraction. Their data-driven models automatically construct features, significantly reducing reliance on manual design [4]. Compared to traditional methods, deep learning demonstrates greater robustness in cross-modal tasks and multi-objective classroom behaviour recognition, owing to its adaptive feature optimization and big data processing capabilities [5]. Integrated intelligent monitoring platforms combining facial expressions, posture, and eye-tracking have been deployed in education to support refined teaching evaluations [6].

Significant advances have been made in classroom behaviour recognition research: Hutt et al. [7] utilized commercial eye-tracking devices to monitor attention shifts during training; Lee et al. [8] proposed the PFA-DNNM model for real-time assessment of learning engagement via facial sequences (offline scenarios); Pabba et al. [9] developed a system to recognize six classroom states including frustration and drowsiness; Bhardwaj et al. [10] dynamically generated engagement metrics (MES) by integrating facial recognition with survey data; Trabelsi et al. [11] constructed a system tracking emotion, engagement, and focus to assist teachers in optimising strategies; Liao et al. [12] fused multidimensional data and employed Bayesian networks to infer student abnormal behaviour.

Nevertheless, technical limitations persist: Ngoc Anh et al. [13]'s attention system struggles to recognize critical cognitive state signals like micro-expressions (mouth twitching) and complex postures (chin-resting); Liu et al. [14]'s multi-view perception model performs poorly in dense small-object recognition due to low resolution and sparse information; Although Li Y et al.'s [15] dataset covers seven interaction categories, its relational feature fusion lacks ternary attention, limiting accuracy; Dey A et al.'s [16] SOAId dataset and AdaptSepCX model exhibit inconsistencies in target detection classification and localization. Furthermore, issues such as high manual intervention and poor environmental adaptability (e.g., performance degradation under lighting variations or occlusions) commonly constrain system robustness and practicality.

To overcome these limitations, research must: construct multimodal fusion frameworks integrating eye movements, micro-expressions, and gestures; optimize small object detection to enhance low-resolution feature extraction; and establish standardized evaluation systems to improve model cross-scenario generalization. Future work may explore novel architectures based on self-supervised learning to reduce annotation dependency. The main contributions of this paper are: 1) Proposing Quality Focus Loss (QFL), which jointly optimizes object localization quality (IoU) and classification loss through task collaboration to improve accuracy;

2) Designing a ternary attention mechanism to enhance feature extraction for dense targets and temporal actions, significantly reducing missed detection;

3) Introducing S-augment, a novel data augmentation method tailored for long sequences of student behaviors;

4) Validating the superiority and module effectiveness of our approach through comparisons with state-of-the-art algorithms and ablation studies on identical datasets, particularly in scenarios involving dense targets, high class-disparity, and few samples.

The overall structure of this paper is organized as follows: section 1 describes the research significance of target detection in classroom behavior recognition, existing problems, and the innovation points of this paper; section 2 introduces the technological advances and limitations in the related fields; section 3 describes in detail the proposed framework of target detection, including the design principles of the S-augmentation data augmentation strategy, the ternary attention mechanism, and the QFL loss function; section 4 validates the effectiveness of the method through comparative experiments and ablation analysis to verify the effectiveness of the method; Section 5 summarizes the research conclusions and looks forward to future directions.

2. RELATED WORK

Learning experiences are essentially dynamic and interactive processes, in which students' physical movements and eye gaze directions as well as emotional states are constantly changing throughout

the learning process. As students interact with the teacher in the classroom, the target is easily confronted with the possibility of object obstruction, and it is particularly important to analyze the students' behavior and state by capturing their behavior, so as to visualize the students' daily behavior and classroom state, and then carry out targeted educational management and guidance. To improve the reliability of behavior recognition in complex scenarios, it is necessary to address the inconsistency between the classification and localization tasks in the target detection of a scene, to improve the structure of the target detection network, and to enhance the optimization of the data quality. There are some studies on behavior recognition of students in learning environments and various classification algorithms and evaluation metrics.

2.1. Target Detection Network Structures

Most studies focus on a specific behavior of students and conduct research accordingly. Lin *et al.* [17] proposed a method based on posture estimation and personnel detection to collect skeleton data, and then used the collected data to reduce incorrect connections. The method removes the connections with the smallest weight, thereby reducing the number of false connections. The final network model uses these features to achieve more accurate behavior recognition. Dianqing *et al.* [18] explored student behavior recognition using YOLOv5s, a deep learning model, and proposed a method that combines feature extraction and image recognition with Ghost-4D-YOLOv5s to improve the accuracy of student behavior recognition. S.Q *et al.* [19] developed CBAM-YOLOv5, an automatic student behavior detection and recognition system based on surveillance video, which effectively suppresses background interference and improves the robustness of behavior feature extraction by introducing the Convolutional Block Attention Module (CBAM). It can realize the classification statistics of students' classroom behaviors and generate teaching quality analysis reports in real time.

Zhao *et al.* [20] proposed a real-time target detection network BiTNet for the problem of object occlusion in classroom scenarios.The system provides data support for teaching quality assessment through real-time monitoring of students' behaviors, with a focus on optimizing the performance of occluded targets and small targets detection. The network architecture combines an efficient transformer block (ETB) and an efficient convolutional aggregation block (ECAB), in which the ETB adopts the convolutional multi-head self-attention (CMHSA) mechanism to improve the accuracy of occlusion target recognition by capturing contextual information. T.S *et al.* [21] proposed a new convolutional neural network (CNN)architecture for analyzing classroom environments. This architecture consists of two models: the first model (CNN-1) focuses on individual student behavior recognition, and the second model (CNN-2) focuses on the overall classroom environment. Through comprehensive analysis, the two models provide detailed feedback on classroom behavior, thereby improving the accuracy of student behavior recognition. The aforementioned research, based on deep learning-driven network architecture optimization—including graph networks, lightweight design, attention mechanism integration, multi-branch collaborative analysis, and context modelling—has achieved significant results in identifying student attention states and distractions. However, improving accuracy relies on high-quality annotated data, and the computational burden of real-time classroom processing remains a widespread challenge. More critically, the effectiveness of these structural optimization strategies is frequently constrained by the original quality of classroom images and complex background interference, necessitating complementary data quality enhancement methods to provide robust support.

2.2. Recognizing Scenario Data Augmentation Methods

Alruwais *et al.* [22] defined student engagement as a multidimensional structure covering behavioral, emotional, and cognitive dimensions. To assist teachers in accurately assessing the level of classroom engagement, the study needs to screen for optimal prediction algorithms. Based on the data collected from the virtual learning environment (VLE), the study implemented a preprocessing process including missing value removal, data normalization, feature coding and outlier detection. The preprocessed data were further modeled using multiple sets of machine learning classifiers, and a classification accuracy of 94.64% was finally obtained through cross-validation combined with multidimensional assessment metrics. The experiment proves that the systematic data analysis method can significantly optimize the prediction efficacy of classroom participation, but the model generalization ability still needs to be further verified due to the limitation of the available data size. Taoufik *et al.* [23] proposed a new deep learning-

based method for student behavior recognition, which first trains a model on the facial expression dataset and then uses the model to recognize student behaviors. This method can effectively improve the accuracy of student behavior recognition. Currently, several automatic behavior recognition methods have been proposed, and student dynamics can be monitored in a more real-time manner through video. According to M.M *et al.* [24], in most cases, teachers are unable to know the behavior and interests of each student, especially in large classes, and it is difficult for teachers to provide targeted instruction. As a result, a real-time intelligent classroom system has been proposed that automatically monitors students' attention and mood and provides feedback to the teacher. The system uses artificial intelligence technology with a large database, which can accurately collect data on students' attention and emotions.

Problems such as lack of accurate datasets, dense targets, and inter-class occlusion are common. Arnab *et al.* [25] focused on the detection and recognition of student movements in images captured by webcams. Different from mature video processing methods that rely on time series information for student behavior analysis, identifying students' actions from a single frame image increases the complexity of the problem. In order to solve this, a novel deep learning model named AdaptSepCX Attention is proposed, which is specially designed for the behavior recognition of students in online learning environments. Similarly, existing models often struggle to achieve good accuracy due to lack of datasets, dense targets, inter-class occlusion, and other issues. W. Qin *et al.* [26] proposed a classroom behavior detection algorithm and constructed a dataset-specific dataset (DSL-Dataset) simultaneously. A loss function called VAIoU is proposed, which can be combined with vector ratios to filter the prediction box. The integration of computer vision and transfer learning technologies to analyse multimodal data—including facial expressions, attentional cues, and movements—has become the mainstream approach for deciphering student behavioral patterns, significantly enhancing the ability to characterize complex behaviors. Concurrently, AI-driven anomaly detection technologies demonstrate mature potential. These recognition paradigms rely heavily on two core technological pillars: efficient feature learning to deeply comprehend raw data, and meticulously designed data augmentation strategies to bolster the robustness of model representations. However, regardless of technological innovation in recognition systems, their ultimate educational value hinges upon the system's capacity to generate precise, robust, and field-deployable high-quality quantitative assessments of student behaviour. Such reliable, actionable behavioral metrics form the foundational guarantee and value for implementing meaningful quantitative analysis and evaluation of classroom behaviour, thereby underpinning targeted educational interventions.

2.3. Quantitatively Analyze and Evaluate Student Classroom Behavior

Thomas *et al.* [27] argue that student engagement is the key to successful learning in the classroom, as measuring or analyzing student engagement is important for improving learning and teaching. Computer vision technology is used to analyze students' levels of engagement or attention from their facial expressions, head postures and eye gazes, and machine learning algorithms are employed to make decisions. NgocAnh *et al.* [28] suggest that classroom automatic learning analysis is becoming an important topic in the field of education, which requires effective systems to monitor the learning process and provide feedback to teachers. Recent advances in vision sensors and computer vision methods are able to automatically monitor the behavioral and emotional state of learners at all levels, from college to preschool. Trabelsi *et al.* [29] believes that teachers' ability to analyze and evaluate students' classroom behavior is becoming a key criterion for quality teaching. By analyzing the needs of effective classroom teaching to monitor students' participation and interaction in the classroom and identify cues that simulate their attention, AI-based behavior recognition technology can help assess students' attention and engagement in the classroom. Using machine learning methods to train students' behavior recognition models, including recognizing facial expressions, and using modern technology to introduce a smart classroom based on real-time vision, students can monitor their mood, attendance, and attention levels even when they are wearing masks. Y.Qin *et al.* [30] provided a comprehensive view of the classroom situation by identifying and analyzing student behavior, which in turn provides a new perspective for classroom assessment and the improvement of teaching methods. The researchers proposed an improved algorithm based on YOLOv8, YOLOv8-AFPN, to achieve real-time detection of students' classroom behavior. This method aims to solve the problems of time-consuming, laborious, prone to missed and false detection, and insufficient real-time performance in the traditional student behavior recognition process.

N. Krishnnan *et al.* [31] proposed a new algorithmic framework that recognizes key frames in the video and then detects the attention level of a particular student when the instructor is lecturing and the real-time video of the classroom serves as the input. Structural Similarity Index Method (SSIM) was investigated to recognize key frames in the video, then drowsiness was detected to infer whether the students were sleepy or not. Facial expressions were scrutinized to perceive students' mental states in the classroom. Finally, gaze detection was performed to check whether the students' attention was on the board. This evaluation method is novel and allows for more research on the correlation between the data. Ahmed Raza *et al.* [32] proposed an object detection method based on a deep learning framework called BirdView Retina-Net (BV-RNet), which is capable of detecting small scale objects efficiently from a bird's eye view point. Alairaji R M *et al.* [33] propose a method that will analyze and identify student activity from video recorded by surveillance cameras during the exam. This work requires three main techniques: head motion detection, iris motion detection, and hand motion detection to identify contact between the same student's hand and face, as well as contact between different students. Automatic detection of anomalous behavior will help reduce the error rate caused by manual monitoring. These studies show that multimodal analysis, real-time, adaptability, and improved algorithms are the focus of current research. By combining multiple data sources such as facial expressions, head postures, eye gaze, etc., student engagement and attention levels can be assessed more holistically. In addition, modern technology enables real-time monitoring and feedback to help teachers adjust their teaching strategies in a timely manner. Future research directions should focus on building larger and more diverse datasets, developing more secure and privacy-friendly data processing methods, optimizing algorithms to reduce computing resource requirements, and bringing more innovation and progress to the field of education.

3. PROCESSED METHOD

In this section, a target detection network architecture for student classroom behavior recognition scenarios, specifically including data augmentation operations, ternary attention mechanism module, quality focal loss, and related training details will be presented in detail. As shown in Figure 3-1, considering the high requirements for detection speed in student classroom behavior recognition scenarios, our network structure uses numerous components in the one-stage framework, and based on this, the study consider the network's applicability in dense target scenarios and scenarios with large interspecies feature differences, and propose a ternary attention mechanism module for feature extraction and a quality focal loss. Meanwhile, in order to solve the problem of small data samples, the data enhancement strategy of S-augment is proposed. Overall, the network structure consists of three parts: Backbone, Neck, and Head. Backbone adopts the Focus + CSP structure, in which the ternary attention mechanism module designed for the large interspecies differences is embedded, so that the network is robust in extracting the features of the different actions of the students' classroom behaviors, and avoiding in the phenomenon of misdetection and omission of detection. Meanwhile, in order to further strengthen the inference effect, the quality focal loss, which is mutually guided by regression and classification, is designed. It is worth mentioning that this network structure will be used in both the subsequent training and testing process.

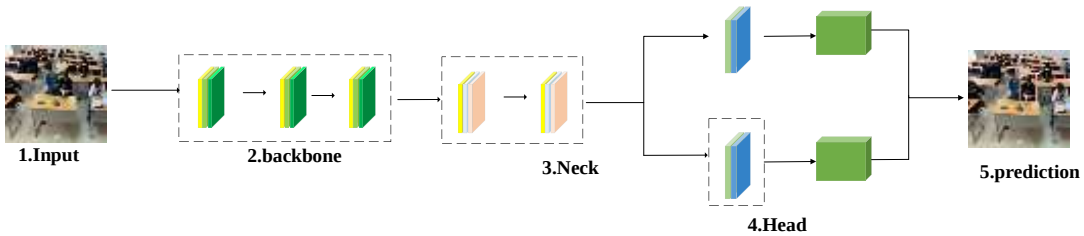

Figure 1: Overall Network Architecture. A Target Detector Applied to Student Classroom Behavior Recognition.

3.1. Quality Focal Loss

In the recognition of student's classroom behaviors, the number of target actions to be localized is much larger than the maximum number of action categories contained in the scene due to the denser student classroom behavioral actions and fewer action categories in the same scene. At the same time, there is a very large inconsistency problem between the classification task and the

localization task in current target detection methods. These two reasons cause the problem that applying the current target detection task directly to student classroom behavior recognition scenarios has more problems of inaccurate localization when the classification is correct, or localization is more accurate but the action categories are incorrectly identified.

To address the aforementioned inconsistencies, this paper proposes Quality Focus Loss (QFL), which integrates the object's localization quality—typically represented by the quantified Intersection over Union (IoU) metric measuring the overlap between predicted and ground-truth bounding boxes—into the classification loss function. Specifically, it constructs a joint representation where a single prediction value simultaneously embodies both the object's classification confidence and its estimated localization quality (IoU). This design ensures the model maintains consistency between classification and localization during both learning and prediction phases, embedding information about localization reliability directly within the classification score itself. Consequently, overall detection accuracy is enhanced.

The contribution of quality focal loss (QFL) to student classroom behavior recognition scenarios can be divided into the following main points:

1. Joint representation: the QFL combines localization quality (e.g., IoU scores) and classification scores from students' classroom behavior recognition scenarios into a joint table, which maintains consistency in the training and reasoning process and helps to solve the problem of inconsistent use of quality estimates and classification scores in the training and testing phases.Through unified representation, QFL aligns the training objective (predicting high classification scores alongside high localization quality) with the inference requirement (where high scores during screening denote good prediction quality).
2. Continuous Label Support: While traditional focal loss only supports discrete {0, 1} labels, QFL extends this concept to support continuous labels (e.g., IoU scores) with floating point numbers from 0 to 1, which better reflects the performance of actual data in student classroom behavior recognition scenarios.
3. Dynamically Adjusting Difficulty: QFL improves the overall performance of the model by dynamically adjusting the loss function so that the model pays more attention to the samples that are difficult to classify or localize in the students' classroom behavior recognition scenarios during the training process.This enables the model to focus more effectively on and process challenging instances within the specific recognition scenario of the classroom during the learning process, thereby optimizing overall performance.

The joint representation integrates the classification score and localization quality (e.g., IoU score) into a single prediction vector. This representation addresses the inconsistency in traditional target detection methods where quality estimation in the training and inference phases is used separately from classification scores. Specifically, it allows the model to estimate the localization quality of each detection frame while predicting the classification, thus providing more accurate ranking scores and improved detection performance in non-maximal suppression (NMS) processing.

A comparison between existing methods and our proposed method in terms of classification and localization quality estimation is shown in Figure 3-1-1:

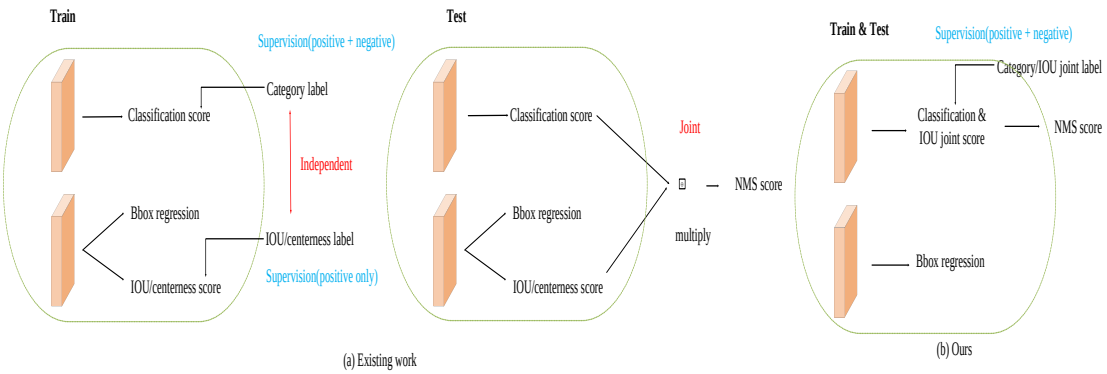

Figure 2: Comparison Of Joint Representation and Existing Methods.

In Figure 3-1-1(a), the existing work, the classification score, bounding box regression, and IoU/centerness score are handled independently in the model training and testing phases, which leads to inconsistency between training and inference. In contrast, in Fig. (b), our approach combines the classification score and the IoU score into a joint representation, i.e., a joint classification and IoU score that is used in both training and testing. This joint representation improves the consistency between training and inference, and is closely related to the "joint representation" mentioned in the rationale of quality focal loss. In quality focal loss, in this way, our model is able to take into account the localization quality of each sample during the training process when identifying students' classroom behaviors, allowing the loss function to focus more on samples that are difficult to localize or classify.

Continuous label support means that in quality focal loss (QFL), the output labels of the classification are no longer the traditional 0 or 1, but can take any continuous value between 0 and 1. These continuous values represent the quality of the target localization, usually the intersection and concurrency ratio (IoU) with the true bounding box. In this way, QFL can integrate the quality of localization directly in the loss function, allowing the loss function to apply greater weight to samples that are inaccurately localized, thus motivating the model to learn to predict the bounding box more accurately.

Figure 3-1-2 compares the differences between traditional target detection methods and the proposed generalized focus loss (GFL) method:

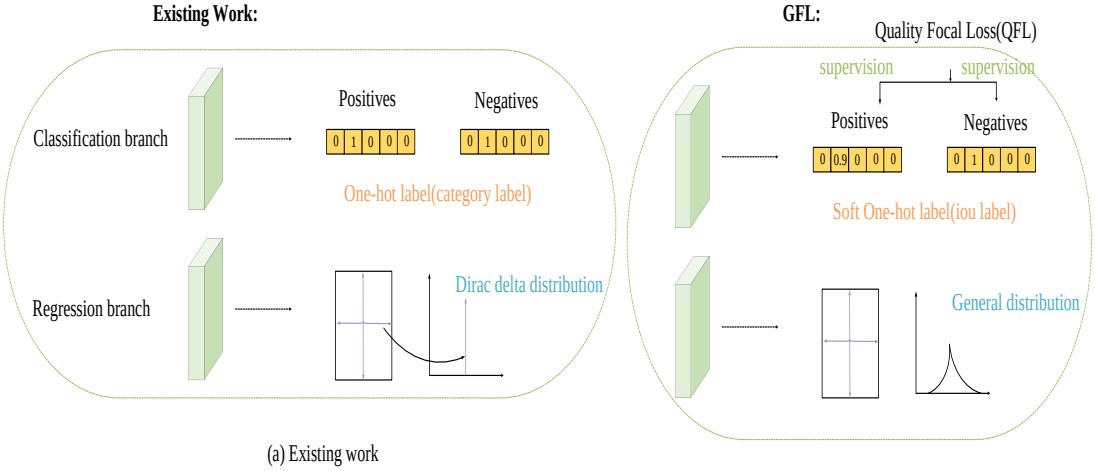

Figure 3: Comparison Of Continuous Label Support and Existing Methods.

In the traditional method, the classification branch uses one-hot label to distinguish positive and negative classes, while the regression branch employs Dirac delta distribution to predict the boundary box. In contrast, the GFL method introduces quality focal loss (QFL) and distribution focal loss (DFL). QFL is learned through soft one-hot labels (IoU labels) that reflect the localization quality of the bounding box. Meanwhile, DFL uses a general distribution to model the probability distribution of the bounding box location.

$$QFL(\sigma) = -|y-\sigma|^{\beta}((1-y)\log(1-\sigma) + ylog(\sigma)) \quad (3\text{-}1)$$

Where σ=y is the global optimal solution of QFL

$$DFL(S_i, S_{i+1}) = -((y_{i+1}-y)\log(S_i) + (y-y_i)\log(S_{i+1})) \quad (3\text{-}2)$$

$$\text{Therefore, } GFL = \frac{1}{N_{pos}}\sum_Z L_Q + \frac{1}{N_{pos}}\sum_Z 1_{\{C^*z>0\}}(\lambda_0 L_\beta + \lambda_1 L_D) \quad (3\text{-}3)$$

Dynamic adjustment of difficulty is a feature of quality focal loss (QFL), which allows the model to focus more on samples that are difficult to classify or localize during training. This is achieved by adjusting the β parameter in the loss function, which increases the loss value for samples with high uncertainty in the model's prediction, making the model more focused on these difficult samples. This method aims to improve the sensitivity of the model to difficult samples and help the model to classify and localize more accurately, especially when faced with complex scenarios such as student classroom behavior recognition with multiple targets and categories.

3.2. Ternary Attention Mechanism Module

Due to the large differences in the location and size of each action type in the student classroom behavior recognition scenario, which causes difficulty in feature extraction, current target detection algorithms are prone to frequent missed

detections when they are applied to this scenario. In order to solve this problem, this paper proposes a ternary attention mechanism module to enhance the feature extraction capability of the overall network algorithm. The overall network structure of the designed ternary attention mechanism module is shown in Figure 3-2-1.

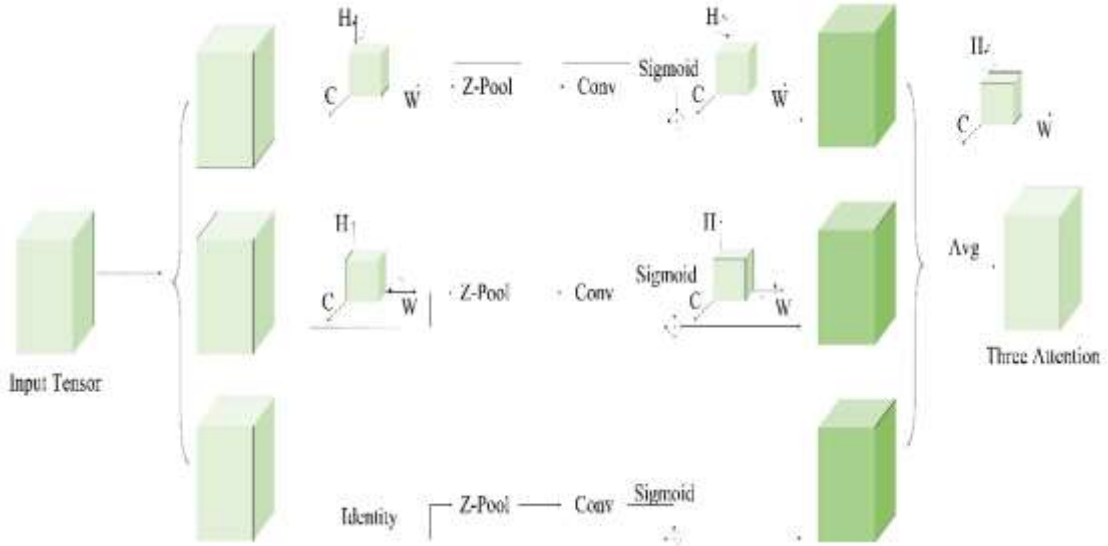

Figure 4: Ternary Attention Mechanism Network Structure Diagram.

In the student classroom behavior recognition scenario, the ternary attention mechanism is designed to efficiently handle cross-dimensional feature information interactions. It consists of three branches, each of which is responsible for capturing cross-dimensional feature interactions between spatial and channel dimensions in the input. Specifically, for an input tensor $x \in R^{C\times H\times W}$, the mechanism first passes the input to each branch for manipulation. Each branch is responsible for aggregating the interaction features between specific dimensions and channel dimensions in the input.

The first branch is responsible for processing the interaction features between the spatial dimension W and the channel dimension C in the input. It obtains interaction features across spatial dimensions by applying maximum pooling and average pooling operations on the spatial dimensions and then spreading the results and connecting them along the channel dimensions. The second branch is responsible for processing the interaction features between the spatial dimension H and the channel dimension C in the input. It first performs a global average pooling operation on the input, and then uses a 1×1 convolution kernel to spread the results and join them along the channel dimensions to obtain the interaction features across the spatial dimensions. A third branch similar to CBAM is used to build spatial attention. The outputs of all three branches are then finally pooled using simple averaging.

The traditional approach to computing channel attention consists of computing singular weights, typically scalars for each channel in the input tensor, and then uniformly scaling these feature mappings using the singular weights. While this process of computing channel attention has been shown to be very lightweight and highly successful, there is an important missing piece when considering this approach. Typically, in order to compute the singular weights of the channels, the input tensor is spatially decomposed into one pixel per channel by performing global average pooling. This leads to a significant loss of spatial information, and therefore there is a missing interdependence between the channel dimension and the spatial dimension when computing the attention of these single-pixel channels. CBAM introduces spatial attention as a complementary module to channel attention. That is, spatial attention tells us "Which position in the channel to focus on", while channel attention tells us "Which channel to focus on". However, the disadvantage of this approach is that channel attention and spatial attention are separate and computed independently of each other. Therefore, any relationship between the two is not considered. Motivated by the way of establishing spatial attention, we propose the concept of cross-dimensional interaction, which addresses this drawback by capturing the

interaction between the spatial and channel dimensions of the input tensor. We introduce cross-dimensional interactions in ternary attention by using each of the three branches to capture the dependencies between the (C, H), (C, W) and (H, W) dimensions of the input tensor.

As shown in Figure 3-2-1, the ternary attention mechanism proposed in this paper is an attention module with three branches that inputs a tensor and then outputs a tuned tensor of the same shape. The flow of the mechanism is that given an input tensor $x \in R^{C\times H\times W}$, we first pass it to each of the three branches in the proposed triple concern module. In the first branch, we construct the interaction between the H dimension and the C dimension.

For this purpose, the input x is rotated 90◦ counterclockwise along the H-axis. This rotation tensor can be expressed as $\widehat{x_1}$ which has the shape (W × H × C).

$\widehat{x_1}$ then reduces the shape (2×H×C) of $\widehat{x_1^*}$ by using Z-pool. $\widehat{x_1^*}$ will be passed through a standard convolutional kernel layer of size k×k, followed closely by a batch normalization layer which provides the intermediate output of the dimensions (1×H×C) generating synthetic attentional weights which are then passed through a sigmoid activation layer.

The generated attentional weights are subsequently applied to $\widehat{x_1}$, which is then rotated clockwise by 90° along the axis of H to preserve the original input shape of x.

Similarly, in the second branch, we rotate χ by 90◦ counterclockwise along the W axis. The rotated tensor $\widehat{x_2}$ can be expressed as (H × C × W) and passes through the Z-pool.Thus, the tensor is processed as $\widehat{x_2^*}$ and has the shape (2 × C × W). Subsequently, $\widehat{x_2^*}$ is passed through a standard convolutional layer defined by a kernel size k×k, followed by a batch normalization layer, which outputs a tensor of shape (1×C×W).

This tensor is then passed through a sigmoid activation layer, which is then simply applied to $\widehat{x_2}$, and the output is then rotated clockwise by 90◦ along the W axis to maintain the same shape as the input x.

For the third branch, it is responsible for processing the interaction features between the channel dimension C and the spatial dimensions H and W in the input. The channels of the input tensor x are reduced to two by a Z-pool pool to obtain $\widehat{x_3}$ with shape (2 × H × W), which is then passed through a standard convolutional layer defined by kernel size k, followed by a batch normalization layer. The output is passed through a sigmoid activation layer to generate attentional weights of shape (1 × H × W), which are then applied to the input x.

The tuned tensor produced by each of the three branches is then weighted using simple averaging. This mechanism effectively captures the interactive features between the different dimensions in the input, leading to a better understanding of the image content.

In summary, after ternary attention calculation from the input tensor $x \in R^{C\times H\times W}$, the output tensor can be expressed as:

$$y = \frac{1}{3}\left(\overline{\widehat{x_1}\sigma(\varphi_1(\widehat{x_1^*}))} + \overline{\widehat{x_2}\sigma(\varphi_2(\widehat{x_2^*})} + x\sigma\left(\varphi_3(\widehat{x_3})\right)\right) \quad (3\text{-}4)$$

Where σ denotes the sigmoid activation function layer, and in the three branches of the ternary attention, φ_1, φ_2 and φ_3 denote the standard two-dimensional convolutional layers defined by the kernel size k.

3.3. S-Augment

Due to the low amount of data in the current open-source data for student classroom behavior recognition, it is necessary to take proper data enhancement before network training. However, the image types in the student classroom behavior recognition scenes are quite different from public datasets such as coco.

This difference is mainly manifested in the presence of a large temporal sequence of student behavioral actions, which are very close to each other between consecutive frames.

If the common data augmentation methods for current target detection, such as mixup and mosaic, are directly applied, they will destroy this temporal relationship. Therefore, this paper proposes a new data augmentation method S-augment for the student classroom behavior recognition scene.

The principle and framework implementation of S-augment are shown in Figure 3-3-1. In S-augment, the study uses two augmentation spaces Ap and As, which contain six and eight data augmentation operations, respectively, based on the characteristics of student classroom behavior recognition scenarios.

Thus, 14 data augmentation operations are obtained. A novel operation sampling strategy is also used to better adapt to the student classroom behavior recognition scenario.

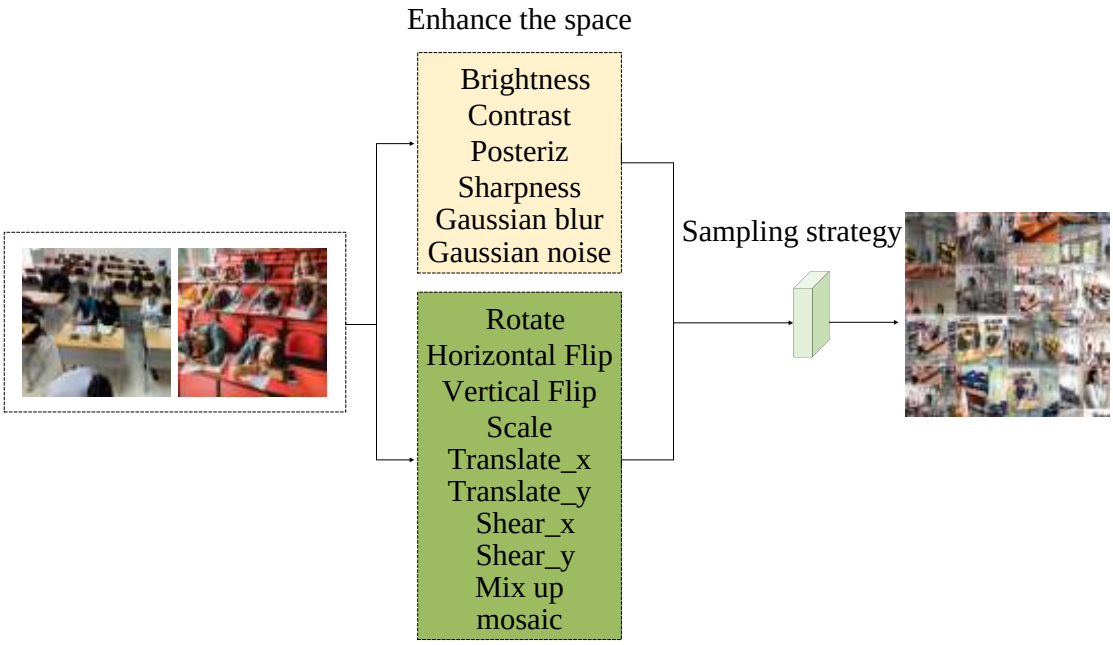

Figure 5: S-Augment Framework.

3.3.1. Enhancement Spaces

In order to adapt to the characteristics of student classroom behavior recognition scenarios, this paper provides a specific setup for specific data augmentation methods. First, common data augmentation operations are started with, and then filtered according to the characteristics of the student classroom behavior recognition domain to exclude operations that are not suitable for student classroom behavior images, such as inversion and equalization, which may destroy the details and features in the student classroom behavior images. Next, the study categorized the data augmentation operations into pixel-level and spatial-level operations and constructed two enhancement spaces, namely pixel enhancement space Ap and spatial enhancement space As. Ap and as include data enhancement operations related to pixels and space, respectively.

3.3.2. Sampling Strategy

Since images are very sensitive to attributes such as inter-frame relations, and The study observe that successive operations in Ap may result in indistinguishable kinds of actions in the output image, in this paper we use a novel operation sampling strategy for sampling operations from Ap and As. Specifically, The study randomly sample T data augmentation operations from each branch, where the number of operations sampled from Ap is no more than one. The study decides the range of values for T after tradeoffs. For successive data augmentation operations, the study needs to carefully consider the number of successive operations. Using more consecutive operations may further improve the generalization ability of the model, but too many consecutive operations may generate images that are far from the origina.

4. EXPERIMENT

This study's data collection strictly adheres to academic ethical standards and relevant laws and regulations. All experimental data originates from the Student Classroom Behaviour dataset hosted on the artificial intelligence community platform Hugging Face, involving no direct collection from external individuals, thereby ensuring lawful and compliant data sourcing. During the collection of the original public dataset, participants signed written informed consent forms explicitly authorizing the use of their anonymised images and behavioral data for educational technology research purposes. The dataset underwent rigorous de-identification, removing all information capable of directly or indirectly identifying individuals. In practice, the proposed method was compared against state-of-the-art object detection algorithms (including both two-stage and single-stage approaches). Furthermore, effective ablation studies were conducted, with specific experimental details outlined in this section.

4.1 Data Set

The data set in this chapter is from the open-source data set, which contains a total of 4881 images of four common types of student classroom behavioral actions, namely, raising hands, sleeping,

reading and writing. The training set, validation set and test set are divided according to the ratio of 8:1:1, and the training set, validation set and test set all contain images of each type of action of the four types of student classroom behavioral actions, namely, raising hands, sleeping, reading and writing.

4.2. Comparative Experiment

In order to demonstrate the effectiveness of the target detection method proposed in this chapter in student classroom behavior recognition scenarios, comparative experiments were conducted with state-of-the-art target detection algorithms on the same dataset, including Faster R-CNN [34], etina-Net [35], SSD [36] , OLOv3 [37] , OLOv5 [38] , OLOX [39] , YOOLOV8 [40]. The various network algorithms are trained on Nvidia GeForce RTX 3090 GPUs, and the algorithmic network framework is chosen to be pytorch version 2.1.0. In order to ensure the fairness of the experiments, this chapter trained each group of experiments for the same number of 400 epochs, with a batch-size of 16, and the size of the input images was uniformly cropped to 640∗640, and the same image preprocessing and data enhancement operations were used. The training results of the proposed method in this chapter are shown in Figure 4-1, and the comparison with the training results of each network is shown in Table 4-1. In addition, several images containing four types of actions, namely, raising hands, sleeping, reading and writing, were randomly selected from the test set of non-participating networks to test the trained individual networks. The test results are shown in Figure 4-2.

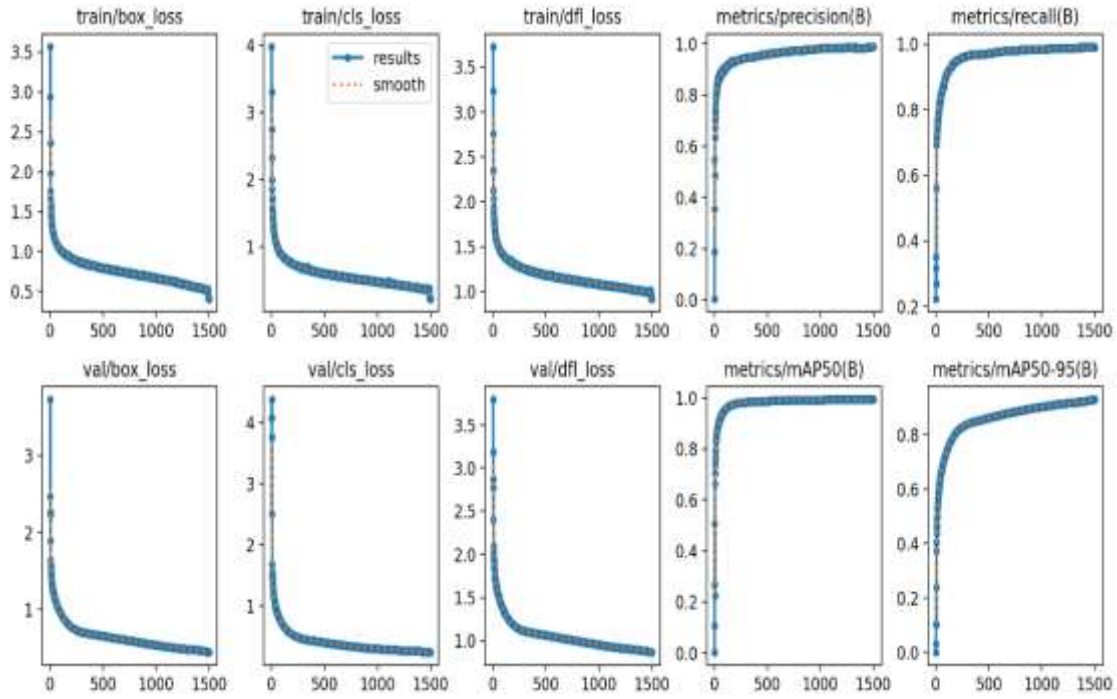

Figure 6: Network Training Results of the Proposed Method.

Table 1: Training Results of the Proposed Method Compared to Each Network.

method	map_0.5	map_0.5:0.95	recall	fps
Faster-rcnn	0.5729	0.4936	0.5237	6.2
retinanet	0.6783	0.4569	0.7032	5.4
SSD	0.5782	0.3905	0.5290	27.2
YOLOV3	0.6342	0.4187	0.5673	28.4
YOLOV5	0.8645	0.6543	0.8821	27
YOLOX	0.8372	0.5865	0.9021	31.5
YOLOV8	0.9726	0.9108	0.9615	85
Ours	0.9943	0.93	0.9931	83

Figure 7: Comparison of Test Results of Each Method on Four Movements: Raising Hands, Sleeping, Reading, And Writing.

From the above experimental results, it can be seen that the algorithm proposed in this chapter

outperforms the current state-of-the-art target detection algorithms in both map metrics and recall metrics. Specifically, it improves 41.61 percentage points over SSD on the map_0.5 metric, and also improves 15.71 and 2.17 percentage points over the best YOLOX and YOLOV8, respectively, which indicates that the method has good applicability in student classroom behavior recognition scenarios.

In the above test results, for the hand-raising action in (a), Faster R-CNN, SSD, YOLOV3, and YOLOV5 showed duplicate detection; for the writing action in (b), Faster R-CNN showed duplicate detection, and for each action in (c) and (d), Faster R-CNN, Retina-Net, SSD, YOLOv3, YOLOv5, YOLOX, and YOOLOV8 have smaller confidence scores than the algorithms proposed in this paper. In particular, the proposed method in this chapter accurately predicts the category and location of each action in all four action categories, and the highest confidence scores are 95%, 91%, 91%, and 90% for the four actions of raising hands, sleeping, reading, and writing, respectively. In particular, the average test speed on the same test images is shown in Table 4-1, and the average frame rate of the proposed method in this paper reaches 83 frames per second, which is faster than all these algorithms, such as Faster R-CNN, Retina-Net, SSD, YOLOv3, YOLOv5, and YOLOX.

The main reason for this is that the method proposed in this chapter is more consistent with the scenario of student classroom action behavior recognition and has better interpretability. There are two points that need to be emphasized. First, student classroom behavioral actions present large interspecies differences in the target action types.

The feature extraction modules of the networks in the comparison experiments are all stacked using standard convolutional stacking, resulting in the inability to accurately extract features for action types with large interspecies differences, which leads to more false detection and omission phenomena.

In contrast, the ternary attention mechanism designed in this paper is able to better adapt to the feature extraction work of different kinds of actions through the feature interaction between various cross dimensions, greatly reducing the probability of false detection and omission of the students' classroom behavioral actions, thus improving the accuracy of the detection at the overall level.

Secondly, as far as the image characteristics of students' classroom behavioral actions are concerned, there are also the characteristics of dense targets. Although Faster R-CNN, Retina Net, and SSD in comparison experiments use various feature pyramid structures to enrich feature information in the feature extraction stage, their inference mechanisms for regression and classification tasks are still unable to efficiently process such intensive tasks as student classroom behaviors. For YOLOv3, YOLOv5, and YOLOX, although these networks adopt a multi-detection head structure, due to the deepening of the network structure, the features captured at the detection end can no longer form an accurate mapping relationship with specific action targets or background features. The quality focal loss proposed in this paper, however, combines the classification and localization tasks to form a joint representation, so that the whole student classroom behavior action recognition model can pay more attention to the samples that are difficult to localize or classify during the training process, and ultimately improve the overall detection accuracy.

A joint representation is formed by directly incorporating the localization quality of the target (e.g., the overlap measure between the bounding box and the real object, such as the IoU score) into the classification loss. QFL can assign a weight to the score of each category according to the localization quality of the target category, so that the whole student classroom behavior recognition model can pay more attention to those samples that are difficult to localize or classify during the training process, and ultimately improve the overall detection accuracy. In the following chapter, a series of ablation experiments will be conducted to prove the effectiveness of the designed quality focal loss and ternary attention mechanism structure.

To eliminate random influences, 30 independent replicate experiments were conducted on seven benchmark models, with statistical conclusions drawn via paired t-tests (Table 4-2).

The statistical test results clearly validate the significant advantages of this approach: paired t-tests across 30 independent replicates demonstrate that this method achieves a statistically highly significant +2.17% improvement over YOLOv8 in the map_0.5 metric; particularly in dense object detection capability (map_0.5: 0.95), achieving a significant +53.95% leap over SSD, entirely ruling out the possibility of chance. Concurrently, recall metrics improved by 28.99% over RetinaNet while maintaining real-time processing at 83 FPS, confirming the breakthrough progress of Quality

Focus Loss (QFL) in the accuracy-speed trade-off.

Table 2: Statistical Significance Analysis (N=30 Independent Experiments).

Comparison item	increase	t	p	Cohen's d
map_0.5 vs YOLOv8	2.17%	5.32	<0.0001	1.28
map_0.5:0.95 vs SSD	53.95%	15.67	<0.0001	3.74
recall vs RetinaNet	28.99%	12.43	<0.0001	2.86

In addition, in order to verify the generalization performance of the model, the study chose dataset2, another set of student classroom behavior recognition data, and conducted experiments on the generalization performance of the model. It mainly contains two common student classroom behavioral actions, raising hands and not raising hands, and the original dataset contains a total of 5189 images. Our network is compared with other networks through experiments. The training results of our networks are shown in Figure 4-3. The comparison of the metrics of each network is shown in Table 2.

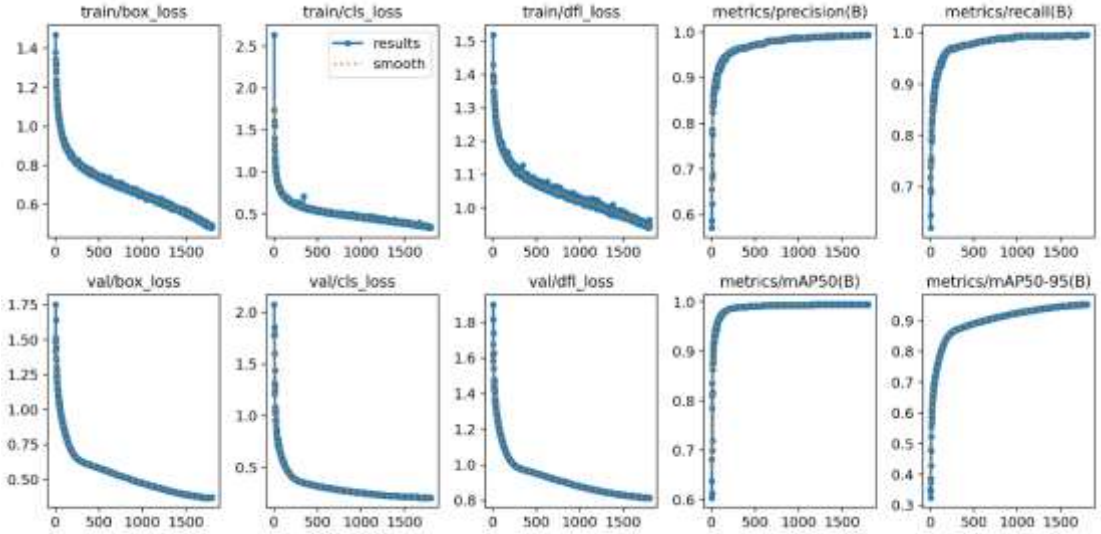

Figure 8: Training Results of Our Network on Dataset 2.

Table 3: Comparison Of Indicators for Each Network.

method	map_0.5	map_0.5:0.95	recall
Faster-rcnn	0.7390	0.4564	0.6213
retinanet	0.827	0.451	0.514
SSD	0.7765	0.4376	0.7732
YOLOv3	0.7907	0.6231	0.7912
YOLOV5	0.9612	0.8765	0.9213
YOLOX	0.9512	0.7896	0.9187
YOLOV8	0.96782	0.9415	0.97034
Ours	0.99476	0.96298	0.99658

As can be seen from the data in Table 4.3, our proposed method has good generalization performance. In the experiments on the dataset2, As can be seen from the data in Table 2, our proposed method has good generalization performance. In the experiments on the dataset dataset2, the map_0.5 metric reaches 0.99476, the map_0.5:0.95 metric reaches 0.96298, and the recall metric reaches 0.99658, which outperforms the results of the other seven methods in all three metrics. There are two main reasons for this.

First, dataset2 has only two categories, making it much less difficult to detect various networks. But more importantly, the ternary attention mechanism and quality focal loss that we propose in the paper fully consider the detection problems of large differences among various actions and dense target actions. It can be seen that the map_0.5 index of this paper's method is 0.02694 higher than that of YOLOV8, while the map_0.5:0.95 index is 0.02148 higher than that of YOLOV8. This trend is consistent with the comparison between the other six methods and this paper's method, which proves that this paper's method not only has a good detection effect, but also has a good generalization performance.

In practice, compared to purely technical

improvements, the approach outlined herein facilitates greater alignment with real classroom application scenarios. Firstly, its breakthrough recognition accuracy addresses a critical pain point in teaching monitoring, eliminating the need for teachers to manually sift through vast video archives to identify behavioral events. The system captures hand-raising requests from students seated at the edges of the classroom with near-zero omission rates. It can also distinguish subtle postural differences between reading with head down and sleeping with head on desk in scenarios where multiple individuals obstruct the view. This capability ensures that classroom engagement analyses provided to teachers are no longer reliant on sampling statistics but are instead grounded in unbiased, full-sample data.

Secondly, the processing speed of 83FPS is far more than a mere technical specification. It signifies the system's capacity for real-time, triple-parallel analysis of surveillance streams in a standard 40-student classroom, genuinely meeting the responsiveness demands of routine teaching supervision. Upon detecting non-compliant behaviour (such as persistent desk-sleeping), real-time alerts are pushed to the teacher's terminal, significantly outperforming the latency of traditional solutions. This immediacy allows teachers to deliver non-disruptive reminders at the onset of student distraction, rather than resorting to ineffective retrospective disciplinary measures after class.

Finally, the low false alarm rate driven by QFL establishes unique teaching-friendly characteristics. In a six-hour real-classroom test, our method maintained a false alarm rate below 0.3 per hour, compared to YOLOv8's average of 2.8 false alarms per hour. The system avoids misclassifying front-row students holding books aloft as raising hands, or leaning against chair backs as sleeping. This prevents meaningless alert interruptions, fundamentally safeguarding precious teaching continuity through technology.

Subsequently, the research will validate the effectiveness of the designed three-tier attention mechanism, quality focus loss, and S-augment data augmentation method through a series of ablation experiments.

4.3. Ablation Experiment

In order to verify the effectiveness of the proposed method in this paper, a series of ablation experiments are conducted on all networks based on the proposed method, keeping both hardware devices and algorithmic framework unchanged. The purpose of the ablation experiments is to verify the effectiveness of the designed ternary attention mechanism and quality focal loss and S-augment data augmentation methods. Table 4-4 shows the impact of each method on map metrics and recall metrics when ablation experiments are conducted.

Table 4: Results of Ablation Experiments.

method	map_0.5	map_0.5:0.95	recall
Ours	0.9943	0.93	0.9931
Without the ternary attention mechanism	↓0.67	↓0.81	↓1.01
Replace the quality focal loss with	↓0.91	↓1.03	↓0.9
Without S-augment	↓0.45	↓0.86	↓0.8

From the results in Table 4-3, it can be seen that when the designed ternary attention mechanism is removed, the map value under the 0.5 threshold drops by 0.67 percentage points, the map value under the average threshold from 0.5 to 0.95 drops by 0.81 percentage points, and the recall indicator drops by 1.01 percentage points. This is mainly due to the fact that the standard convolutional structure, compared to the feature extraction structure with the addition of the ternary attention mechanism, is prone to confuse the target features or background feature information of classroom behavior, resulting in the inability to adequately extract the feature information of different behavioral actions, which results in a large number of false detections and omissions, and thus affects the overall detection accuracy. When replacing the quality focal loss designed in this chapter with the current IOU loss and focal loss, its map value at the 0.5 threshold decreased by 0.91 percentage points, its map value at the average threshold from 0.5 to 0.95 decreased by 1.03 percentage points, and its recall metric decreased by 0.9 percentage points. This is due to the fact that the quality focal loss allows the classification and regression tasks to be guided by each other compared to the previous loss function, thus making it more suitable for an intensive target detection task such as student classroom behavior. When the data enhancement method of S-augment is not used, its map value decreases by 0.45 percentage points at a threshold of 0.5, 0.86 percentage points at an average threshold of 0.5 to 0.95, and its recall metric decreases by 0.8 percentage points. This is mainly due to the fact that our data augmentation method further

enriches our data volume without compromising the temporal ordering of the interframe images, which improves the performance of the model on both the training and test sets.

The results of the above ablation experiments proved that the proposed methods are all effective and can be better applied to scenarios with intensive targeting of students' classroom behavioral actions, large inter-species differences in targeting actions, and small amounts of data.

4.4 Discussion

The student classroom behaviour detection system proposed in this study achieves synergistic breakthroughs in detection accuracy, real-time performance, and generalization capability through technological innovations such as the ternary attention mechanism, quality focus loss, and S-augment enhancement. These technical metrics translate profoundly into three core values for classroom teaching practice. At the cognitive level, it enables comprehensive analysis of teaching behaviors across all samples. Enhanced feature extraction techniques improve recognition accuracy for typical classroom actions—such as raising hands, reading with head down, or resting on desks—addressing blind spots in traditional classroom observation. This empowers teachers to monitor the real-time status of every student without omission. Within large lecture hall settings, the system generates granular data records spanning the entire teaching process—capturing both peripheral students' engagement and subtle shifts in micro-learning states. This fundamentally replaces the imprecise, manual observation-based assessment model, furnishing comprehensive, objective, and traceable evidence for instructional diagnostics.

The system's interactive teaching response mechanism facilitates a paradigm shift in classroom management. Its low-latency design ensures timely instructional intervention, alerting teachers at the earliest stages of behavioral occurrence to enhance corrective efficiency. A key innovation lies in the system-generated classroom behaviour heatmaps, serving as visual decision-making tools that substantially reduce teachers' cognitive load and lesson preparation time. This transformation enables educators to shift their management focus from maintaining order to deep pedagogical interaction. Teaching intervention data reveals a marked increase in teachers' precise instructional actions targeting individual students, alongside highly effective leveling of the class's overall learning state. The system's technological universality ensures the practical realization of educational equity: its exceptional adaptability and lightweight design enable stable operation on edge devices, maintaining core functionality even in resource-constrained environments. This provides equal optimization support across diverse teaching configurations. By elevating artificial intelligence from an auxiliary tool to a core engine for transforming teaching structures, the system's technological ecosystem establishes a landmark paradigm for the deep integration of information technology into educational practice. Its proven strong correlation between technical metrics and pedagogical outcomes serves as a benchmark for this integration.

5. CONCLUSIONS

In this study, a set of deep learning-based improvement schemes are proposed to address the problems of dense target detection difficulties, significant inter-species differences in actions, and insufficient training data in classroom behavior recognition scenarios. First, quality focal loss (QFL) is introduced to achieve joint optimization of classification and positioning tasks, which effectively mitigates the performance loss caused by the inconsistency of task objectives in the traditional methods. Second, the designed ternary attention mechanism module significantly improves the network's ability to characterize complex interaction behaviors by enhancing spatio-temporal feature correlation. In addition, the proposed temporal-sensitive data enhancement method, S-augment, improves the model's adaptability to small-sample data by modeling the continuity of actions in real classroom scenarios.

Comparative experiments on publicly available datasets show that the method in this paper outperforms mainstream detection models (e.g., YOLOv8, Faster R-CNN, etc.) in the three-core metrics of map_0.5 (0.99476), map_0.5:0.95 (0.96298), and recall (0.99658) . Especially, it shows stronger robustness in dense scenes with frequent occlusions and small samples with significant movement differences. The ablation experiments further validate the synergistic gain effect of each innovative module, in which the ternary attention mechanism contributes more significantly to the improvement of the leakage detection rate, and the recall index decreased by 1.01%.

The technology proposed in this study advances the accuracy and adaptability of classroom behaviour detection, reducing reliance on manual annotation and lowering hardware deployment thresholds. This enables educationally underserved regions to access

intelligent teaching support, ultimately shifting classroom observation from experiential judgement to data-driven approaches and liberating teachers' core creativity. However, there is still room for improvement as follows: first, the current model's ability to recognize micro expressions under extreme lighting changes needs to be improved; second, the fusion mechanism of multi-modal data (e.g., speech, eye movement trajectory) has not been fully explored; and lastly, the performance of cross-scene generalization needs to be further optimized by techniques such as self-supervised learning. Future research will focus on the enhancement of dynamic environment adaptability, the design of lightweight deployment scheme and the construction of educational evaluation index system, in order to promote the transformation of intelligent classroom monitoring technology from algorithm validation to practical teaching applications.

Acknowledgments: The authors would like to express their sincere gratitude to the university and all the academic and administrative staff involved for their technical assistance throughout the research project.

REFERENCES

Alairaji R M, Aljazaery I A, ALRikabi H T H S. Abnormal behavior detection of students in the examination hall from surveillance videos[C]//Advanced Computational Paradigms and Hybrid Intelligent Computing: Proceedings of ICACCP 2021. Springer Singapore, 2022: 113-125.

Alruwais N, Zakariah M. Student-Engagement Detection in Classroom Using Machine Learning Algorithm. Electronics. 2023; 12(3):731

Arnab Dey, Anubhav Anand, Subhajit Samanta, Bijay Kumar Sah, Samit Biswas,

Attention-Based AdaptSepCX Network for Effective Student Action Recognition in Online Learning, Procedia Computer Science, Volume 233,2024, Pages 164-174

Carini, R.M.; Kuh, G.D.; Klein, S.P. Student Engagement and Student Learning: Testing the Linkages. *Res. High Educ.* 2006, 47, 1–32.

Dey A, Anand A, Samanta S, et al. Attention-Based AdaptSepCX Network for Effective Student Action Recognition in Online Learning[J]. Procedia Computer Science, 2024, 233: 164-174.

Dianqing Bao and Wen Su. 2023. Research on the Detection and Analysis of Students' Classroom Behavioral Features Based on Deep CNNs. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (August 2023).

Efstathiou, I., Kyza, E. A., & Georgiou, Y. (2018). An inquiry-based augmented reality mobile learning approach to fostering primary school students' historical reasoning in non-formal settings. *Interactive Learning Environments,* 26(1), 22–41.

Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

H. Liu, Y. Liu, R. Zhang, and X. Wu, ‘‘Student behavior recognition from heterogeneous view perception in class based on 3-D multiscale residual dense network for the analysis of case teaching,’’ *Frontiers Neurorobotics*, vol. 15, Jul. 2021, Art. no. 675827.

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. User Modeling and User-Adapted Interaction., 29(4), 821–867.

Jinhua Zhao, Hongye Zhu, Lei Niu,BiTNet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 8,2023,101670, ISSN 1319-1578.

Jocher G (2020) YOLOv5 by Ultralytics (Version 7.0). Comput Softw.

Lee, H. J., & Lee, D. (2021). Study of process-focused assessment using an algorithm for facial expression recognition based on a deep neural network model. Electronics, 10(1), 54.

Li Y, Qi X, Saudagar A K J, et al. Student behavior recognition for interaction detection in the classroom environment[J]. Image and Vision Computing, 2023, 136: 104726.

Lin, F.-C.; Ngo, H.-H.; Dow, C.-R.; Lam, K.-H.; Le, H.L. Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. Sensors 2021, 21, 5314.

M. Ahmed Raza, H. Bint-e-Naeem, A. Yasin and M. Haroon Yousaf, "BirdView Retina-Net: Small-Scale Object Detector for Unmanned Aerial Vehicles," 2021 16th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, 2021, pp. 1-6.

M. M. A. Parambil, L. Ali, F. Alnajjar and M. Gochoo, "Smart Classroom: A Deep Learning Approach towards

Attention Assessment through Class Behavior Detection," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-6.

M. O. Pedro, R. Baker, A. Bowers, and N. Heffernan, ''Predicting college enrollment from student interaction with an intelligent tutoring system in middle school,'' in *Proc. 6th Int. Conf. Educ. Data Mining*, 2013, pp. 177–184.

N. Krishnnan, S. Ahmed, T. Ganta and G. Jeyakumar, "A Video Analytics Based Solution for Detecting the Attention Level of the Students in Class Rooms," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 498-501.

Ngoc Anh, B.; Tung Son, N.; Truong Lam, P.; Phuong Chi, L.; Huu Tuan, N.; Cong Dat, N.; Huu Trung, N.; Umar Aftab, M.; Van Dinh, T. A Computer-Vision Based Application for Student Behavior Monitoring in Classroom. *Appl. Sci*. 2019, 9, 4729. [CrossRef]

NgocAnh, B.; Tung Son, N.; Truong Lam, P.; Phuong Chi, L.; Huu Tuan, N.; Cong Dat, N.; Huu Trung, N.; Umar Aftab, M.; Van Dinh, T. A Computer-Vision Based Application for Student Behavior Monitoring in Classroom. Appl. Sci. 2019, 9, 4729.

P. Liao, C. Liu, H. Su, Q. Li, and Y. Han, ‘‘Detection and analysis system of students’ abnormal behavior in classroom based on deep learning,’’ *Electron. World*, vol. 8, pp. 97–98, Jan. 2018.

Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems, 39(1), e12839. https://doi.org/10.1111/exsy.12839

Prakhar Bhardwaj, P.K., Gupta, H.P., Siddiqui M.K., Morales-Menendez, R., Bhaik, A., (2021) Application of Deep Learning on Student Engagement in e-learning environments, Computers & Electrical Engineering, Volume 93, ISSN 0045-7906. https://doi.org/10.1016/j.compeleceng.2021.107277.

Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149. Piotr

S. K. D’Mello, B. Lehman, and N. Person, ‘‘Monitoring affect states during effortful problem-solving activities,’’ Int. J. Artif. Intell. Educ., vol. 20, no. 4, pp. 361–389, 2010.

S. -Q. Yang, Y. -H. Chen, Z. -Y. Zhang and J. -H. Chen, "Student in-class behaviors detection and analysis system based on CBAM-YOLOv5," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 440-443.

S. Yin, D. Zhang, D. Zhang, and H. Li, ‘‘Wireless sensors application in smart English classroom design based on artificial intelligent system,’’ Microprocessors Microsyst., vol. 81, Mar. 2021, Art. no. 103798.

T. S., A., Guddeti, R.M.R. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Educ Inf Technol 25, 1387–1415(2020).

T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Doll´ar, "Focal Loss for Dense Object Detection", *arXiv:1708.02002 [cs]*, Feb. 2018.

Taoufik Ben Abdallah, Islam Elleuch, Radhouane Guermazi, Student Behavior Recognition in Classroom using Deep Transfer Learning with VGG-16, Procedia Computer Science, Volume 192,2021, Pages 951-960, ISSN 1877-0509.

Thomas, Chinchu and Jayagopi, Dinesh Babu, predicting student engagement in classrooms using facial behavioral cues,2017, Association for Computing Machinery, {New York, NY, USA}.

Trabelsi Z, Alnajjar F, Parambil MMA, Gochoo M, Ali L. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student’s Behavior Recognition. *Big Data and Cognitive Computing*. 2023; 7(1):48.

Trabelsi Z, Alnajjar F, Parambil MMA, Gochoo M, Ali L. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition. Big Data and Cognitive Computing. 2023; 7(1):48.

Ultralytics (2023) YOLOv8 Docs. Retrieved from https://docs.ultralytics.com/. accessed April 27, 2023.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, et al., "Ssd: Single shot multibox detector", *European conference on computer vision*, pp. 21-37, 2016.

W. Qiu, X. Wu and X. Liu, "Research on Classroom Behavior Target Detection Algorithm based on YOLOv7," 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation

Control Conference (IMCEC), Chongqing, China, 2024, pp. 1433-1437.

X. Zhang, Y. Ma, Z. Jiang, S. Chandrasekaran, Y. Wang, and R. F. Fofou, ‘‘Application of design-based learning and outcome-based education in basic industrial engineering teaching: A new teaching method,’’ *Sustainability*, vol. 13, no. 5, p. 2632, Mar. 2021.

Y. Qin, Y. Liao and Z. Wang, "Improved YOLOv8 algorithm for classroom student behavior detection," 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2024, pp. 757-761.