applied sciences

MDPI

Article

SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms

Zhifeng Wang [1,2,*], Minghui Wang [1], Chunyan Zeng [3,*] and Longlong Li [2]

[1] CCNU Wollongong Joint Institute, Central China Normal University, Wuhan 430079, China
[2] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[3] Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China
* Correspondence: zfwang@ccnu.edu.cn (Z.W.); cyzeng@hbut.edu.cn (C.Z.)

Abstract: Detecting student behavior in smart classrooms is a critical area of research in educational technology that significantly enhances teaching quality and student engagement. This paper introduces an innovative approach using advanced computer vision and artificial intelligence technologies to monitor and analyze student behavior in real time. Such monitoring assists educators in adjusting their teaching strategies effectively, thereby optimizing classroom instruction. However, the application of this technology faces substantial challenges, including the variability in student sizes, the diversity of behaviors, and occlusions among students in complex classroom settings. Additionally, the uneven distribution of student behaviors presents a significant hurdle. To overcome these challenges, we propose Student Behavior Detection Network (SBD-Net), a lightweight target detection model enhanced by the Focal Modulation module for robust multi-level feature fusion, which augments feature extraction capabilities. Furthermore, the model incorporates the ESLoss function to address the imbalance in behavior sample detection effectively. The innovation continues with the Dyhead detection head, which integrates three-dimensional attention mechanisms, enhancing behavioral representation without escalating computational demands. This balance achieves both a high detection accuracy and manageable computational complexity. Empirical results from our bespoke student behavior dataset, Student Classroom Behavior (SCBehavior), demonstrate that SBD-Net achieves a mean Average Precision (mAP) of 0.824 with a low computational complexity of just 9.8 G. These figures represent a 4.3% improvement in accuracy and a 3.8% increase in recall compared to the baseline model. These advancements underscore the capability of SBD-Net to handle the skewed distribution of student behaviors and to perform high-precision detection in dynamically challenging classroom environments.

Keywords: student behavior; multi-level features; focal modulation; detection network

Citation: Wang, Z.; Wang, M.; Zeng, C.; Li, L. SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms. *Appl. Sci.* **2024**, *14*, 8357. https://doi.org/10.3390/app14188357

Academic Editor: Andrea Prati

Received: 2 August 2024
Revised: 8 September 2024
Accepted: 14 September 2024
Published: 17 September 2024

1. Introduction

The rapid development of Artificial Intelligence (AI) has greatly promoted its application in various fields [1,2]. In the field of education, the indicators and methods of learning evaluation are a noteworthy research direction [3–5], which plays a crucial role in improving the quality of education and the learning effect of students [6]. A comprehensive and effective learning evaluation method can help educators understand and analyze the various aspects of the teaching process more accurately, so as to make timely improvements and optimizations [7–9]. Various methods in artificial intelligence can make the learning evaluation system more intelligent and refined. For example, we can collect educational data and use machine learning models such as clustering to predict students' scores [10]. We can also use natural language processing (NLP) technology to analyze students' essays and open-ended questions, assess their language skills, logical thinking, and knowledge mastery, and deeply understand their thinking process and emotional state [11]. In addition, a deep learning-based student behavior recognition model is a refined assessment

model [6,12]. In this way, educators can obtain a more detailed and accurate picture of student behavior and carry out more profound observations and assessments. For example, SK Mahapatra et al. used the Internet of Things (IoT) to gather information to improve student learning using a novel gamified educational approach after establishing a communication network and an educational network [13], while Daumiller et al. constructed a model of overall teaching quality for student evaluation from a psychological perspective [14].

In traditional classroom settings, educators and teaching staff rely heavily on a direct observation of students' learning behaviors to understand their learning trajectories. In addition, indirect assessment methods, such as grade records, homework completion, and class attendance, are utilized to measure students' learning processes. Of course, it also involves assessing students' status from a psychological perspective [15]. With the development of technology, real-time smart classroom student behavior detection systems [16,17] as well as psychological assessment systems are gradually emerging [18,19]. However, due to the complexity and high cost of these techniques, they are not widely used. Therefore, what we need should be a lightweight and low-cost target detection system. With the continuous development of deep learning technology, lightweight target detection models are emerging, and the single-stage target detector represented by YOLO is driving the real-time detection system by virtue of its extremely high inference speed as well as low deployment costs. Therefore, we constructed a lightweight SBD-Net framework and SCBehavior dataset for students' classroom behavior detection on this basis. The main goal of current behavioral recognition tasks in education is to use statistical and generalized information about students' behavior to understand their learning, personality, and psychological characteristics in order to assess and improve teaching methods.

This study introduces several significant advancements in the field of classroom behavior detection using computer vision, which are outlined as follows:

(1) Development of SCBehavior Dataset: We constructed the SCBehavior dataset, which encompasses six prevalent student behaviors such as reading, writing, and raising a hand. This dataset is crucial not only for training and validating our model but also as a valuable asset for the broader research community. It supports the development of more nuanced behavior detection systems and aids in further empirical studies. The SCBehavior dataset is publicly available at https://github.com/CCNUZFW/SCBehavior (accessed on 10 July 2024).

(2) Enhanced Feature Fusion: We replaced the traditional SPPF with the Focal-Modulation module to implement a multilevel feature fusion mechanism. This enhancement significantly improves the model's ability to detect and accurately classify students' behavior across different scales, particularly in complex scenarios involving hard-to-detect samples.

(3) ESLoss Function: To address the challenge of sample imbalance that typically skews the model's performance, we designed the ESLoss function. This function is particularly effective in increasing the detection accuracy for infrequent behaviors by assigning higher weights to difficult samples, thus refining the model's sensitivity and precision.

(4) Dyhead Detection Head: Our introduction of the Dyhead detection head, which incorporates multiple attention mechanisms, represents a pivotal improvement. This design enhances the model's capability to discern complex student behaviors with minimal increase in computational load, ensuring efficiency without compromising on performance.

The rest of this paper is structured as follows: Section 2 briefly reviews the current related works. In Section 3, we introduce the concepts that appear in this paper and define the related problems of this research. Section 4 introduces the proposed method. Section 5 introduces the dataset and evaluation metrics, and it provides the details of the experiments and analysis of the results. Finally, Section 6 summarizes the entire paper.

2. Related Work

The recognition of student behavior in educational environments is a growing area of research, leveraging advancements in computer vision and deep learning. This section reviews the key developments and methodologies in the related fields.

2.1. Student Behavior Recognition

Tasks related to classroom behavioral recognition are having a profound impact on the field of education, and work in this area aims to enhance teaching and learning. For example, Sharma et al. determined the level of student engagement in e-learning and distance learning tasks by combining information about students' eye and head movements as well as facial emotions [20]. Jisi A et al. combined spatial affine transform networks with convolutional neural networks for the behavioral recognition of students in education to extract more detailed features for better detection [21]. Wang et al. designed and modified an efficient student behavior detection model based on yolov7, which improves the recognition of student behavior by embedding an attention mechanism and combining it with an augmented dataset [22]. In addition, Lu Shi et al. proposed a learning behavior recognition method for an online English classroom based on feature data mining, which achieved learning behavior recognition in the online English classroom by mining feature data [23]. From these recent studies, we find that many models have been applied to classroom student behavior detection, and all of them have achieved good results.

For example, Cao et al. proposed a student behavior detection model based on an improved SSD algorithm that integrates the Mobilenet architecture, designed for real-time behavior detection in dynamic classroom environments [24]. Their model combines two different approaches to enhance detection capabilities, particularly for classroom behavior. Additionally, Li et al. introduced an innovative method using an attention mechanism and relational reasoning module specifically suited for complex classroom environments, improving the detection of human–object interaction behaviors [25]. Furthermore, Chen et al. presented a model based on an improved YOLOv8, incorporating a multi-head self-attention mechanism, which showed greater robustness and accuracy in classroom scenarios [26].

Summary: While these models have improved performance through structural innovations, they lack targeted optimizations for the specific challenges in student behavior detection, such as occlusion, small-object recognition, and behavior class imbalance. These issues are critical in the classroom environment, where traditional models may struggle with detecting small, occluded objects and underrepresented behaviors. In addition, these studies underscore the significant advancements in student behavior recognition, yet they highlight the need for lightweight and deployable models, prompting our development of an efficient solution tailored to educational environments. So, it is important to build such a particular lightweight model for student behavior detection .

2.2. Object Detection

Object detection has been a cornerstone in the field of computer vision, driving advancements in various applications such as autonomous driving, surveillance, and, importantly, educational environments. The primary goal of object detection is to identify and localize objects within an image, which is crucial for tasks like student behavior recognition where precise detection of students' actions and interactions is needed. Traditional object detection methods, such as Viola-Jones [27] and Histogram of Oriented Gradients (HOG) [28], laid the groundwork by introducing robust feature extraction and classification techniques. However, these methods often struggled with real-time performance and handling diverse, complex scenes.

The advent of deep learning significantly transformed object detection with the development of Convolutional Neural Networks (CNNs). The introduction of Region-based CNN (R-CNN) by Girshick et al. [29] marked a pivotal shift by leveraging CNNs for region proposal and feature extraction. R-CNN generated region proposals using selective

search and then applied a CNN to extract features from each proposed region, followed by a classifier to determine the presence of objects. Although this approach significantly improved detection accuracy, it was computationally expensive due to the repeated CNN computations for each region proposal. To address these limitations, Fast R-CNN [30] was introduced, which significantly improved the efficiency of the R-CNN framework. Fast R-CNN incorporated a single-stage training algorithm that performed region proposal and classification simultaneously. It used a Region of Interest (RoI) pooling layer to extract a fixed-length feature vector from the feature map for each region proposal, thereby reducing redundant computations and speeding up the detection process. Building on the advancements of Fast R-CNN, Faster R-CNN [31] integrated a Region Proposal Network (RPN) directly into the CNN architecture. This end-to-end trainable network generated region proposals more efficiently and accurately than the selective search method used in earlier models. The RPN shared convolutional features with the detection network, which further reduced computational overhead and improved detection speed and accuracy.

The next significant advancement in CNN-based object detection came with the development of single-stage detectors such as the Single Shot MultiBox Detector (SSD) [32] and You Only Look Once (YOLO). Unlike the two-stage R-CNN variants, single-stage detectors eliminated the region proposal step, allowing for real-time object detection. The SSD predicted bounding boxes and class scores for multiple objects directly from feature maps at different scales, improving the detection of objects with varying sizes. YOLO, on the other hand, framed object detection as a single regression problem, predicting bounding boxes and class probabilities simultaneously from the entire image in one evaluation. This approach significantly increased detection speed, making YOLO suitable for real-time applications. In practice, target detection tasks in different fields may require specific models to adapt, which requires us to optimize the original target detection algorithms, such as formulating a specific backbone network to adapt to the specific task, so as to better extract the upstream features, but also through the addition of the attention mechanism or the optimization of the loss function to improve the detection results. The optimization of loss functions directly impacts model performance, convergence speed, and robustness. For instance, Focal Loss [33] was introduced to address the class imbalance problem by down-weighting the loss assigned to well-classified examples, thereby focusing the training on hard negatives. To better capture the quality of bounding box predictions, GIoU loss [34] extended the Intersection over Union (IoU) metric to provide a more comprehensive evaluation of overlap between predicted and ground-truth boxes, addressing shortcomings like the inability to capture the distance between non-overlapping boxes. CIoU loss [35] further improves on GIoU by considering the aspect ratio and distance between the center points of the bounding boxes, ensuring better convergence and accuracy in bounding box predictions.

Summary: Overall, target detection technology is in a phase of continuous innovation and rapid development, and it often requires specific improvements in practical applications, which is a research problem we need to be aware of.

2.3. Attention Mechanism

Traditional deep learning models, such as Recurrent Neural Networks (RNNs) [36] and convolutional neural networks, often struggled with long-range dependencies and efficiently processing relevant features in the input data. This led to the development of attention mechanisms, which were initially introduced in the context of machine translation. Bahdanau et al.'s Additive Attention [37] addressed the limitations of fixed-length context vectors in sequence-to-sequence models by dynamically computing alignment scores between the encoder and decoder hidden states. This allowed the model to focus on relevant parts of the input sequence during translation, significantly improving translation quality.

As the concept and technology of attention mechanisms continued to evolve, a variety of attention mechanisms emerged. Attention mechanisms have not only achieved significant progress in natural language processing but have also been successfully applied to

computer vision tasks, enhancing models' ability to focus on spatial and channel features. For example, the Convolutional Block Attention Module (CBAM) [38] sequentially applies channel and spatial attention to input feature maps, enabling the model to focus on the most informative parts of the feature maps, which has shown significant performance improvements in image classification and object detection tasks. The innovation of CBAM lies in its dual attention mechanism: the channel attention module focuses on "what" is important, while the spatial attention module focuses on "where" it is important, ensuring the network can effectively emphasize critical features. Similarly, Squeeze-and-Excitation Networks (SENets) [39] introduced a channel attention mechanism that adaptively recalibrates channel-wise feature responses by modeling interdependencies between channels. This approach significantly improves the performance of various convolutional neural network architectures on image recognition tasks. SENet's innovation lies in its squeeze operation, which compresses the spatial dimensions of the input, and the excitation operation, which learns the importance of each channel through fully connected layers, significantly boosting performance.

Summary: These advancements underscore the versatility and efficacy of attention mechanisms in deep learning, highlighting their critical role in the ongoing development of more sophisticated and capable models. By allowing models to dynamically focus on the most relevant parts of input data, attention mechanisms have become indispensable tools for achieving high performance in various applications.

3. Preliminary

In this section, we introduce some key shorthand technical terms in Table 1, as well as a detailed description of the terms Focal Modulation and Exponential Moving Average SlideLoss.

Table 1. Symbols and notations.

Number	Notation	Description
1	$\mathbb{R}^{L\times S\times C}$	3D tensor with dimensions of height, width, and channel
2	M	Aggregation operation on attention score matrix
3	T	Interaction between query and targets
4	$\mathcal{F}$	Feature map
5	π_L	Scale-aware attention
6	π_S	Spatial-aware attention
7	π_C	Task-aware attention
8	$f(x)$	Sliding function operation
9	EMA	Exponential Moving Average

Definition 1 (Student Behavior Detection Task)**.** *The task of detecting students' classroom behavior focuses on identifying the input classroom sensor data D_{input} and finding the most closely matching behavior from a set of predefined behavior categories $\{C_i \mid i = 1, 2, 3, \ldots, M\}$. Here, i denotes the index of the specific classroom behaviors recognized in this study. This problem can be mathematically described as follows:*

$$C^* = \arg\max_{C_i} \{g(D_{input}, C_1; \theta), g(D_{input}, C_2; \theta), \ldots, g(D_{input}, C_M; \theta)\} \quad (1)$$

In this context, $g(\cdot)$ is a function that measures the similarity between behaviors, and θ represents the parameters of this function. By assessing the input classroom sensor data D_{input}, we determine the similarity degree with various behavior categories, and the behavior C^* with the highest similarity score is identified as the recognized behavior.

Definition 2 (Focal Modulation)**.** *Focal Modulation is an advanced mechanism designed to enhance neural network efficiency by replacing traditional self-attention modules. Unlike self-attention, which computes the similarity between query tokens and their surrounding tokens*

with high computational complexity, focal modulation aggregates contextual information at each query position and modulates based on the aggregated context. This process involves encoding visual contexts at different granularities through deep convolutional implementations, selectively aggregating these features, and fusing them into the query. In summary, Focal Modulation is a more lightweight feature aggregation mechanism that we use in the backbone module to enhance model performance.

Definition 3 (Exponential Moving Average Slide Loss)**.** *Exponential Moving Average (EMA) is a statistical measure that applies weighting factors to a time series, giving more significance to recent data points while smoothing out fluctuations and reducing the impact of outliers. The formula for EMA is*

$$\mu_t = \beta \times \mu_{t-1} + (1 - \beta) \times \theta_t \tag{2}$$

where μ_t is the current EMA value, μ_{t-1} is the previous EMA value, θ_t is the current data point, and β is the smoothing factor between 0 and 1.

In this study, we apply the concept of EMA to the Slide Loss function to address sample imbalance in student classroom behavior datasets. The threshold parameter μ, which differentiates between simple and difficult samples based on the Intersection over Union (IoU) between predicted and ground truth boxes, is optimized using EMA. This moving average approach helps smooth out μ over time, reducing jitter and avoiding significant fluctuations caused by outliers.The improved Slide Loss function, termed ESLoss, categorizes samples with IoU values below the EMA threshold as negative and those above as positive. This weighted approach emphasizes difficult-to-detect samples, enhancing the model's performance and robustness in behavior detection tasks.

4. Proposed Method

In this section, we present the improvements we made for one specific task of classroom student behavior recognition. We cover the theoretical foundations and the specific process of improving the algorithm. Our model involves five main components: input, backbone, neck, head, and output. We first input a single frame image in the classroom context into the model, resize it to 640 × 640, and then perform initial feature extraction via the Backbone, followed by further extraction and fusion of features at the Neck layer to provide richer information for subsequent prediction, and finally, the output of seven different behaviors via the Dyhead target detection head. The overall structure of the model is shown in Figure 1.

Current models [20–26] mostly rely on general attention mechanisms and conventional modules, lacking deep optimization for the specific challenges posed by student behavior detection, such as occlusion, small-object recognition, and class imbalance in classroom environments. While these models have improved performance through structural innovations, they are not specifically designed to address the unique challenges of student behavior recognition, including occlusion, small-target detection, and behavior class imbalance. As a result, they perform well in general object detection tasks but may struggle to handle the complexities of student behavior detection in a classroom setting. In contrast, our model is designed to specifically tackle these challenges, particularly in handling occlusion and small-object detection. We developed a Focal Modulation module combined with a Dyhead architecture to address the occlusion and small-target detection issues in classroom environments. Additionally, we introduced the ESLoss function, specifically tailored to address the imbalance in student behavior datasets, further enhancing the model's performance. These improvements, designed specifically for student behavior detection, allow our model to outperform existing models that rely on general-purpose modules, especially in the complex classroom environment.

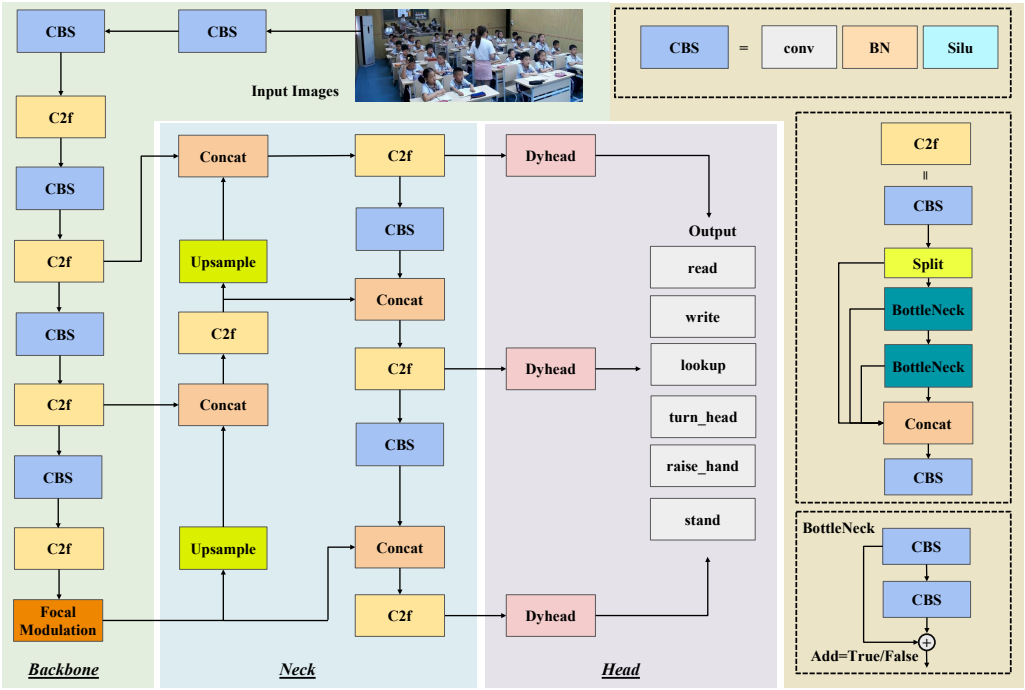

Figure 1. The overall structure of our model.

4.1. Focal Modulation Network

The original backbone of YOLOv8 builds upon its predecessors by incorporating advanced architectural designs such as convolutional layers for hierarchical feature extraction, residual blocks to enhance gradient flow, bottleneck layers for computational efficiency, and CSPNet to balance computational load and improve performance. One of the critical components in this backbone is the Spatial Pyramid Pooling Fast (SPPF) module, which captures multi-scale information by pooling features at multiple scales.

We found that in tasks such as student classroom behavior detection, there are a variety of scenarios with varying sample sizes of different student behaviors as well as more complex classroom environments, which often lead to less effective detection models. Many researchers consider introducing self-attention mechanisms [40] to alleviate this problem in such situations. Self-attention mechanisms are widely used in visual tasks, but their computational complexity is high, especially when dealing with high-resolution inputs. Therefore, we decided to introduce the Focal Modulation technique [41], a relatively lightweight network structure, as a direct replacement for the traditional SPFF module to further limit the complexity of the model, thus guaranteeing the real-time performance of our detector and ensuring that it can be applied to real classroom environments.The Focal Modulation is mainly a multilevel feature fusion mechanism that is incorporated into the module to learn both coarse-grained spatial information and fine-grained feature information in the classroom environment, and acquiring student behavior information from far and near, thus improving the model performance, as shown in Figure 2. Focal Modulation can also increase the focus on hard-to-detect targets (behaviors with small sample sizes in the dataset) and achieve focus on difficult samples, i.e., special behaviors, thus improving the detection accuracy of the model.

Traditional self-attention mechanisms aggregate contextual information by computing the similarity between query tokens and their surrounding tokens, with a computational complexity of quadratic order. Focal modulation, however, redefines this process by first aggregating contextual information at each query position and then modulating based on the aggregated context. Specifically, the process of self-attention can usually be expressed as follows:

$$y_i = M_1(T_1(x_i, X), X) \tag{3}$$

where M_1 represents the aggregation operation performed on the attention score matrix between the query and its targets, and T_1 is the interaction between the query and targets. In contrast, the process of focal modulation is expressed as follows:

$$y_i = T_2(M_2(i, X), x_i) \tag{4}$$

In this case, M_2 is the operation of aggregating the context at each position, and T_2 is the interaction between the query and the aggregated features.

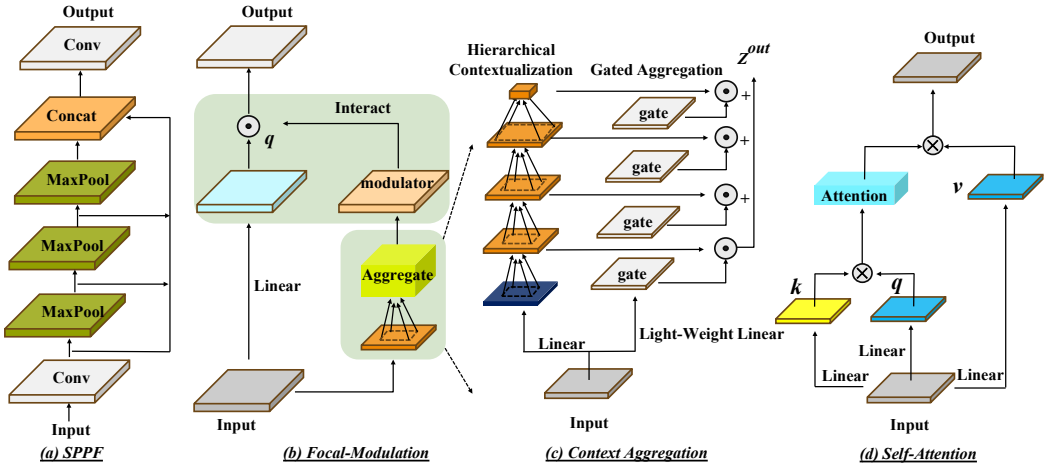

Figure 2. This is a description of the framework for focal modulation, and we compare it with both Self Attention and the SPPF module to show the differences between the three in a more graphic way.

Specifically, as shown in Figure 3 and Algorithm 1, focal modulation uses a set of deep convolutional implementations to encode short- to long-range visual contexts at different granularities, selectively aggregates the contextual features of each marker according to its content, and fuses the aggregated features into the query, greatly simplifying the computational process compared to traditional self-attention modules. The specific computation of focus modulation is shown below:

$$y_i = q(x_i) \odot h\left(\sum_{\ell=1}^{L+1} g_i^{\uparrow \ell} \cdot z_i^{\uparrow \ell}\right) \tag{5}$$

where $g_i^{\uparrow \ell}$ and $z_i^{\uparrow \ell}$ are the gating values and visual features, and $q(x_i)$ is a query projection function.

Figure 3. The main difference between Self Attention (SA) and Focal Modulation lies in the way contextual information is processed. SA captures local contextual information around the query tokens through a fixed-size window, whereas Focal Modulation encodes spatial contextual information at different levels of granularity and adaptively fuses this information according to the query content, making the interaction and aggregation process more lightweight and efficient.

Algorithm 1 Pseudo code for Focal Modulation

Input/Output Shape: (B, H, W, C) ▷ Batchsize B; height H, width W, dim C
Focal Levels: L; **Conv Kernel Size at Level** ℓ**:** k^ℓ
function INIT
    $pj\_in, pj\_cxt \leftarrow \texttt{Linear}(C, 2*C+(L+1)), \texttt{Conv2d}(C, C, 1)$
    $hc\_layers \leftarrow [\texttt{Sequential}(\texttt{Conv2d}(C, C, k^\ell, \texttt{groups} = C), \texttt{GeLU}())$ **for** ℓ **in range**$(L)]$
    $pj\_out \leftarrow \texttt{Sequential}(\texttt{Linear}(C, C), \texttt{Dropout}())$
end function
function FORWARD$(x, m = 0)$
    $x \leftarrow pj\_in(x).permute(0, 3, 1, 2)$
    $q, z, gate \leftarrow \texttt{split}(x, (C, C, L+1), 1)$
    for ℓ **in range**(L) **do**
        $z \leftarrow hc\_layers[\ell](z)$ ▷ Equation (4), hierarchical contextualization
        $m = m + z \times \texttt{gate}[:, :, \ell : \ell + 1]$ ▷ Equation (5), gated aggregation
    end for
    $m = m + \texttt{GeLU}(z.mean(\texttt{dim} = 2, 3)) \times \texttt{gate}[:, :, : L]$ ▷ Equation (6), focal modulation
    $x \leftarrow q + pj\_cxt(m)$
    return $pj\_out(x.permute(0, 2, 3, 1))$
end function

4.2. Dyhead Module

After the backbone and neck structure further refine these features by aggregating information from different scales, the head of YOLOv8 is where the final detection occurs. It processes the aggregated features from the neck to predict bounding boxes and class probabilities. To enhance this process, we use a dynamic head module to unify the attention mechanisms, as shown in Figure 4. The dynamic head [42] combines multiple attention mechanisms at the feature level, spatial location, and output channel. This enables scale-awareness, spatial-awareness, and task-awareness, making the model particularly effective in dealing with complex student behaviors in a classroom setting. The dynamic head module enhances the robustness of student behavior detection, and its design allows for high-performance target detection with relatively low computational resources. Even with limited computational resources, it still provides satisfactory results, aligning with our goal of a lightweight model.

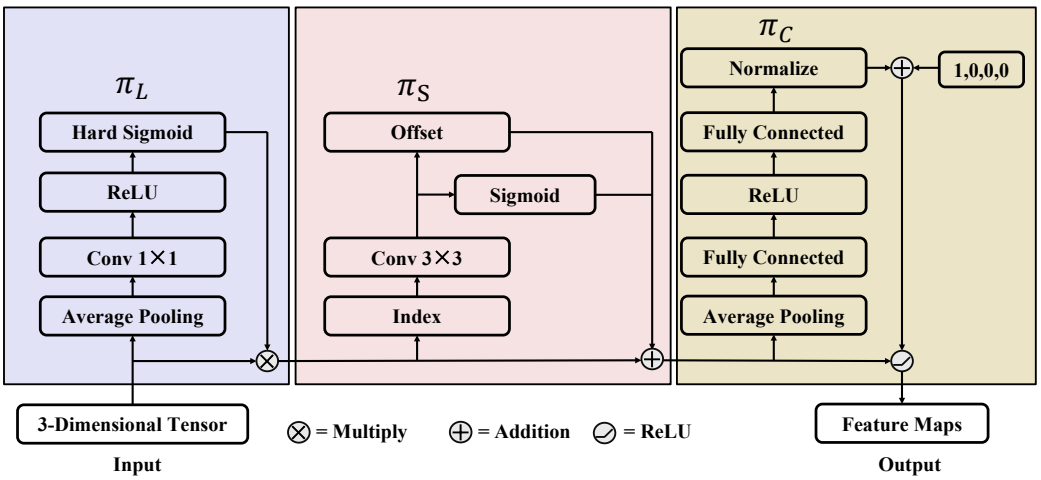

Figure 4. This image shows the execution flow of the three attention modules of Dyhead; it is worth noting that we can stack more than one Dyhead module to achieve better results, but this often leads to performance loss.

To be more specific, given a feature map $\mathcal{F} \in \mathbb{R}^{L \times S \times C}$, the broad definition of attention can be described as follows:

$$W(\mathcal{F}) = \pi(\mathcal{F}) \cdot \mathcal{F} \tag{6}$$

One optimal solution is to adopt global attention, but the attention function at all scales will lead to excessive computational overhead and is not practical due to high-dimensional problems. Instead, we convert the attention function into a series of cascaded attentions, each focusing only on one scale:

$$W(\mathcal{F}) = \pi_C(\pi_S(\pi_L(\mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F} \tag{7}$$

Scale-aware Attention π_L: We first introduce scale-aware attention based on the importance of different scales:

$$\pi_L(\mathcal{F}) = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C}\mathcal{F}\right)\right) \cdot \mathcal{F} \tag{8}$$

where $f(\cdot)$ is a linear function, using a 1×1 convolution, and $\sigma(x)$ is a hard-sigmoid activation function.

Spatial-aware Attention π_S: Next, we introduce spatial-aware attention to focus on different spatial positions of the feature map. Considering the high dimensionality S, we decompose it as follows: first, use convolution to learn the feature transformation, and then aggregate across scales:

$$\pi_S(\mathcal{F}) \cdot \mathcal{F} = \frac{1}{L}\sum_{k=1}^{L} w_{i,j,k} \cdot \mathcal{F}(l; p_k + \Delta k; c) \cdot \Delta m_k \tag{9}$$

where K is the sampling depth, and by using position shifts $(p_k + \Delta k)$ and importance factors Δm_k, the attention mechanism can adaptively focus on regions with high discriminative power.

Task-aware Attention π_C: Finally, we introduce task-aware attention to enhance the learning of task-specific characteristics. It can dynamically open channels to help distinguish different tasks:

$$\pi_C(\mathcal{F}) \cdot \mathcal{F} = \max(\alpha_1^1(\mathcal{F}) \cdot \mathcal{F}_1 + \beta_1^1(\mathcal{F}), \alpha_2^1(\mathcal{F}) \cdot \mathcal{F}_C + \beta_2^1(\mathcal{F})) \tag{10}$$

where $\{\alpha_1^1, \alpha_2^1, \beta_1^1, \beta_2^1\}^T = \theta(\mathcal{F})$ are hyperparameters for parameterizing $\mathcal{F}$. $\max(\cdot)$ is similar to DyReLU.

4.3. ESLoss Function

Datasets of student classroom behavior often suffer from sample imbalance issues. In most cases, students' behaviors are concentrated on simple samples such as read and lookup, resulting in a large number of samples for these two behaviors. In contrast, other difficult samples such as stand, raise hand, and turn head are relatively sparse and easily occluded by other behaviors, leading to poor model training performance. This issue caught our attention. In this work, we designed an improved version of the Slide Loss [43] function to address this problem, which we call ESLoss (EMASlideLoss). The distinction between simple and difficult samples is based on the IoU between the predicted box and the ground truth box. To reduce hyperparameters, the average value of the IoU values of all bounding boxes is used as the threshold μ. Samples with an IoU of less than μ are considered negative samples, while those with an IoU that is greater than μ are considered positive samples.

However, due to unclear classification, samples near the boundary often suffer significant loss. We hope the model can learn to optimize these samples and make full use of them to train the network. However, the number of such samples is relatively small.

Therefore, we attempt to assign higher weights to difficult samples. First, the samples are divided into positive and negative samples using the parameter μ, as shown in Figure 5. Then, the boundary samples are emphasized through the weighting function Slide. The specific process of the Slide weighting function is as follows:

$$f(x) = \begin{cases} 1, & x \leq \mu - 0.1 \\ e^{1-\mu}, & \mu - 0.1 < x < \mu \\ e^{1-x}, & x \geq \mu \end{cases} \tag{11}$$

The function $f(x)$ represents the sliding function operation, where x denotes the Intersection over Union (IoU) between the predicted box and the ground truth box, and μ denotes the weighting threshold. Specifically, the SlideLoss method uses the average IoU of all bounding boxes as the threshold, considering values below μ as negative samples and values above μ as positive samples. In this study, we adopt the concept of Exponential Moving Average (EMA) to optimize the parameter μ of the model. The specific optimization method is as follows:

$$\mu_t = \beta \times \mu_{t-1} + (1 - \beta) \times \theta_t \tag{12}$$

Here, θ_t represents the parameter weight obtained in the t-th update, and μ_t represents the moving average of all parameters in the t-th update. β is the weighting parameter. The moving average can be regarded as the average value over a certain period of time. Compared with exponential functions, the moving average smooths the path of values, reduces jitters, and avoids significant fluctuations caused by occasional outliers. The moving average can enhance the robustness of the current model's performance in detecting student behaviors.

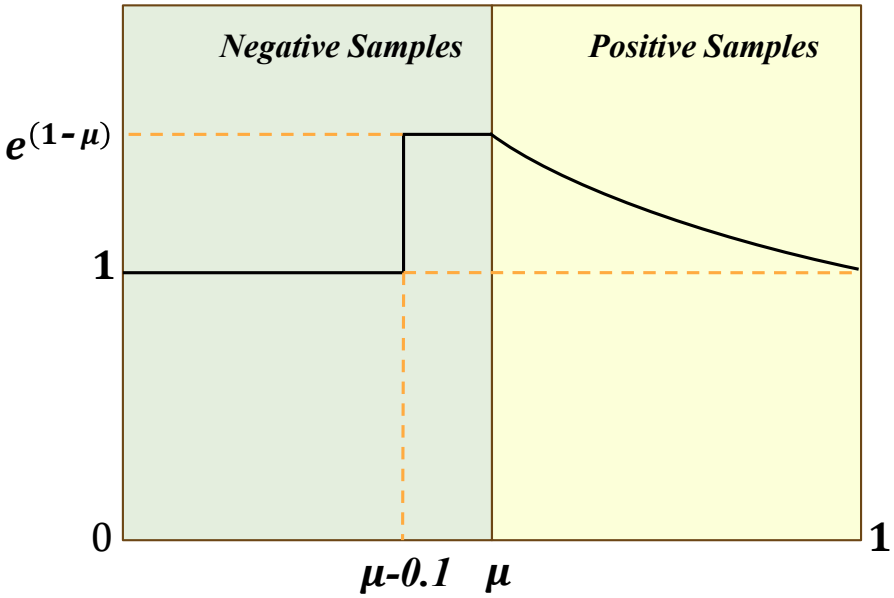

Figure 5. The description of Slide Loss.

5. Experimental Results and Analysis

5.1. Experimental Dataset

In the field of computer vision, the quality and relevance of datasets are crucial for the effective training of models. Researchers across various domains have meticulously constructed numerous datasets to meet specific needs. Our research on classroom behavior recognition similarly requires a corresponding dataset for training and evaluation purposes. Therefore, we decided to utilize our self-constructed SCBehavior dataset and a public dataset SCB-U [44].

5.1.1. SCBehavior Dataset

This dataset comprises 1346 high-resolution images of classroom scenes, each with varying distributions of student behaviors. The SCBehavior dataset encompasses seven different types of student behaviors: read, write, lookup, raise_hand, turn_head, stand, and discuss. These categories essentially cover the typical behaviors observed in daily classroom activities. The diverse distribution of these behaviors across different scenes enhances the dataset's robustness and applicability in real-world scenarios.

In order to maintain the integrity and applicability of the SCBehavior dataset, we conducted a series of rigorous experimental studies on it. Our goal was to unambiguously affirm its value as a powerful educational resource that can significantly contribute to the development of automated systems for student behavior analysis, thereby improving teaching strategies and educational outcomes. Below are some example images from the SCBehavior dataset, illustrating the various student behaviors it captures, as shown in Figure 6.

Figure 6. The seven types of student behaviors are well-documented and displayed, demonstrating that our dataset encompasses a diverse range of classroom environments and student actions. This dataset can be effectively utilized for tasks related to classroom behavior recognition.

More specifically, the number of labels in our dataset and the division of the dataset into training, testing, and validation sets are shown in Table 2 below.

Table 2. SCBehavior Dataset. The dataset includes the number of labels and the division into training, validation, and testing sets.

Number	Behaviors	Labels	Train	Val	Test
1	write	1025	452	491	82
2	read	1075	810	139	126
3	lookup	5725	3620	1656	449
4	turn_head	1025	748	117	160
5	raise_hand	725	561	82	82
6	stand	94	50	30	14
7	discuss	242	172	50	20

The table summarizes the SCBehavior dataset, including the number of labels and the distribution across training, validation, and test sets.

5.1.2. SCB-U Dataset

The SCB-U dataset [44] encompasses six common classroom behaviors: 'raising hand', 'reading', 'writing', 'using phone', 'bowing head', and 'learning'. The SCB-U dataset was compiled from actual classroom surveillance footage and expanded using "frame interpolation" techniques to ensure diversity and precise annotation. Despite the relatively singular

perspective of the dataset images, its high annotation density and accurate simulation of real classroom environments make it invaluable for model training and evaluation.

5.2. Experimental Details

The experiments were conducted in an environment configured with both advanced software and hardware components to ensure optimal performance. The software environment included Ubuntu 20.04 as the operating system, Python 3.8 for programming, PyTorch 1.10.0 as the deep learning framework, and CUDA 11.3 for GPU acceleration. On the hardware side, the setup featured an RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA) with 24 GB of memory, coupled with an AMD EPYC 9654 96-Core Processor (AMD, Santa Clara, CA, USA), providing substantial computational power for efficient data processing and model training.

Setting of HyperParamerters: For the training phase, we initialized the model with a learning rate of 0.01, which was progressively reduced during training, with the final learning rate set to 1% of the initial rate (0.0001). The stochastic gradient descent (SGD) optimizer was employed with a momentum of 0.937 and a weight decay of 5×10^{-4} to help prevent overfitting. A warm-up strategy was applied over the first 3 epochs, during which the momentum started at 0.8, and the bias learning rate was set to 0.1. This warm-up period helped the model stabilize before reaching the main training phase. The batch size was set to 64 to ensure both training efficiency and stability. Additionally, several loss functions were used to enhance detection accuracy: a box loss gain of 7.5, classification loss gain of 0.5, DFL (distribution focal loss) gain of 1.5, and a keypoint object loss gain of 1.0. These carefully tuned hyperparameters ensured robust performance in detecting student behaviors across various classroom settings.

Data Augmentation: We employed several data augmentation techniques during training, including HSV space transformation (Hue ± 1.5%, Saturation ± 70%, Value ± 40%), horizontal and vertical translation (±10%), scaling (±50%), 50% probability of horizontal flip, and Mosaic augmentation applied at 100%. These techniques enhanced the model's robustness to varying target scales, orientations, and class imbalance.

5.3. Evaluation Metrics

To comprehensively evaluate the performance of our model, we employed several key metrics, including mean average precision, precision, recall, and floating point operations (FLOPs). These metrics provide a detailed insight into the effectiveness and efficiency of our model. First, precision measures the accuracy of the positive predictions made by the model. It is defined as in the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

TP (true positive) means the number of correctly detected instances of the target behavior, and FP (false positive) means the number of incorrectly detected instances, i.e., detections that were predicted as the target behavior but are actually not.

Recall measures the model's ability to detect all relevant instances in the dataset. It is defined as the ratio of true positive detections to the total number of actual positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

FN (false negative) means the number of instances that were not detected by the model, i.e., actual instances of the target behavior that the model missed. A high recall indicates that the model is effective at finding all relevant instances, which is important in applications where missing a positive instance can have significant consequences.

We also used mAP to evaluate our model. Mean average precision (mAP) is a standard metric used in object detection tasks to evaluate the accuracy of a model. It considers both

the precision and recall across different thresholds to give a single performance score. Mathematically, the mAP is defined as follows:

$$AP = \int_0^n \text{Precision}\, d(\text{Recall}) \tag{15}$$

$$\text{mAP} = \frac{1}{j} \sum_{i=0}^{j} \text{AP}_i \tag{16}$$

Floating point operations (FLOPs) is a metric used to measure the computational complexity of a model. It indicates the number of floating-point operations required to process a single forward pass through the model. FLOPs is a critical measure of the efficiency of a model, particularly when deploying it in resource-constrained environments such as real-time systems or edge devices.

$$FLOPs = \sum_{l=1}^{L} FLOPs_l \tag{17}$$

For a deep learning model, FLOPs can be computed by summing the number of multiplications and additions performed in each layer. Lower FLOPs generally imply a faster and more efficient model, which is essential for real-time applications like classroom behavior recognition.

5.4. Baselines

To validate the effectiveness of our model, we compare it with several baseline models. These baseline models include a number of detection algorithms that are widely recognized in academia and industry, and this comparative analysis allows us to comprehensively assess the performance of our model in different environments and tasks to ensure its usefulness and reliability in student behavior recognition applications.

YOLOv5 [17]: YOLOv5 is the fifth iteration of the YOLO (You Only Look Once) family, known for its real-time object detection capabilities. It is designed to perform both detection and classification in a single pass through the network, making it highly efficient. YOLOv5 improves upon its predecessors with better accuracy, speed, and a more modular architecture. YOLOv5 uses CSP-Darknet53 as its backbone, which integrates Cross Stage Partial (CSP) connections to improve gradient flow and reduce computational cost. And it also employs a Path Aggregation Network (PANet) in the neck, which enhances information flow between layers and helps in detecting objects at various scales. It offers several variants like YOLOv5s (small), YOLOv5m (medium), YOLOv5l (large), and YOLOv5x (extra-large), allowing users to choose based on their specific needs and computational resources. In our study, we chose YOLOv5s as one of the baseline models because YOLOv5-s is a smaller and more efficient variant of the YOLOv5 model family, designed for real-time target detection, with a similar number of parameters and computational effort to our proposed model, which allows for a better comparison.

YOLOv7 [45]: YOLOv7 introduces several architectural innovations and enhancements that improve its performance compared to its predecessors. Developed by the original creators of YOLO, it focuses on maintaining a balance between speed and detection accuracy, making it suitable for a wide range of applications. First, it introduces ELAN to improve the network's ability to learn complex features by integrating more efficient feature aggregation strategies. And YOLOv7 also integrates RepVGG blocks, which streamline the network architecture, reduce memory usage, and improve inference speed without sacrificing accuracy. We also chose the yolov7-tiny version, which has a similar amount of parameters to our model, as one of the baseline models.

SSD [32]: SSD is a powerful and efficient object detection model that strikes a good balance between speed and accuracy. Its one-shot detection method, use of multi-scale feature maps, and real-time performance make it a popular choice for many real-world

applications, so we chose it as the baseline model to compare with ours, and despite some challenges in achieving the highest accuracy for small or dense objects, it is still a fairly important representative model.

Faster R-CNN [31]: Faster R-CNN consists of a two-stage approach where the first stage generates region proposals using a Region Proposal Network, and the second stage classifies these proposals and refines their bounding boxes. This architecture enables Faster R-CNN to achieve a high detection accuracy, particularly in complex scenes. In the context of classroom student behavior recognition, Faster R-CNN was selected as one of the baseline models for comparison due to its proven effectiveness in detecting and localizing objects accurately.

Deformable-DETR [46]: Deformable-DETR is an advanced object detection model that builds upon the original DETR framework by incorporating deformable attention mechanisms. It addresses some of the limitations of standard DETR, such as slow convergence and difficulties in handling small objects, by enabling more efficient and flexible feature extraction. In the context of classroom student behavior recognition, Deformable-DETR is chosen as one of the baseline models due to its superior capability to capture detailed and varied student actions and postures. By comparing our model with Deformable-DETR, we can assess our model's effectiveness in handling complex and dynamic student behaviors, while also evaluating its performance in terms of computational efficiency and real-time application potential in classroom settings.

DETR with Improved deNoising anchor boxes (DINO) [47]: DINO is a significant enhancement of the original DETR (Detection Transformer) framework, focusing on improving the convergence speed and accuracy of object detection tasks. DINO introduces denoising techniques and enhanced anchor box mechanisms to address the slow convergence and small-object detection challenges commonly associated with DETR models. By refining the object query process and improving the training efficiency, DINO achieves faster and more accurate detection, particularly in scenarios involving complex and dynamic environments. In the context of student behavior recognition, DINO is included as a baseline model due to its robustness in handling diverse and nuanced student actions, offering a strong comparison point for our model's performance in both accuracy and detection efficiency.

EfficientNet [48]: EfficientNet utilizes a compound scaling method that balances network depth, width, and resolution. It achieves state-of-the-art performance while keeping the number of parameters and computational cost relatively low. Its highly optimized architecture allows for real-time inference, making it a strong baseline for tasks requiring speed and accuracy. In our study, we included EfficientNet as a baseline model because its balance of performance and computational efficiency is comparable to the goals of our proposed student behavior detection system, allowing for a meaningful evaluation of detection capabilities in resource-constrained environments like classrooms.

RTMDet [49]: RTMDet is a lightweight object detection model designed for high efficiency and speed, and it is optimized for real-time detection tasks. Its architecture incorporates several innovations, including efficient feature extraction modules and optimized anchor-free detection strategies, which reduce computational complexity while maintaining a strong detection accuracy. In our comparison, RTMDet serves as a baseline model because its real-time detection focus aligns closely with our model's objectives, allowing us to assess both the speed and detection accuracy of our approach in dynamic educational settings.

5.5. Performance

5.5.1. Comparison Study on SCBehavior Dataset

Based on the data presented in Table 3 and Figure 7, it is evident that SBD-Net outperforms the compared models in terms of precision and recall, achieving the highest values of 0.804 and 0.763, respectively. SBD-Net also achieves notable mAP@0.5 and mAP@0.5-0.95 of 0.824 and 0.619, demonstrating its effectiveness in accurately detecting and classifying student behaviors in classroom settings. The integration of focal modulation mechanisms

and the incorporation of the DyHead module enhance SBD-Net's ability to detect small objects and handle complex patterns of student behavior. To be specific, the implementation of the DyHead module contributes to the model's superior feature extraction capabilities, enabling it to process high-resolution feature maps while maintaining a manageable number of parameters. SBD-Net's parameter count of 36.5 M strikes a balance between model complexity and computational efficiency, making it suitable for deployment in resource-constrained environments such as classroom cameras. Furthermore, the modification of the loss function to ES Loss improves the model's robustness and accuracy, particularly in handling imbalanced datasets often encountered in real-world classroom scenarios. And in terms of detection speed, SBD-Net demonstrates competitive performance with a FLOPs value of 9.8 G, making it an efficient choice for real-time applications. This efficiency is crucial for practical deployments where both high accuracy and low latency are required. The model's ability to achieve high precision and recall while maintaining computational efficiency underscores its potential for widespread adoption in educational settings for monitoring and analyzing student behavior.

Figure 7. From left to right are the original image, the YOLOv8 detection image, and the SBD-Net detection image. We can intuitively observe that SBD-Net achieves better results while assigning more weights to the difficult samples, and it obtains better results.

Overall, SBD-Net's enhancements, including focal modulation, the DyHead module, and the ES Loss function, contribute to its superior performance in classroom student behavior recognition tasks. Its ability to accurately detect and classify various student behaviors with high precision and recall, combined with its efficient computational requirements, makes SBD-Net a robust and practical solution for educational applications.

Table 3. Comparison of different models for classroom student behavior recognition.

Model	Precision	Recall	mAP@0.5 (%)	mAP@0.5-0.95 (%)	Params (M)	Flop (G)
YOLOv5 [17]	0.741	0.642	0.651	0.486	13.9 M	15.9
YOLOv7 [45]	0.589	0.576	0.543	0.348	11.7 M	13.2
YOLOv8 [50]	0.799	0.725	0.781	0.595	30.0 M	**8.9**
SSD [32]	0.767	0.733	0.736	0.501	24.5 M	-
Faster-RCNN [31]	0.787	0.687	0.804	0.576	41.7 M	243
Deformable-DETR [46]	0.698	0.722	0.654	0.544	41.1 M	284
DINO [47]	0.773	0.716	0.766	0.564	47.0 M	279
RTMDet [49]	0.793	0.696	0.755	0.511	**8.99 M**	14.8
EfficientNet [48]	0.797	0.716	0.750	0.535	-	-
SBD-Net	**0.804**	**0.763**	**0.824**	**0.619**	36.5 M	9.8

Bold indicates the best performance, while underline represents the second-best performance.

5.5.2. Comparison Study on SCB-U Dataset

To validate the scalability of our method and reduce experimental bias, we introduced the other dataset specifically designed for senior student behavior in classrooms, named SCB-U [44]. This dataset is tailored to high school environments under the K12 educational framework and includes six common and representative student behaviors: 'raising_hand', 'reading', 'writing', 'using_phone', 'bowing_head', and 'learning' (studying at a desk). These behaviors are frequently observed in typical classroom settings, making the dataset highly relevant for research in this field.

The SCB-U dataset is constructed using footage captured from real classroom surveillance, ensuring the authenticity and relevance of the data. To expand the dataset, we employed advanced frame interpolation techniques. This not only increases the diversity of the data but also mitigates potential overfitting issues caused by sparse data, enhancing the generalization ability of our model. Each behavior sample is meticulously annotated, ensuring high annotation density and accuracy, which provides a solid foundation for training and evaluating deep learning models.

Although the dataset is captured from a fixed camera angle, its precise simulation of real classroom scenarios and high-quality annotations make it particularly valuable for model training. The SCB-U dataset not only enriches our research but also effectively tests the robustness and accuracy of the model in handling diverse student behaviors within dynamic classroom settings. This allows for our approach to be applicable in a wide range of real-world educational environments and serves as a strong benchmark for future studies in the field.

As shown in Table 4, the experimental results on the SCB-U dataset demonstrate the robustness and effectiveness of our proposed SBD-Net model. SBD-Net consistently outperformed several state-of-the-art models, including YOLOv5, YOLOv7, YOLOv8, and DINO, across key metrics such as precision, recall, and mAP scores. Notably, SBD-Net achieved the highest mAP@0.5 and mAP@0.5-0.95, reflecting its superior ability to accurately detect and classify students' behaviors in diverse classroom scenarios.

Table 4. Performance on the SCB-U dataset for SBD-Net and other models.

Model	Precision	Recall	mAP@0.5 (%)	mAP@0.5-0.95 (%)	Params (M)	Flop (G)
YOLOv5	0.782	0.524	0.566	0.401	13.9 M	15.9
YOLOv7	0.911	0.57	0.714	0.551	11.7 M	13.2
YOLOv8	0.855	0.705	0.74	0.57	30 M	**8.9**
RTMDet	0.746	0.557	0.679	0.427	**8.99 M**	14.8
DINO	0.821	**0.746**	0.716	0.560	47 M	279
EfficientNet	0.838	0.614	0.744	0.527	-	-
SBD-Net	**0.868**	0.701	**0.745**	**0.577**	36.5 M	9.8

Bold indicates the best performance, while underline represents the second-best performance.

The model's high precision (0.868) and balanced recall (0.701) indicate that SBD-Net not only minimizes false positives but also reliably detects a wide range of student behaviors, even in challenging conditions such as occlusion or subtle gestures, as shown in Figure 8. The combination of advanced data augmentation techniques, meticulous annotations, and the integration of task-specific modules in SBD-Net significantly contributed to its enhanced performance. Moreover, the model's ability to generalize well across different classroom settings, despite being trained on a dataset with a fixed camera angle, highlights its robustness and adaptability. This further confirms SBD-Net's potential to be deployed in real-world educational environments, where diverse behaviors and interactions need to be accurately monitored and analyzed. Overall, the consistent performance across metrics reaffirms the effectiveness of SBD-Net as a highly capable model for student behavior detection and classification.

Figure 8. Performance on SCB-U dataset. (The Chinese text in the top left corner represents the recording time, while the text in the bottom right corner represents the recording location).

5.5.3. Ablation Study

After conducting several tests, we obtained the ablation experiment results shown in Table 5 above. Compared with the benchmark model YOLOv8n, our method shows significant improvements in various metrics.

First, the base YOLOv8n model, without any modifications, achieved a precision of 0.799, recall of 0.725, mAP@0.5 of 0.781, and mAP@0.5-0.95 of 0.595, with 30 M parameters and a computational complexity of 8.9 G FLOPs.

Table 5. Ablation experiment.

Model	Precision	Recall	mAP@0.5 (%)	mAP@0.5-0.95 (%)	Params (MB)	Flop (G)
YOLOv8n	0.799	0.725	0.781	0.595	30 M	8.9
v8n + FM	0.792	0.731	0.799	0.603	31.8 M	8.2
v8n + FM + Dyhead	0.783	0.737	0.808	0.599	36.5 M	9.6
Our Method	**0.804**	**0.763**	**0.824**	**0.619**	36.5 M	9.8

Bold indicates the best performance.

When we introduced the Focal Modulation (FM) module, replacing the SPPF module in YOLOv8, the mAP@0.5 increased by 1.8% to 0.799, and the mAP@0.5-0.95 improved by 0.8% to 0.603. The parameter count increased to 31.8 M, and the computational complexity reduced to 8.2 G FLOPs. This indicates that the Focal Modulation module enhances recall and detection performance while maintaining a lower computational complexity. Further, we incorporated the Dyhead detection head, which combines multiple attention mechanisms to enhance the detection capability of complex behaviors. With Dyhead, the precision reached 0.783, a decrease of 1.6% compared to YOLOv8n, but the recall increased by 1.2% to 0.737. The mAP@0.5 was 0.808, an increase of 2.7%, and mAP@0.5-0.95 slightly increased by 0.4% to 0.599. The parameter count rose to 36.5 M, and the computational

complexity was 9.6 G FLOPs. The inclusion of the Dyhead detection head significantly improved the model's behavior detection performance in complex environments. Finally, in our complete method, in addition to the aforementioned improvements, we optimized the classification loss function by adopting the improved ESLoss function, as shown in Figure 9. The optimized model achieved the best results: precision increased by 0.5% to 0.804, recall increased by 3.8% to 0.763, mAP@0.5 improved by 4.3% to 0.824, and mAP@0.5-0.95 improved by 2.4% to 0.619. Although the parameter count increased to 36.5 M and the computational complexity was 9.8 G FLOPs, the overall performance of the model improved significantly.

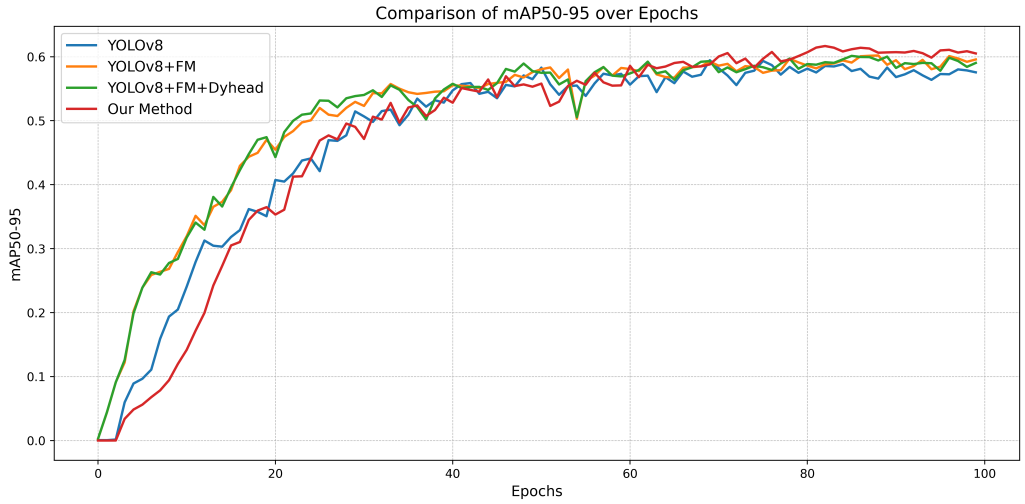

Figure 9. The graph illustrates the comparison of mAP50-95 over epochs for different methods, including YOLOv8, YOLOv8 with Focal Modulation (FM), YOLOv8 with FM and Dyhead, and our proposed method. The mAP50-95 metric, which indicates the mean Average Precision across various IoU thresholds, is used to evaluate the performance of the models, and it shows that incorporating Focal Modulation and dyhead while optimizing the loss function significantly improve the detection performance of the model over time.

These improvements validate the importance of each module in enhancing the model's performance. In particular, the optimization of the classification loss function greatly enhanced the model's precision and recall, proving the effectiveness of our method.

We used a confusion matrix to demonstrate that SBD-Net effectively detects various student behaviors with high accuracy in key categories, as shown in Figure 10. The model excels in recognizing 'write', 'lookup', and 'discuss' behaviors, which are critical for understanding student engagement and classroom dynamics. These high accuracies highlight the model's robustness in identifying distinct and prominent classroom activities. However, there are areas for improvement, particularly in distinguishing 'read', 'turn_head', and 'stand' behaviors from similar actions. The moderate performance in these categories suggests that the model faces challenges in differentiating these behaviors due to their visual similarities with other actions, such as 'lookup' and 'raise_hand'. Enhancing feature extraction and incorporating more training samples for these specific behaviors could help in reducing the confusion and improving the overall detection accuracy. Overall, SBD-Net shows promising results, but further refinements are needed to achieve consistently high accuracy across all behavior categories.

In the training process of machine learning and deep learning models, the performance of the loss function is an important indicator of the training effect of the model, as shown in Figure 11. In this paper, we analyze the effect of the optimized ES Loss and the classification loss (cls loss) that comes with YOLOv8 by comparing the changes in the two over 100 training cycles (epochs). The graphs show the trend of the two loss functions, where the red dashed line represents the optimized ES Loss and the blue realization represents

the original classification loss of YOLOv8. We can clearly see that the optimized ES Loss shows obvious advantages in both the stability of model training and the final results, and through further optimization and tuning, the ES Loss has the potential to show stronger competitiveness and better classification performance in more practical applications.

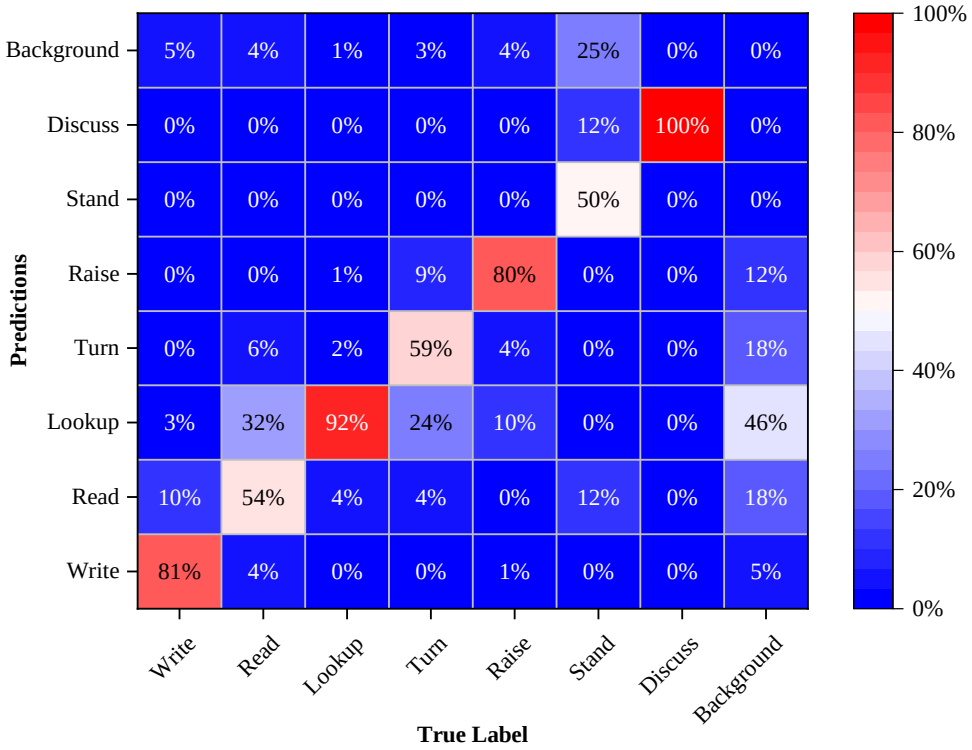

Figure 10. The confusion matrix of SBD-Net in the analysis of different behavioral detection effects.

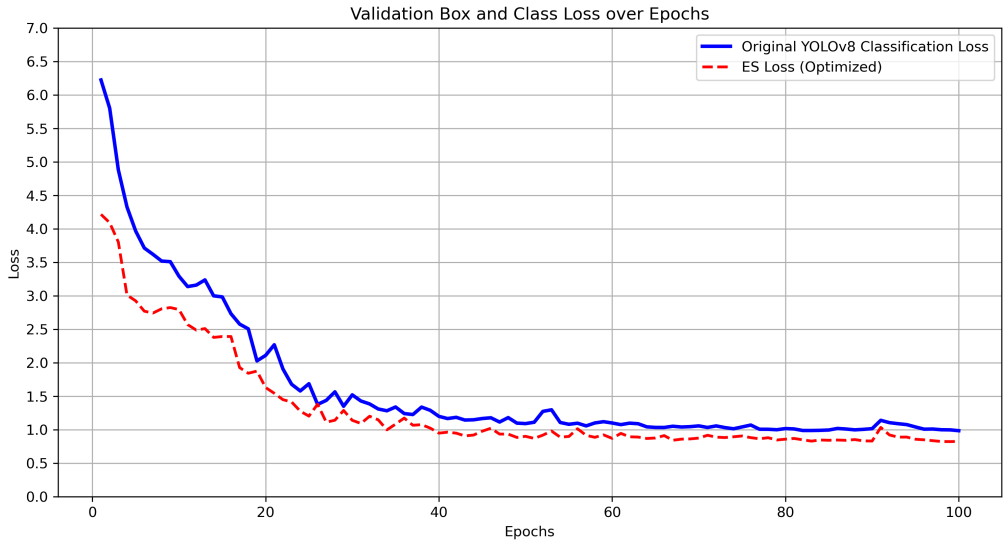

Figure 11. Comparison of loss functions.

5.5.4. Discussion

In our research, to enhance the model's performance in detecting small targets, such as the heads of students seated at the back, we incorporated a combination of the Focal Modulation module and the Scale-aware Attention from the Dyhead module. This integration significantly boosts the model's capability in feature extraction and fusion under complex classroom environments, thereby enhancing the detection accuracy for occlusion and small

objects. Specifically, Focal Modulation employs a multi-level feature fusion mechanism that aggregates context information from short to long distances across different positions in the feature map. Through deep convolutional encoding, the FM module focuses and merges features from various scales at each query location, capturing fine local details. Subsequently, when the model makes inputs into the Dyhead module, its Scale-aware Attention dynamically adjusts the importance of different scale features. Using 1×1 convolution operations, Dyhead's Scale-aware Attention allocates attention weights based on the prominence of different scales in the feature maps, enabling the model to focus more on extracting features of occlusion and small objects.

To validate the scalability of our proposed method in student behavior detection, we conducted experiments using two public datasets. The SCBehavior dataset includes younger students in K-12 education, while SCB-U involves older students. The results demonstrate that our proposed method exhibits strong scalability in recognizing student behaviors across different grade levels.

6. Conclusions

In this paper, we propose SBD-Net, a lightweight and efficient model for detecting students' behaviors in classroom environments. By leveraging advanced computer vision techniques, including the FocalModulation module for multi-level feature fusion, the ESLoss loss function for addressing sample imbalance, and the Dyhead structure to incorporate multiple attention mechanisms without increasing computational complexity, SBD-Net excels in the real-time monitoring and analysis of students' behaviors. These innovations enable teachers to adapt their instructional strategies more effectively, improving the overall quality of education.

Our experimental results on the SCBehavior and SCB-U datasets demonstrate the superiority of SBD-Net. The model achieved a mAP of 0.824 on SCBehavior, outperforming the baseline model YOLOv8 by 4.3%, while maintaining a low computational complexity of 9.8 GFLOPs. In addition, it demonstrated a 3.8% improvement in recall, highlighting its ability to handle unevenly distributed behaviors and perform high-precision detection in complex classroom scenarios. SBD-Net also showed strong performance on the newly introduced SCB-U dataset, validating its generalizability to different classroom environments.

However, SBD-Net has some limitations. It is primarily designed for specific classroom settings, which may limit its adaptability to varied educational contexts. Additionally, while its computational demands have been reduced, further optimization is needed for its deployment on low-power devices. Moreover, the current focus on overt behaviors overlooks subtler psychological and emotional states, which require further exploration.

Future research will focus on optimizing the model's architecture for enhanced accuracy and efficiency. Expanding the dataset to cover a wider range of classroom scenarios, age groups, and cultural backgrounds will further enhance SBD-Net's generalizability. Additionally, exploring multimodal data fusion by integrating visual, audio, and textual inputs will provide a more holistic analysis of student behaviors and emotional states. Techniques such as model compression, quantization, and edge computing will be explored to enable efficient deployment on resource-constrained devices. Collaborating with educators to refine behavior classification and ensure practical applicability will be a key focus, ultimately contributing to the development of intelligent educational systems and the widespread adoption of smart classrooms.

Author Contributions: Conceptualization, Z.W. and C.Z.; methodology, Z.W.; software, Z.W. and C.Z.; validation, Z.W., M.W., L.L., and C.Z.; formal analysis, Z.W.; investigation, Z.W., M.W., L.L., and C.Z.; resources, Z.W.; data curation, Z.W.; writing—original draft preparation, Z.W. and L.L.; writing—review and editing, Z.W. and M.W.; visualization, Z.W. and M.W.; supervision, Z.W.; project administration, Z.W.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research work in this paper was supported by the National Natural Science Foundation of China (No. 62177022, 61901165, 61501199), Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU24JC033), AI and Faculty Empowerment Pilot Project (No. CCNUAI&FE2022-03-01), Collaborative Innovation Center for Informatization and Balanced Development of K-12 Education by MOE and Hubei Province (No. xtzd2021-005), and Natural Science Foundation of Hubei Province (No. 2022CFA007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this paper:

SBD-Net	Student Behavior Detection Network
mAP	mean Average Precision
SCBehavior	Student Classroom Behavior
AI	Artificial Intelligence
NLP	Natural Language Processing
IoT	Internet of Things
HOG	Histogram of Oriented Gradients
CNNs	Convolutional Neural Networks
R-CNN	Region-based CNN
RoI	Region of Interest
RPN	Region Proposal Network
SSD	Single Shot MultiBox Detector
YOLO	You Only Look Once
IoU	Intersection over Union
RNNs	Recurrent Neural Networks
CBAM	Convolutional Block Attention Module
SENet	Squeeze-and-Excitation Networks
EMA	Exponential Moving Average
SPPF	Spatial Pyramid Pooling Fast
FLOPs	Floating Point Operations
CSP	Cross Stage Partial
PANet	Path Aggregation Network
FM	Focal Modulation

References

1. Messeri, L.; Crockett, M.J. Artificial Intelligence and Illusions of Understanding in Scientific Research. *Nature* **2024**, *627*, 49–58. [CrossRef] [PubMed]
2. Wang, Z.; Yao, J.; Zeng, C.; Li, L.; Tan, C. Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network. *Systems* **2023**, *11*, 372. [CrossRef]
3. Uttl, B. Student Evaluation of Teaching (SET): Why the Emperor Has No Clothes and What We Should Do About It. *Hum. Arenas* **2024**, *7*, 403–437. [CrossRef]
4. Li, D. An Interactive Teaching Evaluation System for Preschool Education in Universities Based on Machine Learning Algorithm. *Comput. Hum. Behav.* **2024**, *157*, 108211. [CrossRef]
5. Awidi, I.T.; Paynter, M. An Evaluation of the Impact of Digital Technology Innovations on Students' Learning: Participatory Research Using a Student-Centred Approach. *Technol. Knowl. Learn.* **2024**, *29*, 65–89. [CrossRef]
6. Wang, Z.; Yan, W.; Zeng, C.; Tian, Y.; Dong, S. A Unified Interpretable Intelligent Learning Diagnosis Framework for Learning Performance Prediction in Intelligent Tutoring Systems. *Int. J. Intell. Syst.* **2023**, *2023*, e4468025. [CrossRef]
7. Lin, C.C.; Huang, A.Y.Q.; Lu, O.H.T. Artificial Intelligence in Intelligent Tutoring Systems toward Sustainable Education: A Systematic Review. *Smart Learn. Environ.* **2023**, *10*, 41. [CrossRef]
8. Zhong, X.; Zhan, Z. An Intelligent Tutoring System for Programming Education Based on Informative Tutoring Feedback: System Development, Algorithm Design, and Empirical Study. *Interact. Technol. Smart Educ.* 2024, *ahead-of-print*. [CrossRef]

9. Ramadhan, A.; Warnars, H.L.H.S.; Razak, F.H.A. Combining Intelligent Tutoring Systems and Gamification: A Systematic Literature Review. *Educ. Inf. Technol.* **2024**, *29*, 6753–6789. [CrossRef]
10. Chen, Y.; Zhai, L. A comparative study on student performance prediction using machine learning. *Educ. Inf. Technol.* **2023**, *28*, 12039–12057.
11. Al-Azazi, F.A.; Ghurab, M. ANN-LSTM: A deep learning model for early student performance prediction in MOOC. *Heliyon* **2023**, *9*, e15382. [PubMed]
12. Zhao, J.; Zhu, H. CBPH-Net: A Small Object Detector for Behavior Recognition in Classroom Scenarios. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2521112. [CrossRef]
13. Mahapatra, S.K.; Pattanayak, B.K.; Pati, B.; Laha, S.R.; Pattnaik, S.; Mohanty, B. An IoT Based Novel Hybrid-Gamified Educational Approach to Enhance Student's Learning Ability. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 374–393.
14. Daumiller, M.; Janke, S.; Hein, J.; Rinas, R.; Dickhäuser, O.; Dresel, M. Teaching quality in higher education: Agreement between teacher self-reports and student evaluations. *Eur. J. Psychol. Assess.* **2023**, *39*, 176.
15. Mertens, D.M. *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods*; Sage Publications: Thousand Oaks, CA, USA, 2023.
16. Dimitriadou, E.; Lanitis, A. A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learn. Environ.* **2023**, *10*, 12.
17. Wang, Z.; Yao, J.; Zeng, C.; Wu, W.; Xu, H.; Yang, Y. YOLOv5 Enhanced Learning Behavior Recognition and Analysis in Smart Classroom with Multiple Students. In Proceedings of the 2022 International Conference on Intelligent Education and Intelligent Research (IEIR), Wuhan, China, 18–20 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 23–29. [CrossRef]
18. Zhou, J.; Ran, F.; Li, G.; Peng, J.; Li, K.; Wang, Z. Classroom Learning Status Assessment Based on Deep Learning. *Math. Probl. Eng.* **2022**, *2022*, e7049458. [CrossRef]
19. Yin Albert, C.C.; Sun, Y.; Li, G.; Peng, J.; Ran, F.; Wang, Z.; Zhou, J. Identifying and Monitoring Students' Classroom Learning Behavior Based on Multisource Information. *Mob. Inf. Syst.* **2022**, *2022*, e9903342. [CrossRef]
20. Sharma, P.; Joshi, S.; Gautam, S.; Maharjan, S.; Khanal, S.R.; Reis, M.C.; Barroso, J.; de Jesus Filipe, V.M. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In Proceedings of the International Conference on Technology and Innovation in Learning, Teaching and Education, Lisbon, Portugal, 31 August–2 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 52–68.
21. Jisi, A.; Yin, S. A new feature fusion network for student behavior recognition in education. *J. Appl. Sci. Eng.* **2021**, *24*, 133–140.
22. Wang, Z.; Li, L.; Zeng, C.; Yao, J. Student learning behavior recognition incorporating data augmentation with learning feature representation in smart classrooms. *Sensors* **2023**, *23*, 8190. [CrossRef]
23. Shi, L.; Di, X. A recognition method of learning behaviour in English online classroom based on feature data mining. *Int. J. Reason.-Based Intell. Syst.* **2023**, *15*, 8–14.
24. Cao, Y.; Liu, D. Optimization of Student Behavior Detection Algorithm Based on Improved SSD Algorithm. *Optimization* **2024**, *15*, 104.
25. Li, Y.; Qi, X.; Saudagar, A.K.J.; Badshah, A.M.; Muhammad, K.; Liu, S. Student behavior recognition for interaction detection in the classroom environment. *Image Vis. Comput.* **2023**, *136*, 104726.
26. Chen, H.; Zhou, G.; Jiang, H. Student behavior detection in the classroom based on improved YOLOv8. *Sensors* **2023**, *23*, 8385. [CrossRef]
27. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Kauai, HI, USA, 8–14 December 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, p. I.
28. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
29. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
30. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; NeurIPS: La Jolla, CA, USA, 2015; Volume 28.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016, pp. 21–37.
33. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
34. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.

35. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
36. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
37. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NeurIPS: La Jolla, CA, USA, 2017; Volume 30.
41. Yang, J.; Li, C.; Dai, X.; Gao, J. Focal modulation networks. In *Advances in Neural Information Processing Systems*; NeurIPS: La Jolla, CA, USA, 2022; Volume 35, pp. 4203–4217.
42. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
43. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *arXiv* **2022**, arXiv:2208.02019 .
44. Yang, F.; Wang, T. Scb-dataset3: A benchmark for detecting student classroom behavior. *arXiv* **2023**, arXiv:2310.02522.
45. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
46. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
47. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
48. Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
49. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784.
50. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]