systems

MDPI

Article

Students' Classroom Behavior Detection System Incorporating Deformable DETR with Swin Transformer and Light-Weight Feature Pyramid Network

Zhifeng Wang [1,*], Jialong Yao [1], Chunyan Zeng [2,*], Longlong Li [1] and Cheng Tan [3]

[1] CCNU Wollongong Joint Institute, Central China Normal University, Wuhan 430079, China
[2] Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China
[3] School of Marxism, Jilin University, Changchun 130015, China
* Correspondence: zfwang@ccnu.edu.cn (Z.W.); cyzeng@hbut.edu.cn (C.Z.)

Abstract: Artificial intelligence (AI) and computer vision technologies have gained significant prominence in the field of education. These technologies enable the detection and analysis of students' classroom behaviors, providing valuable insights for assessing individual concentration levels. However, the accuracy of target detection methods based on Convolutional Neural Networks (CNNs) can be compromised in classrooms with multiple targets and varying scales, as convolutional operations may result in the loss of location information. In contrast, transformers, which leverage attention mechanisms, have the capability to learn global features and mitigate the information loss caused by convolutional operations. In this paper, we propose a students' classroom behavior detection system that combines deformable DETR with a Swin Transformer and light-weight Feature Pyramid Network (FPN). By employing a feature pyramid structure, the system can effectively process multi-scale feature maps extracted by the Swin Transformer, thereby improving the detection accuracy for targets of different sizes and scales. Moreover, the integration of the CARAFE lightweight operator into the FPN structure enhances the network's detection accuracy. To validate the effectiveness of our approach, extensive experiments are conducted on a real dataset of students' classroom behavior. The experimental results demonstrate a significant 6.1% improvement in detection accuracy compared to state-of-the-art methods. These findings highlight the superiority of our proposed network in accurately detecting and analyzing students' classroom behaviors. Overall, this research contributes to the field of education by addressing the limitations of CNN-based target detection methods and leveraging the capabilities of transformers to improve accuracy. The proposed system showcases the benefits of integrating deformable DETR, Swin Transformer, and the lightweight FPN in the context of students' classroom behavior detection. The experimental results provide compelling evidence of the system's effectiveness and its potential to enhance classroom monitoring and assessment practices.

Keywords: students' classroom behavior; Swin Transformer; feature pyramid network; learning assessment

Citation: 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. https://doi.org/10.3390/systems11070372

Academic Editors: Shixuan Fu, Bo Yang and Alex Zarifis

Received: 29 May 2023
Revised: 15 July 2023
Accepted: 16 July 2023
Published: 20 July 2023

1. Introduction

The field of computer vision has undergone significant transformations due to rapid advancements in deep learning and artificial intelligence technologies [1]. These advancements have enabled swift and precise recognition capabilities, which can be harnessed to capture and analyze students' behavior in educational environments [2]. Such analysis facilitates more effective quantitative assessments of student concentration in the classroom, serving as a valuable metric for evaluating classroom quality.

Traditional methods for assessing student concentration involve human assessment with teacher involvement [3], post-class question-answering tests [4], and brainwave testing [5]. However, these approaches have inherent limitations. Monitoring students'

concentration through classroom teachers is inefficient and fails to provide continuous monitoring of individual students. Questionnaires and post-class assignments are not comprehensive enough, and the grading criteria can be subjective. The use of brainwave sensors, although capable of real-time detection, may disrupt the classroom environment [6]. Another contemporary approach involves evaluating student concentration through the recognition of speech signals, but it lacks real-time interaction between students and instructors during class [7].

In comparison, computer-vision-based methods for student concentration evaluation offer the characteristics of continuous monitoring, objective evaluation, and real-time interaction, making them more convenient for assessing student concentration. However, computer-vision-based detection methods encounter various challenges, including background clutter, obscured student poses, variations in room lighting, multi-point perspectives, and image distortion. To address these challenges, several solutions have been proposed, such as multi-angle optimization [8], video summarization [9], density estimation, and scale-aware context-awareness [10], aiming to enhance the optimization of camera-captured images. Deep learning algorithms, including Convolutional Neural Networks (CNNs) and variants like ResNet [11] and You Only Look Once (YOLO) [12–14], have found extensive applications in target detection within the field of computer vision. While the use of CNNs in educational environments has been explored, transformer algorithms have demonstrated outstanding performance in the vision domain, signaling a new phase of the technological revolution in target detection. However, transformer-based target detection technology has not yet been implemented in educational environments.

Existing target detection algorithms can identify students in images and videos, but they often struggle with missed and false detections in scenarios involving multiple targets. The transformer-based vision transformer excels in image classification tasks, but falls short in feature extraction and learning for local image features when applied to downstream tasks like target detection and image segmentation.

To overcome these limitations, this paper proposes a novel neural network that combines the strengths of the Swin Transformer and the encoder–decoder structure for object classification, specifically targeting students' classroom behavior detection. Our approach leverages the powerful performance of the Swin Transformer as a backbone network within the Deformable DETR framework, enhancing the detection capabilities in classroom environments. Additionally, we introduce the Feature Pyramid Network (FPN) structure to fuse feature maps obtained from the Swin Transformer at different scales, enabling the extraction of robust top-down semantic features. Furthermore, we incorporate the CARAFE lightweight operator [15] to enhance the receptive field of the FPN network during feature vector rearrangement, utilizing input features to guide the reorganization process. To evaluate the performance of our proposed model, we conducted extensive experiments and comparisons with current mainstream target detectors. We employed various optimization methods to assess the model's detection capabilities. The experimental section provides a detailed account of the experimental setup, including the dataset used, evaluation metrics, and a comprehensive analysis of the results. The adoption of these approaches in our proposed model resulted in a notable 6.1% increase in the mean Average Precision (mAP) value, demonstrating the effectiveness of combining the transformer-based architecture with the encoder–decoder structure in students' classroom behavior detection. This paper makes the following main contributions:

1. We propose a novel neural network that combines the powerful performance of the Swin Transformer as a backbone network with the benefits of an encoder–decoder structure in object classification. The Swin Transformer serves as the backbone network in the Deformable DETR framework, providing enhanced capabilities for students' classroom behavior detection and analysis. The source code of this study is publicly available at https://github.com/CCNUZFW/Student-behavior-detection-system (accessed on 1 May 2023).

2. We propose the feature pyramid network (FPN) structure, which effectively fuses feature maps obtained from the Swin Transformer at four different scales: large scale, medium scale, small scale, and extremely small scale. This integration enables the extraction of robust top-down semantic features, leading to improved accuracy in detecting and analyzing students' classroom behavior.
3. We introduce the CARAFE lightweight operator to enhance the receptive field of the FPN network during feature vector rearrangement. By utilizing input features to guide the reorganization process, the CARAFE operator further improves the precision and effectiveness of the student's classroom behavior detection system.
4. The development of a dedicated dataset naming ClaBehavior for detecting students' classroom behavior is a significant contribution of our study. Having access to reliable and annotated datasets is essential for the development and evaluation of machine learning models. Our ClaBehavior dataset, which comprises a diverse collection of classroom images with behavior annotations, addresses a gap in the existing literature and serves as a valuable resource for future research in the field of students' classroom behavior detection. The dataset from our study is publicly available at https://github.com/CCNUZFW/Student-behavior-detection-system/tree/master/dataset/coco (accessed on 1 May 2023).

The remaining sections of the paper are organized as follows: Section 2 provides an overview of the students' classroom behavior detection system and discusses recent advancements in target detection. In Section 3, we present a detailed description of the problem addressed in this paper and demonstrate the network structure and methodologies employed in this study, including the composition of the network structure. The experimental section, presented in Section 4, includes a comparison of our proposed approach with current mainstream target detectors. Various optimization methods are also employed to evaluate the model's detection performance. In Section 5, we systematically discuss the proposed students' classroom behavior detection system and experimental results. Finally, Section 6 concludes the paper and discusses potential future research directions.

2. Related Work

This section discusses existing methods for recognizing students' classroom behavior and provides a comparison between CNNs and transformer structures in terms of their approaches, advantages, and disadvantages.

2.1. Students' Classroom Behavior Detection

Recognizing students' classroom behavior has been a focal point of research aimed at enhancing the quality of teaching and learning [16]. Various approaches have been proposed to improve the accuracy of target detection in this context.

For instance, Lv et al. [17] improved the accuracy of recognizing students' classroom behavior by introducing a feature pyramid into the SSD algorithm. Ren et al. [18] enhanced the feature extraction of the YOLOv4 network by adopting a structure with jumping paths for feature extraction and combining top-down and bottom-up approaches. Tang et al. [19] employed a weighted bi-directional feature pyramid network (BiFPN) along with the feature pyramid structure of YOLOv5, thereby transforming the target detection problem into a fine-grained representation problem. Hu et al. [20] enhanced the detection performance of YOLOv5 by utilizing the power IoU function. Zheng et al. [21] incorporated a CBL module to improve the YOLOv5 network and used the SIoU function as the loss function to enhance convergence speed and detection performance. Zhang et al. [22] employed the YOLOv3 network for identifying and localizing human bodies in the classroom, and they used HRNet for recognizing body poses, thereby improving the detection accuracy for obscured students.

In addition to computer-vision-based techniques for recognizing students' classroom behavior, Lu et al. [23] performed online detection of English online classroom learning behavior using feature data mining methods. Chakradhar et al. [24] trained a convolu-

tional neural network to recognize and classify students' expressions in the classroom, subsequently determining their attentional focus.

These studies have made significant contributions to the field of recognizing students' classroom behavior by employing various techniques and algorithms based on convolutional neural networks. However, the use of transformer structures in this domain has also gained attention and showcased promising results. In the following section, we will discuss CNNs and transformers in more detail and compare their approaches, advantages, and disadvantages.

2.2. Convolutional Neural Networks

Recent advancements in the semiconductor industry have significantly increased the computational power of chips, which has had a profound impact on the development of neural networks [25–27], deep learning, and target detection technology. Neural-network-based target detection techniques can be broadly categorized into two approaches: single-stage and two-stage methods.

The two-stage approach involves using Region Proposal Networks (RPNs) [28,29] to generate candidate frames, followed by filtering techniques like non-maximum suppression to select the best detection frame for the target object. The R-CNN family of algorithms [30] is a notable example of the two-stage approach. While this method achieves relatively high accuracy, the generation and filtering of a large number of candidate frames result in redundant computations, leading to slow detection speed and limited effectiveness for real-time detection tasks.

In contrast, the single-stage approach directly performs classification and regression calculations on the input image, eliminating the need for extensive candidate frame generation and subsequent filtering. This reduction in redundant operations significantly improves efficiency. Representative algorithms in the single-stage approach include the Single Shot MultiBox Detector (SSD) algorithm [31] and the YOLO series [32].

2.3. Transformers

Transformers [32], initially introduced by Google in 2017 for language tasks, have gained attention in the field of computer vision with the advent of Vision Transformer (ViT) [33] in 2020. Unlike traditional convolutional neural networks, transformers excel in handling tasks with varying scales and long sequences of data due to their unique encoder and decoder structures and global attention mechanism. This makes transformers more promising for practical applications compared to CNNs. The success of the Transformer algorithm in the speech domain inspired the development of ViT for image classification tasks. ViT utilizes a parameter-free self-attention module embedded into the encoder layer, enabling the conversion of image data into a fixed-length vector representation for classification. Comparative studies between ViT and CNN models have demonstrated comparable accuracy performance. However, for target detection tasks that require attention to object locations and sizes within the input, different self-attentive units are required for different image regions.

Detection Transformer (DETR) [34] combines the Transformer framework with set prediction, transforming the target detection task into an ensemble prediction problem. Each location in the image is associated with a query vector using self-attention, enabling end-to-end single-step detection and segmentation. Improved versions of DETR, such as Deformable DETR [35], Conditional DETR [36], and DINO [37], have achieved high detection accuracy on the COCO dataset by enhancing the DETR structure. Furthermore, the sliding window-based Swin Transformer [38] has surpassed CNNs in image classification, target detection, and image segmentation, demonstrating the vast potential of transformer networks in computer vision.

Both CNNs and transformers have been extensively employed in computer vision tasks, including students' classroom behavior detection. CNNs excel at capturing local image features through convolutional operations, making them effective for target detection

and classification tasks. They have achieved remarkable success across various applications and are well-established in the field. On the other hand, transformers have exhibited exceptional performance in natural language processing tasks and have recently gained attention in computer vision. Transformers leverage self-attention mechanisms to capture global dependencies and relationships between elements in the input data, making them suitable for modeling long-range dependencies and capturing contextual information in image data. This capability has led to improved performance in tasks such as image classification and object detection.

In the domain of students' classroom behavior detection, CNNs have been widely used and have made significant progress. They have proven effective in capturing local visual features relevant to behavior analysis, such as body poses and facial expressions. However, CNNs may encounter challenges when modeling complex relationships between different body parts and capturing global contextual information. Transformers, with their ability to capture global dependencies and contextual information, hold promise for enhancing students' classroom behavior detection systems. By effectively modeling the relationships between different body parts and considering the overall context of students' classroom behavior, transformers have the potential to improve the accuracy and robustness of behavior detection models.

3. Materials and Methods

3.1. Research Problem

This section provides an overview of the mathematical notations used in this paper and introduces the students' classroom behavior detection system developed in this work. Additionally, we describe the concepts of deep behavioral feature embedding and multiple attention mechanisms that are utilized in our approach. For quick reference, Table 1 lists the important notations used throughout the paper, and Figure 1 illustrates the flowchart depicting the detection process conducted in this study.

Table 1. Symbols used in this paper.

Notation	Description
I	Classroom image signal
B	Student behaviors
i	Number of identified behaviors
$h,\ w$	Size of image
C	Dimension of image
$X = \{x_1, x_2, \ldots, x_l,\},\ x_l \in R^{c_l \times h_l \times w_l}$	Feature map of the original image
$\hat{X} = \{\hat{x_1}, \hat{x_2}, \ldots, \hat{x_l},\},\ \hat{x_l} \in R^{c_l \times h_l \times w_l}$	Feature map of FPN
N	Number of targets in a single image
$q,\ k,\ v$	Parameters of attentional manipulation
D	Dimension of Deformable DETR's attention module
$d, d \in 0, 1, \ldots, \frac{1}{D}$	Dimension of the current detection sequence
$y_{cls}^{item},\ item \in 1, 2, \ldots, i$	Class results detected by model checking
y_{pos}	Bounding box results identified by model detection
α, β, γ	The weight ratio of the loss function

Definition 1 (Students' Classroom Behavior Detection Problem)**.** *The students' classroom behavior detection problem aims to identify the input classroom image signal I_{input} and determine the best matching students' classroom behavior from a set of students' classroom behavior categories $(B_n | i = 1, 2, 3, \ldots, N)$. Here, n represents the index of the students' classroom behaviors identified in this paper. The problem can be formulated as follows:*

$$B^* = \arg\max_{B_n} \{ f(I_{input}, B_1; \varphi), f(I_{input}, B_2; \varphi), \ldots, f(I_{input}, B_N; \varphi) \} \quad (1)$$

where $f(\cdot)$ is the function that calculates the similarity of behaviors, and φ is the function parameter. By evaluating the input classroom image I_{input}, we calculate the degree of matching with various behaviors, and the behavior B^* with the highest similarity is identified as the detected behavior.

Definition 2 (Deep Behavioral Profile Embedding)**.** *To achieve deep behavioral profile embedding, we employ the Feature Pyramid Network and the multi-scale attention mechanism of Deformable DETR. In the encoder part of the transformer, we utilize position embedding and scale-level embedding techniques. These methods enable differentiation of the position information of the target and the various feature layers.*

The position embedding provides a fixed representation that allows for the determination of the relative order of each input token in the query. By applying the same normalization operation to the input query, the relative position of the target in the classroom image can be accurately determined. On the other hand, the scale-level embedding is specifically trained to discern the different feature layers, facilitating multi-scale target detection. Integrating these techniques with the FPN network significantly improves the accuracy of multi-scale students' classroom behavior detection.

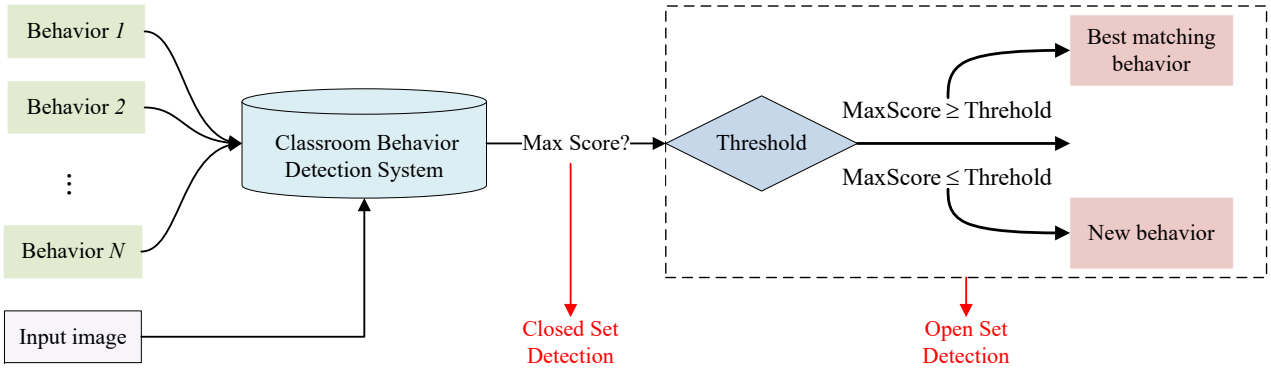

Figure 1. Flowchart of the students' classroom behavior detection system.

Definition 3 (Multiple Attention Mechanisms)**.** *In the backbone network, we utilize both same-window self-attention and cross-window attention to learn and divide the features of the entire classroom image. This enables attention learning in the detection head by employing Deformable attention for local regions around the observation point and a cross-attention mechanism for information exchange across different attention regions. By dividing the computation of the overall task into attention computations into multiple regional parts, we reduce the computational burden while ensuring the effectiveness of the attention mechanism in capturing global information from the entire classroom image.*

3.2. Proposed Method

In this section, we describe the method used for classroom behavior detection, which involves the substitution of the backbone network of Deformable DETR with the Swin Transformer. The extracted classroom image features from the backbone are then passed through the FPN network, and the CARAFE operator is incorporated to enhance the model's operational accuracy. The model architecture used in this research is depicted in Figure 2. During the training phase, the dataset is expanded, and adaptive training is employed to gradually adjust the batch size.

3.2.1. Deformable DETR for Classroom Behavior Detection

Deformable DETR, proposed by Dai et al. at SenseTime [35], is a target detection model that utilizes the Transformer architecture. Unlike traditional target detection models that rely on manually designed object detection frameworks such as anchor boxes [39,40], Deformable DETR transforms the target detection problem into an ensemble problem and leverages the Transformer for ensemble processing.

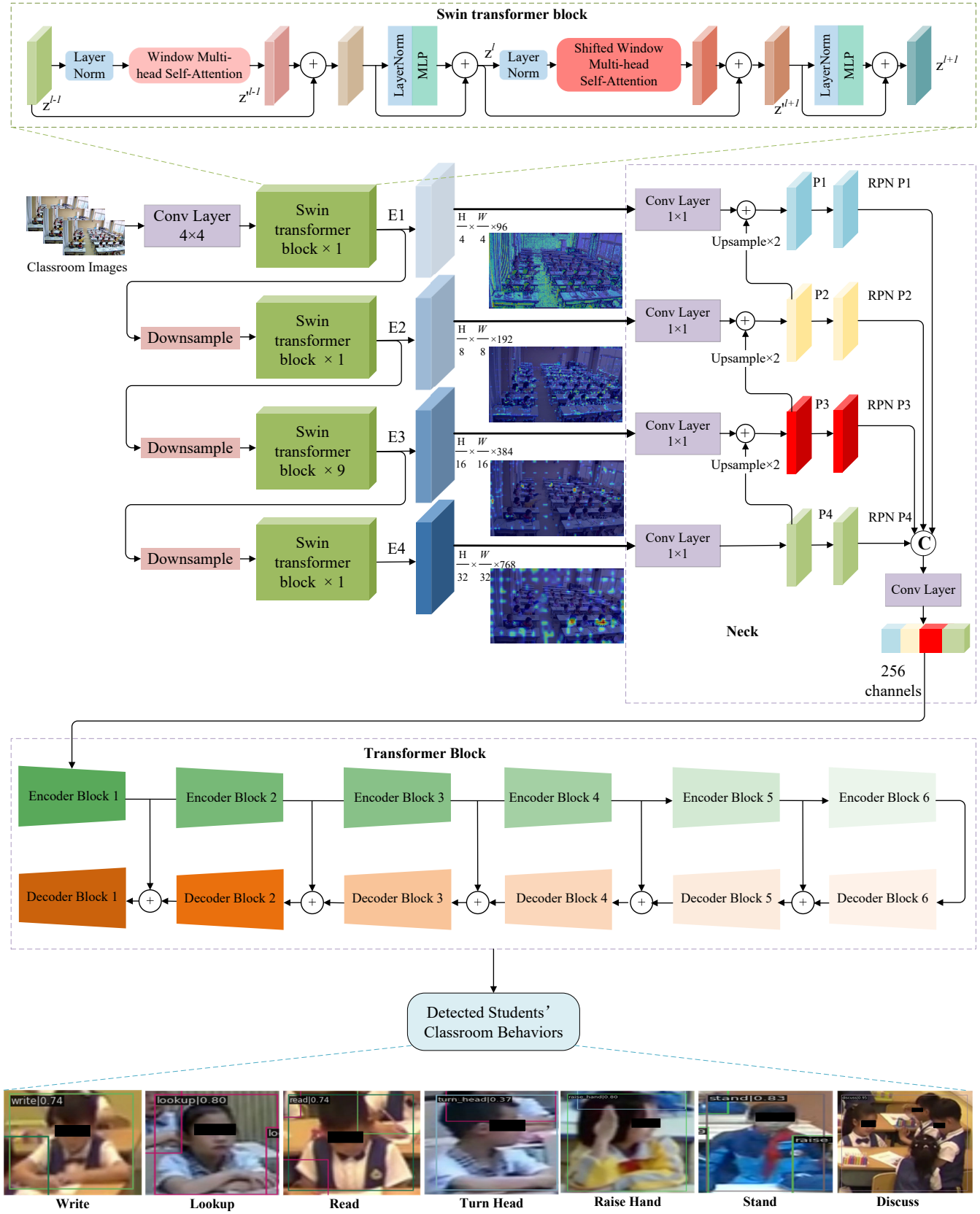

Figure 2. The overall structure of the model. The classroom images of size $H \times W \times 3$ are input into the backbone constructed by the Swin Transformer network, and then the classroom images are formed into 7 patches of $4 \times 4 \times 96$. Each patch is then fed into the Swin Transformer block to perform attentional operations. The feature maps obtained at different levels are fed into the FPN

structure, and the model is able to learn multi-scale features by fusing the deep and shallow features. The output of the FPN structure is transformed into a 256-dimensional sequence and fed into the transformer network, and the final detection result is obtained after 6 layers of encoder and decoder operations.

The architecture of Deformable DETR consists of two main parts: the backbone network for extracting classroom image features and the Transformer module for image classification of each feature map.

The backbone network extracts features from the input classroom image and generates multiple feature maps with different resolutions. These feature maps are then processed by transformers to convert the feature representation from a list format into a global classroom image representation, which is passed on to the decoder.

The detection head of Deformable DETR utilizes a multi-scale deformable self-attention mechanism with local features around key points, followed by cross-attention to link the features between different key points and learn overall features [35].

In the original image detection task using the transformer structure, the attention mechanism assigns different weights to different regions of the input feature map through self-attention in the encoder structure. The resulting features are then fed into the decoder for classification and filtering of the obtained feature vectors. Cross-attention is employed in the decoder to learn features extracted from different encoder layers.

$$Attention(q,k,v) = softmax\left(qk^T\right)v \tag{2}$$

Transformer's global-oriented attention module focuses on all possible spatial locations, but this approach involves redundant and invalid computations despite achieving high accuracy. To address this, Deformable DETR introduces the Deformable Attention mechanism, which enhances the receptive field by utilizing deformable convolution instead of regular convolution, as shown in Figure 3.

$$DeformAttn(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} * W_m^{'} x\left(p_q + \Delta p_{mqk}\right)\right] \tag{3}$$

The Deformable Attention mechanism, represented by Equation (3), takes an input feature map x of size $C \times H \times W$, where each pixel point in the feature map is a vector z_q of C channels. The mechanism uses initial sampled key points p_q, position encoding offsets $pmqk$, and weights W_m, Wm' of the fully connected layer. The dot product of queries and keys $amqk$ represents the weight of multi-headed attention. Both $pmqk$ and A_{mqk} are obtained by passing z_q through the fully connected layer [41].

Deformable convolution allows each convolution kernel to be determined by the offset and deformation of the kernel, capturing the spatial variation between targets of different scales. The Deformation Attention module focuses on a small number of key sampling points around the reference point, irrespective of the spatial size of the feature map. This alleviates the slow convergence of DETR by reducing the number of keys assigned to the query [42].

For target detection, each object is represented as a vector containing the feature and location information. Unlike ViT's application in image classification, determining the location information of each target object in the image is necessary for the target detection task. To accomplish this, location encoding is introduced to assign relative location information between different tokens in the query, enabling the determination of the location and reference point of each token. The feature information around the reference point obtained from attention is then used to predict the relative offset of the reference point, which serves as the predicted bounding box by the detection head. The reference point is initially set as the center of the frame, and the detection head predicts the offset with respect to the reference point.

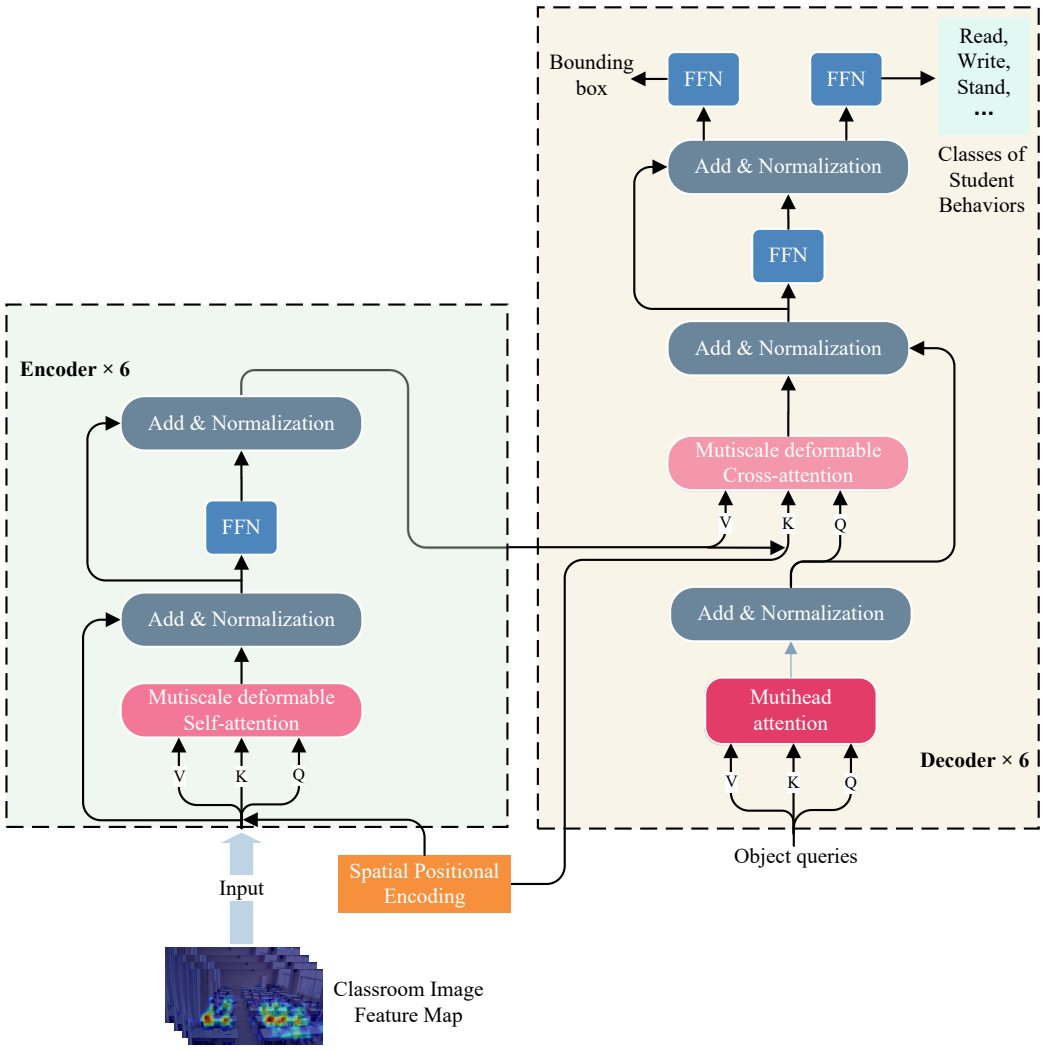

Figure 3. The encoder and decoder structure of Deformable DETR. The attention is computed by a multi-scale deformable self-attention mechanism with local features around key points. After that cross-attention is taken to link the features between different key points to learn the overall features.

3.2.2. Swin-Transformer in Classroom Scenarios

To enhance the performance of classroom behavior detection, we propose replacing the backbone network of Deformable DETR with the Swin Transformer as shown in Algorithm 1. The Swin Transformer has demonstrated impressive capabilities in capturing long-range dependencies and achieving outstanding results across various computer vision tasks [33]. By incorporating the Swin Transformer's hierarchical structure and self-attention mechanism, we aim to improve the representation power of our model.

The Swin Transformer [38] is a variant of the Transformer model that focuses on capturing spatial information in images. It introduces a patch-based processing approach, where a classroom image with dimensions $h \times w$ is divided into m patches of size $n \times n$. Feature extraction is performed on each patch, and the resulting feature vectors are concatenated to form a comprehensive representation of the entire classroom image. This patch-based strategy enables the model to handle objects of different sizes effectively in image detection tasks [38].

Algorithm 1 The proposed model of students' classroom behavior detection

Input: Pictures/videos of teachers in actual educational situations
Output: The behavior of each student in the current classroom

1: Block the original classroom image:

$$x_l = Conv(I_{h\times w\times c})$$

2: Calculate attention within a single patch:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C$$

3: Calculate the cross-attention between two patches:

$$\Omega(SW-MSA) = 4hwC^2 + 2M^2hwC$$

4: Computes the output of the Swin-block. The calculation was repeated for many times, and four feature graphs of different dimensions and sizes were obtained:

$$\begin{aligned} x_l{}' &= W-MSA(LN(x_{l-1})) + x_{l-1} \\ x_l &= MLP\big(LN\big(x_l{}'\big)\big) + x_l{}' \\ x_{l+1}{}' &= SW-MSA(LN(x_l)) + x_{l+1} \\ x_{l+1} &= MLP\big(LN\big(x_{l+1}{}'\big)\big) + x_{l+1}{}' \end{aligned}$$

5: Calculate the multiscale feature map:

$$\hat{X} = \varphi\big(N\big(X, k_{up}\big), W_{l'}\big)$$

6: Calculate deformable attention:

$$DeformAttn\big(z_q, p_q, x\big) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} * W_m^{'} x\Big(p_q + \Delta p_{mqk}\Big)\right]$$

7: Compute classification loss $\mathcal{L}_{cls}$:

$$\begin{cases} y_{cls}^{item} = \underset{item=b^*}{argmaxsoftmax}(FFN(h_l w_l \times Dmodel)) \\ \mathcal{L}_{cls} = -(1-p_t)^{\gamma}\log p_t \end{cases}$$

8: Compute bounding box loss $\mathcal{L}_{bbox}$:

$$\begin{cases} PE_{(y_{pos},2d)} = sin\left(\frac{y_{pos}}{10{,}000^{\frac{2d}{Dmodel}}}\right) \\ PE_{(y_{pos},2d+1)} = cos\left(\frac{y_{pos}}{10{,}000^{\frac{2d}{Dmodel}}}\right) \\ \mathcal{L}_{bbox} = \frac{1}{n\sum|x_i-y_i|} \end{cases}$$

9: Compute IoU loss $\mathcal{L}_{IoU}$:

$$\mathcal{L}_{IoU} = 1 - \left(IoU - \frac{A^c - u}{A^c}\right)$$

10: Compute the total loss L_{total}:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{bbox} + \gamma\mathcal{L}_{IoU}$$

11: Predict the students' classroom behavior B^* and the position of the student y_{pos}.

The Swin Transformer block comprises two main components: the Windowed Multi-head Self-Attention (W-MSA) mechanism and the Sliding-Window Multihead Self-Attention (SW-MSA) mechanism. Additionally, it incorporates a Multilayer Perceptron (MLP) and Layer Normalization (LN) operations. The connectivity between different layers is established using a residual network [43]. The Swin Transformer block structure involves normalizing the input with LN before the W-MSA module and MLP, and the residual network connects the information across layers.

The structure of the Swin Transformer block is illustrated in Figure 4. In each block, the input image of size $(H, W, 3)$ is partitioned into patches, which are then processed by the W-MSA mechanism. After normalization and MLP operations, a sliding-window attention operation is performed to establish feature associations between different patches [43].

$$\begin{cases} x_l' = W\text{-}MSA(LN(x_{l-1})) + x_{l-1} \\ x_l = MLP(LN(x_l')) + x_l' \\ x_{l+1}' = SW\text{-}MSA(LN(x_l)) + x_{l+1} \\ x_{l+1} = MLP(LN(x_{l+1}')) + x_{l+1}' \end{cases} \tag{4}$$

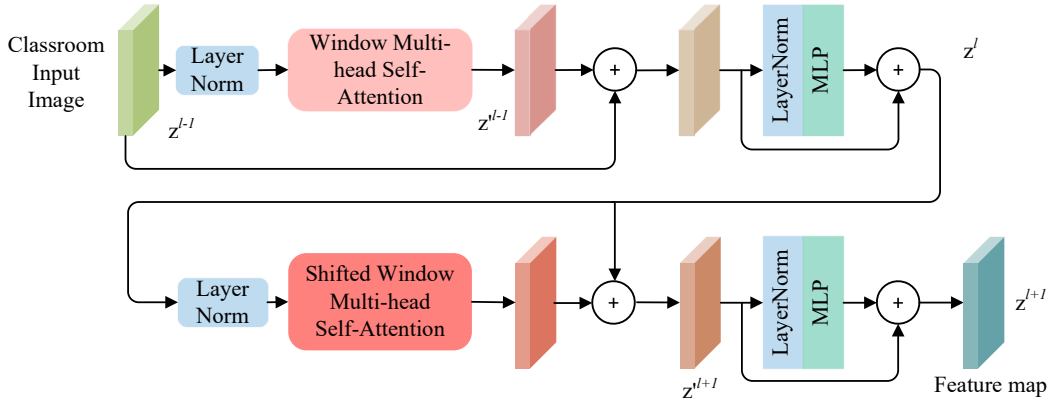

Figure 4. The Swin Transformer block structure is used to divide the incoming (H, W, 3) images into patches, and then each patch is processed by the Window Multihead Self-Attention mechanism (W-MSA). After normalization and fully connected network operations, a sliding-window attention operation is performed to establish feature associations between different patches.

$$\begin{cases} \Omega(MSA) = 4hwC^2 + 2(hw)^2C \\ \Omega(SW\text{-}MSA) = 4hwC^2 + 2M^2hwC \end{cases} \tag{5}$$

The computational complexity of the W-MSA module and SW-MSA mechanism can be approximated as Equation (5). These window-based self-attention modules are more efficient for feature extraction from large visual datasets. Unlike the conventional self-attention mechanism that involves computations across the entire input sequence, W-MSA divides the input sequence into windows, allowing elements within each window to interact only with elements within the same window. This significantly reduces the computational cost. Additionally, SW-MSA facilitates the exchange and learning of attention information between windows, enabling the model to capture global attention information about the entire image. This approach reduces computation while enhancing speed.

A key design element of the Swin Transformer is the shifting of window partitions between consecutive self-attentive layers. This shifting mechanism establishes connections between windows from different layers, significantly enhancing the model's modeling capability. To ensure connectivity, the windows at the edges undergo cyclic leftward and upward shifts through masking operations, adjusting the relative positions between windows and limiting self-attention calculations to each sub-window.

In the context of classroom behavior detection, integrating the Swin Transformer into Deformable DETR brings several advantages. First, the Swin Transformer's patch-based processing enables the model to effectively handle objects of various sizes, which is crucial in classroom scenarios where students' behavior may vary in scale. Second, the hierarchical structure and self-attention mechanism of the Swin Transformer enhance the model's ability to capture long-range dependencies, facilitating the detection of complex behavior patterns. Finally, the optimization techniques employed in the Swin Transformer improve computational efficiency, making it feasible to apply the model to real-time classroom monitoring.

By substituting the backbone network of Deformable DETR with the Swin Transformer, we expect to achieve improved accuracy and robustness in classroom behavior detection. The Swin Transformer's architectural design and efficient feature extraction mechanisms make it a suitable choice for capturing meaningful representations of classroom images and facilitating subsequent detection and classification tasks.

3.2.3. FPN-CARAFE in a Classroom Environment

To further enhance the operational accuracy of the model, we incorporate the Content-Aware ReAssembly of Features (CARAFE) operator [38]. The CARAFE operator [38] is a pixel-adaptive convolutional operator that performs upsampling on low-resolution feature maps while preserving and enhancing fine-grained details. By adaptively gathering and redistributing information from neighboring pixels, the CARAFE operator refines the feature representations and improves the model's ability to capture fine-grained information [44].

In the context of a classroom environment, the FPN-CARAFE model proves valuable for behavior detection tasks due to its capacity to detect and classify targets based on multi-scale feature maps. The Feature Pyramid Network architecture enables the model to extract contextual and semantic information from different scales, while the CARAFE operator enhances the feature recombination process.

FPN constructs a bottom-up feature pyramid by connecting multiple backbone convolutional networks. This pyramid consists of feature maps at different resolutions, with each layer providing varying levels of contextual and semantic information. Through merging these feature maps, FPN creates a more comprehensive representation of the input, thus improving the accuracy of detecting targets at multiple scales. The deep-level features capture the overall shape and structure of an object, while the shallow-level features capture local details such as edges and corners. The fusion of these features enhances the model's ability to handle objects of different sizes.

During the feature recombination process, the CARAFE operator is employed to upsample the feature maps. CARAFE predicts an upsampling kernel for each location based on the input features, enabling a larger field of perception during recombination and guiding the recombination process. Utilizing predicted upsampling kernels, CARAFE recombines the features, resulting in an upsampled feature map. Notably, CARAFE achieves this with fewer parameters and lower computational complexity compared to alternative methods.

The CARAFE process can be represented as follows:

$$\begin{cases} W_{l'} = \psi(N(X_l, k_{encoder})) \\ X'_{l'} = \varphi(N(X_l, k_{up}), W_{l'}) \end{cases} \tag{6}$$

Here, $W_{l'}$ represents the position information obtained by predicting the kernel using the content of each target location. The content-aware module φ reorganizes the input feature map X_l using $W_{l'}$, resulting in the new feature map $X'_{l'}$ or X'.

By combining FPN and CARAFE, the FPN-CARAFE model significantly enhances the detection and classification of behaviors in a classroom setting. The multi-scale feature maps extracted by FPN provide rich contextual information, while CARAFE improves the feature recombination process, enabling accurate behavior detection. This approach aids in monitoring classroom dynamics, analyzing student engagement, and facilitating effective classroom management.

3.2.4. Training Phase and Adaptive Training

During the training phase, we expand the dataset to enhance the model's performance. Adaptive training [45] is employed to gradually adjust the batch size. Adaptive training involves dynamically adjusting the batch size during the training process based on training progress and resource constraints. The batch size refers to the number of training samples processed by the training algorithm in each iteration. A larger batch size can lead to more efficient training as it allows for parallel processing and better utilization of computational resources. However, using a larger batch size may result in finding local optima rather than global optima, while a too small batch size can slow down training and lead to non-convergence. The idea behind adaptive training is to find an optimal batch size that balances the benefits of a larger batch size with the limitations imposed by available resources.

Initially, a smaller batch size can be used to explore the training landscape and expedite the model training process. As training progresses and model robustness improves, the batch size can gradually increase to leverage larger batch sizes and improve convergence speed. By adopting adaptive training, the model can effectively utilize computational resources and achieve improved performance compared to using a fixed batch size throughout the entire training process.

4. Experimental Results and Analysis

This section presents the experimental results and analysis of the proposed method. We begin by providing an overview of the training dataset used in this study. Next, we compare our method with existing approaches that utilize the same dataset to demonstrate its superiority. Furthermore, we conduct a series of intersection experiments to evaluate the performance of our method under different optimization strategies.

4.1. Dataset

In the field of computer vision, various datasets, such as ImageNet [44], COCO [46], and ADE20K [2], have been established for different vision tasks. However, there is currently no comprehensive and authoritative dataset specifically designed for students' classroom behavior detection in classroom environments. To address this gap, we created a dataset, named ClaBehavior, by using screenshots from videos of elementary school language classes in the 2019 Ministerial Excellence Lessons. These videos are publicly available on the China National Resources Public Service Platform [47]. We captured screenshots at one-second intervals, resulting in a total of 1346 images. We annotated these images using the Labelme software [48]. The dataset consists of four classroom scenes with different layouts, each containing 24–36 students. It includes approximately 120 distinct student objects, covering seven different classroom behaviors: reading, writing, raising hands, listening, standing up, group discussion, and turning head to talk. The objects in this dataset are predominantly medium-sized targets due to the dense distribution of characters in the classroom environment. Students in the foreground are considered large target objects, while students located in the corners and edges of the classroom are mostly small target objects.

Table 2 provides an overview of the labels and their corresponding frequencies in the ClaBehavior dataset. The labels are formatted according to the COCO dataset format.

Table 2. Number of labels for different behaviors in ClaBehavior dataset.

Behaviors	Write	Read	Lookup	Turn_Head	Raise_Hand	Stand	Discuss
Statistics	1025	1075	5725	1025	725	94	242

4.2. Evaluation Metrics

To evaluate the effectiveness of our model, we employ various evaluation metrics that measure its performance in accurately identifying and classifying objects. The following metrics are utilized:

- True Positive (*TP*): The model correctly identifies the location and type of the object in the classroom-behavior-detection task.
- False Positive (*FP*): The model correctly identifies the location of the object in the classroom-behavior-detection task, but misidentifies its type.
- False Negative (*FN*): The model fails to identify the correct position and type of the object in the classroom-behavior-detection task.

By calculating the values of *TP*, *FP*, and *FN*, we can derive the precision and recall of the model. These metrics provide insights into the model's ability to correctly detect objects in the classroom behavior detection task. Precision is defined as the ratio of true positives to the sum of true positives and false positives, while recall is the ratio of true positives to the sum of true positives and false negatives:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

Furthermore, to conduct a comprehensive evaluation of precision and recall, we introduce the concepts of Average Precision (*AP*) and mean Average Precision (*mAP*). The *AP* and *mAP* values provide a visual representation of the model's detection accuracy for various target types in the classroom behavior detection task, with the *mAP* value reflecting the overall accuracy by averaging the *AP* values across different targets. This approach allows for a comprehensive assessment of the model's performance in the classroom behavior detection task.

$$AP_i = \int_0^1 P(r)dr \tag{9}$$

$$mAP = \frac{1}{n}\sum_{i}^{n}(AP_i) \tag{10}$$

Another metric we consider is Floating Point Operations (FLOPs), which provides insights into the computational complexity of the proposed model. FLOPs quantifies the number of operations (multiplications and additions) required for data processing and prediction. By measuring FLOPs, we can estimate the computational requirements and model complexity, providing guidance for effective training and deployment.

4.3. Baseline Models

To assess the effectiveness of our proposed method, we compare it to several existing models in the field of students' classroom-behavior-detection systems. We have selected the following baseline models for comparison:

- **Faster R-CNN** [29]: This method utilizes the ResNet network for feature extraction and incorporates the Region Proposal Network (RPN) to generate bounding boxes. It employs a k-means algorithm for post-processing and filtering.
- **SSD** [18]: This method also utilizes the ResNet network for feature extraction and improves the accuracy of students' classroom behavior detection by integrating the Feature Pyramid Network (FPN). The FPN enhances the detection accuracy for small target objects, improving overall performance.
- **YOLOv3** [23]: This method adopts the Darknet-53 network as the backbone for feature extraction. It incorporates the FPN for multi-scale task detection. The classification part utilizes an SVM classifier for behavior classification.
- **YOLOv5** [20]: This method utilizes the CSPDarknet network to extract features from input images. It introduces the Spatial Intersection over Union (SIoU) loss function in the Convolutional Block Layer (CBL) module and employs the GELU function as the activation function.

- **YOLOv7** [14]: This method employs CBSDarknet as the backbone network and utilizes the Path Aggregation Network (PAN) and Feature Pyramid Network (FPN) for feature fusion across different levels of hierarchy in the images. The detection results are obtained using the Efficient Local Attention Network (ELAN) and Category-aware Transformation (CAT) modules as the detection heads.

By comparing our proposed method with these existing algorithms, we can evaluate and demonstrate its effectiveness and performance in the context of students' classroom behavior detection.

In the next section, we will present the experimental results and analysis, including a comparison of the proposed method with the baseline models, further demonstrating its superiority and efficacy.

4.4. Experimental Settings

The experiments were conducted on an Ubuntu 20.04 system equipped with four NVIDIA RTX A5000 GPUs. We implemented our method using the PyTorch-GPU 1.10.3 deep learning framework and CUDA version 11.3.

To address the limited amount of data, we performed data augmentation on the dataset using the Albumentations open-source library [49]. This involved applying various augmentation operations, such as flipping, panning, zooming, and blurring, to the images, effectively expanding the dataset. During the training phase, we utilized the AdamW optimizer [50] with an initial learning rate of 2×10^{-5} and a batch size of 1.

4.5. Comparison Experiments with the State-of-the-Art Methods

To evaluate the effectiveness of our proposed method, we compared it with five state-of-the-art methods on the ClaBehavior dataset. Table 3 provides a comprehensive comparison of different models, including their FLOPs per billion (G), mean Average Precision for object detection, precision, recall, and image size.

As shown in Table 3, YOLOv7 obtained the best performance in mAP, precision, and recall among all baseline methods. In addition, Faster R-CNN has the best performance in FLOPs among all the traditional methods. Compared with the five state-of-the-art methods, the proposed method achieves the best performance in mAP, precision, and recall, and its performance in FLOPs is comparable to that of Faster R-CNN. The above experimental results demonstrate that the proposed method can achieve the best performance in students' classroom behavior detection with high computational efficiency.

Table 3. Comparison of the performance of the proposed method with five state-of-the-art methods on the ClaBehavior dataset. Bolded indicates the best performance, and underlined indicates the second-best performance.

Models	FLOPs/G	mAP(0.50:0.95)	Precision	Recall	Image Size
Faster R-CNN	**23.38**	0.491	0.617	0.643	224 × 224
SSD	34.59	0.430	0.524	0.612	300 × 300
YOLOv3	77.1	0.378	0.378	0.572	640 × 640
YOLOv5	77.6	0.455	0.544	0.607	640 × 640
YOLOv7	104.7	<u>0.583</u>	<u>0.707</u>	<u>0.722</u>	640 × 640
Proposed	<u>33.21</u>	**0.605**	**0.738**	**0.751**	224 × 224

Additionally, Figure 5 compares the model's detection accuracy with other algorithms after an equal number of training sessions, further highlighting the superior performance of our proposed model.

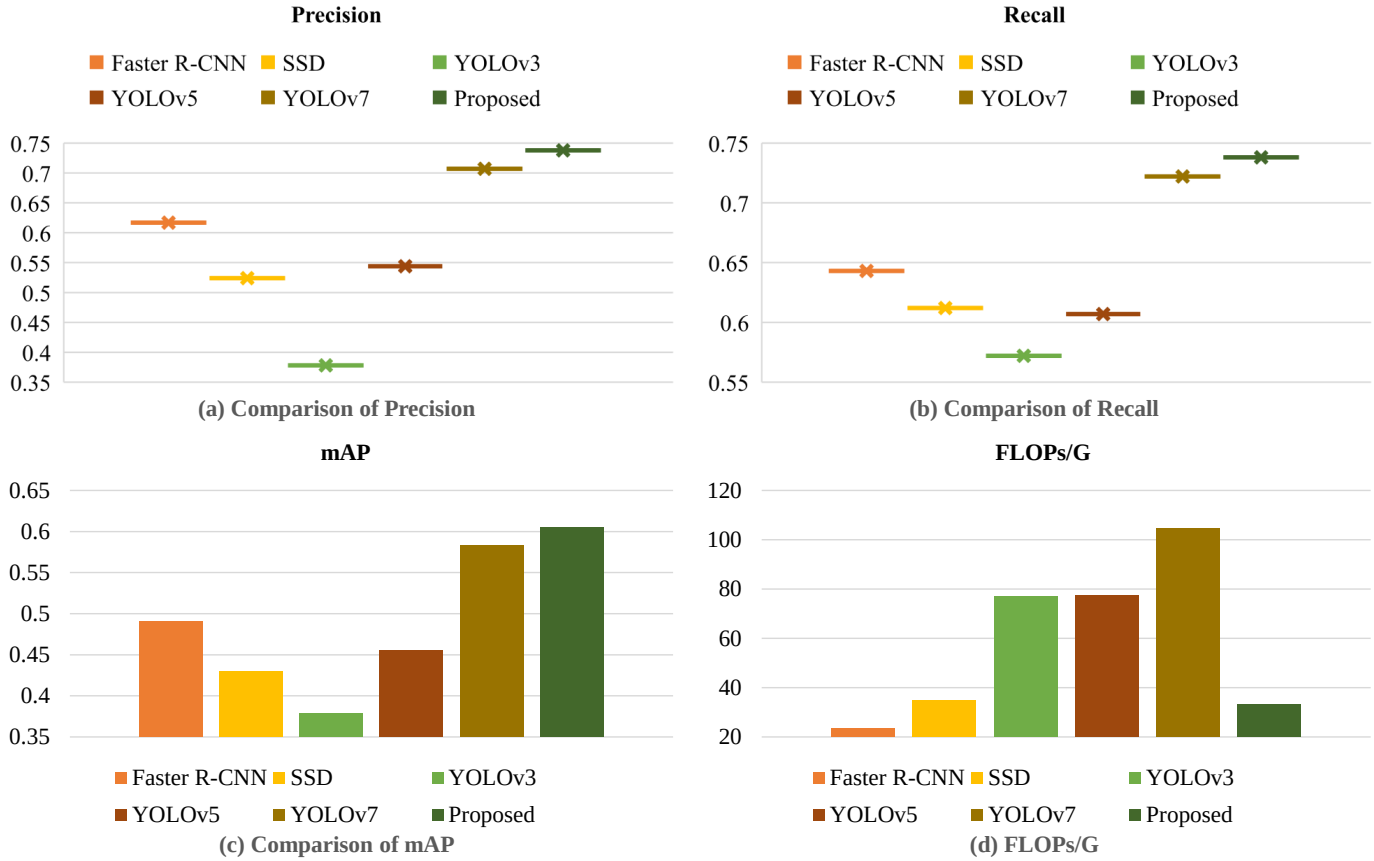

Figure 5. The performance of this method is compared with the baseline method on four metrics. Larger values of mAP, Precision, and Recall indicate better performance, and lower values of the FLOPs metric indicate better performance.

4.6. Ablation Experiments

Table 4 presents the results of the ablation experiments, where different optimization strategies were applied to the DeformableDETR model. The models include DeformableDETR alone, DeformableDETR with CSPDarknet as the backbone network, DeformableDETR with Swin Transformer as the backbone network, DeformableDETR with Swin Transformer and the FPN structure, and our proposed model. In this experiment, we investigated the combination of Deformable DETR and Swin Transformer, as well as the use of CSPDarknet from the YOLO series as the backbone network for Deformable DETR. However, the experimental results showed that this configuration did not yield satisfactory detection performance.

Table 4. Ablation experiments with 5 different combinations. Bolded indicates the best performance.

Models	FLOPs/G	mAP (0.50:0.95)	Precision	Recall	Image Size
DeformableDETR	**11.01**	0.544	0.654	0.722	224 × 224
DeformableDETR+CSPDarknet	85.49	0.488	0.606	0.663	640 × 640
DeformableDETR+Swin	26.51	0.566	0.703	0.725	224 × 224
DeformableDETR+Swin+FPN	28.97	0.593	0.717	0.736	224 × 224
Proposed	33.21	**0.605**	**0.738**	**0.751**	224 × 224

Further integration experiments were conducted to optimize our model. The results are presented in Table 4. When using Swin Transformer as the backbone network, the average accuracy for detecting different behaviors improved by 2.2%. Moreover, with the introduction of the Feature Pyramid Network structure in the neck, the detection performance showed a significant improvement, with a 4.9% increase in detection accuracy

compared to the original network. The FPN structure contributed to enhanced detection accuracy, particularly for small- and medium-sized objects. It is worth noting that this improvement was achieved with only a slight increase in the number of parameters, as indicated by the FLOPs. DeformableDETR+CSPDarknet utilizes an input image size of 640×640 because the student behavior recognition results are better compared to using an input size of 224×224. By adopting a more optimal baseline in this experiment, it demonstrates that the proposed method achieves higher recognition accuracy with smaller input images.

The results demonstrate that the integration of the FPN structure with the CARAFE operator in our proposed model leads to a significant improvement in accuracy while maintaining a manageable increase in operational complexity. This highlights the effectiveness of this combination in enhancing the overall performance of the model. The findings emphasize the importance of selecting appropriate techniques to achieve a balance between accuracy and computational efficiency.

On the other hand, when utilizing CSPDarknet from YOLOv5 as the backbone network, both the number of model parameters and detection accuracy were negatively affected. However, by enhancing the FPN with the lightweight CaraFE operator, the detection accuracy of the model reached 60.5%, representing a 6.1% improvement compared to the original model, as shown in Figure 6. Nevertheless, the number of model parameters increased significantly in this case.

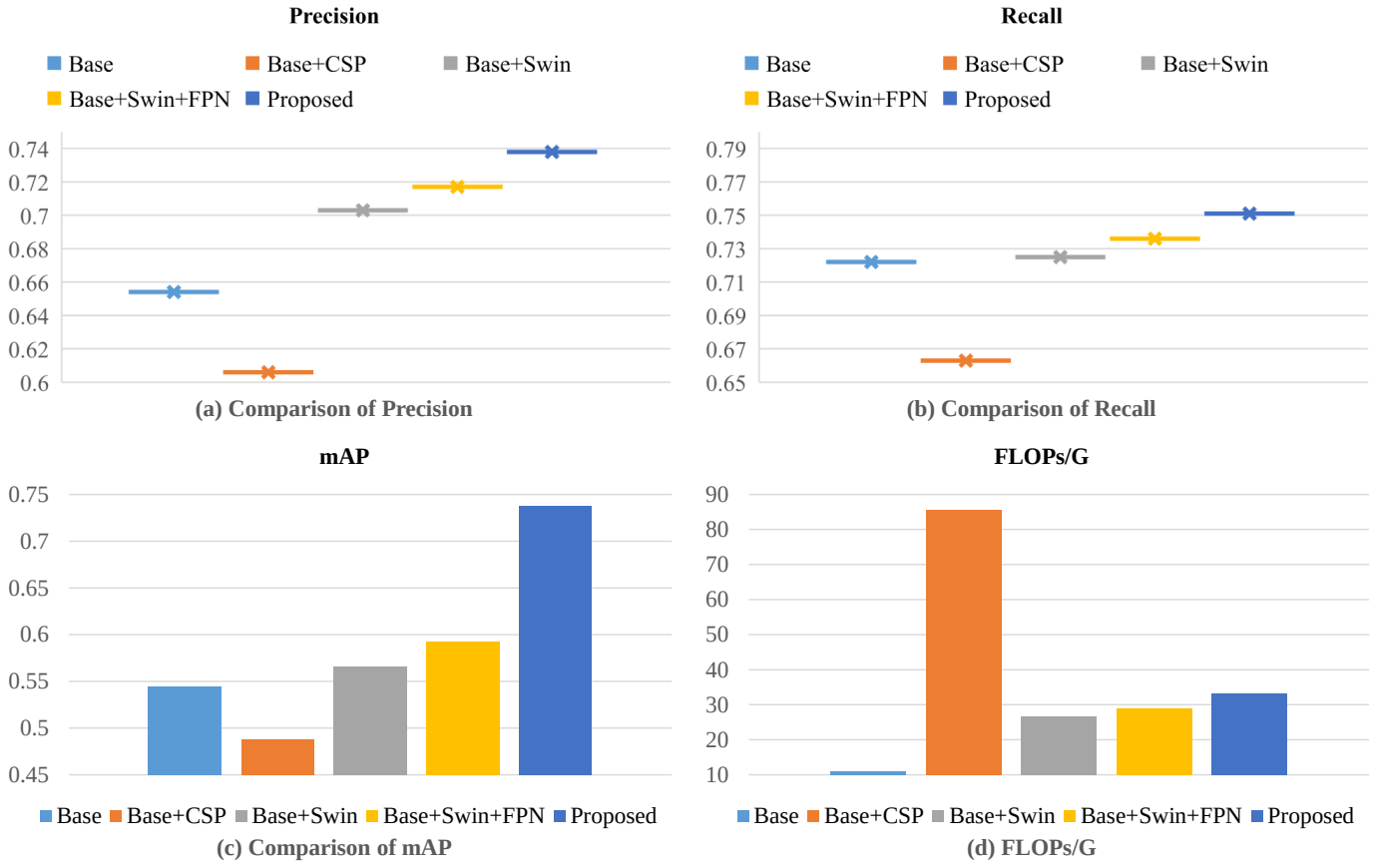

Figure 6. Comparison of the performance of 5 different combination methods on 4 metrics. Larger values of mAP, Precision, and Recall indicate better performance, and lower values of the FLOPs metric indicate better performance.

Figure 7 visualizes the accuracy variation of the model during the training process with different backbone strategies. All four models, employing different strategies, exhibit

stabilization and convergence after the 40th round of training, indicating effective learning from the training data and consistent performance.

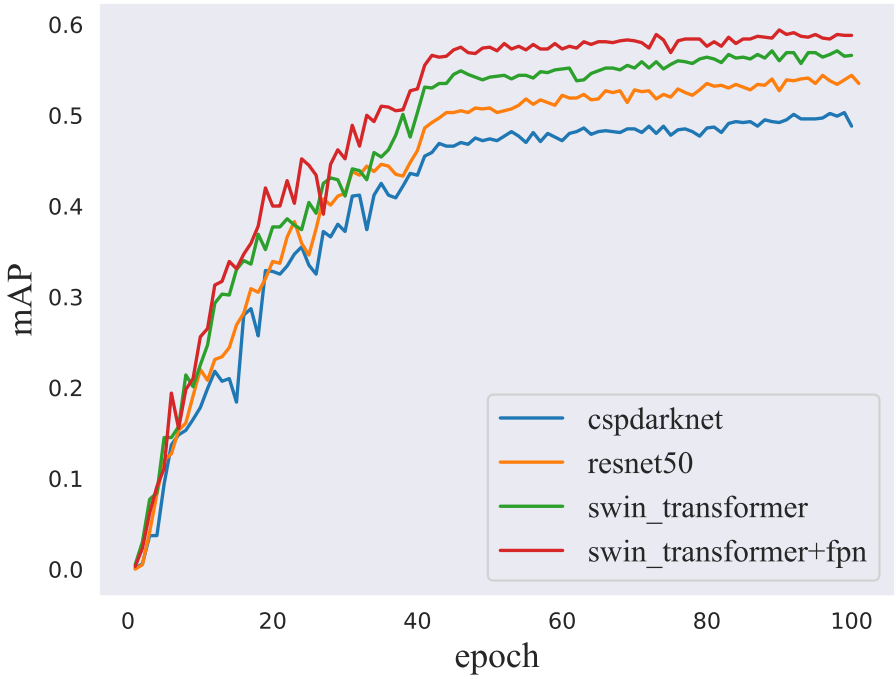

Figure 7. Comparison of model accuracy for different optimization strategies.

These findings emphasize the effectiveness of using the Swin Transformer as the backbone network and highlight the positive impact of incorporating the FPN structure in improving detection accuracy while keeping the model complexity and number of parameters relatively low.

4.7. Case Study

This section analyzes the performance of the proposed method on learner behavior recognition in three parts: representation capability of the proposed method, sensitivity analysis of the proposed method, and multiple case studies.

4.7.1. Representation Capability of the Proposed Method

To analyze the ability of the proposed method to represent students' classroom behavior, we extracted the feature maps of the network for analysis. Figure 8 provides a visualization of the feature maps generated by the FPN network at different layers, highlighting the hierarchical representation of the input data. The visualization results show that the proposed method has a very good capability to represent students' classroom behaviors.

4.7.2. Sensitivity Analysis of the Proposed Method

To assess the performance of our method, we present the detection results for seven distinct classroom behaviors, and these behaviors include writing, reading, lookup, turning heads, standing up, raising hands, and engaging in group discussions. Furthermore, Figure 9 showcases the recognition outcomes for various target scenarios within the classroom from multiple perspectives.

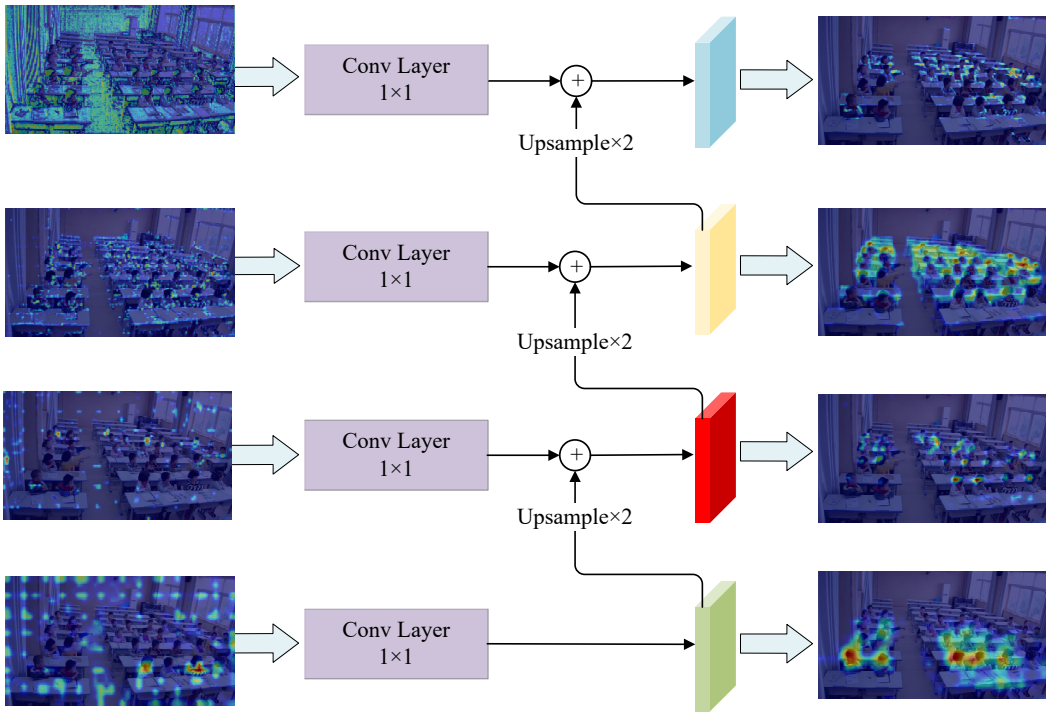

Figure 8. The FPN network outputs feature map results. Based on the feature maps of different sizes obtained from different layers of the backbone as input, the feature maps of the previous layer are upsampled and summed by a top-down process. The multi-scale detection task for targets of different sizes is realized.

Figure 9. The recognition outcomes for various target scenarios within the classroom.

The training progress and convergence of the model can be observed in Figure 10, which depicts the variation of the loss function throughout the training process.

The paper acknowledges that the misdetection of standing and discussion patterns is primarily attributed to the limitations of the dataset. The dataset used in the study mainly consists of elementary school classroom data, which poses challenges when testing the model on high school classroom scenarios. The model tends to misidentify sitting and standing postures as standing behavior due to the relatively higher sitting height of high school students, as depicted in Figure 9.

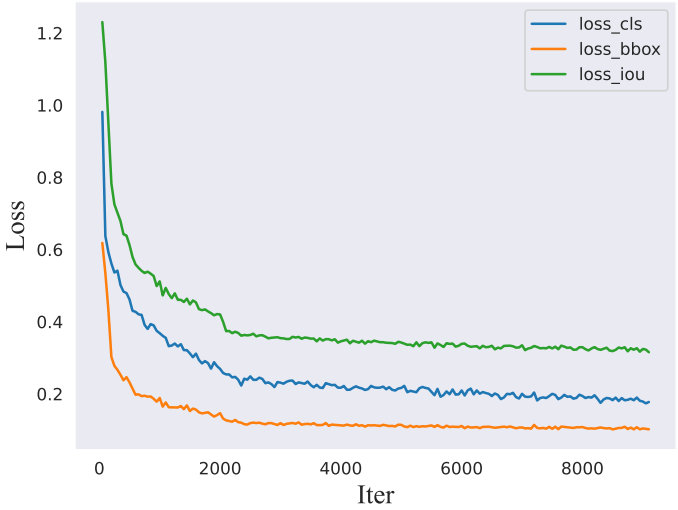

Figure 10. The loss throughout the training process.

Similarly, the dataset includes labels for group discussions that typically occur in a front-to-back table placement pattern, where students gather around two tables for discussion. Consequently, in classrooms with different table configurations, such as the scenarios depicted in Figure 11, there may be instances where the model fails to properly recognize discussion behaviors.

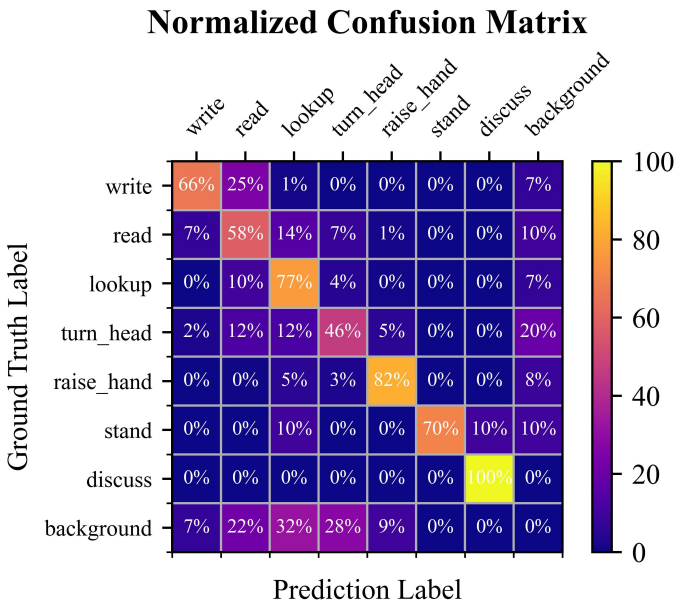

Figure 11. The confusion matrix of the proposed method.

To illustrate the detection accuracy of the model for different students' classroom behaviors, we present a confusion matrix in Figure 11. The confusion matrix highlights the model's performance in recognizing seven distinct classroom behaviors. We observe that the model achieves higher accuracy in recognizing behaviors such as raising hands, standing up, and group discussions, which involve larger body movements. However, it exhibits relatively lower accuracy for behaviors like reading, writing, and listening, which share similar body movements.

Overall, the results and observations presented in this section demonstrate both the strengths and limitations of the SeDetr model for detecting students' classroom behaviors. The model showcases promising generalization capabilities, but certain behavioral differentiation and environmental factors may still pose challenges.

4.7.3. Multiple Case Studies

Figure 12 showcases the identification of seven common behaviors in student classrooms using the SeDetr model. The images used for detection include those from the test dataset, validation dataset, as well as scenarios not present in the dataset. The model demonstrates good generalization ability by successfully recognizing most of the behaviors across various scenarios and individuals. However, there are instances of missed and misidentified results, such as in 2(d) and 2(f) of Figure 12, where the model mistakes writing behavior for reading behavior. This ambiguity arises due to the model's reliance on head or body curvature as a distinguishing factor between the two behaviors. A similar challenge is observed in 1(d) of Figure 12, where the model exhibits low confidence in distinguishing between writing and reading. Additionally, the detection of head-turning behavior can be influenced by the camera angle, as exemplified in 3(a) of Figure 12, where a false detection occurs due to the image's resemblance to head-turning and talking. Hence, the impact of camera angles should be considered in real-world scenarios.

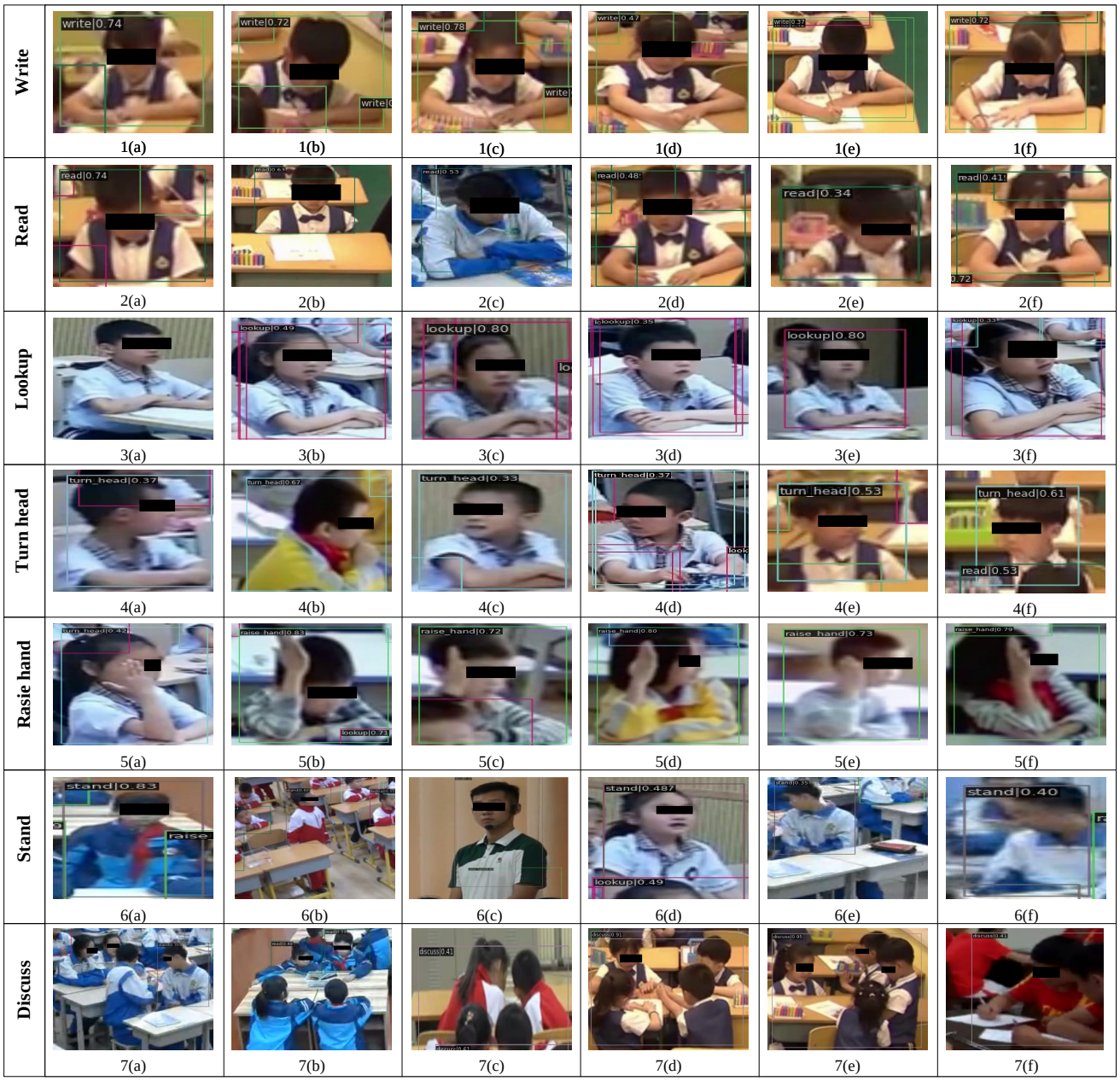

Figure 12. The effectiveness of identifying seven classroom behaviors in different contexts.

5. Discussion

The proposed method for students' classroom behavior detection in the classroom, utilizing a transformer-network-based approach and incorporating advanced techniques, has shown promising results. In this section, we discuss the implications of our findings, highlight the strengths and limitations of our approach, and propose potential areas for future research.

One of the key strengths of our proposed method is the use of transformer networks, specifically the Swin Transformer, as the backbone network for feature extraction. The transformer architecture has demonstrated remarkable success in various computer vision tasks, and our study extends its application to students' classroom behavior detection. By leveraging the self-attention mechanism, transformers capture long-range dependencies and contextual information, enabling accurate behavior detection in complex classroom scenarios.

In addition, the incorporation of the feature pyramid structure and CARAFE operator has contributed to improved detection accuracy for objects with different feature sizes. The feature pyramid structure allows the model to effectively handle objects at multiple scales, while the CARAFE operator enhances the resolution of feature maps, enabling more precise localization and detection of students' classroom behaviors. These techniques address the challenge of detecting small- and medium-sized objects in the classroom and enhance the overall performance of the proposed method.

The construction of a dedicated dataset for students' classroom behavior detection is another significant contribution of our study. The availability of reliable and annotated datasets is crucial for the development and evaluation of machine learning models. Our dataset, consisting of a diverse set of classroom images and behavior annotations, fills a gap in the existing literature and provides a valuable resource for future research in the field of students' classroom behavior detection.

However, our study also has certain limitations that should be acknowledged. Firstly, the size of our dataset is relatively small, which may restrict the generalizability of our findings. Expanding the dataset to include a larger number of classrooms, diverse educational environments, and various age groups would enhance the robustness and applicability of the proposed method. Additionally, the performance of our method on low-resolution images is not satisfactory. Future research should explore techniques to improve the detection accuracy for such images, as they are common in real-world classroom scenarios.

Furthermore, the computational complexity and training time of the proposed method are relatively high due to the size of the model. This poses practical challenges, especially in real-time applications. To address this limitation, future studies should focus on developing lightweight models that maintain high detection accuracy while being more computationally efficient. This would make the proposed method more accessible and applicable in real-world classroom environments.

In terms of future research directions, there are several avenues to explore. Firstly, the inclusion of temporal information can provide valuable insights into students' classroom behavior dynamics and improve the accuracy of behavior detection. Incorporating video data and leveraging techniques such as temporal convolutional networks or recurrent neural networks can enable the modeling of temporal dependencies and capture the temporal evolution of students' classroom behaviors.

Secondly, investigating transfer learning and domain adaptation techniques would be beneficial to address the challenges of deploying the proposed method in different educational environments. Adapting the model to new classrooms or different cultural contexts could improve the generalizability and effectiveness of behavior detection systems.

Moreover, integrating multimodal information, such as audio and text data, can enrich the understanding of students' classroom behaviors and enable more comprehensive analysis. The fusion of visual cues with audio signals or text transcripts can provide a holistic view of classroom interactions and facilitate more accurate behavior detection.

Lastly, conducting user studies and evaluations in real-world classroom environments would be valuable to assess the practical implications and impact of the proposed method. Understanding the perspectives and experiences of teachers, students, and other stakeholders can guide the refinement and optimization of the system for real-world deployment.

In conclusion, our study presents a transformer-network-based method for students' classroom behavior detection in the classroom, demonstrating improved accuracy and performance compared to existing approaches. By leveraging advanced techniques and constructing a dedicated dataset, we contribute to the field of students' classroom behavior analysis and open up new possibilities for automated behavior detection in educational environments. The use of transformer networks, combined with the feature pyramid structure and CARAFE operator, showcases the potential of deep learning techniques in accurately identifying and understanding students' classroom behaviors.

However, it is important to note that automated behavior detection systems should not replace the role of teachers or human observers in the classroom. Instead, they can serve as valuable tools to support educators in their efforts to monitor and manage classroom dynamics. The insights provided by these systems can assist teachers in identifying patterns, assessing student engagement, and informing instructional strategies. The combination of human expertise and automated analysis can lead to more effective and personalized educational experiences.

Ethical considerations should also be taken into account when deploying behavior detection systems in educational environments. Privacy concerns, data security, and informed consent are critical aspects that must be addressed to ensure the responsible and ethical use of such technologies. Transparency and clear communication about the purpose and functionality of these systems are essential to build trust among teachers, students, and parents.

In summary, our study demonstrates the potential of transformer-network-based methods for students' classroom behavior detection in the classroom. The incorporation of advanced techniques and the construction of a dedicated dataset contribute to improved accuracy and performance. While there are limitations and challenges to overcome, such as dataset size, low-resolution images, and computational complexity, future research can focus on addressing these issues and exploring additional avenues, including temporal modeling, transfer learning, multimodal integration, and real-world evaluations. By continuously advancing the field of automated behavior detection, we can support teachers in creating engaging and inclusive learning environments that promote positive educational outcomes for students.

6. Conclusions

In this paper, we have proposed a transformer-network-based method for the accurate detection of seven different students' classroom behaviors in the classroom. By replacing the backbone network of Deformable DETR with the Swin Transformer network, we have improved the feature extraction performance of input images. Additionally, the introduction of the feature pyramid structure and CARAFE operator has enhanced the detection accuracy for objects with different feature sizes. To address the lack of reliable datasets for students' classroom behaviors, we have constructed a dataset consisting of 1342 images and 9911 annotations.

Our proposed method has achieved a significant improvement of 6.1% in detection accuracy compared to the original Deformable DETR network. This demonstrates the effectiveness of our approach in accurately identifying student behaviors in the classroom.

However, we acknowledge certain limitations in our study. The dataset for students' classroom behavior detection is not sufficiently large, and the results for low-resolution images are not satisfactory. Moreover, the model used in our study is large, requiring substantial training time and hardware resources. To overcome these limitations, future research should focus on obtaining higher detection accuracy with lightweight models. Both real-life cases and experimental results indicate that the reading and writing behaviors

are indeed quite similar, posing significant challenges for recognition algorithms. In future research, local feature learning can be employed to focus on the hand movements of students and differentiate between reading and writing behaviors. These directions of improvements would address the mentioned limitations and make the proposed method more accessible and efficient.

By developing a robust dataset and employing advanced techniques such as transformer networks, feature pyramid structures, and CARAFE operators, our study contributes to the field of students' classroom behavior detection. The results highlight the potential of using deep learning methods for understanding and analyzing classroom dynamics, which can have significant implications for educational research and student well-being.

Author Contributions: Conceptualization, Z.W. and C.Z.; methodology, Z.W.; software, Z.W. and C.Z.; validation, Z.W., J.Y., L.L., C.Z. and C.T.; formal analysis, Z.W.; investigation, Z.W., J.Y., L.L., C.Z. and C.T.; resources, Z.W.; data curation, Z.W.; writing—original draft preparation, Z.W. and J.Y.; writing—review and editing, Z.W. and J.Y.; visualization, Z.W. and J.Y.; 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), 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 conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FPN	Feature Pyramid Network
CNNs	Convolutional Neural Networks
SSD	Single Shot MultiBox Detector
YOLO	You Only Look Once
ViT	Vision Transformer
DETR	Detection Transformer
W-MSA	Windowed Multihead Self-Attention
SW-MSA	Sliding-Window Multihead Self-Attention
MLP	Multilayer Perceptron
LN	Layer Normalization
CARAFE	Content-Aware ReAssembly of Features
mAP	mean Average Precision
RPN	Region Proposal Network
CBL	Convolutional Block Layer
PAN	Path Aggregation Network
ELAN	Efficient Local Attention Network
CAT	Category-aware Transformation

References

1. Li, L.; Wang, Z.; Zhang, T. GBH-YOLOv5: Ghost Convolution with BottleneckCSP and Tiny Target Prediction Head Incorporating YOLOv5 for PV Panel Defect Detection. *Electronics* **2023**, *12*, 561.
2. 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]
3. Bhanji, F.; Gottesman, R.; de Grave, W.; Steinert, Y.; Winer, L.R. The retrospective pre–post: A practical method to evaluate learning from an educational program. *Acad. Emerg. Med.* **2012**, *19*, 189–194. [CrossRef] [PubMed]
4. Bunce, D.M.; Flens, E.A.; Neiles, K.Y. How long can students pay attention in class? A study of student attention decline using clickers. *J. Chem. Educ.* **2010**, *87*, 1438–1443. [CrossRef]
5. Chang, J.J.; Lin, W.S.; Chen, H.R. How attention level and cognitive style affect learning in a MOOC environment? Based on the perspective of brainwave analysis. *Comput. Hum. Behav.* **2019**, *100*, 209–217. [CrossRef]
6. Kuh, G.D. What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Chang. Mag. High. Learn.* **2003**, *35*, 24–32. [CrossRef]
7. Ashwin, T.; Guddeti, R.M.R. Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. *IEEE Access* **2019**, *7*, 150693–150709. [CrossRef]
8. Jain, D.K.; Zhang, Z.; Huang, K. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognit. Lett.* **2020**, *139*, 157–165. [CrossRef]
9. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2020**, *130*, 370–375. [CrossRef]
10. Sindagi, V.A.; Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognit. Lett.* **2018**, *107*, 3–16. [CrossRef]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
12. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
14. 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, USA, 18–22 June 2023; pp. 7464–7475.
15. Wenchao, L.; Meng, H.; Yuping, Z.; Shuai, L. Research on intelligent recognition algorithm of college students' classroom behavior based on improved SSD. In Proceedings of the 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI), Beijing, China, 6–8 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 160–164.
16. 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*, 4468025. [CrossRef]
17. Ren, X.; Yang, D. Student behavior detection based on YOLOv4-Bi. In Proceedings of the 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), Virtual, 20–22 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 288–291.
18. Tang, L.; Xie, T.; Yang, Y.; Wang, H. Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism. *Appl. Sci.* **2022**, *12*, 6790. [CrossRef]
19. Hu, M.; Wei, Y.; Li, M.; Yao, H.; Deng, W.; Tong, M.; Liu, Q. Bimodal learning engagement recognition from videos in the classroom. *Sensors* **2022**, *22*, 5932. [CrossRef] [PubMed]
20. Zheng, Z.; Liang, G.; Luo, H.; Yin, H. Attention assessment based on multi-view classroom behaviour recognition. *IET Comput. Vis.* **2022** . [CrossRef]
21. Zhang, Y.; Zhu, T.; Ning, H.; Liu, Z. Classroom student posture recognition based on an improved high-resolution network. *EURASIP J. Wirel. Commun. Netw.* **2021**, *2021*, 140. [CrossRef]
22. 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. [CrossRef]
23. Pabba, C.; Kumar, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Syst.* **2022**, *39*, e12839. [CrossRef]
24. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
25. Li, L.; Wang, Z. Calibrated Q-Matrix-Enhanced Deep Knowledge Tracing with Relational Attention Mechanism. *Appl. Sci.* **2023**, *13*, 2541. [CrossRef]
26. Lyu, L.; Wang, Z.; Yun, H.; Yang, Z.; Li, Y. Deep Knowledge Tracing Based on Spatial and Temporal Representation Learning for Learning Performance Prediction. *Appl. Sci.* **2022**, *12*, 7188. [CrossRef]
27. Wang, Z.; Hou, Y.; Zeng, C.; Zhang, S.; Ye, R. Multiple Learning Features–Enhanced Knowledge Tracing Based on Learner–Resource Response Channels. *Sustainability* **2023**, *15*, 9427. [CrossRef]
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]

29. Agrawal, P.; Girshick, R.; Malik, J. Analyzing the performance of multilayer neural networks for object recognition. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VII 13; Springer: Cham, Switzerland, 2014; pp. 329–344.
30. 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: Cham, Switzerland, 2016; pp. 21–37.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.
34. 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.
35. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3651–3660.
36. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
38. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
39. 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.
40. Jin, J.; Feng, W.; Lei, Q.; Gui, G.; Wang, W. PCB defect inspection via Deformable DETR. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 646–651.
41. Shanliang, L.; Yunlong, L.; Jingyi, Q.; Renbiao, W. Airport UAV and birds detection based on deformable DETR. *J. Phys. Conf. Ser.* **2022**, *2253*, 012024. [CrossRef]
42. Gao, L.; Zhang, J.; Yang, C.; Zhou, Y. Cas-VSwin transformer: A variant swin transformer for surface-defect detection. *Comput. Ind.* **2022**, *140*, 103689. [CrossRef]
43. Kim, J.H.; Kim, N.; Won, C.S. Facial expression recognition with swin transformer. *arXiv* **2022**, arXiv:2203.13472.
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Cham, Switzerland, 2014; pp. 740–755.
45. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv* **2017**, arXiv:1706.02677.
46. Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.* **2019**, *127*, 302–321. [CrossRef]
47. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
48. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
49. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
50. Zhu, Z. Recognition and Application of Head-Down and Head-Up Behavior in Classroom Based on Deep Learning. Ph.D. Thesis, Central China Normal University, Wuhan, China, 2019.