A Study of Classroom Behavior Recognition Incorporating Super Resolution and Target Detection

Xiaoli Zhang , Jialei Nie , ShouLin Wei * , GuiFu Zhu , Wei Dai , Can Yang

Posted Date: 26 July 2024

doi: 10.20944/preprints202407.2160.v1

Keywords: super-resolution;target detection;character interaction;classroom behavior recognition

Article

A Study of Classroom Behavior Recognition Incorporating Super Resolution and Target Detection

Xiaoli Zhang [1], Jialei Nie [2], Shoulin Wei [1,*],Guifu Zhu [3],Wei Dai [1] and Can Yang [2]

[1] Key Laboratory of Computer Science, Kunming University of Science and Technology 650500; zxl_km@kust.edu.cn
[2] School of Information Engineering and Automation, Kunming University of Science and Technology 650500; niejialei1127@163.com
[3] Kunming University of Science and Technology Informationization Construction Management Center 650500; zhuguifu@kust.edu.cn
* Correspondence: weishoulin@kust.edu.cn

Abstract: With the development of educational technology, machine learning and deep learning provide technical support for traditional classroom observation assessment. However, in real classroom scenarios, the technique faces challenges such as lack of clarity of raw images, complexity of datasets, multi-target detection errors and complexity of character interactions.Based on the above problems, a student classroom behavior recognition network incorporating super-resolution and target detection is proposed. To cope with the problem of unclear original images in the classroom scenario, SRGAN (Super Resolution Generative Adversarial Network for Images) is used to improve the image resolution and thus the recognition accuracy.To address the dataset complexity and multi-targeting problems, feature extraction is optimized and multi-scale feature recognition is enhanced by introducing AKConv and LASK attention mechanisms into the Backbone module of YOLOv8s algorithm. To improve the character interaction complexity problem, the CBAM attention mechanism is integrated to enhance the recognition of important feature channels and spatial regions. Experiments show that it can detect six behaviors of students raising their hands, reading, writing, playing cell phones, looking down, and lying on the table in high-definition images. And the accuracy and robustness of this network are verified.Compared with small target detection algorithms such as Faster R-CNN, YOLOv5, and YOLOv8s, this network has better detection performance and can efficiently deal with the behavior recognition of multiple students.

Keywords: super-resolution;target detection;character interaction;classroom behavior recognition

1. Introduction

In the context of today's quest to improve the quality of teaching and learning, the understanding of classroom behavior and affective states has a profound impact on student achievement and teaching effectiveness. Students' classroom behavior not only reflects their individual learning status, but is also an important indicator for assessing teachers' teaching effectiveness. Traditionally, teachers assess student performance through manual observation, but this method is inefficient and subjective.The application of computer vision and deep learning technologies provides an efficient and objective means of analyzing classroom behavior, which is crucial for improving teaching efficiency. In the field of deep learning-driven student classroom behavior recognition, current research is mainly based on facial expression, body movement and gesture estimation.However, the complexity of real classroom environments, such as low-resolution small targets, lighting variations, occlusion phenomena, and student overlap, reduces recognition accuracy. Improving the resolution of small targets and effectively extracting interaction features[1]and suppressing irrelevant features are the key ways to improve the precision and accuracy of classroom behavior recognition.Image super-resolution reconstruction [2] is an effective means to enhance the richness of details in digital images, which can reveal the subtle features of objects more accurately. In classroom scenarios, this technique can enhance the accuracy of classroom

 doi:10.20944/preprints202407.2160.v1

behavior recognition by increasing the resolution of small targets to capture the activity state more accurately.In addition, the application of the attention mechanism can enhance the model's focus on key features while reducing the interference of irrelevant backgrounds and optimizing the model's performance in complex interaction and occlusion situations. As a result, this study refers to the current deep learning-based classroom behavior recognition methods, incorporating the super-resolution model SRGAN and the improved target detection algorithm YOLOv8s to form a cascade detection and recognition network, which achieves the detection of six behaviors: raising the hand, reading, writing, playing with a cell phone, bowing the head, and lying down on the table. In this paper, experiments were conducted to evaluate the accuracy and processing speed of the algorithmic model using the common public datasets (SCB-datasets) and the self-constructed classroom video dataset of our university, and better results were obtained.

2. Related Work

2.1. Human Body Posture Estimation Recognition

Currently, various algorithms have been used to detect and identify students' behaviors in the classroom and to correlate and analyze them with other students' data. For example, Wu et al. [1] combined PSO algorithm with KNN algorithm in order to improve the teaching effect, obtained PSO-KNN joint algorithm, and combined with emotional image processing algorithm, constructed the classroom student behavior recognition model based on artificial intelligence.Wang Zejie et al. [2] used OpenPose human key point detection algorithm to obtain students' key point data, input convolutional neural network for learning, get the posture classifier, fused with local features of interacting objects extracted by YOLOv3 algorithm, and carried out recognition and analysis of students' behaviors.Chen et al. [3] collected seven typical classroom behavior images of 300 students and pre-processed the data, extracted the key human feature information using Openpose technology, and then trained a deep learning model VGG16 network model suitable for students' classroom behavior recognition by migration learning ResNet50 and Alex Net network model, and compared the accuracy of classroom behavior recognition.Fu Rong et al.[4] proposed a classroom learning behavior analysis framework using Faster R-CNN to detect the human body, OpenPose was used to extract the key points of the human skeleton, face, and fingers, and finally a CNN based classifier was designed for action recognition.These methods of student behavior research based on human posture estimation are usually directly extracting sparse key point locations on the human skeleton, which is more ideal for classroom behavior recognition under specific environmental experiments, but in actual complex classroom scenarios, such as character occlusion, overlapping, complex character interactions, and students' varying proximity and distance, these objective factors will lead to poor accuracy of behavior recognition based on human posture estimation.

2.2. Recognition Based on Human-Object Interactions

It has been shown that recognizing the classroom action behavior of multiple students, the clarity of image video is an important factor affecting the recognition accuracy, recognizing the interaction activities of the target object with the surrounding objects, and extracting effective student classroom behavioral features in both the channel and spatial dimensions is an important method to enhance the robustness of the model.In recent years, researchers have been working on the detection of character interaction complexity and have made some significant progress.For example, Kolesnikov et al. [5] proposed BAR-CNN network to encode the spatial location relationship between people and objects with the help of chain rule decomposition probabilistic network.The Visual-Spatial-Graph Network (VSGNet) proposed by Ulutan et al. [6] better characterizes spatial relationships by constructing a graph of human-object interactions.Wang et al. [7] proposed a method based on the YOLOv5s network structure to identify and analyze student classroom behavior. The method involves using convolutional layers in YOLOv5s to extract deep features and applying Squeeze-and-Excitation (SE) attention mechanisms to reduce the weight of interpersonal interaction information during recognition. Finally, the extracted features are classified using Feature Pyramid

 doi:10.20944/preprints202407.2160.v1

Networks (FPN) and Path Aggregation Network (PAN) structures.Wang et al. [8] proposed the IPNet network for predicting human-object interaction points and locating and classifying interaction relationships.The above research methods mainly reason about the interaction relationship by extracting the appearance features and spatial relationship between people and objects, but they lack attention to focusing on the important features of people's interaction, suppressing irrelevant information in the background, varying the distance of students, and enriching the degree of detail of small targets, and there is still a large potential to improve the recognition accuracy.Considering that YOLOv8s network has significant advantages in recognition accuracy, detection speed, and character interaction, this study proposes a student classroom behavior recognition model incorporating super-resolution network SRGSN based on YOLOv8, and optimizes and improves the target detection module, and the character interaction relationship construction link, so as to realize classroom behavior recognition. The main contributions of this study are as follows:

- SRGAN (Image Super Resolution Generative Adversarial Network) is used to generate the original image with high resolution, enrich the degree of detail of small targets, enhance the spatial features of human-object interaction relationship, and improve the accuracy of recognition;
- add variable kernel convolution AKConv in the Backbone module of target detection algorithm YOLOv8s, through the variable kernel convolution to adjust the initial rule pattern of adopting the network, according to the actual needs of adjusting the shape and size of the samples, so as to enable the network to adapt to different datasets and detect more targets;
- In the SPPF of the Backbone module of YOLOv8s, the integration of the LASK attention mechanism expands the receptive field and acquires wider contextual information, which significantly improves the feature aggregation capability of the SPPF module at multiple scales. It makes the network more focused on target-related features, which in turn improves the detection accuracy;
- In order to validate the effectiveness of the proposed network, the study evaluates the accuracy and processing speed of the algorithmic model by utilizing the publicly available dataset (SCB-datasets) and the self-constructed classroom video dataset of our university. The results show that the network exhibits higher recognition accuracy with faster processing speed under the challenges of unclear original images, multiple targets, overlapping characters and complex interactions;

3. Recognition Network Model Used in this Study

3.1. System Architecture

The network model proposed in this study is shown in Figure 1, he consists of SRGAN and improved YOLOv8s[9] target detection network in turn. In this figure, first SRGAN converts low resolution classroom video frames into higher resolution classroom video frames to generate clear classroom images, after that the generated high definition classroom images are used as inputs to the improved YOLOv8s by combining the variable kernel convolution and attention mechanism. Finally six behavioral classifications of students are obtained.

 doi:10.20944/preprints202407.2160.v1

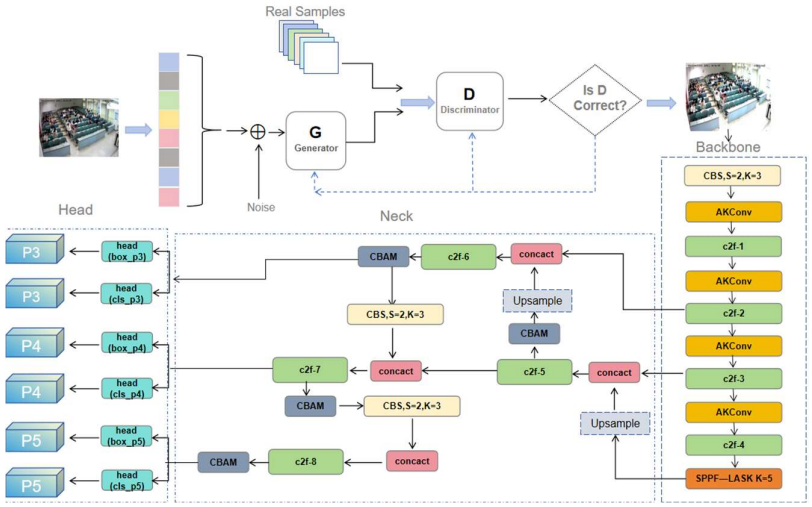

Figure 1. network structure.

Classroom behavior recognition covers character feature extraction and target detection and classification. First, feature extraction is performed on high-resolution images by an improved Backbone network. Second, in Neck network, an optimized FPN structure is used to fuse multi-scale feature maps to enhance semantic representation. Finally, Head network processes the feature maps and decodes them through convolutional and fully connected layers to predict the target location and category.

3.2. Super-Resolution Generative Adversarial Networks

To solve the problem of insufficient image clarity and missing details of small targets in real classroom scenarios, an SRGAN network[10] is used.SRGAN is based on the SRResNet generative network and a novel perceptual loss function that combines adversarial loss with content loss to optimize the image resolution. Compared to the traditional SR model using the MSE loss function that leads to excessive image smoothing and high PSNR values but lack of perceptual quality, SRGAN optimizes the perceptual loss by GAN and trains the generative network G to deceive the discriminator D to produce super-resolution images with high perceptual quality.Therefore in SRGAN network we replace MSE loss with preceptual loss with the following mathematical expression:

$$l^{SR}=\underbrace{l_X^{SR}}+\overbrace{10^{-3}\times l_{GEN}^{SR}} \quad (1)$$

Where l_X^{SR} is the content loss and l_{GEN}^{SR} is the adversarial loss, the content loss is taken as the MSE and a certain percentage of the GAN network's already existing loss function consisting of the expression:

$$l_X^{SR}=l_{MSE}^{SR}=\frac{1}{r^2}\times\frac{1}{W}\times\frac{1}{H}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left(I_{x,y}^{HR}-G_{\theta_G}(I^{LR})_{x,y}\right)^2 \quad (2)$$

The adversarial loss function is the form commonly used in GANs, and after minimization the expression is:

$$l_{GEN}^{SR}=\sum_{n=1}^{N}-\log D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right) \quad (3)$$

The purpose behind this is to allow the image generated by the generative network G to fool the discriminator D in the GAN, where the loss function parameters are solved for:

$$\hat{\theta}_G=\arg\min_{\theta_G}\frac{1}{N}\sum_{n=1}^{N}l^{SR}\left(G_{\theta_G}(I_n^{LR}),I_n^{HR}\right) \quad (4)$$

In the discriminator D, the solution parameters can then be transformed as follows:

$$\min_{\theta_G}\max_{\theta_D}E_{I^{HR}P_{train}(I^{HR})}\left[\log D_{\theta_D}(I^{HR})\right]+E_{I^{LR}P_G(I^{LR})}\left[1-\log\ _{\theta_D}(I^{LR})\right] \quad (5)$$

For GAN networks, they address a min-max type problem. In equation 5, it essentially keeps 'θ_G' unchanged to learn and adjust 'θ_D', with the goal of training a discriminator network 'θ_D'.

 doi:10.20944/preprints202407.2160.v1

3.3. Improved YOLOv8s Network

3.3.1. YOLOv8s Network

The YOLOv8 network[11], developed by Ultralytics in 2023, is an advanced SOTA model of the YOLO family, which improves on previous versions to enhance performance and adaptability.YOLOv8 offers five different sized models, of which YOLOv8s was selected as the base network because of its overall superiority in terms of accuracy and speed. YOLOv8s is divided into Backbone, Neck, and Head.

The Backbone part, including Conv, C2f and SPPF modules, adopts a structure similar to CSPDarknet and is responsible for extracting features from the input image.The Conv module consists of five 3*3 convolutional layers, which reduces the computation and obtains the global information by initial successive downsampling.The C2f module is a customized module for YOLOv8s. Compared with the ELAN of YOLOv7 and the C3 structure of YOLOv5 as shown in Figure 2, C2f provides more hopping layer connections and channel fine-tuning to facilitate feature fusion and extraction.

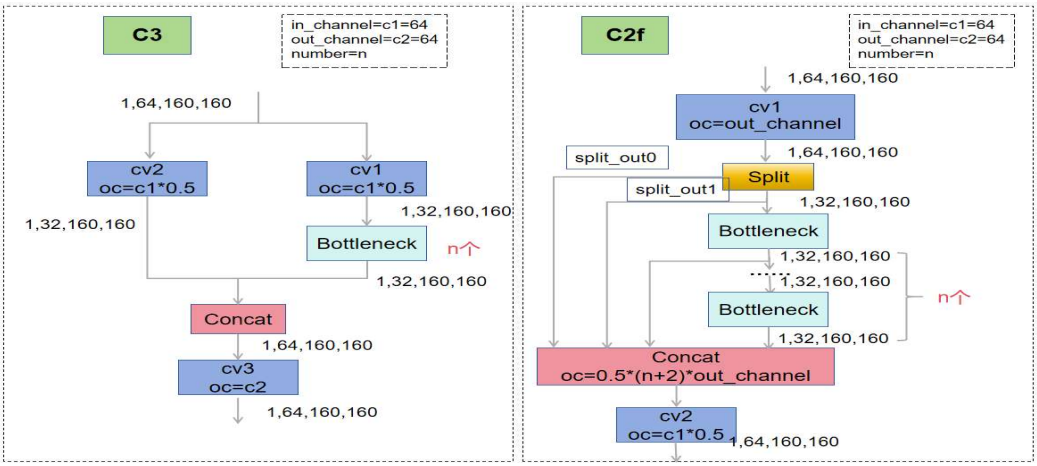

Figure 2. C3 and C2f network structure.

The SPPF module i.e., Spatial Pyramid Pooling is shown in Figure 3, which is used to aggregate features at multiple scales, comparing with SPP[12], SPPF changes the simple parallel max pooling to serial + parallel, which reduces the parameters and increases the speed of computation but does not change the computational results.

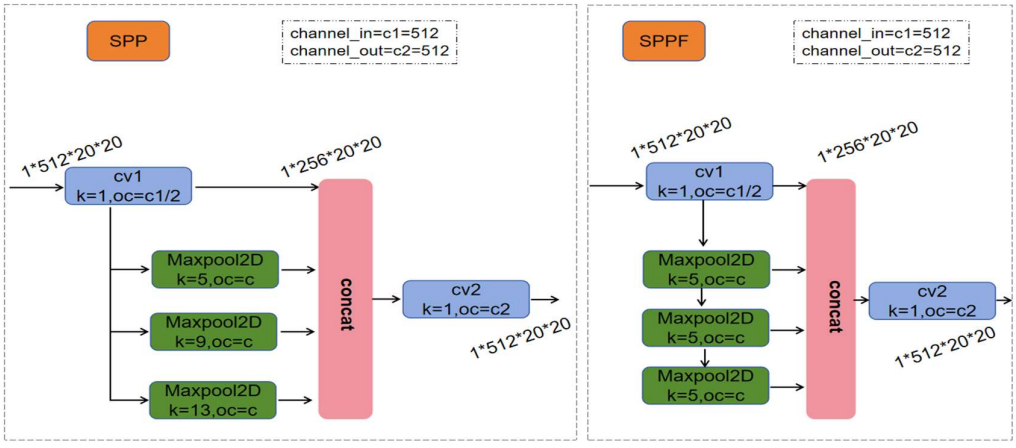

Figure 3. SPP and SPPF network structure.

 doi:10.20944/preprints202407.2160.v1

Neck, as the part that performs feature fusion on the feature maps output from the backbone network, includes two up-sampling, fully connected, C2f modules and two convolution, fully connected, C2f modules.The Neck part concatenates the extracted features from three different scale feature maps obtained by the SPPF module by up-sampling, and outputs the first feature map (80*80) to the Head layer after two up-sampling, and then outputs the remaining two feature maps (40*40 and 20*20) to the Head layer by down-sampling the feature maps through the CBS module.

The Head layer network is responsible for predicting the location and class of the target, and a decoupled head design is used to generate feature maps for Boxes and CLs, respectively.For example, the input P3 feature map (1,256,80,80) is processed through the CBS module and the convolutional layer to generate (1,64,80,80) feature maps for Box prediction and (1,nc,80,80) feature maps for CLs prediction, where nc denotes the number of categories. the Head layer ultimately produces 3 Box and 3 CLs feature maps.

3.3.2. YOLOv8s Network with Variable Kernel Convolution

Compared to standard convolutional operations, the variable kernel convolution AKConv[13] effectively adapts to target variations and captures information from a wider range of locations by endowing the convolution kernel with an arbitrary number of parameters and sampling shapes. The introduction of AKConv in the network as in Figure 4 aims to improve the adaptability to different datasets and target detection, thus improving the accuracy of feature extraction.

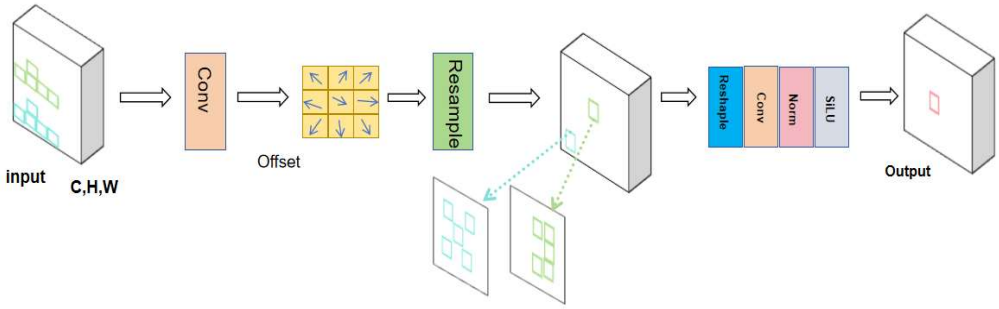

Figure 4. AKConv network structure.

The core of AKConv lies in its initial sampling coordinate algorithm, which dynamically adjusts the sampling position of the convolution kernel to adapt to specific images and targets, thus realizing the resampling of the feature map. This process includes resampling, reshaping, re-convolution, normalization, and finally outputting the results through the SiLU activation function, aiming to improve the accuracy and efficiency of feature extraction.The results show that AKConv can effectively adapt to targets of various sizes and shapes and improve the accuracy and efficiency of feature extraction, thus enabling the convolutional neural network to extract a wider range of features and solving the problem of complex and diverse student classroom behavioral data and multiple detection targets.Therefore, this study combines the Backbone part of AKconv to YOLOv8s.In order to better fit the data and verify the validity, this study also did auxiliary experiments by adding AKConv in different parts of the convolutional layers respectively. Based on the experimental results Table 5,finally all the convolutional layers in the Backbone part of the model were replaced with AKconv as shown in the Backbone part of Figure 1.

Define the initial sampling coordinate algorithm, which is based on convolutional operations to localize the features at the corresponding locations by means of a regular sampling network. The initial sampling coordinate algorithm expression is:

$$L_x = L_0 + \Delta L_x \quad (6)$$

Where $L_x$is the initial coordinate for the irregular convolution, $L_0$is the initial sampling coordinate, and $\Delta L_x$is the adjusted coordinate change according to the offset amount.According to the weights calculated from the adjusted coordinates, the input feature maps are resampled, and finally, the

 doi:10.20944/preprints202407.2160.v1

resampled feature maps are subjected to a convolution operation through the convolution layer to obtain the final output. Then the functional expression of AKConv variable kernel convolution can be defined as:

$$\mathrm{AKConv}(L_0) = \sum \omega(L_0 + L_x) \quad (7)$$

ω denotes the convolution parameter.

3.3.3. A Spatial Pyramid Pooling with Attention (SPPF_LSKA)

Incorporating LSKA[14] attention mechanism improves SPPF and reduces classroom behavior recognition errors.LSKA captures extensive contextual information through a large convolutional kernel and improves multi-scale feature extraction.YOLOv8 handles multi-scale features of SPPF more efficiently, and LSKA weights feature maps so that the network focuses on target-relevant features and improves detection accuracy.

When adding the LSKA attention mechanism to the model, if it is added at the initial stage, although it can enhance the feature expression power, it may prematurely limit the feature learning in the subsequent layers, and it is suitable for the case of small amount of data or single feature. Adding it after intermediate layers such as C2f or Conv module can balance the feature extraction and expression ability, but the features are not sufficiently multiscale processed and may not utilize the potential of LSKA to deal with complex features.The SPPF module provides rich multi-scale features, and the addition of LSKA can enhance the detection accuracy by filtering the reinforced task-related information through the attention mechanism. The experiment compares the effect of adding LSKA in each part, the results are shown in Table 6, and the results show that the best position is shown in Figure 5 after all maximal pooling layer operations and before the fully connected layer in SPPF.

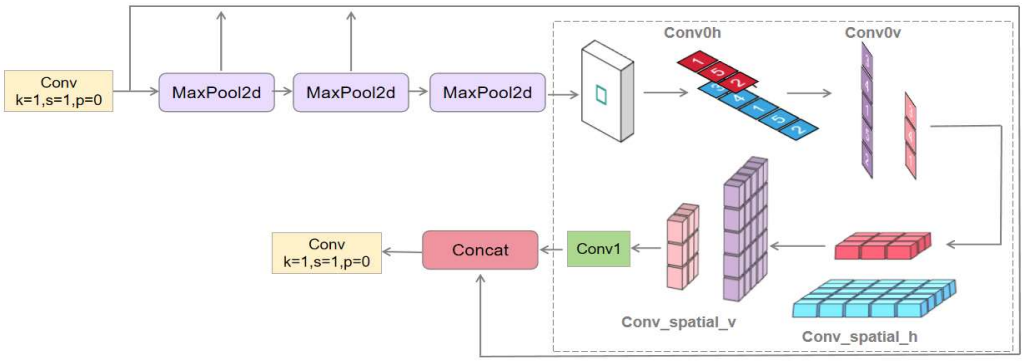

Figure 5. SPPF_LSKA network structure.

The network receives the previous layer of feature maps as input to provide the basis for the LSKA attention mechanism.The input feature map is processed by horizontal and vertical convolutional layers to extract features in horizontal and vertical directions, respectively, to generate a preliminary attention map. Then, LSKA further extracts features by spatially expanding convolution with different expansion rates to cover a larger receptive field and capture a wider range of contextual information. These operations enhance the model's understanding of image spatial relationships.Finally, the features are fused through the Conv1 convolutional layer to generate the final attention map. This attention map is subjected to an element-level multiplication operation with the original input feature map, so that each element in the original feature map is weighted according to the value of the attention map, and the resulting weighted feature map is passed as the output of the LSKA module to the subsequent layers of the network.These feature maps contain richer and more precise information, which helps to reduce the error in model identification.

3.3.4. Improved Feature Fusion Section

 doi:10.20944/preprints202407.2160.v1

The feature fusion component[15] of YOLOv8 is critical to network performance, combining different levels of feature information to account for both detailed and global information. However, in the case of complex classroom behavioral character interactions, fusing this information simultaneously increases computational complexity and may cause feature mismatch problems.To solve this problem, attention mechanisms can be introduced to selectively fuse complex classroom behavioral features. Attention mechanisms are usually classified into four major types: channel, spatial, temporal, and branching, but these unidirectional mechanisms cannot solve the computational and feature matching problems caused by the complexity of classroom behavioral features.Therefore, this study introduces the CBAM[16] attention mechanism in the feature fusion part as shown in Figure 6. CBAM combines spatial and channel attention, and its lightweight design avoids extra computational burden and keeps the model efficient. In the spatial dimension, CBAM pays attention to important spatial locations and suppresses irrelevant information; in the channel dimension, it distinguishes features by calculating channel importance for better feature matching.

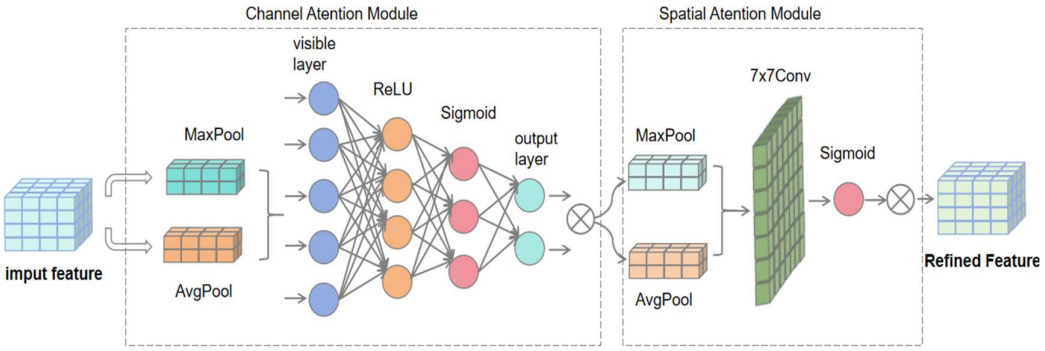

Figure 6. Convolution Block Attention Module.

In this study, the CBAM[17–19] attention mechanism was added after the four C2f layers in the Neck section as shown in Figure 1.CBAM receives the feature maps output from the C2f[20] layer and performs global maximum and average pooling in the channel attention module to capture channel global information and lay the foundation for hierarchical feature matching. The spatial attention module also uses these two pooling operations but performs them channel-by-channel, and the results are concatenated and passed through the convolutional layer to generate the spatial attention weights, adjust the feature map response, and focus on the key regions. Eventually, the adjusted feature map is multiplied element-by-element with the original input to obtain the attention-enhanced feature map, which optimizes the channel and spatial dimensions and enriches the information for subsequent target detection tasks[21].

4. Experimental Results

4.1. Data Set

This study establishes an experimental dataset based on the publicly available dataset SCB-datasetsand the real classroom video dataset of our university. Six videos of 60 minutes each were captured in two classrooms via 720P cameras, converted to static PNG format images using FFmpeg, and filtered for high quality key frames, resulting in 7378 images.The Labelimg tool was used for object annotation, defining six behavioral categories, namely, raising hand, reading, writing, playing cell phone, head down, and lying on the table, with a total of 51,957 labels. The labeling information is saved in TXT files, corresponding to the image names.

After labeling, this experiment divides the labeled 51,957 data into training and validation sets in the ratio of 8:2, where 41,565 data are in the training set and 10,392 data are in the validation set. This is quite challenging to capture the behavior in the dataset due to the complex scene environment and limited camera resolution.In order to enrich the degree of detail of small targets, enhance the

 doi:10.20944/preprints202407.2160.v1

spatial features of human-object interaction relationship, and improve the model recognition accuracy, SRGAN is used to generate high-resolution images by putting the processed dataset into a one-to-one process, as shown in Figure 7, where the left column is the original image captured, and the right column is the high-resolution image generated by the SRGAN model.

Figure 7. Original images and images generated by SRGAN.

4.2. Experimental Environment and Configuration

The hardware platform and software version used in this experiment are shown in Table 1.

Table 1. Experimental environment.

Name	Parameter	
CPU	11th Gen Intel(R) Core(TM) i5-11400F @ 2.60GHz	2.59 GHz
GPU	NVIDIA GeForce RTX 3050	
Memory	16G	
Operating System	Windows11	
PyCharm	2020.1 x64	
Python	3.9.16	
Frameworks	Pytorch 1.12.1+cu113	
CUDA	Version: 12.0	

4.3. Indicators for Evaluation

The evaluation metrics used in this experiment (Table 2) are shown[22–24], including confusion matrix normalization, precision, recall, F1 value, intersection and union ratio (IOU), mAP, mAP50, and mAP50-90.Confusion Matrix Visual Comparison of Classification Accuracy.Precision measures the proportion of true cases among positive cases.Recall measures the proportion of true positive examples that are correctly predicted.The F1 value is the reconciled mean of precision and recall.IOU measures the extent to which the prediction frame overlaps with the labeling frame.mAP denotes multi-category average precision. mAP50 and mAP50-90 denote mAP values in the 50% and 50-95% IOU threshold ranges, respectively.

Table 2. performance metrics.

 doi:10.20944/preprints202407.2160.v1

metrics	description
Precision	How many of the predicted positive samples are correct
Recall	How many of the truly positive samples were correctly predicted as positive
F1-Score	A metric that takes into account both precision and recall
IOU	Assessing the overlap between predicted and actual bounding boxes
mAP	mAP, or mean Average Precision, is the average of AP values calculated under multiple IOU thresholds

4.4. Experimental Results and Analysis

4.4.1. Variable Kernel Convolution Based Ablation Experiments

To cope with the dataset complexity and multi-target problem, this study embeds the variable kernel convolutional AKConv into the backbone network, feature integration layer, and detection head section of YOLOv8s, and analyzes the effect of AKConv position on the evaluation metrics of the recognition task through ablation experiments.The experimental results (Table 3) show that after adding AKConv to the Backbone, SPPF, and Neck sections of YOLOv8s, all evaluation indexes are improved, proving the effectiveness of AKConv.

Table 3. Experimental results of adding AKConv to different locations.

AKConv position	Precision(%)	Recall(%)	F1(%)	mAP50(%)	mAP50-90(%)
YOLOv8s	85.65	84.78	85.35	87.74	67.55
Backbone	88.78	87.42	87.5	89.86	70.56
SPPF	86.11	86.31	87.14	89.52	67.67
Neck	85.68	85.16	87.06	89.24	69.17
Backbone+Neck	84.45	85.21	86.12	88.12	66.78
Backbone+SPPF	85.98	84.88	86.12	88.45	68.12

In this case, when all the convolution kernels in the Backbone structure were replaced with the variable kernel convolution AKConv, the evaluation metrics were significantly improved compared to the original model.This change resulted in a 3.13% increase in precision, 2.64% increase in recall, 2.15% increase in F1 value, 2.12% increase in mAP50, and 3.01% increase in mAP50-90; therefore, based on the experimental results, the present study embedded AKConv into the Backbone part of YOLOv8s.

4.4.2. Ablation Experiments Based on LSKA Localization

After capturing the wide range of contextual information of the image through a separable convolutional kernel, the multi-scale feature extraction capability of the model is effectively improved, and in order to maximize the utilization of these features, LSKA is added on top of the SPPF module, which further filters and strengthens the information that is helpful to the task through the attentional mechanism, thus improving the detection accuracy.In this study, positioning comparison experiments were done for different locations of LSKA added to the SPPF module, and the results of the experiments are shown in Table 4 below.

Table 4. Comparing LSKA positioning in the SPPF module.

LSKA position	Precision(%)	Recall(%)	F1(%)	mAP50(%)	mAP50-90(%)
YOLOv8s(baseline)	85.65	84.78	85.35	87.74	67.55
Conv_LSKA	86.25	85.21	86.35	87.98	68.12
MaxPool2d_LSKA	***89.46***	***88.65***	***87.43***	***90.15***	***69.97***
Concat_LSKA	87.21	86.98	87.25	88.25	68.14

 doi:10.20944/preprints202407.2160.v1

Bulleted

Compared to the YOLOv8s benchmark model, the detection accuracy of the classroom behavior recognition task was improved with the addition of the LSKA attention mechanism to SPPF. Experiments showed that the evaluation metrics improved most significantly when LSKA was added after all maximal pooling layers and before the fully connected layer in the SPPF module (bolded font in Table 6), so this study chose to add LSKA at this location.

4.4.3. Improved Feature Fusion Based Partial Ablation Experiments

In the feature fusion section, features from different network layers are combined to take into account both detailed and global information. However, classroom behavioral features are complex and simultaneous fusion may increase computational complexity and lead to performance degradation, and differences in the statistical properties of features at different layers may lead to mismatch problems.To solve these problems, an attention mechanism is introduced to selectively incorporate important features. In this study, the effectiveness of incorporating the attention mechanism was verified through comparative experiments, and the results are shown in Table 5.

Table 5. Comparing LSKA positioning in the SPPF module.

Attention	Precision(%)	GPU_mem(G)	Params(M)	FLOPs(B)
YOLOv8s(baseline)	85.65	4.1	11.2	28.6
GAM	89.45	4.9	19.8	32.8
ECA	91.3	5.4	25.2	56.9
SENet	84.31	4.5	15.2	30.2
ShuffleAttention	88.90	4.7	17.8	34.5
CBAM	**90.12**	**4.6**	**17.89**	**31.54**

lists look

As seen in the table, this study incorporates GAM[25], ECA[26], SENet[27], ShuffleAttention[28] and CBAM attention mechanisms in the feature fusion section and evaluates them by accuracy, computational resource consumption and model complexity.The results show that the addition of the CBAM attention mechanism improves the accuracy by 4.47% over the YOLOv8s baseline without increasing the model complexity and computational resource consumption, and the GPU resource consumption is relatively small. Although the ECA attention mechanism improves the accuracy to 91.3%, its GPU resource consumption and model complexity are large, increasing the computational cost.After comprehensive comparison, this study chose to add the CBAM attention mechanism after the four C2f layers in the feature fusion section.

In order to further validate the role of the embedded attention mechanism CBAM, this study used the class activation graph method to visualize the attention graph, and Figure 8 shows the results of the CBAM attention visualization.In Figure 8 the original approach focuses more on the whole classroom and ignores the character interactions around the aggregated students due to the aggregation of characters and the larger background area, however, the approach focuses on where the students are aggregated through both spatial and channel dimensions, which makes the network pay more attention to the contextual cues around the students and thus improves the recognition performance.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2024 doi:10.20944/preprints202407.2160.v1

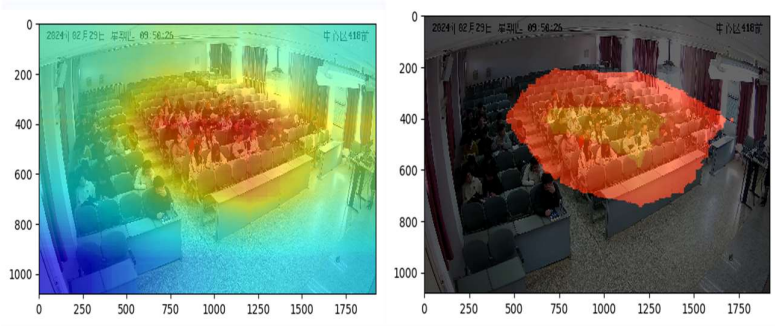

Figure 8. Visualization results of CBAM.

4.4.4. Comparative Experiments with Different Models

To verify the superiority of the improved algorithm, YOLOv8s, Faster R-CNN[29], OpenPose[30], YOLOv5[31], and YOLOv7[32] are selected for comparison experiments in this study, keeping the same base parameters.The experiments were conducted using accuracy and mAP50 metrics, and values were taken at 30 round intervals to plot the curves, and the results are shown in Figures 9 and 10.The improved YOLOv8s outperforms the other models in both accuracy and mAP50, and the accuracy and mAP50 are improved by 2% and 2.3%, respectively, compared with the baseline YOLOv8s model after processing by SRGAN. Combining the experimental results, the model proposed in this study has the best performance, especially 14.2% improvement over YOLOv5 in mAP50.

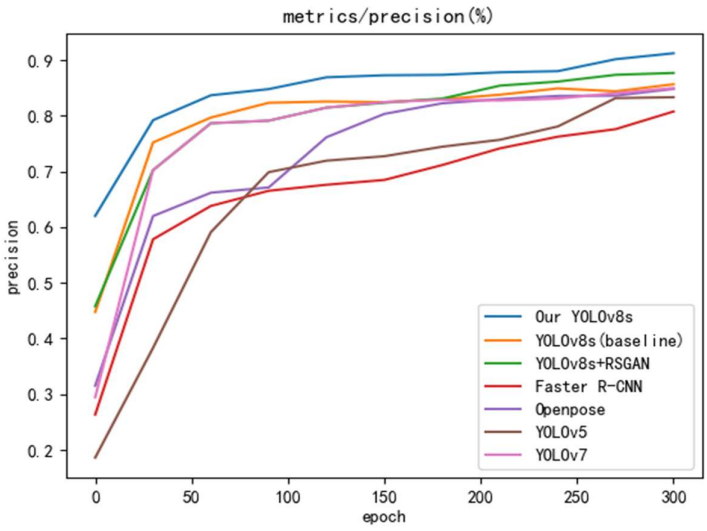

Figure 9. Comparative experimental results of precision.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2024 doi:10.20944/preprints202407.2160.v1

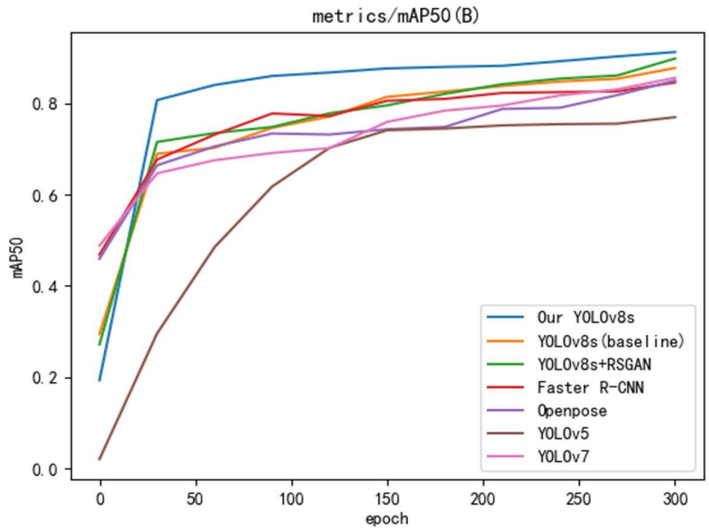

Figure 10. Comparative experimental results of mAP50.

In order to evaluate in depth how the improvement of the models in this study affects the recognition effect on the six categories of behaviors, therefore, the recognition rates of the above mentioned models on the six categories of behaviors were compared separately, and the experimental results are shown in Table 6.

like

From Table 7, it can be clearly seen that the improved method of this study has improved very much in the recognition rate of Hand-raising, Reading, and Writing behaviors, and the average recognition rate of this study's method in the six categories of behaviors has reached more than 90% compared to several other models, and from the point of view of the recognition effect on behaviors affecting the six categories of behaviors, the method has been effective and competitive in performing the multi-target detection tasks with a certain degree of effectiveness and competitiveness.

Table 7. Recognition results of student behavior using different models.

classroom behavior	Our YOLOv8s	Faster R-CNN	OpenPose	YOLOv5	YOLOv7
Hand-raising	**0.995**	0.842	0.836	0.884	0.841
Reading	**0.920**	0.839	0.819	0.888	0.802
Writing	**0.925**	0.814	0.789	0.892	0.782
Using phone	**0.962**	0.959	0.958	0.961	0.966
Bowing the head	**0.894**	0.901	0.888	0.842	0.905
Learning over the table	**0.991**	0.984	0.983	0.980	0.988

4.4.5. Results and Analysis of this Experiment

The results of PR_Curve (plot of Precision vs. Recall) for the student behavioral validation set on the proposed method in this study are presented in Figure 11.

 doi:10.20944/preprints202407.2160.v1

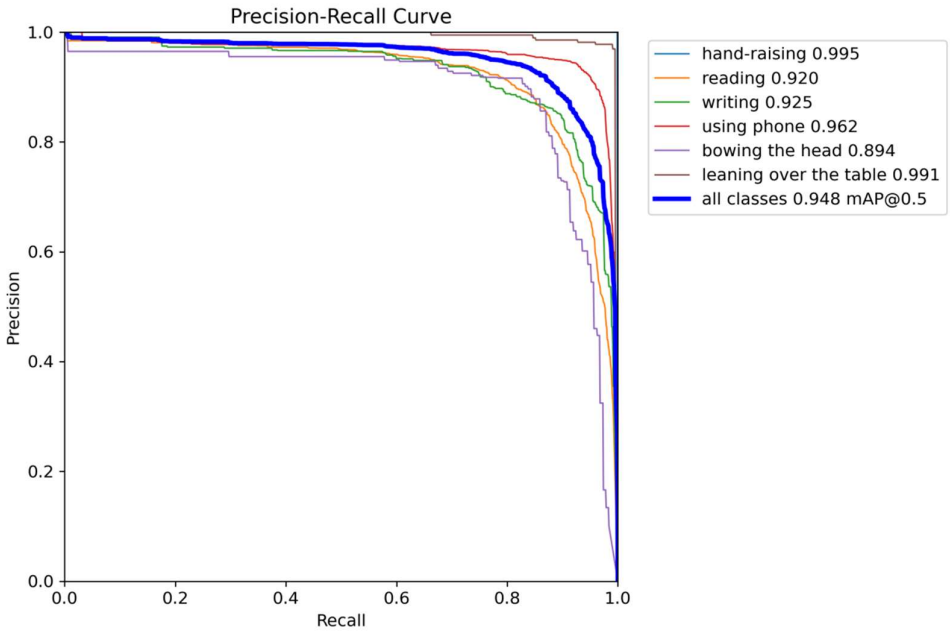

Figure 11. Results of the PR_Curve on the validation set.

In the PR_Curve, the horizontal axis represents recall, and the vertical axis represents precision. They often have a negative correlation; a curve closer to the top right corner indicates that the model has high precision and high recall, meaning accurate predictions. Therefore, as shown in the figure, this research method exhibits high prediction accuracy on the validation set.

Figure 12 shows the confusion matrix of the validation set, and the numbers on the diagonal represent the recall of each category. As can be seen from the figure, the recalls of the combinations of similar behaviors such as “reading” and “learning over the table”, “using phone” and “bowing the head” are higher. “bowing the head”, showing the effectiveness of the proposed method in distinguishing similar behaviors. Therefore, the confusion matrix also reflects the high overall recognition rate of our method on the validation set.

 doi:10.20944/preprints202407.2160.v1

Figure 12. Confusion matrix with proposed method on student behavior validation.

Figure 13 shows the performance of the training and validation sets in terms of localization loss, confidence loss and classification loss, and depicts the performance change curves for precision, recall and average precision during training. As can be seen from the figure, the various losses decrease dramatically after a few iterations and drop to a low level and stabilize after 300 rounds of iterations. The model performance improves with the number of training rounds, stabilizes after 200 iterations, and reaches the best performance at 300 iterations. Label smoothing almost overlaps with the experimental results, indicating that the model structure is well-designed without overfitting or underfitting.

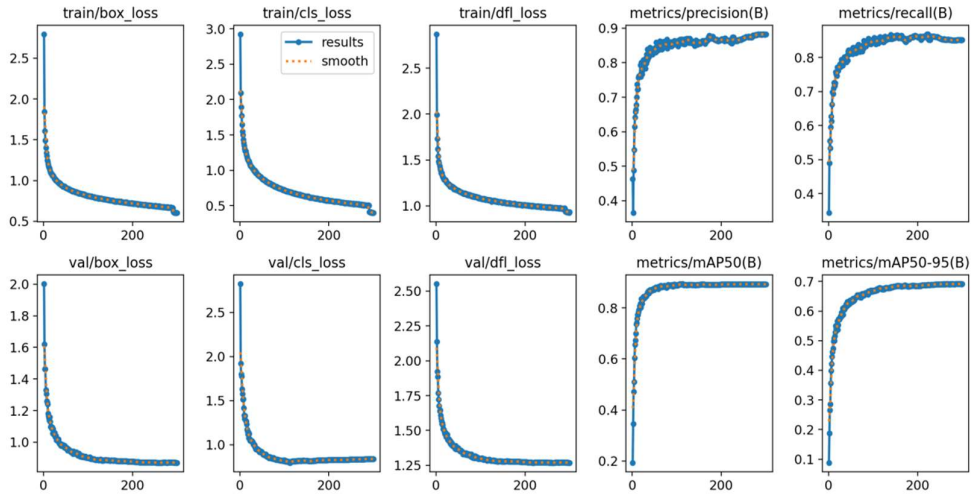

Figure 13. training loss functions and performance curves.

The test results of the optimized YOLOv8s model on another classroom video dataset are shown in Figure 14.Despite the challenges faced by this dataset, such as lack of clarity, target diversity, complex interactions, and object occlusion, the improved model accurately identifies student behaviors, showing excellent robustness. This improvement significantly enhances the performance

 doi:10.20944/preprints202407.2160.v1

of YOLOv8s model in classroom learning behavior detection, capturing student engagement more comprehensively and accurately. This result confirms the feasibility and effectiveness of the improved YOLOv8s model in practice.

Figure 14. Classroom behavior detection results.

5. Discussion

In this paper, we use SRGAN technique to synthesize high-resolution images as input for the recognition task. In order to improve the feature extraction capability and global and local information fusion of YOLOv8s model, we introduce variable kernel convolution and attention mechanism in Backbone and Neck part.These improvements allow the network to extract features more comprehensively and selectively fuse them according to the importance of the information, effectively handling details and global information. In the six student classroom behavior recognition tasks, the model demonstrates high accuracy on three behaviors: hand-raising, using phone, and leaning over the table.It is also effective in recognizing for READING and WRITING. However, its recognition performance is affected by the unbalanced behavioral categories, especially the small sample size of bowing the head. This requires future research to explore data augmentation and category balancing strategies to optimize the minority category recognition rate. The solution strategies include oversampling the minority category or undersampling the majority category, as well as assigning different weights to different categories during training to focus on the minority category and improve the model performance. It is expected that model performance will be further improved by oversampling or undersampling.

6. Conclusion

To address the challenges in real classroom scenarios, this study proposes an improved algorithm based on YOLOv8s.To solve the problem of lack of image clarity, SRGAN technique is used to generate high-resolution images. To face the diversity and complexity of the dataset, variable kernel convolution AKConv is introduced into the Backbone module of YOLOv8s.To enhance the multi-scale feature extraction capability, LASK attention mechanism is integrated into SPPF. To address the complexity of character interaction, CBAM attention mechanism is added.These improvements enhance the feature extraction and recognition ability of the model in complex scenes with an accuracy of 91%, 91.5% and 70% for mAP50 and mAP50-95, respectively.In the future, we will collect more diverse datasets to improve the model generalization ability and applicability, and improve the behavior recognition algorithms, especially the small-sample learning and long-tailed

 doi:10.20944/preprints202407.2160.v1

distribution problems, in order to improve the recognition accuracy of uncommon behavioral categories.

Author Contributions: Xiaoli Zhang conceptualized and designed the experiment; Jialei Nie performed the experiments and analyzed the data, contributing to the materials and analytical tools;Xiaoli Zhang and Jialei Nie wrote the manuscript;Finally Shoulin Wei and Guifu Zhu provided relevant information and valuable comments for this article. All authors have read and agree to the published version of the manuscript.

Funding: This work was supported by the Yunnan Provincial Department of Education Science Research Fund Project "Research on Digital Competency Evaluation and Enhancement of University Teachers Based on Deep Knowledge Tracking", 2024J0105;National Natural Science Foundation of China (No.11903009).

References

[1] WU S. Simulation of classroom student behavior recognition based on PSO-kNN algorithm and emotional image processing [J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(4): 7273-83.
[2] ZEJIE W, CHAOMIN S, CHUN Z, et al. Recognition of classroom learning behaviors based on the fusion of human pose estimation and object detection [J]. Journal of East China Normal University (Natural Science), 2022, 2022(2): 55.
[3] CHEN G, JI J, HUANG C. Student classroom behavior recognition based on openpose and deep learning; proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), F, 2022 [C]. IEEE.
[4] FU R, WU T, LUO Z, et al. Learning behavior analysis in classroom based on deep learning [J]. 2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP), 2019: 206-12.
[5] KOLESNIKOV A, KUZNETSOVA A, LAMPERT C, et al. Detecting visual relationships using box attention; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision workshops, F, 2019 [C].
[6] ULUTAN O, IFTEKHAR A, MANJUNATH B S. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2020 [C].
[7] WANG Z, YAO J, ZENG C, et al. Yolov5 enhanced learning behavior recognition and analysis in smart classroom with multiple students; proceedings of the 2022 International Conference on Intelligent Education and Intelligent Research (IEIR), F, 2022 [C]. IEEE.
[8] WANG T, YANG T, DANELLJAN M, et al. Learning human-object interaction detection using interaction points; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020 [C].
[9] LIU Q, JIANG R, XU Q, et al. YOLOv8n_BT: Research on Classroom Learning Behavior Recognition Algorithm Based on Improved YOLOv8n [J]. IEEE Access, 2024.
[10] LIU B, CHEN J. A super resolution algorithm based on attention mechanism and srgan network [J]. IEEE Access, 2021, 9: 139138-45.
[11] LUO Z, WANG C, QI Z, et al. LA_YOLOv8s: A lightweight-attention YOLOv8s for oil leakage detection in power transformers [J]. Alexandria Engineering Journal, 2024, 92: 82-91.
[12] JOOSHIN H K, NANGIR M, SEYEDARABI H. Inception-YOLO: Computational cost and accuracy improvement of the YOLOv5 model based on employing modified CSP, SPPF, and inception modules [J]. IET Image Processing, 2024, 18(8): 1985-99.
[13] ZHANG X, SONG Y, SONG T, et al. AKConv: Convolutional Kernel with Arbitrary Sampled Shapes and Arbitrary Number of Parameters. arXiv 2023 [J]. arXiv preprint arXiv:231111587.
[14] LAU K W, PO L-M, REHMAN Y A U. Large separable kernel attention: Rethinking the large kernel attention design in cnn [J]. Expert Systems with Applications, 2024, 236: 121352.
[15] CHEVTCHENKO S F, VALE R F, MACARIO V, et al. A convolutional neural network with feature fusion for real-time hand posture recognition [J]. Applied Soft Computing, 2018, 73: 748-66.
[16] YANG K, ZHANG Y, ZHANG X, et al. YOLOX with CBAM for insulator detection in transmission lines [J]. Multimedia Tools and Applications, 2024, 83(14): 43419-37.
[17] JIA L, WANG Y, ZANG Y, et al. MobileNetV3 with CBAM for bamboo stick counting [J]. IEEE Access, 2022, 10: 53963-71.
[18] SHENG W, YU X, LIN J, et al. Faster rcnn target detection algorithm integrating cbam and fpn [J]. Applied Sciences, 2023, 13(12): 6913.

[19] FU H, SONG G, WANG Y. Improved YOLOv4 marine target detection combined with CBAM [J]. Symmetry, 2021, 13(4): 623.
[20] PISCHEDDA V, RADESCU S, DUBOIS M, et al. Experimental and DFT high pressure study of fluorinated graphite (C2F) n [J]. Carbon, 2017, 114: 690-9.
[21] CHEN Y, ZHANG X, CHEN W, et al. Research on recognition of fly species based on improved RetinaNet and CBAM [J]. IEEE Access, 2020, 8: 102907-19.
[22] SUN B, WU Y, ZHAO K, et al. Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes [J]. Neural Computing and Applications, 2021, 33: 8335-54.
[23] WANG Z, YAO J, ZENG C, et al. Learning behavior recognition in smart classroom with multiple students based on YOLOv5 [J]. arXiv preprint arXiv:230310916, 2023.
[24] LIN J, LI J, CHEN J. An analysis of English classroom behavior by intelligent image recognition in IoT [J]. International Journal of System Assurance Engineering and Management, 2022, 13(Suppl 3): 1063-71.
[25] ZAMRI F N M, GUNAWAN T S, YUSOFF S H, et al. Enhanced Small Drone Detection using Optimized YOLOv8 with Attention Mechanisms [J]. IEEE Access, 2024.
[26] JI X, NIU Y. A Lightweight Network for Human Pose Estimation Based on ECA Attention Mechanism [J]. Electronics, 2023, 13(1): 150.
[27] JIA Z, WANG K, LI Y, et al. High precision feature fast extraction strategy for aircraft attitude sensor fault based on RepVGG and SENet attention mechanism [J]. Sensors, 2022, 22(24): 9662.
[28] LIU P, WANG Q, ZHANG H, et al. A lightweight object detection algorithm for remote sensing images based on attention mechanism and YOLOv5s [J]. Remote Sensing, 2023, 15(9): 2429.
[29] LEE H, EUM S, KWON H. Me r-cnn: Multi-expert r-cnn for object detection [J]. IEEE Transactions on Image Processing, 2019, 29: 1030-44.
[30] SAIKI Y, KABATA T, OJIMA T, et al. Reliability and validity of OpenPose for measuring hip-knee-ankle angle in patients with knee osteoarthritis [J]. Scientific Reports, 2023, 13(1): 3297.
[31] LI L, LIU M, SUN L, et al. ET-YOLOv5s: toward deep identification of students' in-class behaviors [J]. IEEE Access, 2022, 10: 44200-11.
[32] YANG F. Student Classroom Behavior Detection based on Improved YOLOv7 [J]. arXiv preprint arXiv:230603318, 2023.