Received: June 29, 2023. Revised: July 23, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Students Behavior Detection Based on Improved YOLOv5 Algorithm Combining with CBAM Attention Mechanism

Muhanad Abdul Elah Alkhalisy[1]* **Saad Hameed Abid[2]**

[1]*Informatics Institute for Postgraduate Studies, Iraq*
[2]*Computer Engineering Department, Al-Mansur University College, Iraq*
* Corresponding author's Email: phd202020557@iips.icci.edu.iq

Abstract: The efficiency of distance learning can be assessed by examining specific student behaviors during online courses and assessments. Detecting abnormal student behavior can guide evaluating the effectiveness of learning. Detection speed, accuracy, and model efficiency are essential considerations in distance learning environments. This study proposes a model for detecting abnormal student behavior based on a modified YOLOv5 object detector. Firstly, to achieve model efficiency, the YOLOv5 feature extraction network is pruned by removing the portion responsible for extracting high-level feature maps of small objects, which consumes significant computing resources and parameters. Also, the model's feature fusion part and prediction layers are adjusted accordingly. A convolutional block attention module (CBAM) is added between the model's neck and prediction head to boost the model's focus on students' areas and detection accuracy. Five behaviors were tested on a recently created dataset. The results show that the suggested SPL-YOLOv5 detection method outperforms the SSD-MobileNet, Faster R-CNN, the original YOLOv5 algorithms and state-of-the-art methods regarding performance. The suggested approach increased mAP@0.5 and mAP@0.5:0.95 by 0.14% and 0.13%, respectively, compared to the original YOLOv5. In addition, the model's inference time decreased by 2.7%, computational resources (FLOPs) by 40%, and parameters by 47%.

Keywords: Distance learning, Online exam, Yolov5, Attention module.

1. Introduction

Distance learning or online education has, without a doubt, become increasingly popular over the years. More than 6 million students in the USA took at least one online course [1]. Online education, like massive open online courses (MOOC), enables educational institutions to operate more cheaply and to reach more students. These online classes allowed campus-based students to further their education, which sparked the localization of online courses and e-learning settings [2]. Observing various student behaviors during online courses and examinations makes it possible to evaluate the effectiveness of distance learning [3]. Finding unusual student behavior can help in determining how well students are learning. Speed of detection, accuracy, and model effectiveness are crucial factors in distance learning contexts [4]. The integrity of online assessments, or the requirement to administer exams using the proper resources and procedures ,is one of the educational system's main issues [5]. Universities have used various proctoring tools to monitor student behavior during online courses in response to the fast rise in online distance learning and the necessity to preserve academic integrity [6].

Incorporating artificial intelligence into E-learning and leveraging deep learning techniques to detect students' behaviors during online exams can provide valuable insights into their learning progress and improve teaching effectiveness.

Two primary categories of deep learning algorithms for object detection exist two-stage and one-stage methods. Two-stage methods, exemplified by the faster R-CNN [7] family of algorithms, involve extracting pre-selected boxes from the image and classifying the targets based on this training. While these methods offer high accuracy, they tend to have longer image processing times. In contrast, the one-stage approach, represented by the you only look once (YOLO) series [8] and single shot multi-box detector (SSD) [9], directly feeds the entire image into the neural network and promptly provides output for both object location and classification. One of the YOLO algorithm family's most recent and most advanced iteration is the YOLOv5. Real-time applications can be met using the YOLOv5 model's strong performance and quick detection speed [10].

Researchers have recently explored different approaches to tackle behavioural detection in learning environments using the YOLOv5 model. In the following, we will examine the relevant start of the art studies categorized based on the methods employed.

Z. Wang proposed a novel YOLOv5s network structure for identifying and evaluating students' behavior. The method utilizes the default YOLOv5 backbone convolutional layers to extract features from input images. To address the issue of background interference, they applied a squeeze-and-excitation (SE) attention detection mechanism to reduce the emphasis on irrelevant information during recognition. The YOLOv5 default feature pyramid networks (FPN) and path aggregation network (PAN) structures were also utilized. However, a limitation of their approach is its reliance on squeeze-and-excite networks (SE), which only perform channel weighting.

Additionally, no modifications were suggested for the feature extraction and feature fusion networks to reduce computational power and achieve a lighter model. Our proposed approach employs the convolutional block attention module (CBAM), which incorporates both channel and spatial attention weighting. This enables it to capture spatial and channel-wise dependencies, making it beneficial for tasks requiring local and global contextual information. Furthermore, we have modified the feature extraction and feature fusion networks, improving model performance [11].

W. Niu and X. Sun proposed an enhanced skeletal recognition system based on YOLOv5 for detecting classroom behavior. The YOLOv5 detection algorithm was improved to address the issue of missed detection by incorporating the coordinate attention (CA) module into the network. The CA module decomposes channel attention into one-dimensional feature encoding and is added after each CSP module in the backbone to enhance its impact on the output accuracy. However, this modification increased the computational effort of the model. To mitigate this, they modified the CSP module by removing one convolutional layer to reduce wasted resources and improve detection speed. Human skeleton information was obtained using the Alphapose framework, and the skeletal data was then fed into a two-stream adaptive graph convolution network (2S-AGCN) to recognize various classroom behaviors accurately. Extensive testing demonstrated that the proposed bone recognition-based detection algorithm improved detection accuracy and reduced the false detection rate, indicating its effectiveness. However, a drawback of their method was its reliance on addressing the partial blockage of numerous human targets, and the inclusion of the skeletal (2S-AGCN) module resulted in slower inference and detection times. In our proposed approach, we aimed to strike a balance between model speed and accuracy by modifying all parts of the model [12].

Z. Zheng employed deep learning techniques to detect and recognize student behavior from multiple perspectives. They introduced an improved detection model based on YOLOv5, focusing on optimizing the CBL module by replacing the default LeakyReLU activation function with the GELU activation function. They also utilized the SIoU loss function to expedite the convergence of the prediction box. The study retained the YOLOv5 feature pyramid structure of FPN + PAN for feature fusion and the same prediction head. According to the experimental findings, the proposed approach surpassed YOLOv5 by achieving a favorable balance between accuracy and identification speed. However, a limitation of their method was its lack of an attention mechanism to enhance performance. Instead, using GELU activation and SIoU loss function added computational complexity. In our proposed approach, we incorporated the convolutional block attention module (CBAM), which is advantageous for capturing local and global contextual information. Furthermore, we modified the feature extraction and feature fusion networks, improving model performance [13].

S. Yang developed a novel student behavior detection model based on the fundamentals of YOLOv5. By integrating multiple convolutional block attention modules (CBAM) between the feature extraction network and the feature fusion network of the original YOLOv5 model structure, the resulting CBAM-YOLOv5 model was able to mitigate the influence of background information and effectively extract robust features related to student behaviors. According to the experimental results, the proposed approach demonstrated efficiency and a solid ability to extract features. The utilization of CBAM improved the algorithm's capability to recognize student behaviors. However, a limitation of their method was its dependence on the number and

Received: June 29, 2023. Revised: July 23, 2023.

placement of the inserted CBAM modules. Adding three modules before the feature fusion network increased the computational load and could potentially result in the loss of significant feature information crucial for the feature fusion process. In our proposed approach, we adopted a single convolutional block attention module (CBAM) at the end of the feature fusion network, connected to a single prediction layer. This modification led to improved model accuracy and performance [14].

Z. Zhang and D. Ao presented an enhanced YOLOv5-based deep learning model for detecting student behavior. In their study, they introduced the CIoU loss function as a replacement for the IoU loss function, addressing the limitation of IoU in predicting the distance between the actual box and the predicted box when the two boxes do not intersect. To overcome the issue of significant neuron loss, they utilized the MetaAcon activation function instead of the Sigmoid activation function, which improved generalization and transfer performance. The experimental results demonstrated that the improved YOLOv5 model was more suitable for capturing the data characteristics of student behavior. However, a drawback of their approach was its reliance on utilizing new activation and loss functions, which increased computational requirements. In our proposed approach, we modified the feature extraction, feature fusion, and prediction layers, resulting in improved model performance while achieving a lightweight model [15].

J. Wena proposed a model for recognizing abnormal behavior in an examination room based on YOLOv5. They significantly improved the YOLOv5 backbone network to reduce the parameters and computation required during model training. To achieve this, they replaced the original YOLOv5 backbone network with MobileNetV3-Large, which allows for more efficient feature extraction by reducing the number of parameters. Additionally, they introduced a cascading attention mechanism that combines spatial and channel attention mechanisms to enhance the model's ability to extract features effectively. The model was optimized using a classification loss and the CIoU loss function. The experimental results demonstrated that the proposed model achieved favorable detection results for abnormal behavior in examinees, with metrics such as precision (P) at 92.53%, mean average precision (mAP) at 93.52%, and frames per second (fps) at 0.547. The proposed algorithm exhibited higher accuracy and faster recognition speed than several classical abnormal behavior detection algorithms. However, a limitation of their approach was the reliance on MobileNetV3-Large as the backbone network, which, although requiring fewer parameters, had lower accuracy in detection. In our proposed approach, we modified the original YOLOv5 backbone feature extraction networks, addressing this limitation [16].

There are few pertinent studies on the detection of student behavior in distance learning environments. Students' behavior detection algorithms still have flaws related to the model’s lightness, detection speed, accuracy, and performance, which should meet online learning and assessment requirements. Therefore, motivated by intelligent education, this paper proposed a lightweight Single Prediction Layer YOLOv5 (SPL-YOLOv5), a student behavior detection model based on the improved YOLOv5 algorithm. The experiment results have shown that the revised model can adapt to the data features of student behavior more successfully. Compared to the original YOLOv5 model, it has much higher precision and performance.

The main objective of this article is to improve the efficiency of the behavioral detection process and enhance the speed and performance of the detection model. This will be achieved by focusing on three key contributions:

1. The feature extraction layers responsible for generating the feature map related to small object detection have been appropriately adjusted in the default network topology of YOLOv5. This adjustment reduces the computational power requirements and the number of parameters and significantly improves the model's speed.
2. A new adaptation of the model's feature fusion network topology and prediction head is implemented, reducing the computational power requirements and significantly improving the model's inference time.

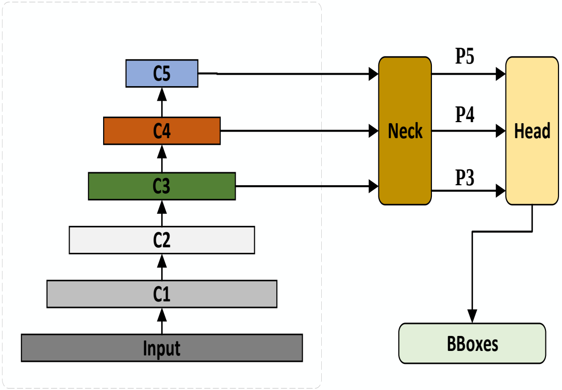


Figure. 1 Illustrates the default inference approach used in YOLOv5

Received: June 29, 2023. Revised: July 23, 2023.

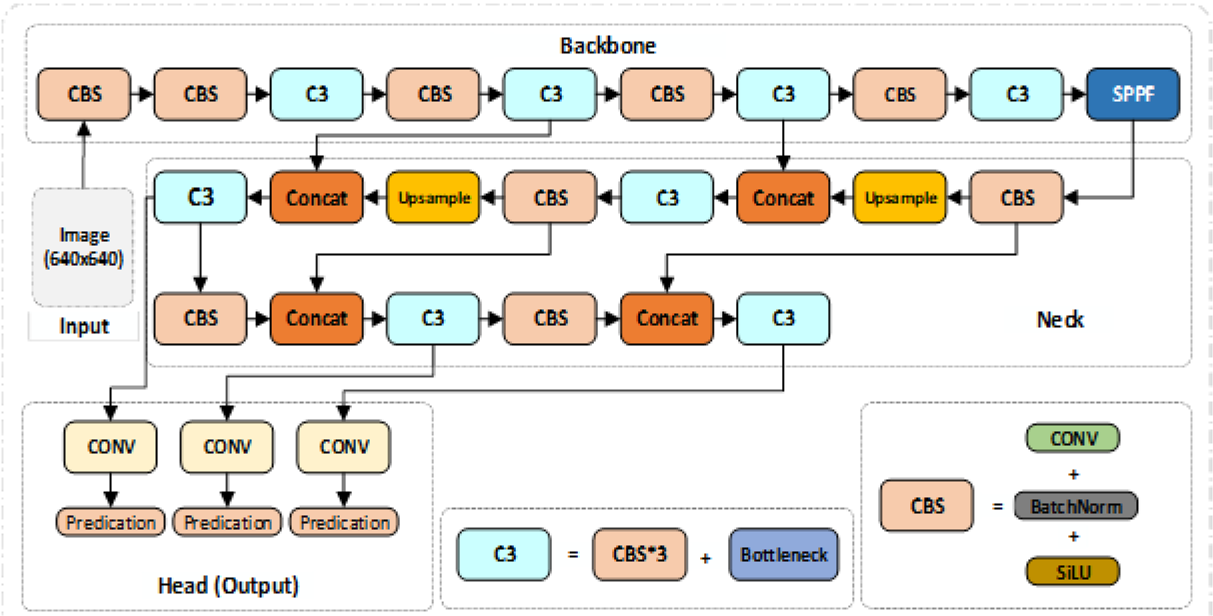


Figure. 2 Depicts the default framework of the YOLOv5 model

3. Incorporating the CBAM module, an attention mechanism, at the end of the feature fusion network amplifies the prominence of the object region within a complex background, thereby significantly enhancing the detector's accuracy.

The structure of the article is as follows. The background and theories are provided in section II. A scientific-technical description of the core modules and proposed methods can be found in section III. Experimental results and analysis can be found in section IV. concludes and offers suggestions for future work in section V.

2. Background and theories

Our approach is based on the YOLOv5 model, and convolutional block attention module (CBAM). Subjects will be briefly covered in this section.

2.1 YOLOv5 overview

The YOLOv5 is the most recent and state-of-the-art iteration of the You Only Look Once (YOLO) algorithm family [17]. With its high performance and quick detection speed, the YOLOv5 model is up to the challenge of real-time applications. YOLOv5 can be divided into three main components: the backbone, the neck, and the head [18].

The YOLOv5 algorithm is based on the same principles as its previous versions. It takes an input layer as its initial step and proceeds with feature extraction using the (backbone) [19]. The backbone generates three feature maps, namely. P_3, P_4, and P_5, which have dimensions of (80 × 80), (40 × 40), and (20 × 20), respectively, and are utilized for detecting multi scales objects in the image. These feature maps are created by combining various-sized features and fusing them via the feature fusion network (neck) [20].

The prediction head (head) receives these three features maps and performs bounding-box regression and confidence calculations using predetermined prior anchors on each feature map pixel. This generates an array with multiple dimensions, including information about the object's class, confidence level coordinates in the bounding box, and width and height [21]. The final detection data is obtained using a non-maximum suppression (NMS) [19] technique and setting appropriate thresholds to eliminate unnecessary data from the array. The conversion of the input image to a bounding box (BBoxes) is known as the inference process. Fig. 1 illustrates the default inference methodology utilized in YOLOv5 [22].

2.1.1. Model structure

The YOLOv5 network structure is typically referred to as the neck and backbone. Fig. 2 illustrates the default framework of the YOLOv5 model.

A. The backbone

The backbone core structure is composed of several CBS ($Conv + BatchNorm + SiLU$) modules, C3 ($CBS^*3 + Bottleneck$) modules, and one SPPF ($CBS^*2 + MaxPool^*3$) module. While the SPPF module increases the backbone's

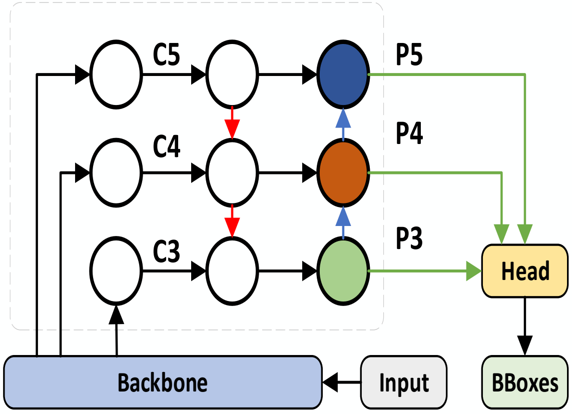


Figure. 3 The YOLOv5 feature fusion path

capabilities for feature expression, CBS module aids C3 module in feature extraction [23]. Fig. 2 shows the default framework of the YOLOv5 model.

As a result, the C3 module is the most significant layer in the YOLOv5 backbone. Cross-stage partial networks (CSPNet) is where the basic concept of the C3 module originated [24].

B. The neck

As shown in Fig. 3, YOLOv5 uses the path aggregation network (PANet) [25] and feature pyramid network (FPNet) [26] to build multiple additional feature maps for distinguishing targets of various scales from the output feature mapo.

While the FPNet module uses a top-down feature fusion approach to combine combining high level semantic data with low level semantic data, the PANet module uses a bottom-up feature fusion approach to combine low-level semantic information with high-level semantic information [27].

2.2 Convolutional block attention module

An attention module in computer vision is a kind of architecture that combines an attention mechanism to selectively focus on significant features of an image for a particular purpose.

Convolutional block attention module (CBAM), an attention module introduced by S. Woo et al. [28], highlights essential features along the channel and spatial axes, just like convolutions operations do. Every convolutional block in deep neural networks can use this type of attention, explicitly designed for feed-forward convolutional neural networks. The channel attention module (CAM) and the spatial attention module (SAM) are two successive sub-modules that makeup CBAM.

As can be seen in Fig. 4, CBAM takes as input an intermediate feature map (F) and progressively infers a 1D channel attention map (Mc) and a 2D spatial attention map (Ms). Equations (2) and (3) can summarize the entire attention process.

$$F' = M_c(F) \otimes F \qquad (2)$$

$$F'' = M_s(F') \otimes F' \qquad (3)$$

Where ⊗ represents element-by-element (element-wise) multiplication. The attention values are propagated similarly during multiplication, channel attention values are propagated along the spatial dimension, and vice versa. The final refined output is F" [28].

3. Proposed approach

Third-order headings, as in this paragraph, are discouraged. However, if you must use them, use 11-point Times New Roman, boldface, initially capitalized, flush left.

The YOLOv5 Algorithm demonstrates impressive performance in object detection tasks. However, the detection results may not be optimal in scenarios with numerous multiscale targets. Therefore, there is a need for further improvement and customization to address these challenges specific to our problem. To tackle these issues and make the algorithm more suitable for our needs, several questions were raised:

1. The YOLOv5 set three alternative feature map output sizes and three predictor heads for detection scenes of varying scales objects. Multiple convolutions down sampling procedures are required to obtain the feature map, and prediction process, which uses a lot of processing power and parameters.

The question: Are all feature maps output and the multi-prediction heads for high-level different scale targets required to handle our problem?

2. To improve the detection performance for objects of various scales, a mix of a top down path and a bottom up path was included in the YOLOv5 feature fusion network. The algorithm addresses the difficulties presented by objects of different sizes in the detection process by combining these two paths.

The question: Is it possible to design a new feature fusion network or reorganize the existing network structure to achieve a dense output feature map capable of detecting large-scale objects?

3. Usually, there are several objects in the actual room, including chairs, walls, windows, and other background surroundings, while looking at the area of untargeted regions in a picture.

Received: June 29, 2023. Revised: July 23, 2023.

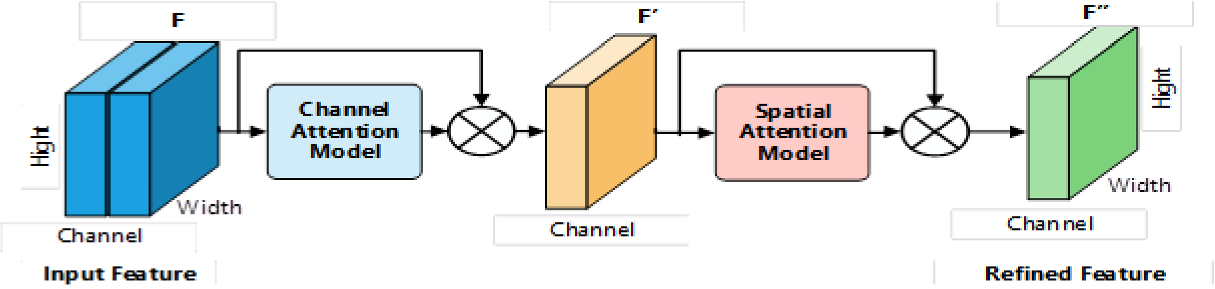


Figure. 4 CBAM modules architecture

The question: Can we use an attention mechanism that helps the model better concentrate on the data from the region of interest (ROI), the students?

In order to address the three questions above and to achieve the paper-suggested model (SPL-YOLOv5), we will do:

The last convolution down-sampling operation layer, which creates the high-level feature map C_5 for small-scale objects, was removed from the backbone section of the algorithm in order to make it lighter. Instead, we only kept the 3th C_3 and 4th C_4 feature maps. Additionally, we decreased the number of detection layers in the prediction layer component (head) from the original three to one, which minimized the convolution down-sampling processes.

Second, based on the modification to the backbone and head sections, we performed the appropriate adjustments to the feature fusion network's route (neck) and layers for significant object detection based on the top-down path (FPN) and bottom-up path (PAN) feature fusion techniques.

Finally, to enhance the model's attention toward the regions of interest (ROI) in the students' images, we incorporated the Convolutional Block Attention Module (CBAM) between the neck and prediction networks. This module aids the model in sharpening its focus on the data that matters in the students' areas, enhancing the functionality of the entire system.

3.1 Feature map pruning

By default, YOLOv5 generates feature maps of different sizes to detect objects of varying scales. This is achieved through consecutive convolution-downsampling operations and feature fusion processes. These feature maps capture information about large, medium, and small objects in the input data. The findings of the default original YOLOv5s network's numerical visualization are shown in Table 1.

According to the information presented in Table 1, in the feature extraction network part (Backbone), the C_5 sample layer is indicated by layers 7 and 8; as observed, the sampling layer takes up a significant number of parameters.

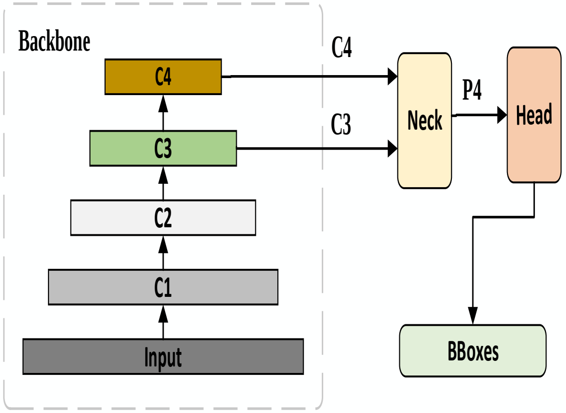


Figure. 5 Model inference diagram after applying the pruning procedure

In the proposed model, we make the architecture lighter by removing the original C_5 feature map of YOLOv5s. Additionally, the feature fusion network's output feature maps (P3, P5) and their associated prediction head are truncated. Fig. 5 illustrates the modified inference diagram of the proposed model after applying these pruning steps.

Accordingly, the neck represents the redesigned feature fusion network, while the head represents the single prediction layer.

3.2 Modification for feature fusion path

A modified feature fusion path was created based on prior alteration and clipping. The previous elimination of the C5 feature map that was done during feature map pruning leads to removing any related node and connection. The feature pyramid and path aggregation networks continue to utilize a top-down and bottom-up approach to integrate multiscale features for prediction. This method takes into account both the bottom layer location data and top layer semantic data. Fig. 6 depicts the standard YOLOv5 feature fusion approach.

Table 1. The original YOLOv5s network numerical visualization

	Module location	**Connect From**	**number of modules**	**Number of Parameter**	**Module Name**	**Module Arguments [In, Out, kernel]**
Backbone	0	-1	1	3520	Conv	[3, 32, 6, 2, 2]
	1	-1	1	18560	Conv	[32, 64, 3, 2]
	2	-1	1	18816	C3	[64, 64, 1]
	3	-1	1	73984	Conv	[64, 128, 3, 2]
	4	-1	2	115712	C3	[128, 128, 2]
	5	-1	1	295424	Conv	[128, 256, 3, 2]
	6	-1	3	625152	C3	[256, 256, 3]
	7	-1	1	1180672	Conv	[256, 512, 3, 2]
	8	-1	1	1182720	C3	[512, 512, 1]
	9	-1	1	656896	SPPF	[512, 512, 5]
Neck	10	-1	1	131584	Conv	[512, 256, 1, 1]
	11	-1	1	0	Upsample	[None, 2, 'nearest']
	12	[-1, 6]	1	0	Concat	[1]
	13	-1	1	361984	C3	[512, 256, 1, False]
	14	-1	1	33024	Conv	[256, 128, 1, 1]
	15	-1	1	0	Upsample	[None, 2, 'nearest']
	16	[-1, 4]	1	0	Concat	[1]
	17	-1	1	90880	C3	[256, 128, 1, False]
	18	-1	1	147712	Conv	[128, 128, 3, 2]
	19	[-1, 14]	1	0	Concat	[1]
	20	-1	1	296448	C3	[256, 256, 1, False]
	21	-1	1	590336	Conv	[256, 256, 3, 2]
	22	[-1, 10]	1	0	Concat	[1]
	23	-1	1	1182720	C3	[512, 512, 1, False]
Head	24	[17, 20, 23]	1	26970	Yolo.detect	[5, [[10, 13, 16, 30, 33, 23], [30, 61, 62, 45, 59, 119], [116, 90, 156, 198, 373, 326]], [128, 256, 512]]

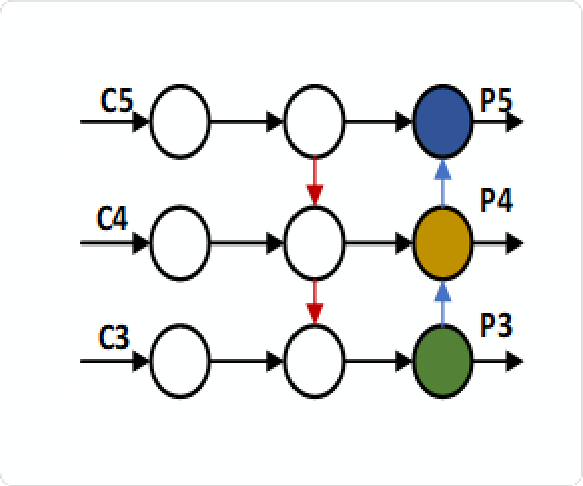


Figure. 6 Default YOLOv5 feature fusion path

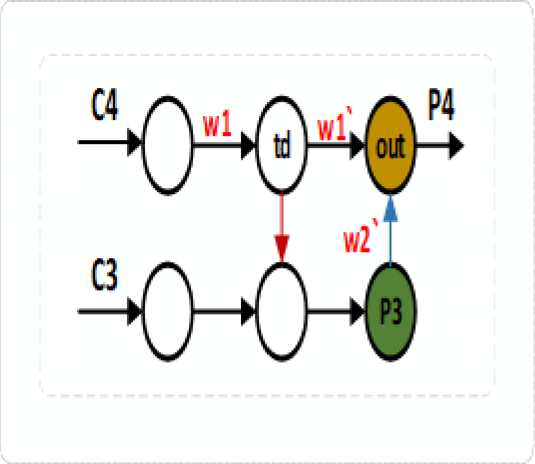


Figure. 7 Proposed model feature fusion path

Fig. 7 illustrates the proposed model feature fusion path after modification and clipping.

If we consider, for example, a node P4 in Fig. 1, the two-feature fusion process formed using Eqs. (6) and (7).

Received: June 29, 2023. Revised: July 23, 2023.

Table 2. The proposed SPL-YOLOv5 network numerical visualization

	Module location	Connect From	number of modules	Number of Parameter	Module Name	Module Arguments [In, Out, kernel]
Backbone	0	-1	1	3520	Conv	[3, 32, 6, 2, 2]
	1	-1	1	18560	Conv	[32, 64, 3, 2]
	2	-1	1	18816	C3	[64, 64, 1]
	3	-1	1	73984	Conv	[64, 128, 3, 2]
	4	-1	2	115712	C3	[128, 128, 2]
	5	-1	1	295424	Conv	[128, 256, 3, 2]
	6	-1	3	625152	C3	[256, 256, 3]
	7	-1	1	656896	SPPF	[256, 512, 5]
Neck	8	-1	1	131584	Conv	[512, 256, 1, 1]
	9	-1	1	0	Upsample	[None, 2, 'nearest']
	10	[-1, 6]	1	0	Concat	[1]
	11	-1	1	361984	C3	[512, 256, 1, False]
	12	-1	1	33024	Conv	[256, 128, 1, 1]
	13	-1	1	0	Upsample	[None, 2, 'nearest']
	14	[-1, 4]	1	0	Concat	[1]
	15	-1	1	90880	C3	[256, 128, 1, False]
	16	-1	1	147712	Conv	[128, 128, 3, 2]
	17	[-1, 14]	1	0	Concat	[1]
	18	-1	1	296448	C3	[256, 256, 1, False]
	19	-1	1	590336	CBAM	[256, 256, 3, 2]
Head	21	[20]	1	7710	Yolo.detect	[5, [[116, 90, 156, 198, 373, 326]], [128, 256, 512]]

$$P_4^{td} = \mathrm{Conv}\left(\frac{w_1 . p_4^{in}}{w_1+\epsilon}\right) \quad (6)$$

$$P_4^{out} = \mathrm{Conv}\left(\frac{w_1' . p_4^{td}+w_2' . \mathrm{Resize}(p_3)}{w_1'+w_2'+\epsilon}\right) \quad (7)$$

A p_4^{in} and p_4^{out} in the formula stand for C4 and P4, respectively, whereas p_4^{td} a top down intermediate feature map between C4 and P4. Conv stands for the convolution process, resize for up- or down-sampling, W for Weight, and ε is a tiny number to prevent numerical instability, often set at 0.0001.

3.3 Add the attention module

The primary objective of the attention mechanism is to focus on the input items' crucial areas to gather essential information. By selectively emphasizing specific elements, the attention mechanism can extract important details from large amounts of data while disregarding irrelevant information. In this work, the CBAM attention module was incorporated at the end of the feature fusion network to facilitate this enhancement. Fig. 8 can express the proposed altered network in this section.

The results of the proposed SPL-YOLOv5 network's numerical visualization are shown in Table 2.

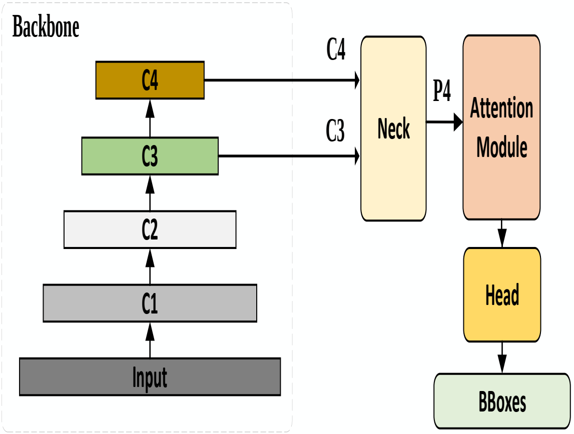


Figure. 8 Altered networks of proposed SPL-YOLOv5

3.4 The grad-CAM model analysis

Gradient-weighted class activation mapping (Grad-CAM) [29] is one of the explainable artificial intelligence (XAI) [30] techniques used in CNNs to

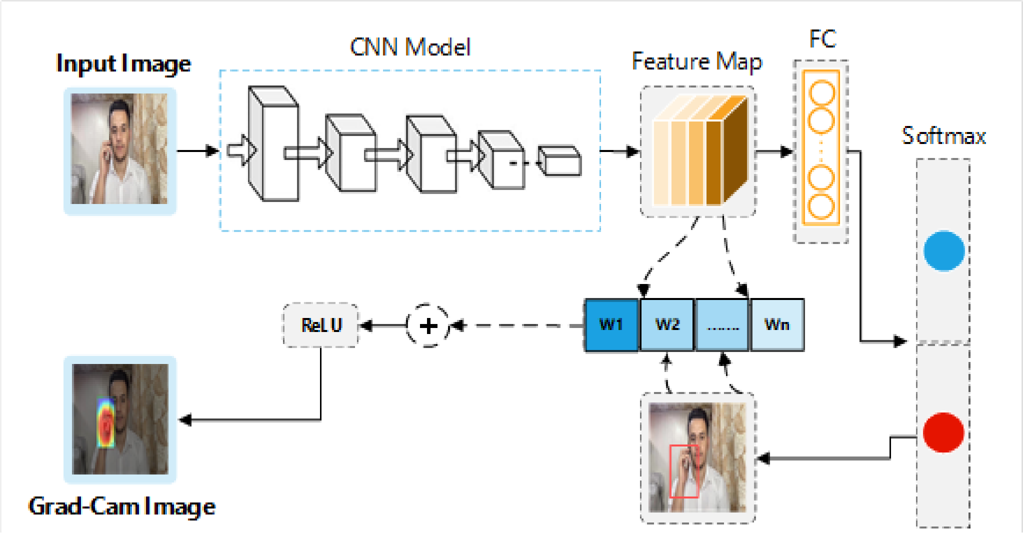
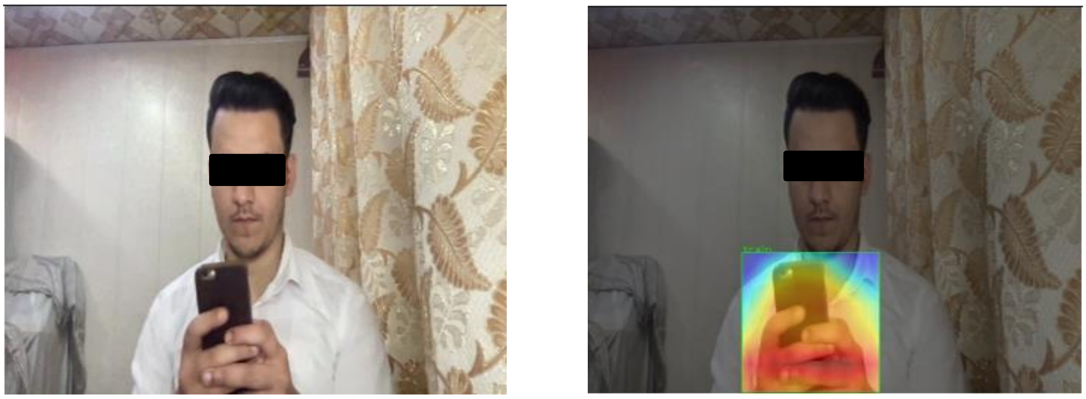


Figure. 9 Architecturallayut of the grad-CAM



(a) (b)

Figure. 10 An image generated using the grad-CAM technique

produce a class-specific heatmap based on a particular input picture using a trained CNN model. The suggested SPL-YOLOv5 model detection transparency is determined using the Grad-CAM method. By emphasizing the areas of the input picture that the model concentrates on during the classification process, this method suggests that the feature maps created in the final convolution layer include the spatial information necessary to capture the visual pattern effectively. This visual pattern helps the ability to identify across classes. The layers and extracted features from the trained model are used to apply the grad-CAM approach. Fig. 9 depicts the architectural layout of the grad-CAM method used in this work.

Fig. 10 displays an example of an image produced using the Grad-CAM technique. Fig. 10.a displays the input image, whereas Fig. 10.b displays the input image with an overlay heatmap. The suggested model contains an attention mechanism that increases the focus on the goal and decreases the focus on other items, enhancing the training impact.

4. Experimental results and analysis

4.1 Dataset development

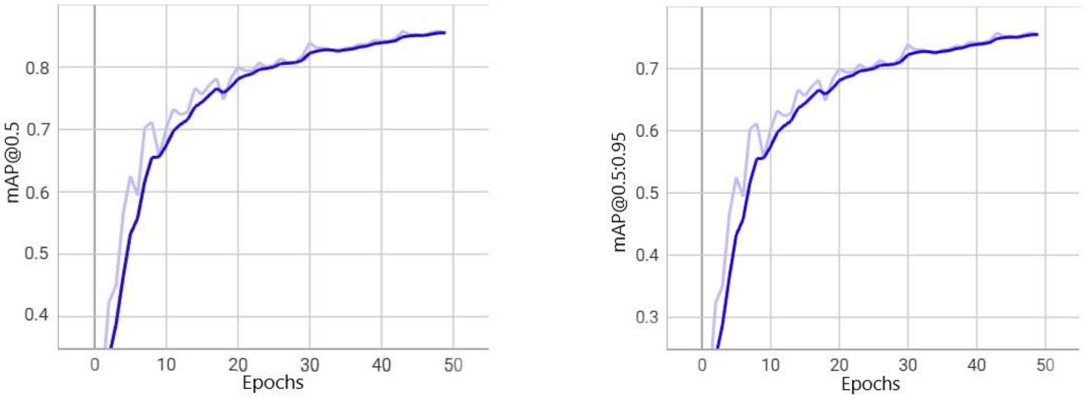
Due to the lack of a publicly accessible dataset of videos taken during actual online tests, an online exam student behavior dataset was manually created for this paper. Twenty-four videos with a frame rate of 15 frames per second were acquired using a webcam. Images for the dataset were extracted from recorded videos at predetermined frame intervals; there are 18,520 labelled images in total in the dataset. The dataset contains five classes: mobile using, hand moving, eye moving, mouth opening, and looking side. Our dataset can be downloaded from: https://doi.org/10.7910/DVN/WUWRAB.

4.2 Experimental configurations and training

This article utilized a Windows 10 operating system for the experimental environment, an Intel(R)

Table 3. Performance of YOLOv5 and YOLOv5-C4 in experiments

Model	Parameter	GFLOPs	Inference Time (ms)	mAP@0.5	mAP@0.5:0.95
YOLOv5	7,033114	16.0	7.3	0.83	0.80
YOLOv5-C4	2,110454	12.0	4.8	0.85	0.77



(a) (b)

Figure. 11. The training mAP@0.5 and mAP@0.5:0.95 curves

Table 4. Experimental performance of YOLOv5-C4 and SPL-YOLOv5

Model	Parameter	GFLOPs	Inference Time (ms)	mAP@0.5	mAP@0.5:0.95
YOLOv5-C4	2,110454	12.0	4.8	0.85	0.77
SPL-YOLO5	2,241624	12.2	5.0	0.97	0.93

Core i7 CPU@2.8HZ for the CPU configuration, and a GeForce GTX 3030Ti graphics card with a video memory size of 16 GB for computing. The runtime environment was established using Pytorch a deep learning framework, and the Python environment.

Before beginning the training process, the dataset must be categorized and divided into training and validation components. We divide the dataset as test 10%, valid 20%, and train 70%. We begin training with 2750 picture patches of entities from four different classes, starting learning rate of 0.01, anchor-multiple thresholds of 5.0, SGD momentum of 0.936, batch size of 16, and epochs of 50. The weights of the model are enhanced using the Adam optimizer.

4.3 Evaluation metrics

To assess the effectiveness of the object detection algorithm, matrices were employed. Commonly used metrics for evaluating performance include precision (P), recall (R), average precision (AP), mean average precision (mAP), the number of parameters, and floating-point operations per second (FLOPs).

4.4 Experimental results

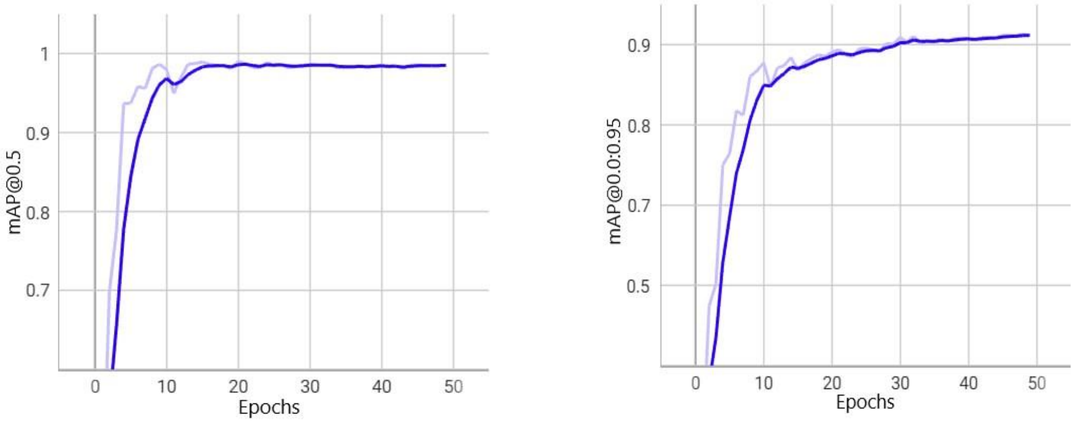
The original YOLOv5 and proposed SPL-YOLOv5 are contrasted in this section based on how well they performed in experiments. All comparisons are based on parameter, GFLOP, inference time, mAP@0.5, and mAP@0.5:0.95 metrics.

YOLOv5-C4 is a new algorithm generated by pruning the feature extraction layers and corresponding modification on the nick and head layers of the original YOLOv5 model. Table 3 compares the experimental performance of the original YOLOv5 and the proposed YOLOv5-C4 models.

As a result of this modification, mAP@0.5 has gone up by 0.03, while mAP@0.5:0.95 has gone down by 0.03, the Parameters and GFLOPs have decreased, and inference time has gone down. According to the results, removing the high-level feature map can significantly enhance the model's performance, even though the detection accuracy may be slightly impacted. Fig. 11 displays the mAP@0.5 and mAP@0.5:0.95 curves on the dataset using the proposed YOLOv5-C4 model.

The results indicate an improvement in the mAP@0.5 but a little decrease in mAP@0.5:0.95.

An enhanced version of the YOLOv5-C4 algorithm, known as the SPL-YOLOv5 algorithm, was created by adding the CBAM attention module to the first modification method suggested. This



(a) (b)

Figure. 12 The training mAP@0.5 and mAP@0.5:0.95 curves

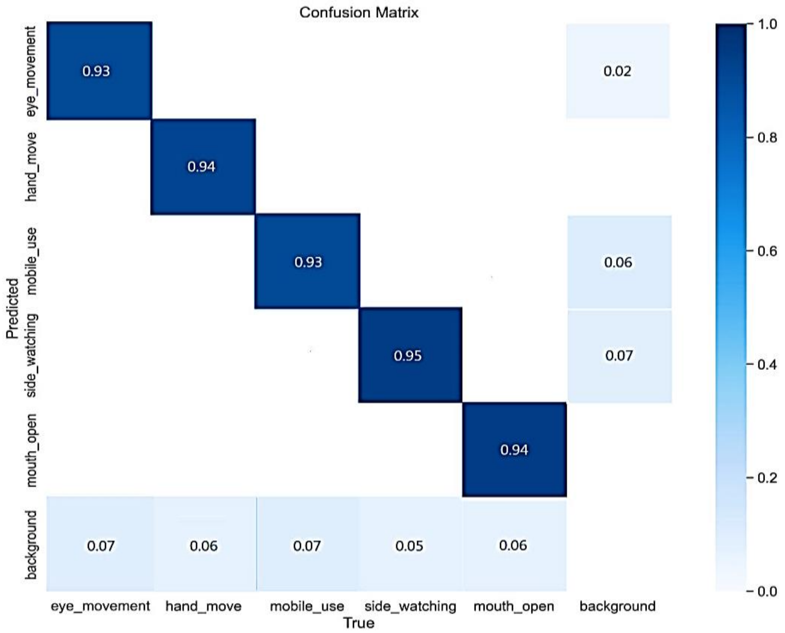


Figure. 13 Validation confusion matrix

Table 5. The outcomes of the proposed SPL-YOLOv5 model

Class Name	Precision	mAP@0.5	mAP@0.5:0.95
Mobile_Use	0.97	0.96	0.93
Hand_Move	0.96	0.97	0.94
Eye_Movement	0.95	0.98	0.92
Side_Watching	0.98	0.97	0.95
Mouth_Open	0.97	0.98	0.94
Average for all categories	**0.96**	**0.97**	**0.93**

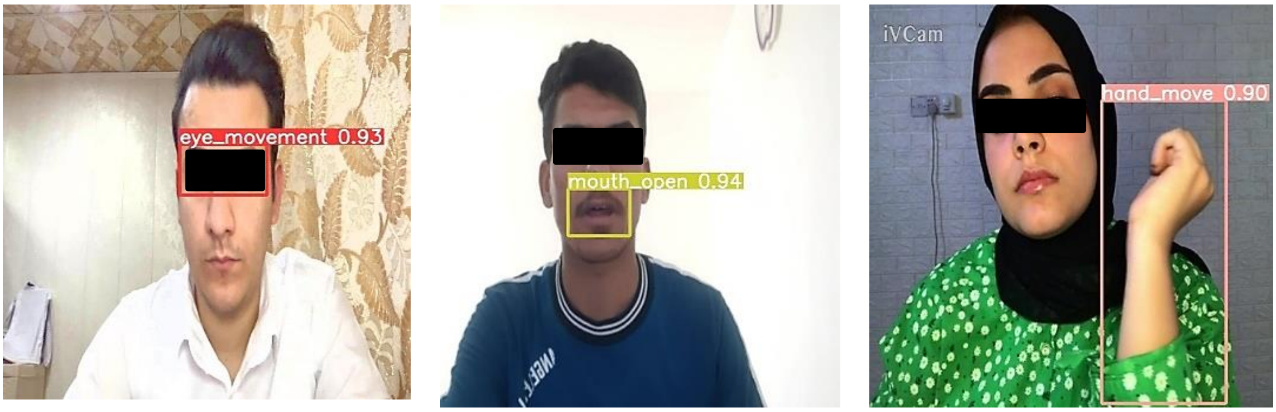
Table 6. Contrasting the addition of various modules before and after

Model	Parameter	GFLOPs	Inference Time(ms)	mAP@0.5	mAP@0.5: 0.95
YOLOv5	7,033,114	16.0	7.3	0.83	0.80
YOLOv5-C4	2,110,454	12.0	4.8	0.85	0.77
SPL-YOLO5	2,241,624	12.2	5.0	0.97	0.93
Improvement	**4,791,490**	**13.3**	**-2.3**	**0.14**	**0.13**

Received: June 29, 2023. Revised: July 23, 2023.

Table 7. Results comparison in different algorithms

Model	Parameter	GFLOP	Inference Time(ms)	mAP@0.5	mAP@0.5:0.95
YOLOv5	7,033114	16.0	7.3	0.83	0.80
Faster R-CNN	41,319661	134.3	16.4	0.75	0.47
SSD-MobileNet	18,950729	35.3	8.5	0.65	0.57
SPL-YOLOv5	2,406232	12.7	5.0	0.97	0.93



[redacted]
[redacted]
[redacted]

(a) (b) (c)

Figure. 14 Some of the visual scenes resulting from the proposed SPL-YOLOv5

approach is slightly more computationally intensive and has few more parameters than YOLOv5-C4, but it dramatically increases the detection accuracy by 0.12 and 0.16 in mAP@0.5 and mAP@0.5:0.95, respectively. Table 4 compares the experimental performance of the YOLOv5-C4 and SPL-YOLOv5 models. This modification shows that adding a more effective attention mechanism to the algorithm would be possible and improve its performance.

Fig. 12 displays the mAP@0.5 and mAP@0.5:0.95 curves on the dataset using the proposed SPL-YOLOv5 model.

The results indicate a noticeable improvement in the mAP@0.5 and mAP@0.5:0.95. Fig. 13 depicts the confusion matrix of the testing dataset using the proposed SPL-YOLOv5 model.

Table 5 displays the final test result of the proposed SPL-YOLOv5 model.

The original YOLOv5 and the proposed modified models are contrasted in Table 6 based on how well they performed in experiments.

Table 6 shows that SPL-YOLOv5 performs better at detection than YOLOv5s, with an increase of 1.8 in mAP@0.5 and 0.9 in mAP@0.5:0.95. The fact that SPL-YOLOv5 reduces the number of parameters (M) and FLOPs (G) by 60.3% and 14.5%, respectively, shows that the algorithm uses fewer processing resources while boosting detection accuracy. The model can process more pictures and videos because the inference time (ms) decreases. So, the total detection accuracy of the method proposed in this paper has increased by 5.6% when compared to the previous YOLOv5 algorithm. The visual test results are displayed in Fig. 14.

The test findings show that the upgraded algorithm used in this research can recognize objects more accurately than the original YOLOv5 model. The algorithm can process features more efficiently and gather high-level semantic data.

This study compares the algorithm described to other widely used object detection algorithms to assess it. All the algorithms were trained using the same parameters and samples to compare fairly.

The experiment results are listed in Table 7 to objectively evaluate the algorithm's benefits.

Table 7 results demonstrate that the improved model (SPL-YOLOv5) described in this dissertation outperforms the other three algorithms in terms of parameters, GFLOPs, Inference time, and average accuracy; this is because attention mechanism and spatial pyramid pooling modules were included, along with modifications to the feature extraction, feature pyramid, and prediction layers. The algorithm's inference time is 5.0 (ms), and the average accuracy is 0.97% which is faster and more accurate than the other selected model, including the original YOLOv5.

Table 8 contrasts the suggested approach and different cutting-edge methods, explicitly emphasizing the modified module in backbone, attention module used, the inference time per second, and the mean average precision (mAP) attained by

Table 8. Compared the proposed method with other state-of-the-art methods

Ref.	Module/ Backbone	Attention module used	FPS	mAP
[11]	DarkNet	SE	55	0.88
[12]	CSP/DarkNet	CA	43	0.72
[13]	CBL/DarkNet	NO	59	0.93
[14]	DarkNet.	CBAM	58	0.75
[15]	Meta/DarkNet	NO	55	0.95
[16]	MobileNetV3	NO	47	0.92
Our	Proposed	CBAM	62	0.97

each technique.

The results emphasize the exceptional performance of the suggested approach when compared to alternative methods, primarily due to its remarkable capability to achieve both rapid inference time and accurate detection.

5. Conclusions

This study utilizes an improved YOLOv5 algorithm to detect abnormal students' behavior in the distance learning environment. The number of computational resources required by the original YOLOv5 model was significantly reduced, and the model was made lighter by removing the part of the feature extraction network responsible for high-level feature maps for small objects. Also, the model's feature fusion and prediction layers are modified accordingly. The model's focus on relevant information in students' regions was enhanced by incorporating the CBAM attention module, which improves the detector's accuracy. The experiments' findings show that, in comparison to the original YOLOv5 method, the improved algorithm can precisely detect a variety of student behaviors with more accuracy and less computation. Also, the algorithm suggested in the study exhibits better performance than other state of arts methods and other algorithms like faster R-CNN and SSD concerning mean average precision (mAP) and inference time. In comparison to the original YOLOv5, the proposed algorithm demonstrates an enhancement of 0.14% in mAP@0.5 and 0.13% in mAP@0.5:0.95. Additionally, it significantly reduces the number of parameters by 47%, computational resources (FLOPs) by 40%, and inference time by 2.7%.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author contributions

In this research article, the author's contributions are as follows: "conceptualization, Saad H. Abid ; methodology, Muhanad A. Alkhalisy; software, Muhanad A. Alkhalisy ; validation, Muhanad A. Alkhalisy ; formal analysis, Saad H. Abid ; investigation, Muhanad A. Alkhalisy ; resources, Muhanad A. Alkhalisy ; data curation, Saad H. Abid ; writing— original draft preparation, Muhanad A. Alkhalisy ; writing—review and editing, Muhanad A. Alkhalisy ; visualization, Muhanad A. Alkhalisy ; supervision, Saad H. Abid ; project administration, Saad H. Abid ; funding acquisition, Muhanad A. Alkhalisy .".

References

[1] J. Seaman, I. Allen, and J. Seaman, “G RADE I NCREASE Grade Increase :”, p. 49, 2018, [Online]. Available: https://files.eric.ed.gov/fulltext/ED580852.pdf.

[2] M. Abisado, B. Gerardo, L. Vea, and R. Medina, “Towards academic affect modeling through experimental hybrid gesture recognition algorithm”, *ACM Int. Conf. Proceeding Ser.*, No. July, pp. 48–52, 2018, doi: 10.1145/3239283.3239305.

[3] M. Labayen, R. Vea, J. Florez, N. Aginako, and B. Sierra, “Online Student Authentication and Proctoring System Based on Multimodal Biometrics Technology”, *IEEE Access*, Vol. 9, pp. 72398–72411, 2021, doi: 10.1109/ACCESS.2021.3079375.

[4] F. Mahmood and J. Arshad “Implementation of an Intelligent Exam Supervision System Using Deep Learning Algorithms”, *Sensors*, Vol. 22, No. 17, 2022, doi: 10.3390/s22176389.

[5] M. Masud, K. Hayawi, S. Mathew, T. Michael, and M. E. Barachi, “Smart Online Exam Proctoring Assist for Cheating Detection”, *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Vol. 13087 LNAI, No. January, pp. 118–132, 2022, doi: 10.1007/978-3-030-95405-5_9.

[6] M. Hussein, J. Yusuf, A. Deb, L. Fong, and S. Naidu, “An Evaluation of Online Proctoring Tools”, *Open Prax.*, Vol. 12, No. 4, p. 509, 2020, doi: 10.5944/openpraxis.12.4.1113.

[7] D. M. Seo, H. J. Woo, M. S. Kim, W. H. Hong, I. H. Kim, and S. C. Baek, “Identification of Asbestos Slates in Buildings Based on Faster Region-Based Convolutional Neural Network (Faster R-CNN) and Drone-Based Aerial Imagery”, *Drones*, Vol. 6, No. 8, 2022, doi: 10.3390/drones6080194.

[8] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A Review of Yolo Algorithm Developments”, *Procedia Computer Science*, Vol. 199, pp. 1066–1073, 2021, doi: 10.1016/j.procs.2022.01.135.

[9] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, “SSD: Single shot multibox detector”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.

[10] M. Horvat and G. Gledec, “A comparative study of YOLOv5 models performance for image localization and classification”, In: *Proc. of 33rd Cent. Eur. Conf. Inf. Intell. Syst.*, pp. 349–356, 2022, [Online]. Available: https://github.com/mhorvat/YOLOv5-models-

[11] Z. Wang, J. Yao, C. Zeng, W. Wu, H. Xu, and Y. Yang, “YOLOv5 Enhanced Learning Behavior Recognition and Analysis in Smart Classroom with Multiple Students”, *arXiv preprint arXiv:2303.10916*, 2023.

[12] W. Niu, X. Sun, and K. Yi, “Improved YOLOv5 for skeleton-based classroom behavior recognition”, In: *Proc, of the third International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI 2022)*, Vol. 12509, pp. 107-112, 2023.

[13] Z. Zheng, G. Liang, H. Luo, and H. Yin, “Attention assessment based on multi-view classroom behaviour recognition”, *IET Comput. Vis.*, No. August, 2022, doi: 10.1049/cvi2.12146.

[14] S. Yang, Y. Chen, Z. Zhang, and J. Chen, “Student in-class behaviors detection and analysis system based on CBAM-YOLOv5”, In: *Proc. of 2022 7th Int. Conf. Intell. Comput. Signal Process. ICSP 2022*, No. d, pp. 440–443, 2022, doi: 10.1109/ICSP54964.2022.9778630.

[15] Z. Zhang, D. Ao, L. Zhou, X. Yuan, and M. Luo, “Laboratory Behavior Detection Method Based on Improved Yolov5 Model”, In: *Proc. of 2021 Int. Conf. Cyber-Physical Soc. Intell. ICCSI 2021*, No. 61901059, 2021, doi: 10.1109/ICCSI53130.2021.9736251.

[16] J. Wen, Y. Qin, and S. Hu, “Abnormal behavior identification of examinees based on improved YOLOv5”, In: *Proc. of the International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2022)*, Vol. 12604, pp. 946-953, 2023.

[17] J. Du, “Understanding of Object Detection Based on CNN Family and YOLO”, *J. Phys. Conf. Ser.*, Vol. 1004, No. 1, 2018, doi: 10.1088/1742-6596/1004/1/012029.

[18] D. Thuan, "Evolution of Yolo Algorithm and Yolov5: The State of the Art Object Detection Algorithm", *Oulu University of Applied Sciences*, pp. 1–61, 2021, http s://www.theseus.fi/bitstream/handle/10024/452552/Do_Thuan.pdf?sequence=2 (accessed January 27, 2022)

[19] H. Jung and G. Choi, “Improved YOLOv5: Efficient Object Detection Using Drone Images under Various Conditions”, *Appl. Sci.*, Vol. 12, No. 14, 2022, doi: 10.3390/app12147255.

[20] W. Wu, H. Liu, L. Li, X. Wang, Z. Wang and Y. Long, “Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image”, *PLoS One*, Vol. 16, No. 10 October, pp. 1–15, 2021, doi: 10.1371/journal.pone.0259283.

[21] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, “TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios”, In: *Proc. of IEEE Int. Conf. Comput. Vis.*, Vol. 2021-Octob, pp. 2778–2788, 2021, doi: 10.1109/ICCVW54120.2021.00312.

[22] R. Li and Y. Wu, “Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism”, *Electron.*, Vol. 11, No. 11, 2022, doi: 10.3390/electronics11111673.

[23] L. Tang, T. Xie, Y. Yang, and H. Wang, “Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism”, *Appl. Sci.*, Vol. 12, No. 13, 2022, doi: 10.3390/app12136790.

[24] C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, “CSPNet”, *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, Vol. 2020-June, pp. 1571–1580, 2020.

[25] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “PANet: Path Aggregation Network for Instance Segmentation. (arXiv:1803.01534v3 [cs.CV] UPDATED)”, *Cvpr*, pp. 8759–8768, 2019, [Online]. Available: http://arxiv.org/abs/1803.01534.

[26] Y. Zhao, R. Han, and Y. Rao, “A new feature pyramid network for object detection”, In: *Proc. of 2019 Int. Conf. Virtual Real. Intell. Syst. ICVRIS 2019*, pp. 428–431, 2019, doi: 10.1109/ICVRIS.2019.00110.

[27] Z. Baojun, Z. Boya, T. Linbo, W. Wenzheng, and W. Chen, “Multi-scale object detection by top-down and bottom-up feature pyramid network”, *J. Syst. Eng. Electron.*, Vol. 30, No. 1, pp. 1–12, 2019, doi: 10.21629/JSEE.2019.01.01.

[28] S. Woo, J. Park, J. Lee, and I. S. Kweon, “Convolutional_Block_Attention”, *Eccv*, p. 17, 2018, [Online]. Available: http://files/737/Woo et al. - Convolutional Block Attention Module.pdf.

[29]R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”, *Int. J. Comput. Vis.*, Vol. 128, No. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

[30] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, *Explainable AI Methods - A Brief Overview*, Vol. 13200 LNAI, 2022. doi: 10.1007/978-3-031-04083-2_2.