Zhong *et al.*
*EURASIP Journal on Advances in Signal Processing* (2023) 2023:55
https://doi.org/10.1186/s13634-023-01019-w

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

# Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms

Hongmei Zhong[1], Tingting Han[1*], Wei Xia[1], Yan Tian[1] and Libao Wu[2]

*Correspondence: hanting608@163.com

[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2] Faculty of Education, Tianjin Normal University, Tianjin 300387, China

**Abstract**

Studying the real-time face expression state of teachers in class was important to build an objective classroom teaching evaluation system based on AI. However, the face-to-face communication in classroom conditions was a real-time process that operated on a millisecond time scale. Therefore, in order to quickly and accurately predict teachers' facial expressions in real time, this paper proposed an improved YOLOv5 network, which introduced the attention mechanisms into the Backbone model of YOLOv5. In experiments, we investigated the effects of different attention mechanisms on YOLOv5 by adding different attention mechanisms after each CBS module in the CSP1_X structure of the Backbone part, respectively. At the same time, the attention mechanisms were incorporated at different locations of the Focus, CBS, and SPP modules of YOLOv5, respectively, to study the effects of the attention mechanism on different modules. The results showed that the network in which the coordinate attentions were incorporated after each CBS module in the CSP1_X structure obtained the detection time of 25 ms and the accuracy of 77.1% which increased by 3.5% compared with YOLOv5. It outperformed other networks, including Faster-RCNN, R-FCN, ResNext-101, DETR, Swin-Transformer, YOLOv3, and YOLOX. Finally, the real-time teachers' facial expression recognition system was designed to detect and analyze the teachers' facial expression distribution with time through camera and the teaching video.

**Keywords:** Teachers' facial expression recognition, YOLOv5, Attention mechanism, Education

## 1 Introduction

In recent decades, facial expression recognition (FER) had aroused great interest in the computer field. At the same time, many researchers tried to introduce FER into the field of education and teaching, because teachers' emotions affected the teaching quality and played an important role in classroom teaching activities. How to improve the detection accuracy and detection time of FER was the key problem to be solved when this technology went into practical. Traditional machine learning methods had been used for FER, mainly including hidden Markov model (HMM), fuzzy mathematics, and Bayesian classifier. Filntisis et al. [1] put forward the method of combining

HMM with active appearance model (AAM), which proved that HMM interpolation could effectively classify facial expressions. Halder et al. [2] used fuzzy mathematics to make the classification accuracy reach 87.8%. Sebe et al. [3] put forward a facial expression classification method by Cauchy Bayesian classifier, which used Cauchy distribution instead of Gaussian distribution and achieved better classification results.

With the development of deep learning, the convolutional neural networks (CNN), such as Faster R-CNN [4], R-FCN [5], SSD [6], YOLO [7], VGGNet [8], and GoogLeNet [9], had achieved fruitful results in target detection tasks. Many researchers applied them to FER. In 2018, He et al. [10] recognized facial expressions through VGGNet, and the accuracy reached 73% in Fer2013 dataset. In 2020, Khanzada et al. [11] used ResNet50 network and SGD optimizer for face recognition, achieving the detection time of 40 ms and the accuracy of 75.8%. In 2021, Li et al. [12] proposed a face detection method in natural scenes based on improved Faster-RCNN and used ResNet-50 CNN to extract face features. The average accuracy on the Wider Face dataset was 89.0% and took at least 100 ms for each image. Although the above models got high accuracy in FER, the detection time could not well meet the real-time recognition requirements of teachers in class. In 2016, Redmon et al. presented a unified and real-time detector-YOLO (You Only Look Once) [7] that had the advantages of small computation and fast recognition speed and had been applied to real-time targets detection, such as tomato detection, fine-grain object detection, and real-time growth stage detection [13]. In 2021, Lawal et al. [14] proposed an improved YOLOv3 network for tomato detection, which obtained the accuracy of 99.5% and the detection time of 52 ms. In [15], a high-performance real-time fine-grain object detection network based on YOLOv4 was presented, and the detection time had reduced to 14.29 ms. Hence, YOLO was very suitable for FER in dynamic and changeable real-time scenes. In 2021, Aung et al. [16] combined YOLO algorithm with VGG16 CNN to propose a face detection system with the detection accuracy of 95% and the detection time of 29 ms, which greatly improved the face detection speed in real-time video. As time went on, YOLO versions evolved from v1 to v7 [7, 17–19], which further improved the accuracy and the detection time. YOLOv5 had four models, which could be flexibly selected according to the requirements. In addition, many FER studies focused on the video emotion recognition, which used 3D data combined with the time sequence information. In 2016, Fan et al. [20] combined recursive neural network and C3D convolutional neural network for dynamic video emotion recognition, and the accuracy reached 59.02% on AFEW dataset. In 2021, Liu [21] proposed a network of Capsule-LSTM to better extract the time sequence information about video sequences for video facial expressions recognition, and the accuracy reached 40.16% on AFEW dataset. In 2022, Zhang [22] proposed a double-stream network structure for emotion recognition in complex scenes. According to the current research, the accuracy of video FER was lower than that of image FER.

In addition, the introduction of attention mechanisms (AMs) could focus attention on the important information with high weight, ignore the irrelevant information with low weight, and constantly adjust the weight to select the important information under different situations. In 2020, Qin et al. [23] applied AMs to face key point detection, and solved the problem of network depth and detection time balance. In 2020, Kang et al.

[24] proposed a CNN expression recognition method based on AMs, which effectively alleviated the over-fitting phenomenon, enriched the facial expression features learning, and shifted its attention to the unobstructed facial areas with rich information.

In view of the hybrid algorithm idea proposed by the above scholars, this paper proposed a real-time FER network that mainly put the AMs in the Backbone structure of YOLOv5. RAF-DB dataset was re-screened and labeled to eliminate the poor quality images and equalize the remaining images. First, different AMs were added after each CBS module in the CSP1_X module of the Backbone structure of YOLOv5 (called CSPA), respectively. Then, CAs were incorporated after the Focus (called FA), the CBSs (called CBSA), and the SPP (called SA) in the Backbone of YOLOv5, respectively. In addition, we studied diverse networks by integrating FA, SA, CBSA, and CSPA into each other. The results showed that the CSPA network based on coordinate attentions (CAs) got the best accuracy of 77.1%, increased by 3.5%, compared with YOLOv5, and the detection time was 25 ms. Meanwhile, compared with Faster-RCNN, R-FCN, ResNext-101, DETR, Swin-Transformer, YOLOv3, and YOLOX, its accuracy increased by 5.7%, 4.3%, 3.6%, 3.2%, 3.1%, and 3.6%, and the detection time decreased by 8 ms, 4 ms, 21 ms, 17 ms, 15 ms, and 2 ms, respectively. Then, some teachers' expression pictures with different genders, face sizes, and facial postures were tested. Finally, we built a real-time FER system based on CSPA network to detect and analyze the teachers' facial expression through camera and the teaching video.

## 2 Proposed models

### 2.1 Framework overview

The whole framework of the YOLOv5 model is illustrated in Fig. 1. The whole structure could be divided into three parts: Backbone, Neck, and Head, as shown in Fig. 1a. First, the Backbone was used for extracting rich information features from the input

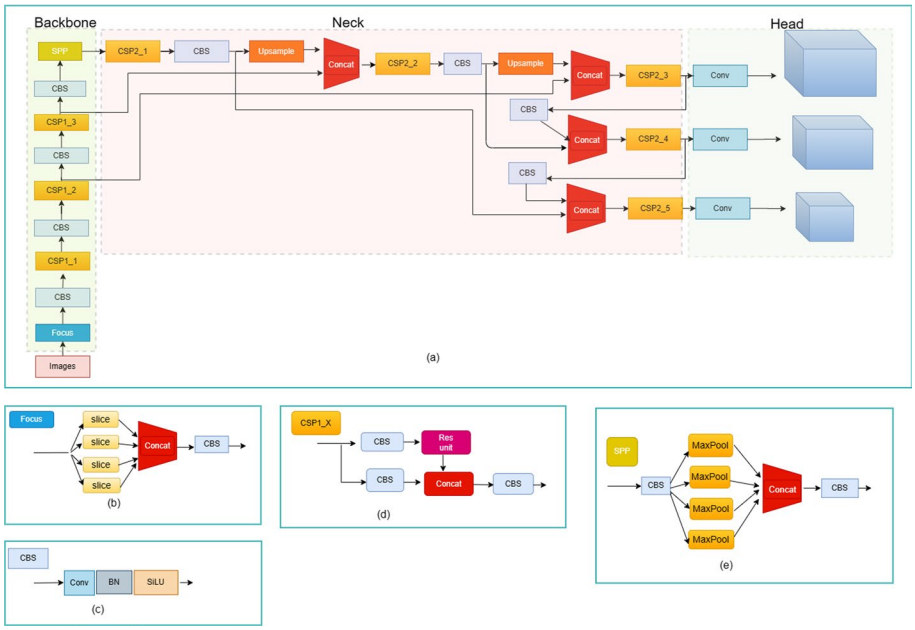


**Fig. 1** The framework of YOLOv5 model

image. For images with different input sizes, Mosaic data enhancement and adaptive anchor box computing were used at the input end. The original images were uniformly scaled to a standard size by random scaling, cropping, and arrangement, and then sent to the Backbone network. The Backbone network contained the Focus, CBS, CSP1_X, and SPP structures. The Focus module used slicing operation to split a high-resolution feature map into many low-resolution feature maps, as shown in Fig. 1b. The CBS module contained the convolution, batch normalization (BN), and SiLU operation to obtain features, as shown in Fig. 1c. Figure 1d shows the CSP1_X module, where X represented the number of residual modules. The input of the CSP1_X module was divided into two branches, in order to further extract the information in the image. One of the branches passed through the CBS first, and then through the residual network to obtain sub-feature map 1. The other branch passed through another CBS to get sub-feature map 2. Finally, the sub-feature map 1 and map 2 were spliced and then input into another CBS. Figure 1e shows the SPP module, which extracted and fused high-level features. In the process of fusion, maximum pooling was used many times to extract as many high-level features as possible. The receptive field was raised almost without reducing the speed, which was helpful to solve the alignment problem between the anchor frame and the feature layer. Second, the Neck network further extracted and fused the image feature information output by Backbone. It contained the CSP2_X, CBS, Upsample, and Concat structures. CSP2_X was different from CSP1_X only in that CSP2_X replaced the residual network with 2 * X CBS. Concat let the model learn more features. Finally, the Head had three detection heads, which could classify and locate the feature information output from Neck, and output the classification probability, confidence, box, and other information of the detected target.

### 2.2 Attention mechanisms

Hence, according to the structural characteristics of YOLOv5, we introduced AMs into the Backbone structure to improve feature extraction capability of the model, thus improving the performance of the model. This paper adopted squeeze-and-excitation (SE) [25], efficient channel attention (ECA) [26], convolutional block attention module (CBAM) [27], and coordinate attention (CA) [28]. SE was a channel attention mechanism. First, the input feature graph was globally averaged and pooled, and the channel compression was proceeded directly. Then, by giving each channel a weight, each feature graph got a different weight and paid attention to more useful features. ECA used 1-D convolution to efficiently realize local cross-channel interaction and extract the dependencies between channels. CBAM allocated attention to both channel dimension and space dimension, which was called mixed attention mechanism, and could further improve the representation ability of the network. In order to integrate more spatial information, CBAM used global average pooling (GAP) and global max pooling (GMP) to compress the spatial information of feature map. CA was also a combination of spatial attention and channel attention. Moreover, CA decomposed spatial attention into two one-dimensional features and aggregated features along two spatial directions ($X$ and $Y$ directions), respectively. This allowed long distance dependencies

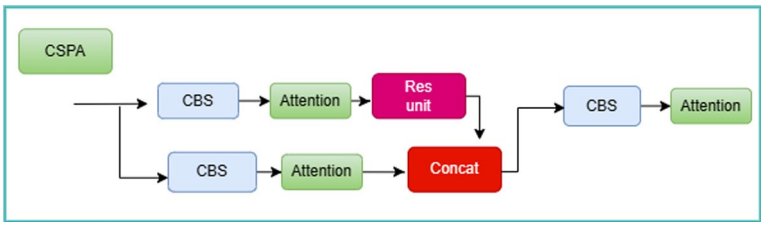


**Fig. 2** The structure chart of CSPA

to be captured in one spatial direction while retaining accurate position information in the other spatial direction. Then, the obtained feature map was encoded into a pair of directional and position sensitive attention maps, respectively, which could be applied to complement the input feature map to enhance the representation ability of the object of interest.

### 2.3 YOLOv5 added by different AMs

To study the effects of different AMs on the model, SE, ECA, CBAM, and CA were added after each CBS module in the CSP1_X of the Backbone structure (called CSPA), respectively, as shown in Fig. 2.

### 2.4 YOLOv5 added by the same AM at different positions.

Then, we studied the effect of the AMs on different modules. The AMs were added after (a) the focus module (called FA), (b) the CBS module (called CBSA), and (c) the SPP module (called SA), as shown in Fig. 3. Meanwhile, we integrated FA, CBSA, SA, and CSPA network into each other to form diverse networks.

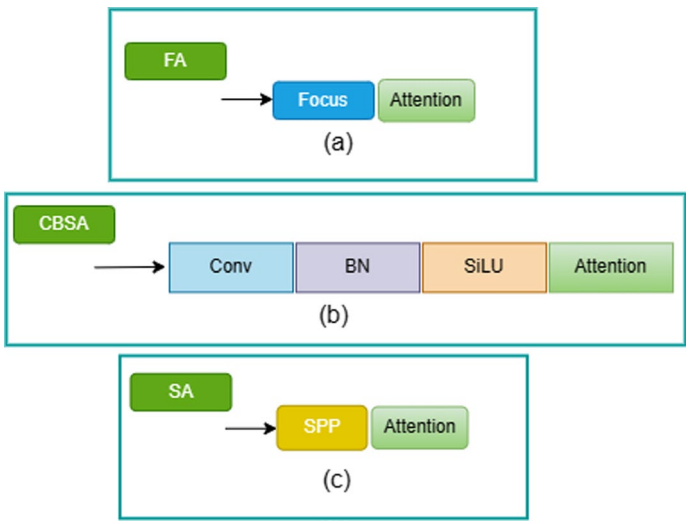


**Fig. 3** The structure chart of **a** FA, **b** CBSA, and **c** SA

## 3 Experiments

### 3.1 Experimental dataset

RAF-DB [29] was a database of real scenes, which contained 29,672 facial images downloaded from Internet search alerts. This dataset was highly diversified with facial expressions of different ages, genders, races, postures, and scenes. Such a rich dataset had a good generalization, which could greatly enhance the robustness of the network. This dataset contained two different sub-datasets: one was a subset of single expression labels, including six basic expressions (surprise, fear, disgust, happiness, sadness, anger) and neutral expressions, and the other was a subset of compound expressions, including 12 kinds of compound expressions.

The dataset was rich in information, but it also had many low-quality images. Therefore, we re-screened RAF-DB dataset by choosing high-quality images. Meanwhile, the number of images of each category was balanced. Figure 4a shows the re-screened RAF-DB dataset, and each category randomly selected about 320 images. Then, the dataset was converted into VOC2007 format. We used LabelImg to mark the labels, as shown in Fig. 4b.

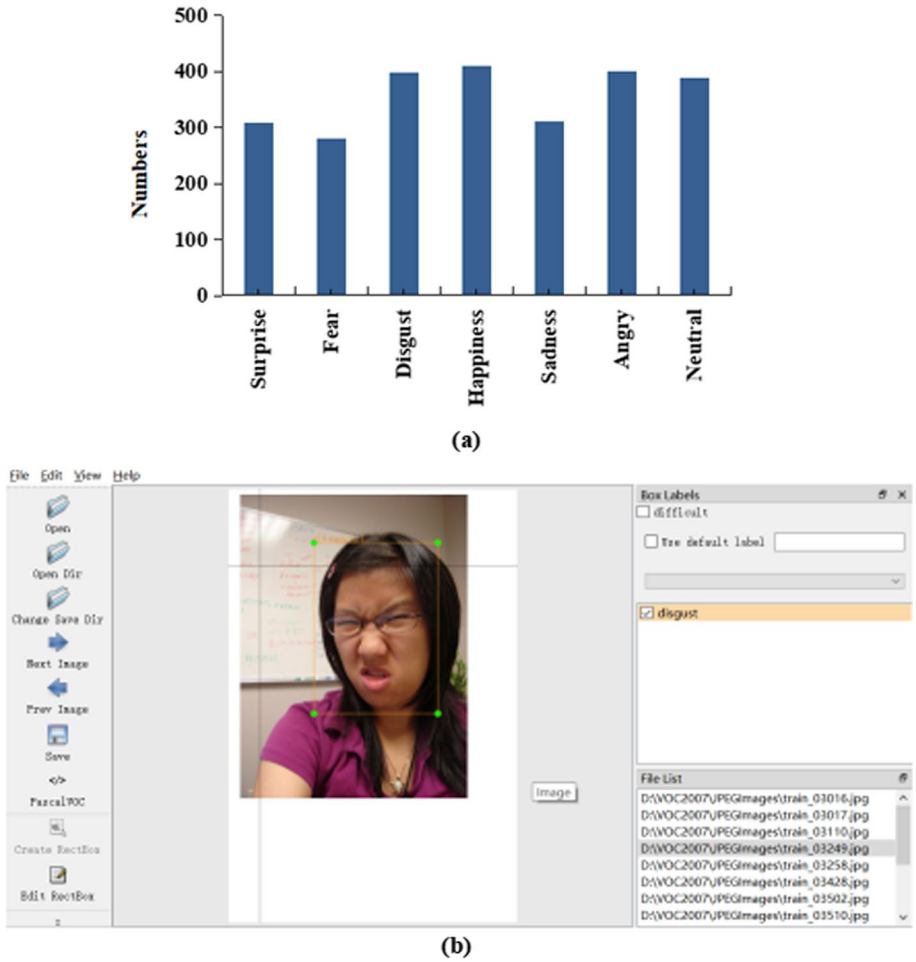


**Fig. 4** **a** The category distribution of the re-screened RAF-DB dataset and **b** labelImg tool

### 3.2 Implementation details

We used Linux as the operating system, JupyterLab as the training environment, and PyTorch as the open-source deep learning framework. Hengyuanyun[1] Server provided NVIDIA Geforce RTX 3090 GPU, which had extremely high computing power and could provide acceleration for PyTorch, ensuring the efficient training of YOLOv5 model under the condition of large data volume. The learning rate *lr* was 0.01, iteration times epochs were 300, and batch size was 16. SGD optimization algorithm was adopted by default in YOLOv5, and cosine annealing strategy was dynamic.

### 3.3 Evaluation index

In the experiments, we judged the FER performance of the model by comparing the detection time, accuracy, and mAP@0.5 of the model. Detection time referred to the time of detecting a frame image.

The accuracy referred to the ratio of the number of correctly classified samples to the total number of samples, as shown in Eq. (1). The precision of one class referred to the ratio of the number of samples predicted correctly for this class to the number of samples predicted as this class, as shown in Eq. (2). The recall of one class referred to the ration of the number of samples predicted correctly for this class to the total number of samples in this class, as shown in Eq. (3).

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

where $n_{\text{correct}}$ was the number of correctly classified samples. $n_{\text{total}}$ was the total number of samples. True positives (TP) was the number of samples identified correctly for this class. False positives (FP) was the number of samples identified incorrectly as this class for other classes. False negatives (FN) was the number of samples identified incorrectly as other classes for this class.

IOU was the intersection and union ratio between the prediction frame and the ground truth. When the threshold of IOU was set to 0.5, we calculated the average precision (AP) of all pictures in each category. AP was the area enclosed by precision and recall as two axes. Then, the mean average precision (mAP), the average AP of all categories, was calculated.

## 4 Results and discussion

### 4.1 Analysis of experimental results

Here, we mainly compared the performance of some mainstream networks, such as Faster-RCNN, R-FCN, ResNext-101, DETR, Swin-Transformer, YOLOv3, YOLOX,

[1] https://www.gpushare.com/.

**Table 1** Different models experiment on RAF-DB dataset

| Method | Model size | Parameters | Accuracy (%) | mAP@0.5 (%) | Time (ms) |
|---|---|---|---|---|---|
| Faster-RCNN | 110.901 M | 53.103 M | 71.4 | 79. 7 | 33 |
| R-FCN | 96.901 M | 51.283 M | 72.8 | 79.3 | 29 |
| YOLOv3 | 95.827 M | 50.853 M | 73.5 | 79.3 | 27 |
| ResNext-101 | 235.310 M | 112.231 M | 73.5 | 80.5 | 46 |
| YOLOX | 92.437 M | 48.569 M | 73.6 | 81.4 | 21 |
| YOLOv5 | 89.624 M | 46.170 M | 73.6 | 81.8 | **15** |
| DETR | 184.714 M | 91.449 M | 73.9 | 82.0 | 42 |
| Swin-Transformer | 131.663 M | 69.406 M | **74.0** | **82.2** | 40 |

Bold meant the best performance of tampering detection in the same experimental setting

and YOLOv5. The results are shown in Table 1, including the model size, parameters, accuracy, mAP@0.5, and detection time.

The results showed that the accuracy of Swin-Transformer was better than other networks, which was 74.0%. The accuracy of YOLOv5 was 0.4% lower than Swin-Transformer. However, the detection time of YOLOv5 was the smallest, which was 25 ms faster than Swin-Transformer. In order to meet the real-time recognition of teachers' classroom, hence, YOLOv5 was used as the baseline model. Based on YOLOv5, different attention mechanisms were introduced into the Backbone structure.

### 4.2 Experiments of different AMs

We added SE, ECA, CBAM, and CA after each CBS in the CSP1_X module of the Backbone to constitute the CSPA structure, as shown in Fig. 2. The performance of YOLOv5, CSPA-SE, CSPA-ECA, CSPA-CBAM, and CSPA-CA on RAF-DB dataset is reported in Table 2. Compared with YOLOv5, the accuracy and mAP@0.5 of the networks added by AMs increased. Meanwhile, the model size, parameters, and detection time also increased. Compared with YOLOv5, the network CSPA-CA had the best accuracy of 77.1% and mAP@0.5 of 83.4%, which increased by 3.5% and 1.6%, respectively. The detection time was 25 ms, which was 10 ms slower than that of YOLOv5. This was because CA focused on both channel dimension and spatial dimension, and used two pool cores to compress the spatial information of the feature maps horizontally and vertically, respectively. Thus, the ability of feature extraction was improved. The proposed model outperformed other detection methods. Compared with Swin-Transformer, its accuracy was improved by 3.1%, and the detection time was reduced by 15 ms.

**Table 2** Model experiments based on YOLOv5 and AMs added in the CSP1_X module on RAF-DB dataset

| Method | Model Size | Parameters | Accuracy (%) | mAP@0.5 (%) | Time (ms) |
|---|---|---|---|---|---|
| CSPA-SE | 92.268 M | 48.608 M | 74.2 | 81.9 | **20** |
| CSPA-ECA | 92.630 M | 48.883 M | 75.0 | 82.0 | 22 |
| CSPA-CBAM | 94.578 M | 49.309 M | 75.4 | 82.2 | 23 |
| CSPA-CA | 94.601 M | 49.974 M | **77.1** | **83.4** | 25 |

Bold meant the best performance of tampering detection in the same experimental setting

### 4.3 Experiments of CA at different positions

According to Table 2, the network with CA had the best accuracy. Then, we conducted a series of ablation experiments to study the effects of AMs on different modules in the Backbone of YOLOv5. First, we integrated CA into different modules of Backbone, such as Focus, CBS, and SPP, as shown in Fig. 3. Table 3 shows the performance of YOLOv5 with SA, FA, CBSA, and CSPA networks. Compared with YOLOv5, the accuracy of the four networks increased by 1.3%, 1.7%, 2.1%, and 3.5% respectively, and the mAP@0.5 increased by 0.3%, 0.5%, 0.7%, and 1.6%, respectively. The results showed the CSPA network also achieved the best accuracy and mAP@0.5. This was because CA was the combination of the spatial and channel attention. The feature was abstracted into a series of point attention weights by GAP and GMP, and then the relationship between these weights was established and attached to the original spatial or channel features. From the extraction position, CA extracted the features of RGB images through Focus module to obtain the attention weights of three channels and the spatial attention of the whole image size. The number of channels was small and the spatial feature map was too large to focus on specific features, which led to the increase in useless information. When CA was added after SPP, the number of channels was too large, which was easy to cause over-fitting. Meanwhile, SA was closer to the classification level. The effect of attention was easy to affect the decision of the classification level. Therefore, considering the channel and spatial information, CA was the most suitable for the middle layer of network with moderate channel and space. Meanwhile, CSPA had the residual structure, which increased the gradient value of back propagation between layers and avoided the gradient disappearance caused by deepening network structure, so that finer-grained features were extracted without worrying about network degradation. Therefore, according to the location and advantages of CSP1_X, the combination of CSP1_X and CA was the best.

**Table 3** Results of CAs in the Focus, CBS, SPP, and CSPA on RAF-DB dataset

| Method | Model size | Parameters | Accuracy (%) | mAP@0.5 (%) | Time (ms) |
|---|---|---|---|---|---|
| SA | 91.827 M | 46.853 M | 74.9 | 82.1 | **18** |
| FA | 91.827 M | 46.853 M | 75.3 | 82.3 | **18** |
| CBSA | 92.490 M | 47.884 M | 75.7 | 82.5 | 22 |
| CSPA | 94.601 M | 49.974 M | **77.1** | **83.4** | 25 |

Bold meant the best performance of tampering detection in the same experimental setting

**Table 4** Integrated results on RAF-DB dataset

| Method | Model Size | Parameters | Accuracy (%) | mAP@0.5 (%) | Time (ms) |
|---|---|---|---|---|---|
| FASA | 92.201 M | 47.295 M | 75.1 | 82.1 | **20** |
| FA + CBSA + CSPA + SA | 112.400 M | 55.735 M | 75.3 | 82.4 | 37 |
| CSPA + SA | 96.900 M | 51.103 M | 75.8 | 82.0 | 29 |
| FA + CSPA | 96.900 M | 51.103 M | 76.5 | 82.7 | 29 |
| CBSA + CSPA | 105.501 M | 52.310 M | **76.8** | **83.0** | 32 |

Bold meant the best performance of tampering detection in the same experimental setting

Second, we integrated SA, FA, CBSA, and CSPA into each other to constitute diverse networks. Table 4 shows the performance of YOLOv5 with FA + SA, FA + CBSA + CSPA + SA, CSPA + SA, FA + CSPA, and CBSA + CSPA networks. Compared with YOLOv5, the accuracy of the five corresponding networks increased by 1.5%,

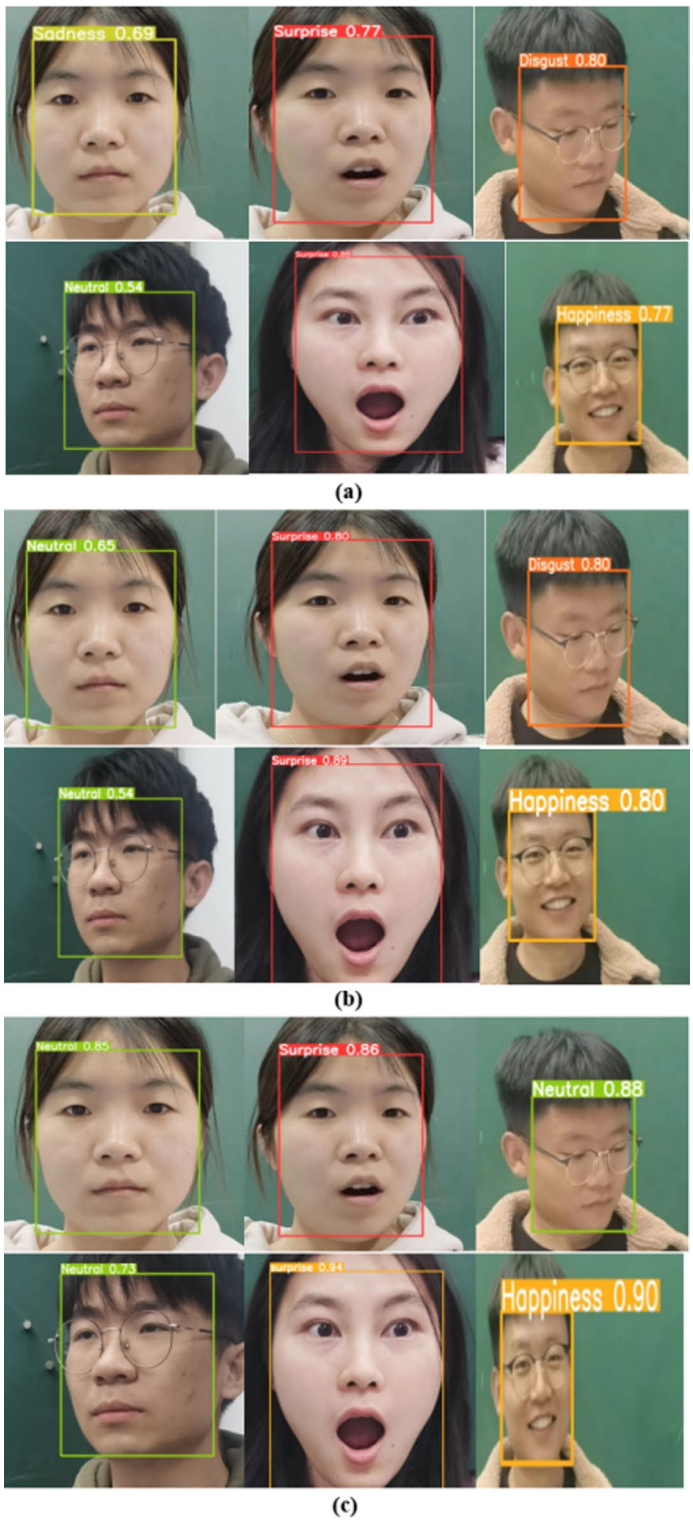


**Fig. 5** The qualitative results under the network of **a** YOLOv5, **b** CBSA + CSPA and **c** CSPA.

1.7%, 2.2%, 2.9%, and 3.2%, respectively. Meanwhile, the mAP@0.5 increased by 0.3%, 0.6%, 0.2%, 0.9%, and 1.2%. It was found that when CSPA was combined with other modules, their performance degrades. This was because the expansion of the attention range led to the increase of useless features, which made the learning ability of the model decline. Meanwhile, the models had more parameters, resulting in over-fitting. In addition, the detection time of these models increased compared with YOLOv5. In view of the above experiments, our model (CSPA) got the best accuracy.

## 5 Teacher's facial expression analysis

### 5.1 Teachers' facial expression analysis by images

Considering both the accuracy and detection time of the network, we selected YOLOv5, CBSA + CSPA, and CSPA to verify the actual recognition effect of the teachers' expression pictures with different gender, face size, and face posture. Figure 5 shows the qualitative results. CSPA had no errors, as shown in Fig. 5c. CBSA + CSPA had one error, which was the third picture, as shown in Fig. 5b. YOLOv5 had two pictures errors detection. The first picture was identified as sadness, and the third picture was identified as disgust, as shown in Fig. 5a.

### 5.2 Real-time teachers' facial expression analysis

Then, the real-time FER system based on CSPA network was designed to detect teachers' classroom expressions through the camera and the local teaching video, as shown in Fig. 6. We simulated 1.5 min teaching video about the advanced mathematics guidance course of freshmen to detect. Meanwhile, the recognition result log was printed to the corresponding window in real time, and saved in the corresponding local folder. Finally, the system showed expression distribution with time by scatter chart and pie chart, as shown in Fig. 7. The scatter chart showed the real-time expression distribution with time

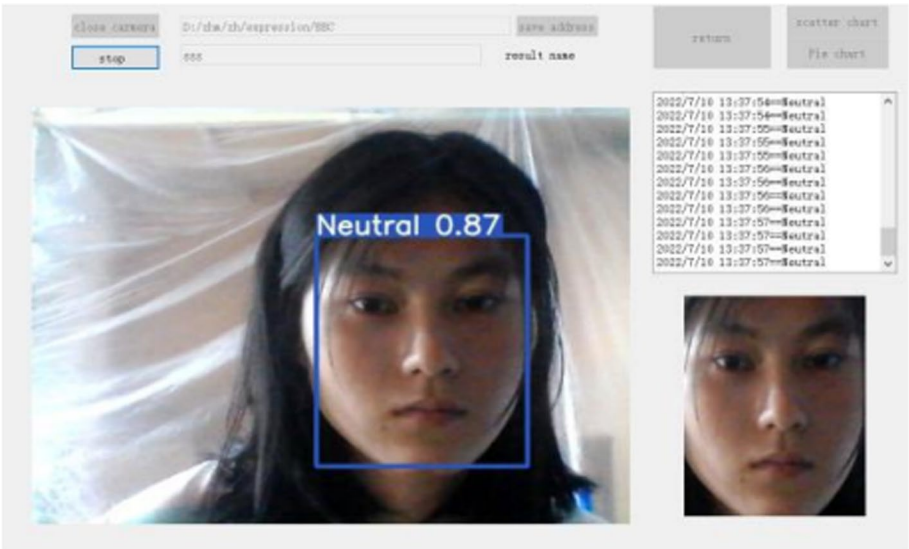


**Fig. 6** The real-time FER system.

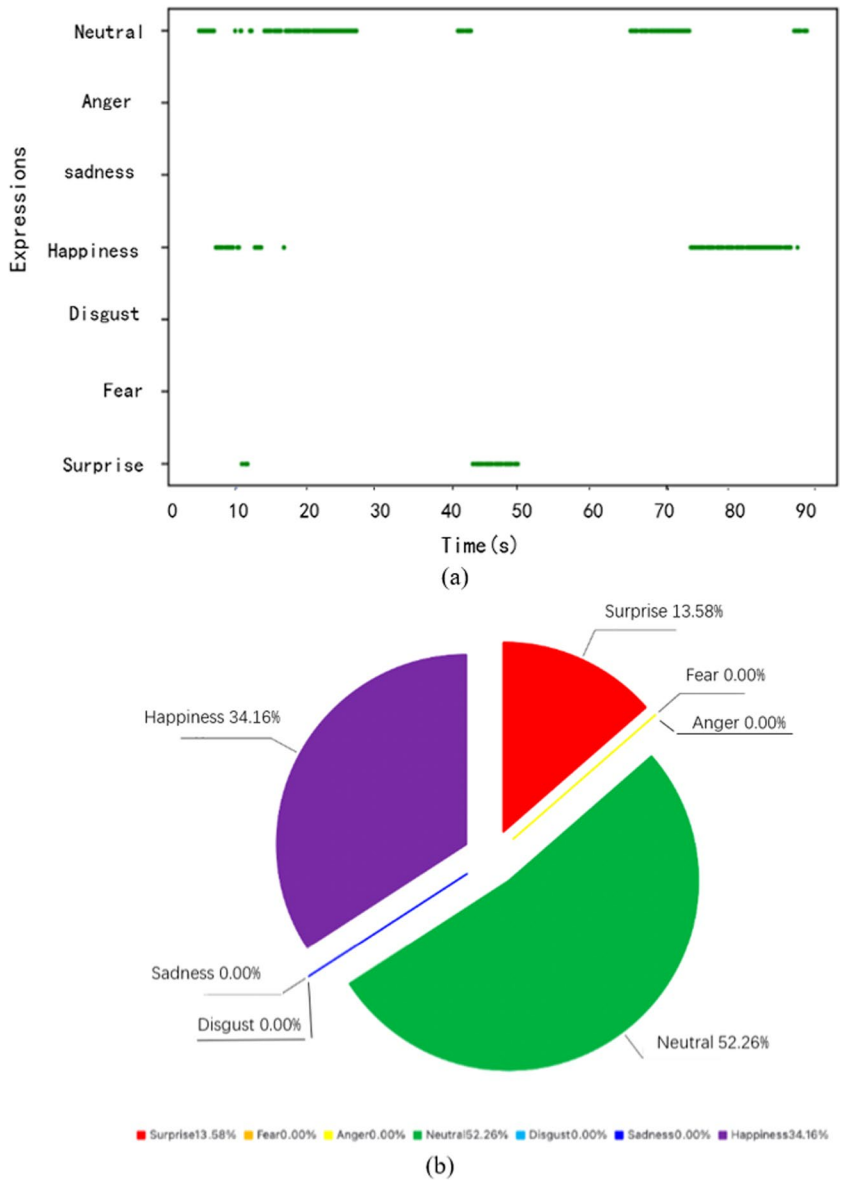


**Fig. 7** Facial expression distribution by the **a** scatter chart and **b** pie chart.

by analyzing a teachers' video. The abscissa indicated the time of real-time detection, and the ordinate indicated the expression, as shown in Fig. 7a. From the scatter points in the scatter chart, we could easily find the expression at every moment. However, there were no facial expressions in the process of recognition from 28 to 40 s and from 50 to 65 s. This was because there was no face during this time. The pie chart analysis clearly showed the real-time proportion of each expression, as shown in Fig. 7b. From the pie chart, we could clearly know the proportion of each expression in the 1.5 min teaching video. For this teacher, most of the expressions were neutral, happiness, and surprise, with no expressions of disgust, anger, sadness, or fear.

## 6 Conclusion

In this paper, we proposed a real-time teachers' expressions recognition network based on YOLOv5 and AMs. We studied the effects of different AMs on the CSP1_X module and CA on different modules in the Backbone structures of YOLOv5. The results showed the network that CA was incorporated after each CBS module of the CSP1_X module (CSPA) achieved the best accuracy of 77.1% and mAP@0.5 of 83.4% on RAF-DB dataset, which increased by 3.5% and 1.6%, respectively, compared with YOLOv5. Meanwhile, the detection time was 25 ms. The proposed model outperformed other detection methods, including Faster-RCNN, R-FCN, ResNext-101, DETR, Swin-Transformer, YOLOv3, and YOLOX. Finally, the real-time teachers' facial expression recognition system was designed based on CSPA to detect and analyze the teachers' facial expression distribution with time through camera and the teaching video. However, the method analyzed teachers' expressions only through images, ignoring other information contained in the real-time video, such as the time sequence information, the audio, and the text. In addition, there were still great challenges in teachers' facial expression recognition, such as lack of datasets and the complexity of teachers' expressions. In the next work, we can enrich teachers' expressions and make a special dataset for teachers. At the same time, we can analyze the teacher's expression from multiple aspects including the audio, text, and posture of teaching to improve the recognition accuracy.

**Abbreviations**

| | |
|---|---|
| AI | Artificial intelligence |
| AMs | Attention mechanisms |
| HMM | Hidden Markov model |
| AAM | Active appearance model |
| Faster-RCNN | Fast region-based convolutional network |
| R-CNN | Region-convolutional neural networks |
| R-FCN | ObjectDetectionvia region-based fully convolutional networks |
| SSD | Single shot multibox detector |
| YOLO | You only look once |
| VGGNet | Visual geometry group network |
| GoogLeNet | Inception network |
| SGD | Stochastic gradient descent |
| SE | Squeeze-and-excitation |
| ECA | Efficient channel attention |
| CBAM | Convolution block attention module |
| CA | Coordinate attention |
| FER | Facial expression recognition |
| CNN | Convolutional neural networks |
| BN | Batch normalization |
| GAP | Global average pooling |
| GMP | Global max pooling |

**Acknowledgements**

The authors would like to express their sincere thanks to the editors and anonymous reviewers.

**Author contributions**

TH and HZ designed the research. HZ conducted numerical and experimental validations and prepared the manuscript. All authors took part in discussing the results. All authors read and approved the final manuscript.

**Funding**

This work was supported by the Tianjin Science and Technology Planning Project under Grant No. 20JCYBJC00300 and the National Education Science Planning Project under Grant No. BHA220139.

**Availability of data and materials**

Please contact the authors for data requests.

## Declarations

**Ethics approval and consent to participate**
The article has been ethically approved and approved.

**Consent for publication**
All presentations of the case report have been agreed for publication.

**Competing interests**
The authors declare that they have no competing interests.

Received: 3 September 2022 Accepted: 3 May 2023
Published online: 13 May 2023

**References**

1. P.P. Filntisis, A. Katsamanis, P. Maragos, Photorealistic adaptation and interpolation of facial expressions using HMMS and AAMS for audio-visual speech synthesis, in *IEEE International Conference on Image Processing (ICIP), Beijing, China*, pp. 2941–2945. https://doi.org/10.1109/ICIP.2017.8296821 (2017).
2. A. Halder, A. Chakraborty, A. Konar, A. K. Nagar, Computing with words model for emotion recognition by facial expression analysis using interval type-2 fuzzy sets, in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India*, pp. 1–8. https://doi.org/10.1109/FUZZ-IEEE.2013.6622543 (2013).
3. N. Sebe, M.S. Lew, I. Cohen, A. Garg, T.S. Huang, Emotion recognition using a Cauchy Naive Bayes classifier, in *International Conference on Pattern Recognition, Quebec City, QC, Canada*, pp. 17–20. https://doi.org/10.1109/ICPR.2002.1044578 (2002).
4. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
5. J. Dai, Y. Li, K. He, et al., *R-FCN: Object Detection via Region-based Fully Convolutional Networks* (Curran Associates Inc, 2016), p. 379–387.
6. W. Liu, et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision. arXiv preprint, arXiv:1512.02325 (2016).
7. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91 (2016).
8. Simonyan, K., & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint, arXiv:1409.1556 (2014).
9. C. Szegedy et al., Going deeper with convolutions, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594 (2015).
10. H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei, J. Peng, Facial expression recognition based on VGGNet convolutional neural network, in *Chinese Automation Congress (CAC), Xi'an, China*, pp. 4146–4151. https://doi.org/10.1109/CAC.2018.8623238 (2018).
11. A. Khanzada, C. Bai, F. T. Celepcikay, Facial Expression Recognition with Deep Learning, arXiv preprint, arXiv:2004.11823 (2020).
12. L. X. bing, C. Lian, Face detection in natural scene based on improved Faster-RCNN. Comput. Eng. **47**(1), 210–216 (2021).
13. A.M. Roy, J. Bhaduri, Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4. Comput. Electron. Agric. **193**(2022), 106694 (2022). https://doi.org/10.1016/j.compag.2022.106694
14. M.O. Lawal, Tomato detection based on modified YOLOv3 framework. Sci. Rep. **11**, 1447 (2021). https://doi.org/10.1038/s41598-021-81216-5
15. A.M. Roy, R. Bose, J. Bhaduri, A fast accurate fine-grain object detection model based on YOLOv4 deep neural network. Neural Comput. Appl. **34**, 3895–3921 (2022). https://doi.org/10.1007/s00521-021-06651-x
16. H. Aung, A. V. Bobkov and N. L. Tun, Face detection in real time live video using yolo algorithm based on Vgg16 convolutional neural network, in *International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Sochi, Russia*, pp. 697–702. https://doi.org/10.1109/ICIEAM51226.2021.9446291 (2021).
17. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pp. 6517–6525. https://doi.org/10.1109/CVPR.2017.690 (2017).
18. J. Redmon, A. Farhadi. YOLOv3: An Incremental Improvement. arXiv preprint, arXiv:1804.02767 (2018).
19. B. Alexey, W. Chien-Yao, Y.M.L. Hong, YOLOv4: optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934v1 (2020).
20. Y. Fan, X. Lu, D. Li, et al., Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445–450 (2016).
21. Liu, Research and Implementation of Face Expression Recognition Algorithm based on Video Image. CQUPT, https://doi.org/10.27675/d.cnki.gcydx.2021.001271 (2021).
22. Zhang, Video Emotion Recognition Based on Dual-stream Network. JLU, https://doi.org/10.27162/d.cnki.gjlin.2022.002308 (2022).
23. Q.X. Fei, S. Kai, Z. Yue, Y. Yong, Z. Gang, J. Cheng, L.C. Ming, L.X. Dong, Z.J. Feng, Detection algorithm for key points on face based on attention model. Opt. Instrum. **42**(2), 45–49 (2020)
24. J. Kang, S. Li, Convolutional neural network face expression recognition based on attention mechanism. J. Shaanxi Univ. Sci. Technol. **38**(04), 159–165+171 (2020)

25. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2011–2023 (2020). https://doi.org/10.1109/TPAMI.2019.2913372
26. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pp. 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155 (2020).
27. S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional Block Attention Module. arXiv preprint, arXiv:1807.06521v2 (2018).
28. C. Xie, H. Zhu, Y. Fei, Deep coordinate attention network for single image super-resolution. IET Image Proc. **16**(1), 273–284 (2022)
29. S. Li, W. Deng and J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in *–IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pp. 2584–2593. https://doi.org/10.1109/CVPR.2017.277 (2017).

## Publisher's Note



**Hongmei Zhong** received the B.S. degree in Communication Engineering from NeiJiang Normal University. She is currently a master's degree candidate in information and communication engineering of Tianjin Normal University. Her current research interests include deep learning and computer vision.

**Tingting Han** received the B.S. degree in the School of Precision Instrument and Opto-Electronics Engineering at Tianjin University, China, in 2008, and the Ph.D. degree in School of Information Science and Technology from Nankai University, China, in 2013. She has been working for Tianjin Normal University, China, as a Lecturer since 2013 and as an associate professor since 2020. Her research interests include artificial intelligence technology in education and intelligent fiber sensing.

**Wei Xia** received the B.S. degree in electronic and information engineering from North China Institute of Aerospace Engineering. He is currently a master's degree candidate in information and communication engineering of Tianjin Normal University. His current research interests include deep learning and computer vision.

**Yan Tian** received the master's degree in artificial intelligence from Tianjin Normal University. His current research interests are deep learning and image recognition.

**Libao Wu** received Ph.D. degree from Beijing Normal University, China. He is a professor in Faculty of Education, Tianjin Normal University, China. In the past years, he has published over 150 academic papers in a variety of education journals in China. His current research interests are teacher education and maths education.