



FER-YOLO: Detection and Classification Based on Facial Expressions

Hui Ma¹(✉) , Turgay Celik^{1,2} , and Hengchao Li¹ 

¹ Southwest Jiaotong University, Chengdu 610031, China

² University of the Witwatersrand, Johannesburg 2000, South Africa

Abstract. Due to the wide application prospect and market value of emotion recognition, it has become an important research topic in today's society. Among them, facial expression recognition (FER) plays an important role in expressing human emotional information. Generally, the FER classification process includes face pre-processing (face detection, alignment, etc.), which adds extra workload. To this end, detection and classification are carried out simultaneously in this paper. We first manually annotated the RAF-DB dataset. We then designed an end-to-end FER network with better performance and applied it to facial expressions called FER-YOLO. FER-YOLO is built on the basis of YOLOv3. We combine the squeeze-and-excitation (SE) module with the backbone network and assign a certain weight to each feature channel so that FER-YOLO can focus on learning prominent facial features. We also discussed the performance changes caused by the lightweight enhanced feature extraction networks. Experimental results show that the proposed FER-YOLO network is 3.03% mAP higher than YOLOv3 on the RAF-DB dataset.

Keywords: Emotion recognition · Facial Expression Recognition (FER) · Detection · Convolutional Neural Network (CNN)

1 Introduction

Emotion recognition is a popular research topic in computer vision and pattern recognition. With the development of artificial intelligence and deep learning technology, as well as the widespread application of emotion recognition in the fields of emotion computing [1], human-computer interaction [2], auxiliary medical [3], intelligent monitoring and security, entertainment industry [4], remote education [5], emotional state analysis [6] and other fields have attracted the attention of many researchers. Emotions play an active and important role in daily human communication. It can be expressed by detecting physiological signals such as breathing, heart rhythm, and body temperature and by detecting emotional behaviors such as facial expressions, language, and posture.

Southwest Jiaotong University.

© Springer Nature Switzerland AG 2021

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 28–39, 2021.

https://doi.org/10.1007/978-3-030-87355-4_3

Among them, facial expressions contain rich human emotion information. Emotional changes in human hearts will lead to different degrees of changes in facial expressions, indicating the daily emotions of human beings as well as subtle and complex emotional changes. In addition, emotional recognition based on facial expression recognition [7–9] is simple, and facial expression images are easy to capture.

The research and development of facial expression recognition can accelerate the advancement of research and technological development in machine vision, human-computer interaction, psychology, etc., and help demonstrate and analyze new interdisciplinary theories and methods. Facial expression recognition methods are divided into traditional methods and deep learning-based methods. The traditional FER method includes three steps: pre-processing (detection, alignment, etc.), feature extraction, and face image classification. Unlike traditional FER, deep learning-based methods allow end-to-end learning directly from the input image, reducing reliance on pre-processing and the cost of manually extracting features. Because deep neural networks' automatic learning has the advantage of discrimination, the FER method based on deep learning is better than the traditional FER method [10–12]. GPU computing technology development further promotes facial expression recognition based on deep learning to become the mainstream of today's research. There are also many research achievements in facial expression recognition based on deep learning.

Zou et al. [7] improved the convolutional neural network by using batch regularization, ReLU activation function, and Dropout technology. Singh et al. [8] studied the static images based on CNN in the cases of pre-processing and without pre-processing, respectively. Mohan et al. [10] proposed a FER-net network, which first automatically learns facial image features through the FER-net network and then recognizes them through the Softmax classifier. Xie et al. [11] designed two CNNs to extract local features and global features of the face images and obtain rich face information by fusing the two features to complete the classification.

The above-mentioned facial expression recognition method based on deep learning separately completes face image detection and classification tasks independently. With the rapid development of object detection technology based on deep learning, it is possible to simultaneously complete detection and classification tasks. In 2015, Redmon et al. [13] proposed YOLO (You Only Look Once), which was the first one-stage detector. Later, YOLOv2(YOLO9000) [14], YOLOv3 [15], and YOLOv4 [16] were extended on the basis of YOLO. Yolo divided the image into different regions and simultaneously predicted each region's bounding box and possibilities. To improve accuracy, YOLOv2 first modified the pre-trained classification network's resolution to 448×448 and then removed the fully connected layer (to obtain more spatial information) and used anchor boxes to predict bounding boxes. YOLOv3 introduces the FPN structure to realize multi-scale prediction. The backbone feature extraction network uses a better classification network, Darknet53, and the classification loss function uses Binary Cross-entropy Loss instead of Softmax. YOLOv4 has been improved

on the basis of YOLOv3, mainly by changing the backbone feature extraction network Darknet53 into CSPDarknet53 and enhancing the feature extraction network using SPP and PANET structures as well as data augmentation techniques. In 2016, Liu et al. [17] proposed the SSD model. SSD adopts the structure of the pyramidal feature hierarchy and introduces the prior box, which has the advantage of mean Average Precision (mAP) compared with YOLO. In 2017, Lin et al. [18] proposed RetinaNet. RetinaNet alleviates the problem of data imbalance by introducing a focal loss function.

Although the above one-stage object detection methods have achieved excellent results in many fields, they are rarely used in FER tasks to the best of our knowledge. In this paper, we have applied FER tasks based on one-stage object detection methods. First, we manually annotated the RAF-DB dataset. Then based on YOLOV3, the SE [19] module is combined with the Backbone module to improve the ability to enhance the feature extraction network. We also discussed the performance changes caused by the lightweight enhanced feature networks.

2 Methodology

2.1 Model Architecture

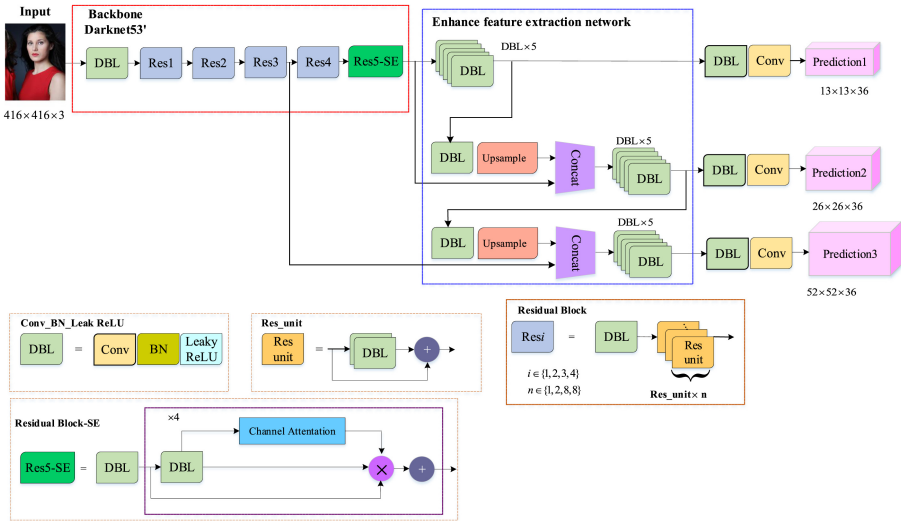


Fig. 1. The architecture of FER-YOLO, which includes the backbone network (red dashed line), enhance feature extraction network (blue dashed line) and result prediction. (Color figure online)

In this paper, the FER-YOLO network model architecture based on FER is shown in Fig. 1. The backbone feature extraction network uses Darknet-53' with

the fully connected layer removed. Enhance feature extraction network fuses three scales feature maps. In order to improve the performance of the feature extraction network, the SE module is introduced into the Res5 residual block (Res5-SE). Res5-SE aims to extract more useful information from the deep features of the backbone network. Compared with datasets dedicated to object detection, although the facial expression dataset has a single characteristic, it still contains many environmental backgrounds. In the facial expression recognition task, pre-processing is usually carried out to remove the image background and reduce the difficulty of recognition. Unlike other general facial expression recognition methods, FER-YOLO input data does not have any pre-processing steps. Our input data contains a lot of background information and the size of the input image is $416 \times 416 \times 3$.

Finally, output the prediction results corresponding to the three feature layers, and the shape of the output layer is $13 \times 13 \times 36$, $26 \times 26 \times 36$ and $52 \times 52 \times 36$. $36 = 3 \times (7 + 4 + 1)$, where, 3 is the three prior boxes, 7 is the number of expression categories, 4 is the parameters of x , y , w , and h , and 1 is the detection of objects.

2.2 Channel Attention

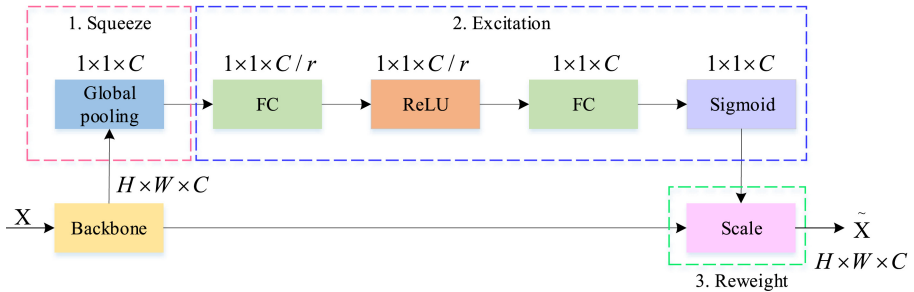


Fig. 2. FER-YOLO channel attention, including squeeze module (red dotted line), excitation module (blue dotted line) and reweight module (green dotted line). (Color figure online)

In order to enable the FER-YOLO network to learn more discriminative features and improve accuracy, this paper uses the channel attention mechanism to recalibrate the depth feature maps extracted by the backbone network. Equivalent to input enhanced feature extraction network information is more discriminant.

As shown in Fig. 2, FER-YOLO channel attention consists of the squeeze module (red dotted line), the excitation module (blue dotted line), and the reweight module (green dotted line). The squeezing operation compresses each feature channel of the output U ($U \in R^{W \times H \times C}$) of the backbone through the global average pooling layer to obtain a $1 \times 1 \times C$ real number sequence. The

excitation module is composed of two fully connected layers, in which the scaling parameter r is used to reduce the number of channels and thus reduce the amount of calculation. Empirically, here we set it to 4. After the channel attention, the input and output shapes are the same, but each position's value has been re-corrected.

2.3 Depth-Wise Separable Convolution

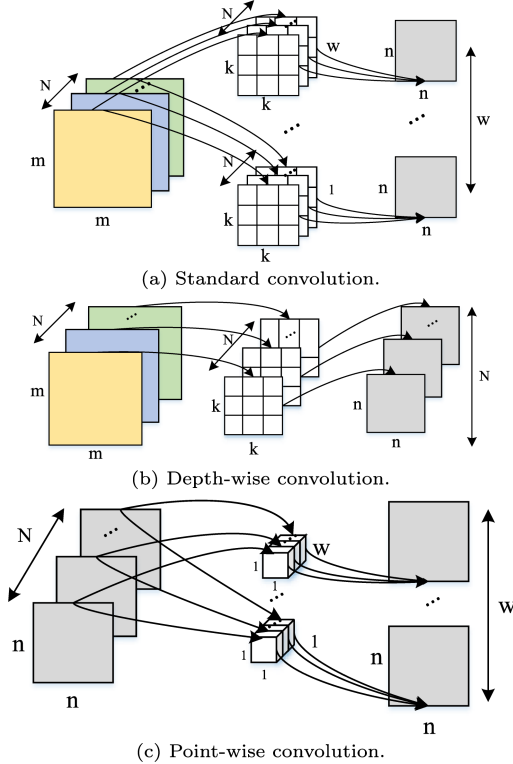


Fig. 3. Standard convolution vs depth-wise separable convolution (a)

In 2012, Mamalet et al. [20] first proposed depth-wise separable convolution. Depth-wise separable convolution (*DSC*) consists of depth-wise convolution (*DW*) and point-wise convolution (*PW*). The standard convolution (*SC*) operation is that all channels in the corresponding image region are considered simultaneously. Different from the standard convolutional network, the convolution kernel is first divided into single channels. The convolution operation is performed on each channel without changing the depth of the input feature image. In this way, the output feature map with the same number of channels as

the input feature map is obtained. However, this operation independently performs convolution operations on each channel of the input layer, and there is no information exchange between channels. Therefore, 1×1 convolution is needed to increase the information exchange between channels. The separation of channels and regions can accelerate the training process, and the trained model has fewer parameters and faster speed, which is suitable for lightweight models.

As shown in Fig. 3, given the input feature maps size is $m \times m \times N$, and the kernel size is $m \times k \times k \times N \times w$, and the stride is 1, the parameters (P) of standard convolution is:

$$P(SC) = k \times k \times N \times w \quad (1)$$

And the floating point operations ($FLOPs$) is:

$$FLOPs(SC) = m \times m \times k \times k \times N \times w \quad (2)$$

For the depth-wise separable convolution, the number of parameters and FLOPs are:

$$P(DSC) = P(DW) + P(PW) = k \times k \times N + N \times w \quad (3)$$

$$\begin{aligned} FLOPs(DSC) &= FLOPs(DW) + FLOPs(PW) \\ &= m \times m \times k \times k \times N + m \times m \times N \times w \end{aligned} \quad (4)$$

$$F_P = \frac{P(DSC)}{P(SC)} = \frac{1}{w} + \frac{1}{k^2} \quad (5)$$

$$F_{FLOPs} = \frac{FLOPs(DSC)}{FLOPs(SC)} = \frac{1}{w} + \frac{1}{k^2} \quad (6)$$

It can be seen from Eq. (6) and Eq. (5) that the depth-wise separable convolution has advantages in terms of the number of parameters and the FLOPs.



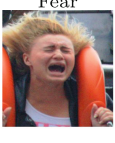




3 Experimental Results and Analysis

3.1 Implementation Details

Our experiment is implemented by using the PyTorch framework on a workstation computer with the following specifications: Intel(R) Core(TM) i9-9980XE CPU @ 3.00GHz, 125 GB RAM, and GeForce RTX 2080 Ti.

- 1) *Facial expression dataset*: The experiment in this paper is based on the RAF-DB dataset [21], which is an unconstrained large-scale facial expression recognition dataset of various sizes. For end-to-end experiment, we manually annotated them. It is worth pointing out that we did not perform any pre-processing such as face alignment. The RAF-DB dataset contains 12,271 training images and 3,068 test images. The details are shown in Table 1.

Table 1. The sample statistics of RAF-DB dataset.

Facial Expression Categories						
Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
						
training						
705	717	281	4,772	1,982	1,290	2,524
testing						
162	160	74	1,185	478	329	680

- 2) *Parameter settings:* We use the Adam optimization method for all models, with 16 batch-size samples for end-to-end training. To be fair, we train all network models for 160 epochs and set the initial learning rate to 0.001, which is multiplied by 0.9 times every 5 epochs. During the training process, we initialize the FER-YOLO model network’s parameters with the weights of DarkNet53 pre-trained on the COCO dataset. It is worth pointing out that all models in this paper only initialize the backbone network’s weights.

3.2 Ablation Experiments

We conducted experiments on the FER-YOLO network model and its deformation. The test results on the test set are shown in Table 2. FER-YOLO* indicates that FER-YOLO removes the first feature fusion channel, which is the prediction result of $52 \times 52 \times 36$. In the same way, YOLOv3* represents YOLOv3 removes the first feature fusion channel. FER-YOLO-L means that the depth-wise separable convolution replaces the standard convolution in the enhanced feature extraction network in FER-YOLO, thereby lightweight FER-YOLO. Similarly, YOLOv3*-L, YOLOv3-L, and FER-YOLO*-L represents the lightweight of YOLOv3*, YOLOv3, and FER-YOLO*, respectively.

Table 2. Ablation experiments for FER-YOLO and some of its deformation. Legend: “Angry” (Ang); “Disgust” (Dis); “Fear” (Fea); “Happy” (Hap); “Neutral” (Neu); “Sad” (Sad) and “Surprise” (Sur)

Method	Ang	Dis	Fea	Hap	Neu	Sad	Sur	mAP(%)	Param(M)
FER-YOLO*	65.71	36.31	54.48	93.14	76.22	81.22	82.81	69.98	61.56
YOLOv3*	77.16	45.43	50.37	92.01	76.73	86.00	83.74	73.06	46.91
FER-YOLO*-L	75.85	51.26	46.94	93.79	77.42	83.19	83.84	73.19	60.51
YOLOv3*-L	80.15	49.92	47.78	94.23	75.99	87.52	83.83	74.20	44.81
YOLOv3-L	80.79	46.66	55.34	93.86	77.98	85.19	83.48	74.76	45.07
FER-YOLO-L	77.55	51.82	54.75	92.96	78.62	84.72	84.93	75.05	47.17
FER-YOLO	80.41	52.44	61.04	94.48	76.46	84.47	88.06	76.77	63.65

We use depth-wise separable convolution instead of standard convolution in the enhanced feature extraction network to explore these two different convolutions' impact on network performance. From the theoretical analysis of Sect. 2 B, it can be seen that compared to standard convolution, depth-wise separable convolution can indeed greatly reduce model parameters. From the experimental results in Table 2, it can be seen that the depth-wise separable convolution reduces the network model parameters used in this paper by approximately 15M. For different network structures, the depth-wise separable convolution replaces the standard convolution, and mAP may not be improved. And our experimental results show that in the RAF-DB dataset, a small receptive field has a large impact on mAP. If the small receptive field is removed, the network will miss small face images, and the mAP of the FER-YOLO network will be reduced from 76.77% to 69.98%.

Figure 4 shows the detection results of the FER-YOLO network model on the RAF-DB test dataset. The corresponding categories and scores are the prediction results and confidence scores of the detected images, respectively. We can see that the FER-YOLO network model can successfully detect facial expressions in images with complex backgrounds and accurately locate and predict them.

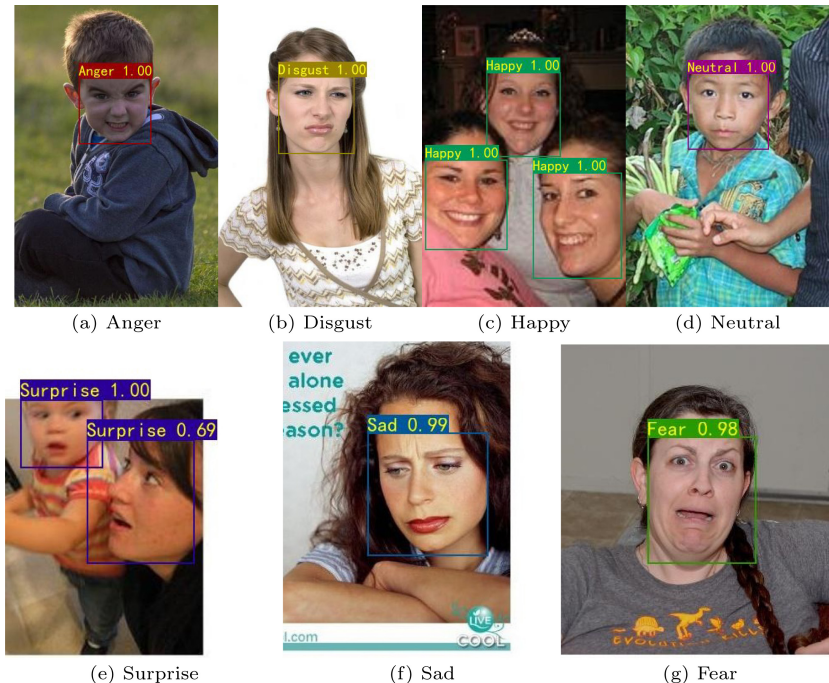


Fig. 4. Test sample detection results.

3.3 Comparisons with State-of the Art Methods

In Table 3, we compare the FER-YOLO network model with one-shot object detection state-of-the-art methods. Black bold font represents the best result.

The results show that our proposed FER-YOLO network model is better than the one-shot object detection model considered in this paper. Compared with YOLOv4 and YOLOv3, mAP has increased by 3.13% and 3.03%, respectively. The Average precision (AP) of the FER-YOLO network model in the categories of “Angry”, “Disgust”, “Fear” and “Surprise” is higher than the other one-shot object detection methods listed in the paper. The AP values of “Happy”, “Neutral”, and “Sad” are lower than the highest RetinaNet by 0.1%, 7.62%, and 2.14%, respectively.

Table 3. Comparisons with one-shot object detection state-of the art methods. Legend: “Angry” (Ang); “Disgust” (Dis); “Fear” (Fea); “Happy” (Hap); “Neutral” (Neu); “Sad” (Sad) and “Surprise” (Sur)

Method	Ang	Dis	Fea	Hap	Neu	Sad	Sur	mAP(%)
RetinaNet [18]	76.58	34.10	38.59	94.58	84.08	86.61	82.16	70.96
SSD [17]	79.99	39.46	50.13	95.19	80.28	84.12	76.24	72.20
YOLOv4 [16]	79.95	42.19	60.39	90.86	79.08	84.34	82.70	73.64
YOLOv3 [15]	79.07	45.55	57.38	93.28	71.69	84.67	84.51	73.74
FER-YOLO	80.41	52.44	61.04	94.48	76.46	84.47	88.06	76.77

3.4 Real-Time Facial Expression Detection via Camera

We use the real-time camera to read the face image for detection and classification directly. The results are shown in Fig. 5. As shown in Fig. 5, the FER-YOLO network model’s performance in real-time detection through the camera is also prominent, which can accurately locate the face image and perform the correct classification.

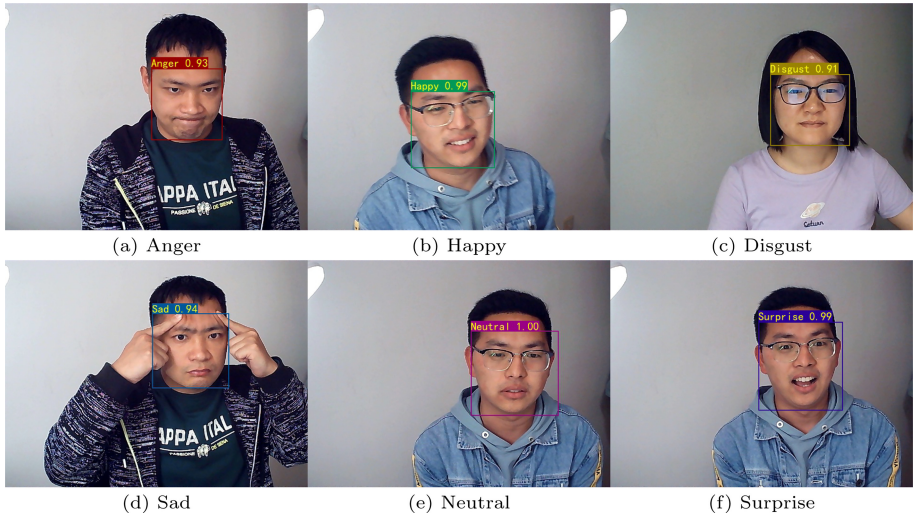


Fig. 5. Detected facial expressions via real-time camera.

4 Conclusion

In this paper, we design a FER-YOLO network model for facial expression recognition. This model's backbone is combined with the SE module to improve the performance of the enhanced feature extraction network. Experiments were performed on the manually annotated RAF-DB dataset. The experimental results show that FER-YOLO achieves better results compared with other one-shot object detection methods. It provides a reference for one-shot object detection methods based on expression recognition.

Acknowledgements. This work was supported by Sichuan Provincial Science and Technology Projects (2019JDJQ0023).

References

1. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. *Inf. Fusion* **37**, 98–125 (2017)
2. Spezialetti, M., Placidi, G., Rossi, S.: Emotion recognition for human-robot interaction: recent advances and future perspectives. *Front. Robot. AI* **7**, 145–155 (2020)
3. Joesph, C., Rajeswari, A., Premalatha, B., Balapriya, C.: Implementation of physiological signal based emotion recognition algorithm. In: *IEEE 36th International Conference on Data Engineering (ICDE) 2020, Dallas, TX, USA*, pp. 2075–2079 (2020). <https://doi.org/10.1109/ICDE48307.2020.9153878>

4. Cosentino, S., Randria, E.I.S., Lin, J.-Y., Pellegrini, T., Sessa, S., Takahashi, A.: Group emotion recognition strategies for entertainment robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018, Madrid, Spain, pp. 813–818 (2018). <https://doi.org/10.1109/IROS.2018.8593503>
5. Li, G., Wang, Y.: Research on Leamer’s emotion recognition for intelligent education system. In: IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, pp. 754–758 (2018). <https://doi.org/10.1109/IAEAC.2018.8577590>
6. Rusli, N., Sidek, S.N., Yusof, H.M., Ishak, N.I., Khalid, M., Dzulkarnain, A.A.A.: Implementation of wavelet analysis on thermal images for affective states recognition of children with autism spectrum disorder. *IEEE Access* **8**, 120818–120834 (2020)
7. Zou, J., Cao, X., Zhang, S., Ge, B.: A facial expression recognition based on improved convolutional neural network. In: IEEE International Conference of Intelligent Applied Systems on Engineering (ICIASE), Fuzhou, China, pp. 301–304 (2019). <https://doi.org/10.1109/ICIASE45644.2019.9074074>
8. Singh, S., Nasoz, F.: Facial expression recognition with convolutional neural networks. In: 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp. 0324–0328 (2020). <https://doi.org/10.1109/CCWC47524.2020.9031283>
9. Ma, H., Celik, T., Li, H.-C.: Lightweight attention convolutional neural network through network slimming for robust facial expression recognition. *Signal Image Video Process.* 1863–1711 (2021)
10. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: FER-net: facial expression recognition using deep neural net. *Neural Comput. Appl.* **33**(15), 9125–9136 (2021). <https://doi.org/10.1007/s00521-020-05676-y>
11. Xie, S., Hu, H.: Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Trans. Multimedia* **21**(1), 211–220 (2019)
12. Ma, H., Celik, T.: FER-Net: facial expression recognition using densely connected convolutional network. *Electron. Lett.* **55**(4), 184–186 (2019)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: 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 (2016). <https://doi.org/10.1109/CVPR.2016.91>
14. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 6517–6525 (2017). <https://doi.org/10.1109/CVPR.2017.690>
15. Joseph, R., Ali, F.: YOLOv3: An Incremental Improvement. [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
16. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020)
17. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
18. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **422**, 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
19. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **428**, 2011–2023 (2020)

20. Mamalet, F., Garcia, C.: Simplifying ConvNets for fast learning. In: International Conference on Artificial Neural Networks (ICANN 2012), Lausanne, Switzerland, pp. 58–65 (2012). https://doi.org/10.1007/978-3-642-33266-1_8
21. Li, S., Deng, W., Du, J.: 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 (2017). <https://doi.org/10.1109/CVPR.2017.277>