Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2022.026813
Article

Tech Science Press

Negative Emotions Sensitive Humanoid Robot with Attention-Enhanced Facial Expression Recognition Network

Rongrong Ni[1], Xiaofeng Liu[1,*], Yizhou Chen[1], Xu Zhou[1], Huili Cai[1] and Loo Chu Kiong[2]

[1]College of IoT Engineering, Hohai University, Changzhou, 213100, China
[2]Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
*Corresponding Author: Xiaofeng Liu. Email: xfliu@hhu.edu.cn
Received: 05 January 2022; Accepted: 12 February 2022

Abstract: Lonely older adults and persons restricted in movements are apt to cause negative emotions, which is harmful to their mental health. A humanoid robot with audiovisual interactions is presented, which can correspondingly output positive facial expressions to relieve human's negative facial expressions. The negative emotions are identified through an attention-enhanced facial expression recognition (FER) network. The network is firstly trained on MMEW macro-and micro-expression databases to discover expression-related features. Then, macro-expression recognition tasks are performed by fine-tuning the trained models on several benchmarking FER databases, including CK+ and Oulu-CASIA. A transformer network is introduced to process the sequential features engineered by the FER network and output a final stable control order. This order is used to control the robot's facial motor units to generate different expressions, e.g., a smile expression. Evaluations on benchmarking databases verify that the proposed method can precisely recognize facial expressions. The joint modulation with the humanoid robot proves that the robot can respond effectively to the user's negative emotions.

Keywords: Humanoid robot; facial expression recognition; attention mechanism; transfer learning; negative emotions

1 Introduction

Intelligent companion robots have been widely used in homes for the elderly. Persons restricted in movements or older adults who live alone tend to have negative emotions in their daily life easily. Emotional interaction with the intelligent companion robot can effectively relieve their negative emotions [1,2]. There are many ways to obtain emotional information in human-computer interaction [3–5]. Former researches show that people of different cultures have the same facial expressions to express their negative emotions [6]. Therefore, it is feasible to recognize their negative emotions based on their facial expressions. However, current mainstream companion robots suffer from recognizing users' expressions precisely. Moreover, they always have simple facial structures, which cannot effectively respond to users' negative emotions, resulting in a lack of user experience.

The premise that the companion robot effectively responds to the user's negative emotions is to recognize the user's facial expressions accurately. Facial expression recognition (FER) is a classic problem in the field of emotion computing. Commonly used FER methods consist of hand-crafted feature-based and deep neural network-based strategies. For the former one, widely used expression-related features consist of geometric and appearance features. For geometric features, informative features vectors [7,8] are calculated based on facial landmark points detected on RGB face images. Further, facial landmark points detected on RGB-D face images [9] can provide more accurate geometric information. For appearance features, holistic spatial analysis [10], local binary pattern (LBP) [11,12], the histogram of oriented gradients (HOG) [13], and Gabor texture [14] are widely used for feature engineering.

Recently booming deep neural networks (DNNs) have achieved great successes in different applications, such as object detection [15], anomaly detection [16], semantic segmentation [17], trajectory prediction [18], wisdom medical [19], and action recognition [20]. Unlike hand-crafted features, DNNs can automatically extract expression-related features in a data-driven manner [21–23]. DNNs-based FER approaches can significantly outperform hand-crafted feature-based approaches with enough training data and a proper training strategy.

Although DNNs-based FER approaches perform well on public datasets, they still suffer from low inaccuracies while detecting users' actual expressions. One reason is the scale mismatch between the network parameters and the data volume in public datasets. This mismatch causes the network's failure to fully exploit the expression-related features, resulting in poor generalization performance. Due to the inability to accurately recognize the user's facial expressions, the robot cannot identify the negative emotions revealed by the user, resulting in poor companionship. Meanwhile, the frame-based FER ignores the temporal correlation between frames, resulting in the inability to send reliable control instructions to the robot, affecting the robot to respond effectively.

To identify users' facial expressions accurately, a shallow attention-enhanced facial expression recognition network (SAFERN) is proposed. A two-stage training strategy is used to force the network to focus better on facial macro-expressions. A lightweight transformer network is introduced to process the sequence features output by the SAFERN. Afterward, it can predict a stable FER result, which is beneficial to sending control orders to the humanoid robot. When the robot detects users' negative emotions such as sadness, frustration, and fear, it controls the facial motor unit to generate a smile expression as a response. Meanwhile, the voice comfort function will be added in future work to improve the robot's company performance. Our main contributions are as follows:

1. A SAFERN is proposed to perform frame-based FER with an attention enhancement to force the network to explore facial details better. Meanwhile, we present a two-stage training strategy, which firstly trains the network to distinguish facial macro-and micro-expressions, and then migrate the network to macro-expression recognition. Such a training strategy can better explore macro-expression-related features.
2. A lightweight transformer network is introduced to process the sequence features output by SAFERN to obtain stable FER results in the temporal domain. Therefore, it can avoid false control instructions caused by occasional wrong recognition results predicted by SAFERN.
3. We design a humanoid robot that can detect users' negative emotions like sadness, frustration, and fear. The robot can generate a smile expression to respond to users' negative emotions.

The rest of this work is organized as follows. Section 2 reviews frame-based and sequence-based FER approaches. Section 3 presents the proposed method. The experimental results are performed in Section 4. Section 5 discusses and concludes the proposed work.

2 Related Work

2.1 Frame-Based FER Approaches

Frame-based FER approaches recognize different facial expressions from static face images. As we mentioned above, early works concentrate on hand-crafted features, including holistic spatial analysis [10], LBP [11,12], HOG [13], depth features [24], and Gabor texture [14]. However, these methods suffer from low accuracy under some challenging cases, such as poor illumination conditions.

To improve the recognition performance, modern FER approaches always resort to DNNs [25–27]. For example, Yang et al. [28] presented multi-channel DNNs to perform FER. Features extracted by different channels are combined in a weighted manner. Li et al. [29] generated 2D facial attribute maps from a 3D scan and fed all maps into a multi-channel convolutional neural network (CNN). Jan et al. [30] proposed a deep fusion CNN to learn from local facial regions. Barros et al. [31] proposed a lightweight FER model named FaceChannel, which contains ten convolutional layers and uses shunting inhibitory fields in the last layer. Zhang et al. [32] proposed an end-to-end deep learning model, exploiting different poses and expressions jointly for simultaneous facial image synthesis and pose-invariant facial expression recognition.

Benefiting from the strong feature engineering power, DNNs-based FER approaches outperform hand-crafted features by a large margin.

2.2 Sequence-Based FER Approaches

Unlike frame-based FER approaches, sequence-based FER approaches recognize different facial expressions from a facial image sequence, for example, a video captured by a web camera [33]. The key to recognizing different facial expressions stably relies on modeling the temporal associations between consecutive frames [34]. For example, the optical flow-based [35], dynamic image-based [36], multiple handcraft features-based [37], multi-signal CNN-based [38] and 3D CNN-based [39,40] methods are proposed to process the image sequence.

Except for the methods mentioned above, Long Short-Term Memory (LSTM) [41] and Gated Recurrent Unit (GRU) [42] are commonly used to model temporal associations. Yu et al. [43] proposed a multi-task global-local FER network based on LSTMs. Kang et al. [44] presented a convolutional gate recurrent unit for video FER in the wild. However, LSTMs and GRUs cannot model long-term sequences. Recently, the transformer network has achieved great successes in natural language processing [45,46] and computer vision [47,48]. With the help of the multi-head attention mechanism, the transformer network is good at modeling long-term sequences. This work introduces a lightweight transformer network to process the facial image sequence to output a stable recognition result.

3 Proposed Method

3.1 System Overview

Fig. 1 illustrates the overview of the proposed human-like robot that can respond to user's negative emotions. Both the software and hardware designs are illustrated in this figure. A landmark-based face detection approach [49] is used in the software design to perform face detection in the captured facial image sequence. Afterward, the detected faces are fed into the frame-based facial expression recognition module, which comprises well-designed convolutional neural networks. This module will output a set of sequence features, which are further fed into the sequence-based facial expression recognition module to generate stable recognition results. Specifically, a shallow transformer network consists of masked multi-head attention, add & norm, and feed-forward neural network (FNN) to process preceding sequence features. Based on the FER results, the human-like robot will generate different expressions to respond to its users. For example, it can generate a smile expression if it detects negative emotions revealed by its uses.

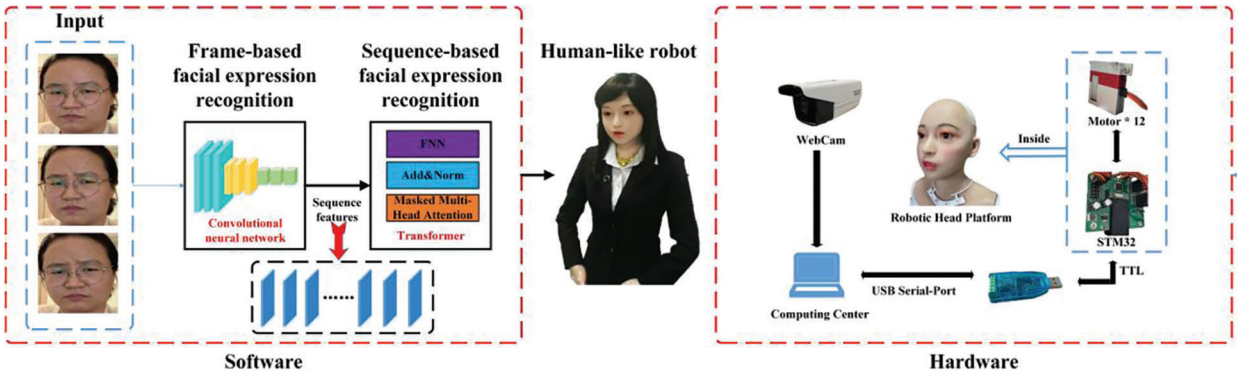


Figure 1: Overview of the proposed companion robot system. Our human-like robot is shown in the middle. The left and right sub-figures present the software and hardware designs, respectively

In the hardware design, the robot webcam collects images of the user's face and transmits them to the computing center through the network. The computing center determines the user's current mood through the facial expression recognition algorithm and generates the robot facial action instructions according to the emotion label, which is sent to the STM32 microprocessor of the robot head via the serial port. The microprocessor controls the motion of 12 micro-servo motors in the robot's headspace, affecting the robot's soft skin through the nylon rope to generate a facial expression.

3.2 Extractions of Frame-Based FER-Related Features

The pipeline to extract frame-based FER-related features is illustrated in Fig. 2. As shown in the figure, a shallow attention-enhanced facial expression recognition network (SAFERN) is proposed to extract FER-related features from a single frame. The structure of SAFERN is given in the bottom blue rectangle. Specifically, it comprises five shallow down-sampling modules (SDMs) and nine shallow residual attention modules (SRAMs). SDM (3, 16, 3) consists of a convolutional layer with input channel 3, output channel 16, and kernel size 3. The convolutional layer is followed by a batch normalization layer, a max-pooling layer, and the P-Relu layer. SRAM (16, 16, 3) consists of two convolutional layers with input channel 16, output channel 16, and kernel size 3. Each convolutional layer is followed by a spatial attention layer, a batch normalization layer, and the P-Relu layer. The spatial attention layer introduces a Softmax enhancement to explore region-related features. Residual connection is used to avoid the gradient vanishing. The adaptive pooling layer is used to convert the 2-dimensional feature map into a 1-dimensional feature vector, which will be used for classification. In the training stage, the cross-entropy loss is used as the loss function to optimize SAFERN.

The cross-entropy loss is replaced by the Softmax function to generate the probability scores in the testing stage. Then, the highest probability score category will be the predicted class during the inference process. Specifically, the predicted class is given by the argmax (Softmax (•)) function as follows:

$$\hat{y}^k = \arg\max \frac{e^{\theta^T x^k}}{\sum_{i=1}^{K} e^{\theta_i^T x^k}} \tag{1}$$

where k denotes the category number, θ represents the network parameter, x^k and $\hat{y}^k$ denote the input feature and predicted class, respectively.

To better explore macro-expression-related features, a two-stage training strategy is proposed. As shown in the figure, in the first stage, SAFERN_v1 (output channel of FC layer is set to two) is pre-trained on MMEW, which consists of both macro and micro-expressions. The discrepancies between macro and

micro-expressions may force the network to distinguish features belonging to different expressions. Afterward, we fine-tune SAFERN_v2 (output channel of FC layer is set to six) on CK+ and Oulu-CASIA to extract macro-expression-related features. Given a facial image sequence, SAFERN_v2 will output a set of sequence features, which will be further processed to extract sequence-based FER-related features.

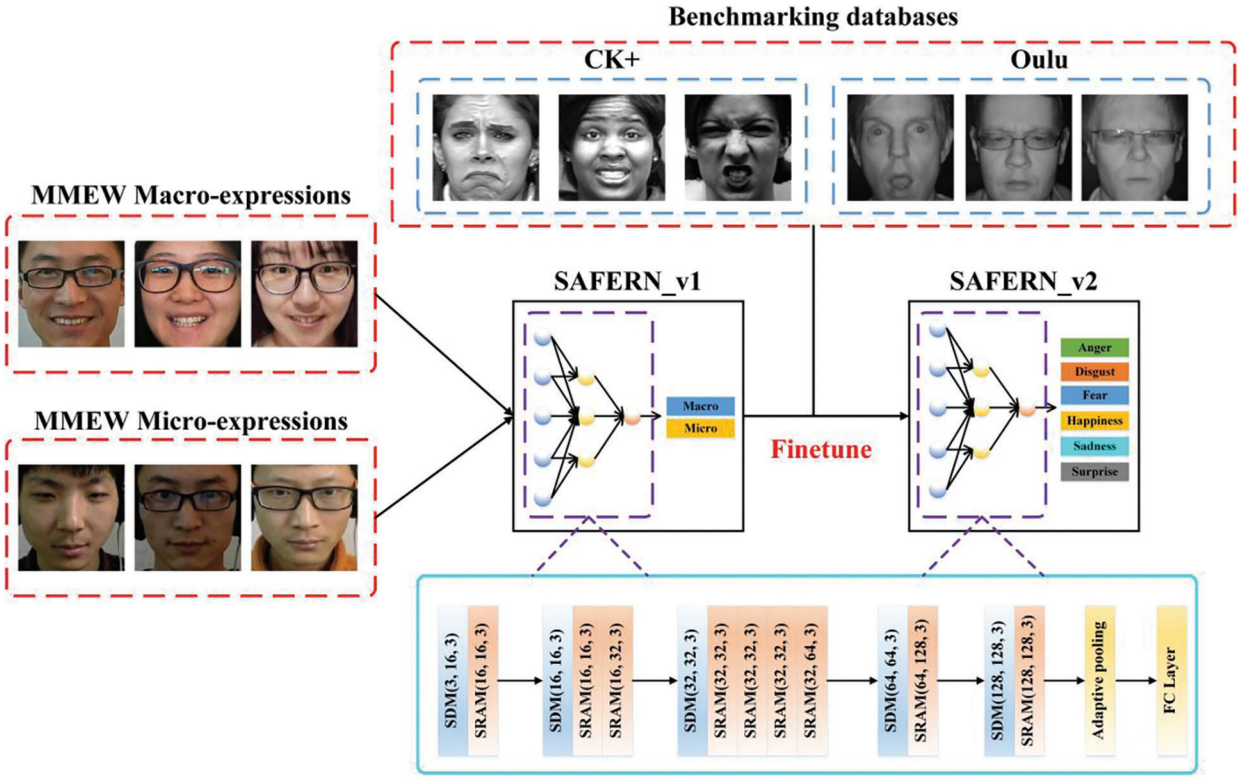


Figure 2: Pipeline to extract frame-based FER-related features. The proposed SAFERN is first pre-trained to distinguish micro-and macro-expressions. Afterward, it is migrated into FER tasks by fine-tuning on CK+ and Oulu-CASIA

3.3 Extractions of Sequence-Based FER-Related Features

Fig. 3 illustrates the pipeline to extract sequence-based FER-related features. To process the sequence features output by the SAFERN_v2, we introduce a shallow transformer network, which repeatedly stacks three transformer blocks. As shown in the figure, the sequence features $\{X_i\}$ are firstly combined with position embedding, which is defined as follows:

$$\tilde{X}_i = PE_{(t,d)} + X_i \tag{2}$$

The position embedding $PE_{(t,d)}$ is used to capture the sequential properties of input sequence features, which is defined as follows:

$$PE_{(t,d)} = \begin{cases} \sin(t/10000^{d/d_{\text{madel}}}) & \text{when } d \text{ is even} \\ \cos(t/10000^{d/d_{\text{mace}}}) & \text{when } d \text{ is odd} \end{cases} \tag{3}$$

where d represents the dimension of the sequence feature at time step t.

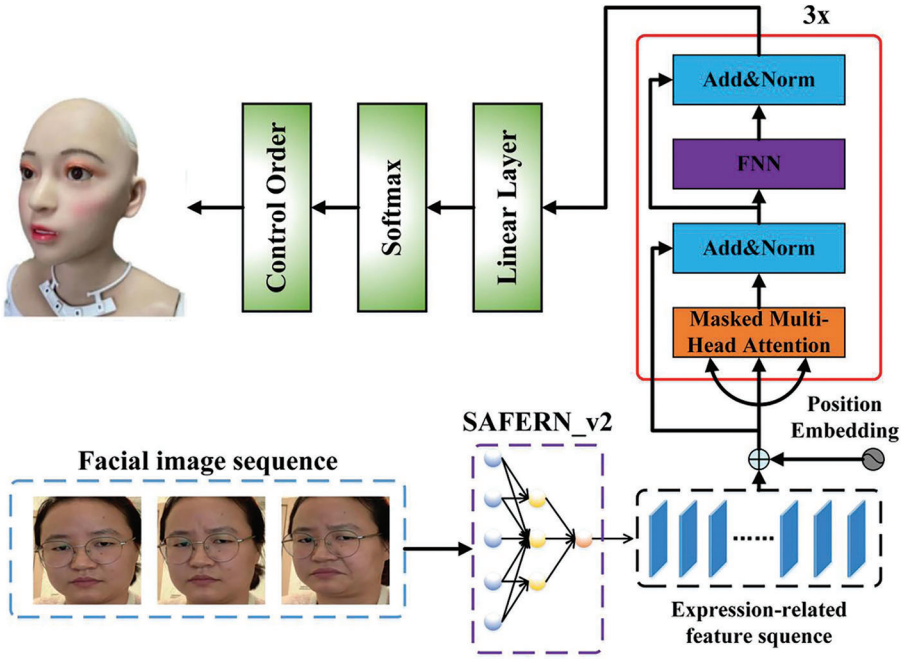


Figure 3: Pipeline to extract sequence-based FER-related features. A shallow transformer network is used to process the sequence features output by the SAFERN_v2. Afterward, a linear layer, followed by a softmax activation, is used to convert the transformer's output into the control order sent to the humanoid robot

Then, the combined inputs are fed into the masked multi-head attention module to explore the correlations between different feature sequences. Specifically, we define three linear transformations, including the query (Q), key (K), and value (V), are calculated as follows:

$$Q = W_Q \tilde{X}_{\mathrm{i}} \tag{4}$$

$$K = W_K \tilde{X}_i \tag{5}$$

$$V = W_V \tilde{X}_{\mathrm{i}} \tag{6}$$

where $\mathrm{W_Q}$, $\mathrm{W_K}$, and $\mathrm{W_V}$ are learnable weights. Afterward, the self-attention calculation can be given as follows:

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

The single-headed self-attention layer exists a constraint on the specific location attention. To improve its performance, the multi-headed attention mechanism forces different sub-regions to focus on multiple specific locations. The definition of the multi-headed attention mechanism with head number N is as follows:

$$MultiHead(\hat{Q},\ \hat{K},\ \hat{V}) = Concat(head_1,\ \ldots\ldots,\ head_N)W^{MH} \tag{8}$$

where $\hat{Q}$, $\hat{K}$, and $\hat{V}$ denote sets of $\{Q_i\}_{i=1}^N$, $\{K_i\}_{i=1}^N$, and $\{V_i\}_{i=1}^N$, respectively. $\mathrm{W^{MH}}$ is a linear projection matrix to calculate the multi-headed attention.

Outputs of the multi-headed attention mechanism are fed into the add&norm, FNN, and add&norm sequentially. Residual connections are shown in the figure. Finally, the output of the transformer network

is fed into the linear layer, followed by a softmax activation to generate a control order that will be sent to the humanoid robot. Such an order is given based on users' stable FER results.

3.4 Hardware Design

The robot head platform equips 12 motors to control the movement of 14 degrees of freedom (DoF). The entire platform can be further divided into the skull, facial motion, and neck motion modules. The skull module is generated by 3d printing technology, with a microprocessor and a motor servo system installed inside, and the skeleton tightly fits with the soft skin. The facial motion module involves the movements of the eye, eyebrow, eyelid, and cheek areas. It pulls the corresponding nylon rope with the corresponding motor to produce different movements. The neck module achieves six DoF rotations of the head, including front flexion and rear extension, left and right rotation, and left and right swing, through the coordinated control of three motors.

As shown in Fig. 1, the robot head platform is driven by a motor servo module controlled by the STM32F103C8T6 microprocessor. The microprocessor and the motor communicate in a question-and-answer manner. Specifically, the controller issues the instruction package, and the steering machine returns to the response package. Multiple motors are allowed in the bus topological control network, each assigned a unique ID number. Given the user's emotion label, the motor control command code is generated based on the expression tag according to the communication instruction package, which is sent to the microprocessor through the serial port. The sent control commands include the motor's ID number, position, and speed. The motor position controls the amplitude of the robot's facial expression movement and neck movement. The motor speed determines the speed of the robot's movement.

4 Experimental Results

4.1 Databases

MMEW [50]**:** This database is collected from 30 Chinese subjects at 90 fps. It provides both macro and micro-expressions. Hence, it is used to preliminary train the proposed SAFERN. We use the middle frame of each sequence as the apex frame for micro-expressions because it has no annotation information. For macro-expressions, we select one expression from each subject. Notably, this database is only used for pre-training.

CK+ [51]**:** This database consists of 593 sequences with seven expressions (happiness, surprise, sadness, fear, disgust, anger, and neutral), sampled from 123 male and female subjects. In the evaluations, six basic expressions (except the neutral) are used, and the last frame of each sequence is chosen as the peak expression. Therefore, roughly 50 to 100 samples are chosen for each expression.

Oulu-CASIA [52]**:** This database consists of 10,800 labeled samples sampled from 80 subjects. Six basic expressions are used in the evaluations. For each expression of each subject, the middle and last frames are chosen. Therefore, there are 160 samples for each expression.

Fig. 4 illustrates Some examples selected from (a) CK+ and (b) Oulu-CASIA databases.

4.2 Implementation Details

Each facial image is re-sized to 168×168 pixels and then is normalized to (0, 1). We randomly crop 148×148 patches and flip them to perform data augmentation to handle the over-fitting issue. We optimize the network with the Adam approach with a learning rate, beta1, and beta2 of 0.001, 0.9, 0.999, respectively. The learning rate is reduced to half for every 100 epochs until the total 300 epochs. A ten-cross validation strategy is used to evaluate the frame-based FER approach on benchmarking databases. The proposed network is built with the Pytorch framework and is trained with an Intel I7 CPU and an NVIDIA GTX-3080 GPU.

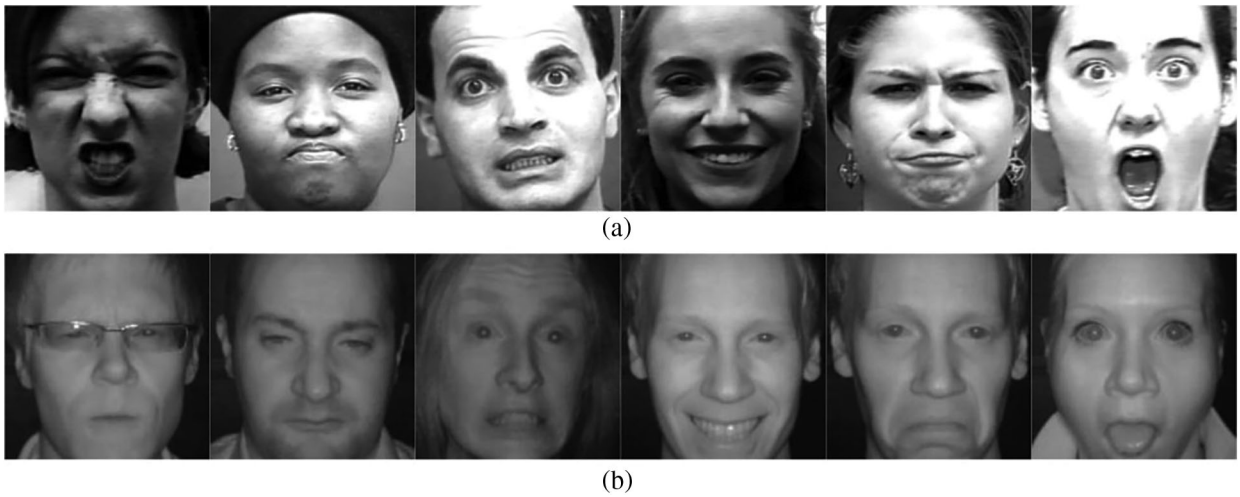


Figure 4: Some examples selected from (a) CK+ and (b) Oulu-CASIA databases. Expressions from left to right are anger, disgust, fear, happiness, sadness, and surprise, respectively

4.3 Evaluation Metric

The frame-based FER approaches are evaluated using accuracy because the used databases are almost balanced. The accuracy is defined as follows:

$$ACC_k = \frac{TP_k}{N_k} \tag{9}$$

where k denotes the category, TP_k is the number of TP (true positive) belonging to class k, and N_k is the number of samples for class k.

The sequence-based FER approaches are evaluated using temporal accuracy. We segment the captured facial image sequence into temporal slices with a sliding window. Each temporal slice contains 25 frames. The used sequential processing method will output one recognition result for each slice. Temporal accuracy is defined as the ratio between the correctly recognized slices and the total slices.

4.4 Ablation Studies

The key to recognizing a user's bad mood is accurate FER. To verify the effectiveness of the proposed method, ablation studies of the frame-based FER approach are performed on CK+ and Oulu-CASIA. Specifically, SAFERN_v2 without attention enhancement is used as a baseline in the evaluations, then the attention mechanism and the two-stage training strategy are added to the baseline, respectively. All methods are trained with the same setting for fair comparisons. As shown in Tab. 1, due to the use of attention enhancement which can better explore expression-related features, the proposed method achieves a tiny increase in accuracy compared with the baseline. The two-stage training strategy can also improve the recognition accuracy compared with the baseline. Using both strategies, the proposed method enhances the baseline accuracy by 4.86 and 3.82 on CK+ and Oulu-CASIA, respectively.

4.5 Quantitative Evaluations of SAFERN

In this section, we mainly evaluate the proposed SAFERN. We compare the proposed baseline with other commonly used backbones, such as VGG16, DenseNet, and Resnet18. We also compare the proposed method with several recent works on CK+ and Oulu-CASIA.

Table 1: Ablation study results of the frame-based FER approach

Attention enhancement	Two-stage training	CK+	Oulu-CASIA
		91.26	80.68
√		92.28	81.42
	√	94.22	83.13
√	√	**96.12**	**84.50**

Tab. 2 reports the comparison results between the proposed baseline and other widely used baselines, including VGG16, Resnet18, and DenseNet. As shown in the table, VGG16 achieves 93.12 and 82.69 on CK+ and Oulu-CASIA. DenseNet is superior to VGG16 because the used dense connection is beneficial for extracting multi-scale features. As a widely used backbone, Resnet18 outperforms DenseNet due to the used skip connection, restraining the vanishing gradient. Our SAFERN is a variant of Resnet by concentrating more on middle-scale features because these features are more discriminative to recognize facial macro-expressions. Therefore, ours achieves the best performance on both databases.

Table 2: Comparison results of different backbones

Backbone	CK+	Oulu-CASIA
VGG16	93.12	82.69
Resnet18	95.88	84.24
DenseNet	94.26	83.37
Ours	**96.12**	**84.50**

Besides, we compare the proposed method with several recent works on CK+ and Oulu-CASIA. For the CK+ database, we compare the SAFERN_v2 with Decoder Regional Adaptive Affinitive Patterns (DRADAP) [53], center loss [54], Inception [55], CNN with the Island Loss (IL-CNN) [54], Identity-aware CNN (IACNN) [56], Deep Locality-preserving CNN (DLP-CNN) [57]. As shown in Tab. 3, SAFERN_v2 achieves the best performance compared with other results, which are reported in their original works. Although tiny differences exist in the data splitting strategy, the comparisons still reveal our superiority in recognizing facial macro-expressions. DLP-CNN introduced a deep locality-preserving learning strategy, exploring details in local facial regions. Therefore, it achieves the second-best performance. SAFERN_v2 introduces the attention enhancement and two-stage training strategy. Therefore, it is good at mining macro-expression-related features, leading to high recognition accuracy.

Table 3: Comparison results between the proposed method and the recent works on the CK+ database

Methods	Strategy	Acc.
DRADAP [53]	10 folds	90.63
center loss [54]	10 folds	92.26
Inception [55]	5 folds	93.20
IL-CNN [54]	10 folds	94.35
IACNN [56]	8 folds	95.37
DLP-CNN [57]	5 folds	95.78
SAFERN_v2	10 folds	**96.12**

For the Oulu-CASIA database, we compare the SAFERN_v2 with AdaBoost LBP (AdaLBP) [58], Spatio-temporal manifold expressionlet (STM-ExpLet) [59], center loss [54], IL-CNN [54], GoogLeNet [60], and Deep temporal appearance-geometry network (DTAGN) [61]. All methods use the 10-fold cross-validation strategy. As shown in the Tab. 4, SAFERN_v2 achieves the best recognition performance. It has a gain of 3.04 over the second-best DTAGN [61], which used joint fine-tuning in the DNNs. The accuracy of other methods is all less than 80. The comparison results indicate the superiority of SAFERN_v2 in performing frame-based FER.

Table 4: Comparison results between the proposed method and the recent works on the Oulu-CASIA database

Methods	Strategy	Acc.
AdaLBP [58]	10 folds	73.54
STM-ExpLet [59]	10 folds	74.59
center loss [54]	10 folds	75.63
IL-CNN [54]	10 folds	77.29
GoogLeNet [60]	10 folds	79.21
DTAGN [61]	10 folds	81.46
SAFERN_v2	10 folds	**84.50**

4.6 Quantitative Evaluations of Shallow Transformer Network

Except for the high accuracy frame-based FER, sequence-based FER is also critical to generate a stable recognition result given a consecutive facial image sequence. To evaluate the shallow transformer network, we compare it with two sequential processing methods, including the temporal voting strategy and LSTM, on practically captured facial image sequences. The temporal voting strategy outputs the recognition result with the most votes in a given sliding window. Similar to the transformer network, LSTM takes the sequence features in the sliding window as input and outputs the recognition result. We sample 528 practical facial image sequences with labels from 23 subjects to evaluate different sequential processing methods. Each sequence lasts for six seconds. Tab. 5 reports the comparison results in which our method achieves the best recognition performance due to the transformer's ability to handle long-term sequences. It leads to the second-best method, the temporal voting method, by 7.54. The temporal voting strategy is superior to LSTM because the latter may tend to sequence features at later steps, therefore ignoring the global dependency in the user's dynamic expression changes.

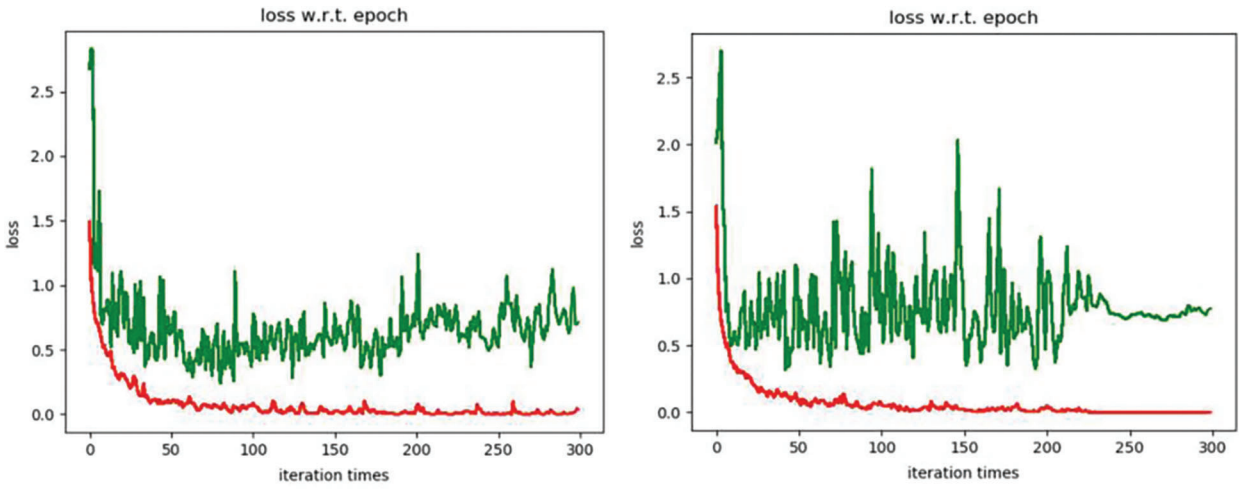
Table 5: Comparison results of different sequential processing methods

Method	Temporal accuracy
Temporal voting	71.32
LSTM	66.82
Transformer	**78.86**

4.7 Qualitative Evaluations

This section mainly illustrates the recognition performance of the proposed frame-based FER approach and shows the robot's responses to different facial expressions. Fig. 5 presents the average training and validation loss curves on CK+ and Oulu-CASIA. Both training losses decay to zero, and validation losses

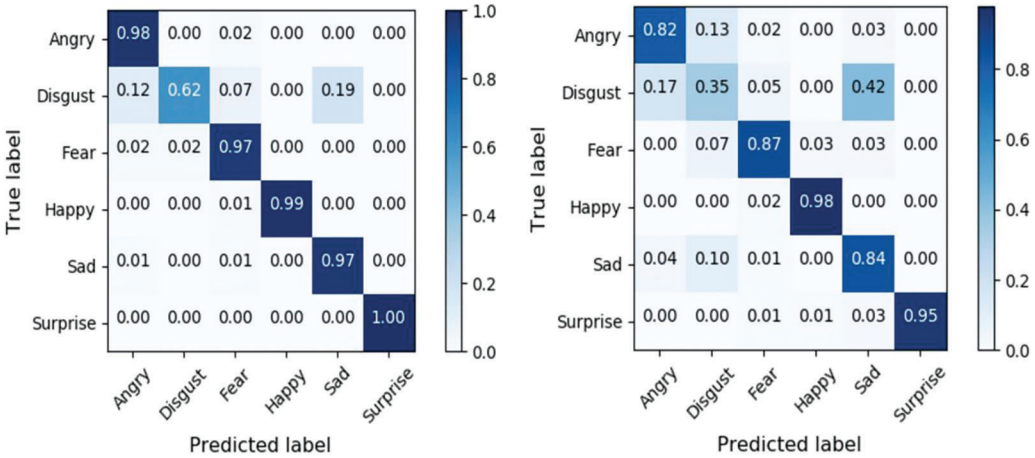
also show a decreasing tendency. Such loss curves indicate that the proposed method has a good generalization ability.



(a) CK+ (b) Oulu-CASIA

Figure 5: The average training and validation loss curves on (a) CK+ and (b) Oulu-CASIA

Fig. 6 shows the confusion matrix of the proposed method on CK+ and Oulu-CASIA. The confusion matrix provides the average ten recognition results because we use ten-cross validation as the training protocol. Although the proposed method achieves satisfactory recognition performance on these two databases, it performs differently in recognizing various expressions. For both databases, it fails to recognize the disgust expression with high accuracy. It easily mistakes disgust expression as sad. As shown in Fig. 4, disgust and sadness expressions present lots of similarities in facial images. For example, their eyebrows and eyes are in a state of tightening and their mouths are zipped. It is difficult to discriminate these facial expressions even for humans. Such a result reveals that we need more powerful discriminative features to distinguish several similar but different expressions.



(a) CK+ (b) Oulu-CASIA

Figure 6: The confusion matrix of the proposed method on (a) CK+ and (b) Oulu-CASIA

In Fig. 7, we provide some practical cases of our robot's responses to different facial expressions. In this design, the robot will try to make smiles or other positive expressions to conciliate users when it detects negative facial expressions from them. The first row shows the different facial expressions generated by our robot. The second row illustrates the robot's responses to recognized negative facial expressions. The robot can respond correctly to users' negative expressions, further generating positive expressions to comfort users. Such success is due to the proposed method that can output a stable and accurate expression recognition result.

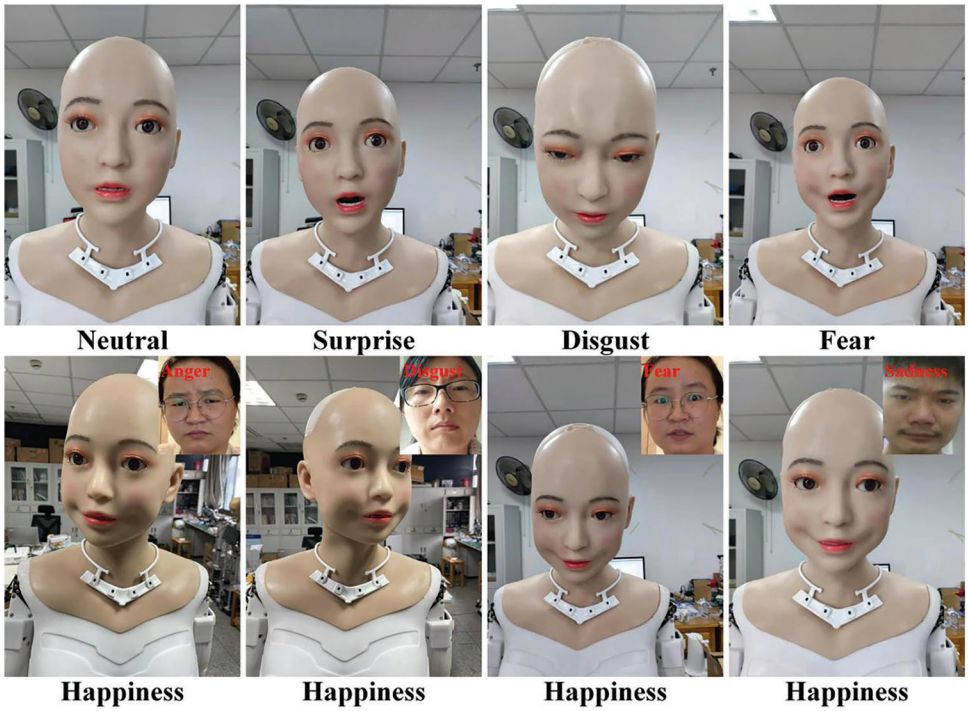


Figure 7: Illustrations of robot's generated expressions and their responses to recognized negative facial expressions, including anger, disgust, fear, and sadness

5 Conclusions and Discussions

A humanoid robot equipped with a modern FER approach is presented. It can focus on users' negative emotions and make smile expressions as a response. In the software part, SAFERN is proposed to perform frame-based FER. A two-stage training is used to improve the recognition performance on macro-expressions by distinguishing macro and micro-expressions. Further, a shallow transformer is introduced to process the facial sequence data to output a stable recognition result. Evaluations on CK+ and Oulu-CASIA indicate that the proposed method has achieved comparative performance compared with recent works. In the hardware part, the robot can generate a smile expression as a response when it detects negative emotions revealed by the users. Therefore, the robot can provide a more comfortable companion by always paying attention to users' bad moods. In particular, the proposed negative emotions sensitive system can be further used in homes for the elderly who are restricted in movements or live alone, to alleviate their negative emotions by emotional interaction with this robot.

In this work, the facial expressions of the robot are generated by hard coding, and the flexibility needs to be improved. In the future, we will delve into robot facial expression generation, so that the robot can make more realistic facial expression.

Acknowledgement: Thanks are due to Dr. Song for guidance in writing standard.

Funding Statement: This work was supported in part by National key R&D program of China 2018AAA0100800, the Key Research and Development Program of Jiangsu under grants BK20192004B and BE2018004-04, Guangdong Forestry Science and Technology Innovation Project under Grant 2020KJCX005, International Cooperation and Exchanges of Changzhou under Grant CZ20200035.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Melkas, L. Hennala, S. Pekkarinen and V. Kyrki, "Impacts of robot implementation on care personnel and clients in elderly-care institutions –ScienceDirect." *International Journal of Medical Informatics*, vol. 134, 2020.

[2] T. Kazue, K. Takahiro and S. Takanori, "Comparison of verbal and emotional responses of elderly people with mild/moderate dementia and those with severe dementia in responses to seal robot, PARO," *Frontiers in Aging Neuroscience*, vol. 6, pp. 257, 2014.

[3] H. Jiang, R. Jiao, D. Wu and W. Wu, "Emotion analysis: Bimodal fusion of facial expressions and eeg," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 2315–2327, 2021.

[4] X. R. Zhang, T. Xu, W. Sun and A. G. Song, "Multiple source domain adaptation in micro-expression recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8371–8386, 2021.

[5] Mustaqeem and S. Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated cnn features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[6] C. F. Benitez-Quiroz, R. B. Wilbur and A. M. Martinez, "The not face: A grammaticalization of facial expressions of emotion," *Cognition*, vol. 150, pp. 77–84, 2016.

[7] S. Jain, C. Hu and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *2011 IEEE Int. Conf. on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, pp. 1642–1649, 2011.

[8] D. Ghimire, J. Lee, Z. N. Li and S. H. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7921–7946, 2017.

[9] T. Huynh, R. Min and J. L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *2012 Asian Conf. on Computer Vision*, Berlin, Heidelberg, pp. 133–145, 2012.

[10] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan and S. Y. Lee, "Depth camera-based facial expression recognition system using multilayer scheme," *IETE Technical Review*, vol. 31, no. 4, pp. 277–286, 2014.

[11] X. H. Huang, S. J. Wang, X. Liu, G. Y. Zhao, X. Y. Feng *et al.,* "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 32–47, 2017.

[12] Y. Sun and J. Yu, "Facial expression recognition by fusing gabor and local binary pattern features," in *Int. Conf. on Multimedia Modeling*, Reykjavik, Iceland, pp. 209–220, 2017.

[13] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh and F. C. Jeffrey, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[14] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779–788, 2016.

[16] B. Yang, J. M. Cao, N. Wang and X. F. Liu, "Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 4, pp. 473–482, 2018.

[17] Y. F. Cai, L. Dai, H. Wang, L. Chen and Y. C. Li, "DLnet with training task conversion stream for precise semantic segmentation in actual traffic scene," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[18] B. Yang, G. C. Yan, P. Wang, C. Y. Chan, X. Song *et al.,* "A novel graph-based trajectory predictor with pseudo-oracle," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[19] X. Zhang, X. Sun, W. Sun, T. Xu, P. Wang *et al.,* "Deformation expression of soft tissue based on bp neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1041–1053, 2022.

[20] B. Yang, W. Q. Zhan, P. Wang, C. Y. Chan, Y. F. Cai *et al.,* "Crossing or Not? context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[21] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. W. Yan *et al.,* "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[22] M. Wu, W. Su, L. Chen, Z. Liu, W. H. Cao *et al.,* "Weight-adapted convolution neural network for facial expression recognition in human-robot interaction," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 3, pp. 1473–1484, 2021.

[23] K. Prabhu, S. SathishKumar, M. Sivachitra, S. Dineshkumar and P. Sathiyabama, "Facial expression recognition using enhanced convolution neural network with attention mechanism," *Computer Systems Science and Engineering*, vol. 41, no. 1, pp. 415–426, 2022.

[24] B. Yang, J. M. Cao, D. P. Jiang and J. D. Lv, "Facial expression recognition based on dual-feature fusion and improved random forest classifier," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 20477–20499, 2018.

[25] M. Alam, L. S. Vidyaratne and K. M. Iftekharuddin, "Sparse simultaneous recurrent deep learning for robust facial expression recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4905–4916, 2018.

[26] A. Fathallah, L. Abdi and A. Douik, "Facial expression recognition via deep learning," in *2017 IEEE/ACS 14th Int. Conf. on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia, pp. 745–750, 2017.

[27] F. Z. Salmam, A. Madani and M. Kissi, "Fusing multi-stream deep neural networks for facial expression recognition," *Signal, Image and Video Processing*, vol. 13, no. 3, pp. 609–616, 2019.

[28] B. Yang, J. M. Cao, R. R. Ni and Y. Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.

[29] H. Li, J. Sun, Xu Z. and L. M. Chen, "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[30] A. Jan, H. Ding, H. Meng, L. Chen and H. Li, "Accurate facial parts localization and deep learning for 3D facial expression recognition," in *2018 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, pp. 466–472, 2018.

[31] P. Barros, N. Churamani and A. Sciutti, "The FaceChannel: A light-weight deep neural network for facial expression recognition," in *2020 15th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2020)*, Buenos Aires, Argentina, pp. 652–656, 2020.

[32] F. F. Zhang, T. Z. Zhang, Q. Mao and C. S. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, SALT LAKE CITY, USA, pp. 3359–3368, 2018.

[33] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing *et al.,* "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, no. 1, pp. 303–311, 2020.

[34] J. Lee, S. Kim, S. Y. Kim and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 6977–6991, 2020.

[35] Q. Li, J. Yu, T. Kurihara, H. Zhang and S. Zhan, "Deep convolutional neural network with optical flow for facial micro-expression recognition," *Journal of Circuits, Systems and Computers*, vol. 29, no. 1, pp. 1–7, 2020.

[36] S. Song, E. Sanchez, L. Shen and M. Valstar, "Self-supervised learning of dynamic representations for static images," in *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, Milan, Italy, pp. 1619–1626, 2021.

[37] J. Chen, Z. H. Chen, Z. Chi and F. Hong, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, 2016.

[38] K. H. Zhang, Y. Z. Huang, Y. Du and W. Liang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[39] H. Li, J. Sun, Z. Xu and L. Chen, "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[40] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Honolulu, USA, pp. 30–40, 2017.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares *et al.,* "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[43] M. Yu, H. Zheng, Z. Peng, J. Dong and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognition Letters*, vol. 131, pp. 166–171, 2020.

[44] K. Kang and X. Ma, "Convolutional gate recurrent unit for video facial expression recognition in the wild," in *2019 Chinese Control Conf. (CCC)*, Guangzhou, China, pp. 7623–7628, 2019.

[45] K. Han, A. Xiao, E. Wu, J. Y. Guo, C. J. Xu *et al.,* "Transformer in transformer," arXiv preprint arXiv:2103.00112, 2021.

[46] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[48] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel *et al.,* "Bottleneck transformers for visual recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 16519–16529, 2021.

[49] H. W. Kim, H. J. Kim, S. Rho and E. Hwang, "Augmented EMTCNN: A fast and accurate facial landmark detection network," *Applied Sciences*, vol. 10, no. 7, pp. 2253, 2020.

[50] X. Ben, Y. Ren, J. Zhang, S. J. Wang, K. Kpalma *et al.,* "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[51] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar *et al.,* "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, USA, pp. 94–101, 2010.

[52] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. Third IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 200–205, 1998.

[53] M. Mandal, M. Verma, S. Mathur, S. K. Vipparthi, S. Murala *et al.,* "Regional adaptive affinitive patterns (RADAP) with logical operators for facial expression recognition," *IET Image Processing*, vol. 13, no. 5, pp. 850–861, 2019.

[54] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly *et al.,* "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, pp. 302–309, 2018.

[55] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, pp. 1–10, 2016.

[56] Z. Meng, P. Liu, J. Cai, S. Han and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, pp. 558–565, 2017.

[57] S. Li, W. Deng and J. P. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2852–2861, 2017.

[58] G. Zhao, X. Huang, M. Taini, S. Z. Li and M. Pietikalnen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[59] M. Liu, S. Shan, R. Wang and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1749–1756, 2014.

[60] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han *et al.,* "Peak-piloted deep network for facial expression recognition," in *European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 425–442, 2016.

[61] H. Jung, S. Lee, J. Yim, S. Park and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 2983–2991, 2015.