Dual attention and element recalibration networks for automatic depression level prediction

Mingyue Niu, Ziping Zhao, Jianhua Tao, Ya Li, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Niu, Mingyue, Ziping Zhao, Jianhua Tao, Ya Li, and Björn W. Schuller. 2023. "Dual attention and element recalibration networks for automatic depression level prediction." *IEEE Transactions on Affective Computing* 14 (3): 1954–65. https://doi.org/10.1109/taffc.2022.3177737.

Dual Attention and Element Recalibration Networks for Automatic Depression Level Prediction

Mingyue Niu, Ziping Zhao, *Member, IEEE*, Jianhua Tao, *Senior Member, IEEE*, Ya Li, *Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

Abstract—Physiological studies have identified that facial dynamics can be considered as biomarkers to analyze depression severity. This paper accordingly develops a Dual Attention and Element Recalibration (DAER) network to extract facial changes to predict the depression level. In this model, we propose two blocks: a Dual Attention (DA) block and Element Recalibration (ER) block. The DA block uses the self-attention to investigate the dynamic changes in the representation sequence of a facial video segment. It further examines the influence of feature components of the representation sequence on depression level prediction through bilinear-attention. Moreover, to improve the representation ability of network, the ER block is used to obtain the global information to recalibrate each element of the tensor. Adopting this approach, for the depression level prediction task, we first divide the long-term video into fixed-length segments and use the trained ResNet50 to encode each frame to generate the representation sequences of video segments. Second, the representation sequences are input into DAER network to obtain the depression level scores. Finally, the average of these scores yields the prediction result corresponding to the long-term video. Experiments on publicly available AVEC 2013 and AVEC 2014 depression databases illustrate the effectiveness of our method.

Index Terms—DAER network, depression level prediction, dual attention block, element recalibration block, facial differences

1 Introduction

As a psychological disease, major depressive disorder causes people to fall into a state of low mood for a long period, resulting in their the inability to participate normally in social life. In addition, depression can reduce self-awareness and even lead to self mutilation or suicide [1]. Data released by the World Health Organization in 2017 shows that around 350 million people globally suffer from depression, with depression predicted to become the second leading cause of death in 2030 [2]. Unfortunately, the process of diagnosing depression is often both laborious and primarily dependent on doctors' clinical experience, which results in many patients being unable to recognize their physiological changes and access timely treatment. It is therefore imperative to develop an automatic depression diagnosis system to assist doctors in improving the current medical conditions.

Related physiological studies [3], [4] reveal differences in facial activities between depressed and healthy individuals. In other words, the pattern of facial activities can be considered as a biomarker for use in analyzing an individual's level of depression, which can be measured using the Beck Depression Inventory-II (BDI-II) score [5] as shown in Table 1. Based on this information, many researchers [6], [7], [8], [9], [10], [11], [12] have proposed different methods for extracting the representations of depression cues in order to predict the BDI-II score. However, in the works of [6], [7], the authors only examine the spatial structure of the facial image while ignoring the facial changes in the video. Moreover, the methods in [8], [9] use the Three Orthogonal Plane (TOP) framework proposed in [13] to calculate the feature histogram of the facial video clip, which is used to obtain the corresponding spatiotemporal representation. Notably, however, the feature histogram is not sensitive to the salient features [14]. Furthermore, hand-crafted features rely heavily on the experience of designers, which can lead to the loss of some useful information. To this end, Shang [15] employ the Convolution Neural Network (CNN) to process the facial images, which is

- *Mingyue Niu and Ziping Zhao are with the School of Computer and Information Engineering, Tianjin Normal University (TJNU), Tianjin 300387, China. E-mail: mniu@tjnu.edu.cn, ztianjin@126.com.*
- *Jianhua Tao is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automatic Chinese Academy of Sciences (CASIA), Bejing 100190, China, with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the CAS center for Excellence in Brain Science and Intelligence Technology, Bejing 100190, China. E-mail: jhtao@nlpr.ia.ac.cn.*
- *Ya Li is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: yli01@bupt.edu.cn.*
- *Björn W. Schuller is with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with GLAM – Group on Language, Audio, & Music, Imperial College London, SW7 2BX London, U.K. E-mail: bjoern.schuller@imperial.ac.uk.*

This work was supported in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202200012, in part by the National Natural Science Foundation of China under Grant 62071330, and in part by the New Talent Project of Beijing University of Posts and Telecommunications under Grant 2021RC37.
(Corresponding authors: Jianhua Tao and Ziping Zhao.)
Digital Object Identifier no. 10.1109/TAFFC.2022.3177737

TABLE 1
BDI-II Scores and Corresponding Depression Severity

BDI-II Score	Depression Degree
0-13	None
14-19	Mild
20-28	Moderate
29-63	Severe

not enough for examining facial dynamics. Zhu *et al.* [10] input facial images and optimal flows into a deep Convolution Neural Network (CNN) to investigate the influence of facial appearance and dynamics on depression level analysis. Similarly, the 3D CNN and attention mechanism are adopted in [16] and [17] to extract spatiotemporal features for predicting the BDI-II score. However, these spatiotemporal feature extraction methods rarely consider the impact of feature components on depression level prediction. In addition, they fail to highlight the effective elements in the tensor, resulting in the annihilation of subtle information related to depression.

In recent years, numerous reports in the fields of speech recognition [18], [19] and video processing [20], [21] have shown that the self-attention mechanism is both feasible and advantageous for modeling the temporal sequence. Moreover, the experimental results presented in [22], [23], [24] confirm the fact that bilinear-attention is able to capture discriminative fine-grained features due to the extraction of second-order statistics. Meanwhile, the popular Squeeze-Excitation (SE) block [25] and its variants [26], [27] illustrate that networks' representation ability can be strengthened by embedding the global information into the feature channels. Inspired by these research achievements, and to alleviate the above issues in the field of automatic depression level prediction, we propose a Dual Attention and Element Recalibration (DAER) network to predict the BDI-II scores of individuals. In our approach, there are two main blocks: the Dual Attention (DA) block and the Elements Recalibration (ER) block. The DA block uses the self-attention and bilinear-attention, from the perspective of temporal changes and feature components, to investigate the facial dynamics of individuals with different depression levels and highlight the effective parts in the feature components. Furthermore, unlike recalibration feature channels in the SE block, the ER block is able to recalibrate each element of the tensor for capturing subtle depression cues. In more detail, our method comprises three steps:

First, we divide a long-term video into fixed-length segments and each frame in the video segment is encoded into a representation vector using the trained ResNet50 with Euclidean loss and "ReLU" rather than "Softmax". In this way, a video segment is encoded into a representation sequence and a long-term video corresponds to multiple representation sequences. Second, these representation sequences are input into the proposed DAER network for BDI-II score prediction. Third, we take the average of BDI-II scores obtained from those representation sequences as the prediction result corresponding to the long-term video. Experimental results on two publicly available Audio/Visual Emotion Challenge (AVEC) 2013 [28] and AVEC2014 [29] depression databases demonstrate the effectiveness of our method.

The main contributions of this paper can be summarized as follows:

1) In this paper, we propose a Dual Attention (DA) block. This block adopts self-attention to investigate the temporal differences in facial activities among individuals with different depression levels. Moreover, our proposed bilinear-attention treats each feature component in the representation sequence as a time series and calculates its second-order statistics to highlight depression-related parts.
2) To capture subtle depression cues, we additionally develop an Element Recalibration (ER) block. This block embeds the global information into the tensor to emphasize the elements that contribute to predicting depression levels.
3) We use the proposed DA and ER blocks to construct a novel deep architecture-i.e., the Dual Attention and Element Recalibration (DAER) network–in order to predict individual depression levels. Experiments are performed on two publicly available databases (i.e., AVEC 2013 and AVEC 2014 depression databases) to verify the contribution of each block and demonstrate the effectiveness of our method.

The remainder of this paper is organized as follows. In section 2, we review the works related to automatic video-based depression level analysis methods. In section 3, we provide a detailed description of the method proposed in this paper. Our experimental results and discussion are presented in section 4. Finally, section 5 concludes the paper.

2 Related Works

Physiological research has shown that facial changes can be used as a biomarker to predict individual depression severity. Many researchers have accordingly attempts to employ machine learning technologies to find the mapping relationship between the facial feature and depression score. Therefore, in this section, we briefly review prior works on this subject.

2.1 Automatic Depression Level Prediction With Hand-Crafted Features

As part of the AVEC challenges held in 2013 and 2014, Valstar *et al.* [28], [29] released two databases for depression level prediction and provided the baseline features. For AVEC2013, these authors used the Local Phase Quantization (LPQ) pattern to extract the feature of each frame, then calculated the mean of these features to yield the representation corresponding to the video in question. For AVEC2014, the baseline feature was the Local Gabor Binary Pattern (LGBP) feature from the XY-T image plane, which was used to characterize the facial changes. In [30], Dhall *et al.* extracted the Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) feature from the non-overlapping blocks obtained from the video, then used Fisher Vector (FV) encoding to aggregate these LPB-TOP features to the spatiotemporal representation of the video. Wen *et al.* [31] adopted LPQ-TOP to represent temporal changes in facial region sub-volumes, after which the sparse coding method and discriminative mapping were used to obtain the visual-based nonverbal behavior descriptor used

to predict the individual depression level. In [9], He *et al.* combined the Median Robust Local Binary Pattern with the TOP framework to obtain the MRLBP-TOP, which was used capture the spatiotemporal information of facial microstructure and macrostructure in the video segment. Furthermore, to automatically learn the number of mixtures in the Gaussian Mixture Model, these authors proposed a Dirichlet Process Fisher Vector encoding scheme to aggregate the features extracted from the video segments in order to obtain the video representation. Niu *et al.* [8] developed a novel local pattern named Local Second-Order Gradient Cross Pattern (LSOGCP) to extract the facial detailed texture and generated the LSOGCP-TOP to capture the subtle facial dynamics. Moreover, they adopted a hierarchical method of between-group classification and within-group regression to predict the BDI-II score.

Notably, the hand-crafted features used in above works rely heavily on the experience of the designers, meaning that some effective information related to depression is lost under these approaches. In addition, the TOP framework is essentially a histogram feature generation method and ignores the fact that the salient features are distributed unevenly in the temporal space [14].

2.2 Automatic Depression Level Prediction With Deep Neural Networks

In recent years, deep neural network-based methods have achieved good performance in the fields of image [32], [33] and video [34], [35] processing. Some researchers have accordingly constructed various models to extract the representations of depression cues for estimating depression severity from facial activities. Zhou *et al.* [7] presented a deep CNN equipped with a Global Average Pooling (GAP) layer, termed DepressNet, to process video frames. More specifically, the full face and three overlapping facial regions (top, central, and bottom) were input into DepressNet to yield the complementary information used to predict the depression score of each frame. The average of the scores corresponding to video frames was taken as the prediction result of the video. Similarly, the works of [15], [46] also predicted the individual BDI-II score through examining the different facial regions. Jan *et al.* [36] extracted the deep features from AlexNet and VGG-Face. Then, the Feature Dynamic History Histogram was constructed to characterize facial dynamics to predict BDI-II score.

Zhu *et al.* [10] designed a two-stream framework to capture both facial appearance and dynamics. In this framework, the face frames and facial optical flow images were input into two deep CNNs to extract the appearance representation and dynamic features. Moreover, the two deep CNNs were integrated by joint-tuning layers to predict individuals' BDI-II scores. Uddin *et al.* [37] also proposed a two-stream method for depression level estimation. These authors used the Inception-ResNet-v2 network [38] to extract facial spatial features, after which a combination of Temporal Median Pooling and Bidirectional Long Short-Term Memory was adopted to capture the spatiotemporal information. Moreover, the dynamic feature map generated by the Volume Local Directional Number pattern was processed by the CNN, TMP layer and Bi-LSTM in turn, yielding the high-level spatiotemporal feature and accordingly the prediction result. Furthermore, to directly generate the spatiotemporal feature, Jazaery [12] *et al.* used 3D CNN to process tightly aligned and loosely non-aligned face clips to estimate the BDI-II score. In [11], the authors considered that the full-face and eye regions were helpful in improving the prediction accuracy. They accordingly integrated the 3D-GAP layer into the 3D CNN to capture the depression cues from global and local facial regions. Similarly, He *et al.* [47] proposed a 3D NN equipped with a spatiotemporal feature aggregation module to characterize the depression cues in the video segment.

The methods discussed above extract the high-level representations of depression cues through the use of CNN-based network models. However, these methods rarely consider the influence of feature components on depression level prediction. In addition, failure to investigate elements of the tensor will also lead to the annihilation of subtle depression-related information.

In a departure from the above-mentioned works, we examine the facial changes and highlight the depression-related feature components in the representation sequence using the DA block. In addition, the ER block expands the concept of SE block and improves the network representation ability by recalibrating each element of the tensor. The experiments are conducted on AVEC 2013 and AVEC 2014 depression databases. The results demonstrate the effectiveness of our proposed method.

3 DAER Network Framework for Automatic Depression Level Prediction

Physiological studies [3], [4] have shown that facial activities can be reviewed and used as a biomarker to analyze individual depression levels. Accordingly, in this paper, we use the proposed DAER network to capture differences in facial dynamics among individuals with different depression levels. In our model, we first divide the long-term video into fixed-length segments and obtain the representation sequences of these segments via the trained ResNet50. Second, the prediction scores corresponding to these representation sequences are obtained via the proposed DAER network. Third, the average of these prediction scores is taken as the depression severity estimate corresponding to the long-term video. Fig. 1 illustrates the whole flow of our proposed framework.

3.1 Representation Sequence Generation of Video Segments Using the Trained ResNet50

Recently, CNN-based models have demonstrated strong representation ability and achieved good performance in a large number of computer vision tasks [32], [33], [34], [35]. Therefore, in order to represent the facial information, we use ResNet50 trained with the ImageNet dataset to extract the facial encoding vector. In addition, it is important to note that depression level prediction is essentially a regression problem; thus, we employ the Euclidean loss as the objective function and replace "Softmax" with "ReLU" in the process of training the ResNet50, as in [10], [12], [37].

In this paper, for the trained ResNet50, we input the face video frame and take the output of the "avg_pool" layer as

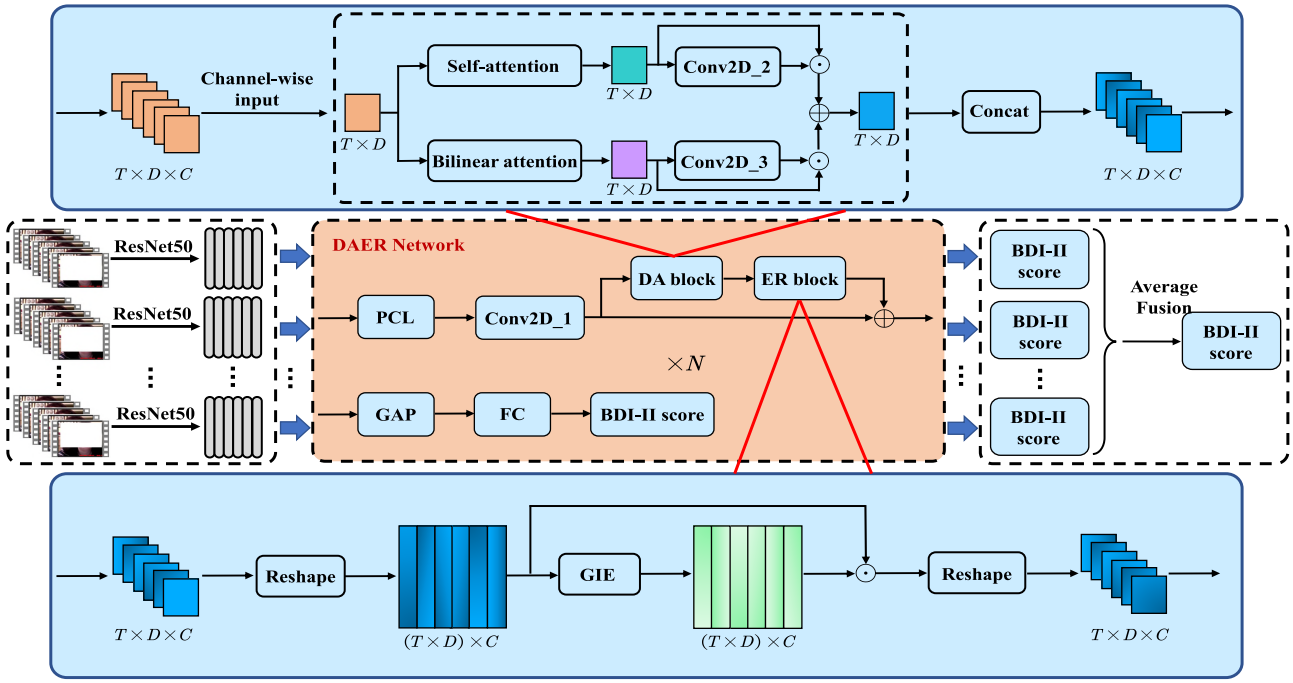


Fig. 1. The proposed Dual Attention and Element Recalibration (DAER) (as shown in the brown part) network is used for automatic depression level prediction. The "ResNet50" refers to a model that has been trained using the ImageNet dataset and is used to encode a video frame into a representation vector. Thus, a video segment can be encoded into a representation sequence, which is input into the DAER network to predict the BDI-II score. T, D and C denote the rows, columns and the number of channels of the input tensor, respectively. "$\oplus$" and "$\odot$" refer to matrix addition and element-wise multiplication. "$\times N$" means that the enclosed part is performed N times. The "PCL" refers to Progressive Conv1D Layers, used to reduce the dimension of representation sequence. "DA" and "ER" are the abbreviations for the Dual Attention block and Element Recalibration block, respectively. "GIE" refers to Global Information Extraction. "GAP" and "FC" refer to the Global Average Pooling and Full Connection. The loss function for the DAER network is Root Mean Square Error (RMSE), as shown in Eq. (12).

the frame encoding vector. Through this operation, a video segment containing T frames can be encoded into a representation sequence denoted as $\mathbf{S}' \in \mathbb{R}^{T\times D'}$, where D' is the dimension of the encoding vector. Furthermore, we adopt the progressive Conv1D layers to reduce the dimension of $\mathbf{S}' \in \mathbb{R}^{T\times D'}$ to $\mathbf{S} \in \mathbb{R}^{T\times D}$, where $D < D'$.

3.2 Dual Attention Block for Facial Dynamic Extraction

As Fig. 1 shown, the facial representation sequence $\mathbf{S}$ is input into the "Conv2D_1" layer. We denote the output result as $\mathbf{X} \in \mathbb{R}^{T\times D\times C}$, where C is the number of channels. It can be readily observed that each channel $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ $(i = 1, 2, \ldots, C)$ of $\mathbf{X}$ remains temporal. In view of this characteristic of $\mathbf{X}_i$, we propose a DA block to capture the differences of facial activities among individuals with different depression levels. More concretely, the DA block uses self-attention to investigate the temporal changes of the facial representation sequence. Moreover, a novel bilinear-attention is constructed to examine the influence of each feature component of the representation sequence on the depression level prediction.

3.2.1 Self-Attention for Facial Temporal Changes

Self-attention has certain advantages when it comes to process the sequence data in the fields of speech recognition [18], [19] and video processing [20], [21]. Inspired by these promising performance, we use the self-attention to capture the contextual temporal relations for depression level prediction. In this way, for a given sequence $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ $(i = 1, 2, \ldots, C)$, the self-attention can be expressed using Eq. (1). Note that, the $\mathrm{Softmax}(\cdot)$ is applied along each row of the matrix

$$\mathbf{Y}_i^{\mathrm{SA}} = \mathrm{Softmax}\left(\frac{\mathbf{K}_i\mathbf{Q}_i^{\mathsf{T}}}{\sqrt{D}}\right)\mathbf{V}_i \in \mathbb{R}^{T\times D},\ i = 1, 2, \ldots, C, \tag{1}$$

where "T" refers to matrix transposition. $\mathbf{K}_i, \mathbf{Q}_i, \mathbf{V}_i$ can be obtained through Eq. (2)

$$\begin{cases} \mathbf{K}_i = \mathcal{F}_K(\mathbf{X}_i) \in \mathbb{R}^{T\times D} \\ \mathbf{Q}_i = \mathcal{F}_Q(\mathbf{X}_i) \in \mathbb{R}^{T\times D} \\ \mathbf{V}_i = \mathcal{F}_V(\mathbf{X}_i) \in \mathbb{R}^{T\times D} \end{cases}. \tag{2}$$

In this paper, for each feature channel $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ $(i = 1, 2, \ldots, C)$, we adopt Conv2D layers to produce the $\mathbf{K}_i, \mathbf{Q}_i$, and $\mathbf{V}_i$ in Eq. (2). The temporal information is then extracted via Eq. (1); the result is denoted as $\mathbf{Y}_i^{\mathrm{SA}}$. Fig. 2 presents the process of capturing facial temporal changes through self-attention.

As described above, the self-attention takes each frame as a token and processes the representation sequence of the facial video segment from the perspective of temporal changes. Therefore, the self-attention is conductive to capture the facial dynamic patterns of individuals with different depression levels.

3.2.2 Bilinear-Attention for Examining the Importance of Feature Components

As noted in [22], [23], [24], the second-order statistics are helpful in boosting the model's representation ability. To

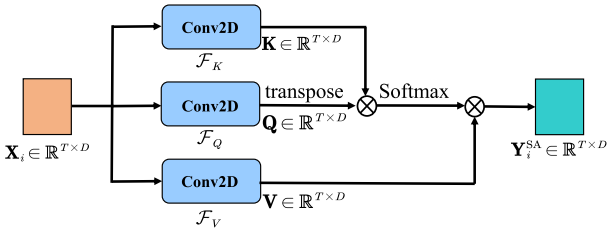


Fig. 2. The temporal information is extracted from each channel of tensor using the self-attention. $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ $(i = 1, 2, \ldots, C)$ is the i-the channel and $\mathbf{Y}_i^{\mathrm{SA}}$ is the corresponding result. "$\otimes$" denotes the matrix multiplication.

this end, in this paper, we propose a novel bilinear-attention method to examine the influence of each feature component on depression level prediction.

In more detail, for the ith channel $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ $(i = 1, 2, \ldots, C)$, we take each feature component as a time series and denote as $\mathbf{x}_i^d \in \mathbb{R}^{T\times 1}$ $(d = 1, 2, \ldots, D)$. The second-order statistics are then extracted via Eq. (3). Furthermore, the embedding $\mathbf{e}_i^d \in \mathbb{R}^{T\times 1}$ is generated using Eq. (4). The result of bilinear-attention can therefore be obtained through Eq. (5). The corresponding process is presented in Fig. 3

$$\mathbf{r}_i^d = \mathcal{C}_1\big(\mathbf{x}_i^d\big) \odot \big(\mathcal{C}_1\big(\mathbf{x}_i^d\big)\big)^{\mathsf{T}} \in \mathbb{R}^{T\times T}, \tag{3}$$

where $\mathbf{x}_i^d \in \mathbb{R}^{T\times 1}$ $(d = 1, 2, \ldots, D)$ is the dth column of $\mathbf{X}_i$, while "$\otimes$" denotes the matrix multiplication. $\mathcal{C}_1(\cdot)$ is a one-dimensional convolution (i.e., a Covn1D layer)

$$\mathbf{e}_i^d = \mathrm{Softmax}\big(\mathcal{C}_1\big(\mathbf{r}_i^d\big)\big) \in \mathbb{R}^{T\times 1}. \tag{4}$$

$$\mathbf{Y}_i^{\mathrm{BA}} = \mathbf{E}_i \odot \mathbf{X}_i \in \mathbb{R}^{T\times D}, \tag{5}$$

where $\mathbf{E}_i$ is constructed as shown in Eq. (6), while "$\odot$" denote element-wise multiplication

$$\mathbf{E}_i = \big[\mathbf{e}_i^1, \mathbf{e}_i^2, \ldots, \mathbf{e}_i^D\big] \in \mathbb{R}^{T\times D}. \tag{6}$$

From the above process, it can be readily observed that the second-order statistics are obtained through quadratic linear transformation on each feature component of the representation sequence in this paper. Moreover, we use the attention mechanism to embed these second-order statistics into the representation sequences in order to highlight those feature components related to depression in the representation sequence of a video segment.

In this way, on the one hand, we use the self-attention to extract the facial activities of individuals with different depression levels from the perspective of temporal changes. On the other hand, our proposed bilinear-attention is able to highlight the useful parts associated to depression from the perspective of feature components. To ensure that we make full use conferred by the advantages of these two attention mechanisms, we fuse their results with Eq. (7) to form the DA block for realizing the joint extraction of depression cues contained in the representation sequence

$$\mathbf{Y}_i^{\mathrm{DA}} = (\mathcal{C}_2(\mathbf{Y}_i^{\mathrm{SA}}) \odot \mathbf{Y}_i^{\mathrm{SA}}) \oplus (\mathcal{C}_2(\mathbf{Y}_i^{\mathrm{BA}}) \odot \mathbf{Y}_i^{\mathrm{BA}}) \in \mathbb{R}^{T\times D}, \tag{7}$$

where $\mathcal{C}_2(\cdot)$ is a two-dimensional convolution (i.e., a Covn2D layer). "$\odot$" and "$\oplus$" denote element-wise multiplication and element-wise addition, respectively.

3.3 Element Recalibration Block for Boosting Network Representation Ability

Recently, the successful application of SE block in [39], [40], [41] has shown that global information embedding can effectively improve networks' representation ability. Generally speaking, the SE block comprises two parts: global information extraction and adaptive recalibration. Mathematically, these two parts can be expressed as Eqs. (8) and (9), respectively

$$\mathbf{w}^{\mathrm{SE}} = \sigma(\mathbf{W}_2^{\mathrm{SE}}(\delta(\mathbf{W}_1^{\mathrm{SE}}(\mathcal{G}^{\mathrm{SE}}(\mathbf{X}))))) \in \mathbb{R}^{1\times C}, \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{T\times D\times C}$ is the input tensor and $\mathcal{G}^{\mathrm{SE}}(\cdot)$ is the GAP layer. $\mathbf{W}_1^{\mathrm{SE}} \in \mathbb{R}^{C\times \frac{C}{r_C}}$ and $\mathbf{W}_2^{\mathrm{SE}} \in \mathbb{R}^{\frac{C}{r_C}\times C}$ are two linear transformations. Here, r_C is the reduction rate, while $\delta(\cdot)$ and $\sigma(\cdot)$ refer to the "ReLU" and "Sigmoid" functions

$$\mathbf{R}_i^{\mathrm{SE}} = w_i^{\mathrm{SE}} \cdot \mathbf{X}_i \in \mathbb{R}^{T\times D},\ i = 1, 2, \ldots, C, \tag{9}$$

where w_i^{SE} is the ith element of $\mathbf{w}^{\mathrm{SE}}$ and $\mathbf{X}_i$ is the ith channel of $\mathbf{X}$.

From the above process, it can be clearly seen that the SE block is insufficient to recalibrate the elements of input tensor $\mathbf{X}$, because it treats all elements in one channel equally. To this end, we propose a novel block (i.e., ER block) to alleviate this limitation for further improving the representation ability of network.

In the proposed ER block, we reshape each feature channel into a column vector and calculate the autocorrelation matrix along the channel axis through Eq. (10). The corresponding result is considered as the global information of input tensor $\mathbf{X}$ and denoted as $\mathbf{I}^{\mathrm{G}}$. The elements in feature channels are then recalibrated using Eq. (11). Fig. 4 illustrates the complete process

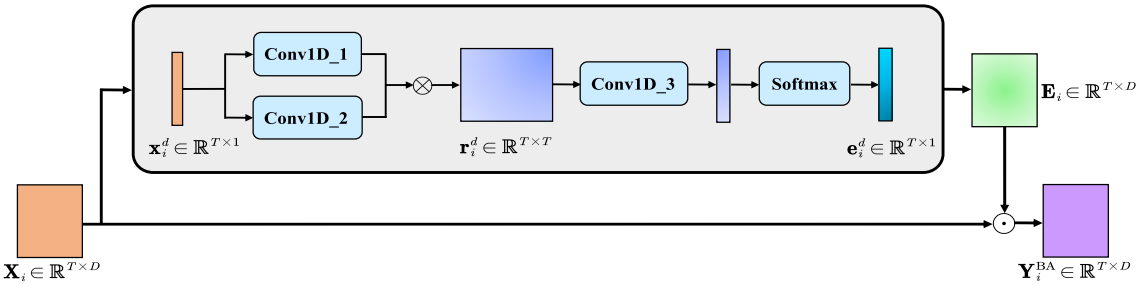


Fig. 3. Bilinear-attention for examining the influence of each feature component on depression level prediction. $\mathbf{X}_i$ $(i = 1, 2, \ldots, C) \in \mathbb{R}^{T\times D}$ is the ith channel. $\mathbf{x}_i^d \in \mathbb{R}^{T\times 1}(d = 1, 2, \ldots, D)$ is the time series composed of the dth feature component of $\mathbf{X}_i$. "$\otimes$" and "$\odot$" denote the matrix product and element-wise multiplication, respectively. $\mathbf{Y}_i^{\mathrm{BA}}$ $(i = 1, 2, \ldots, C) \in \mathbb{R}^{T\times D}$ is the corresponding result.

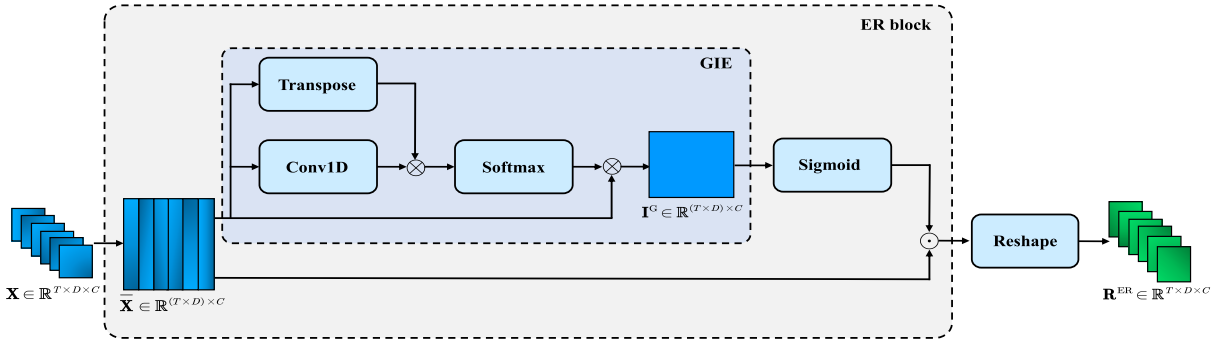


Fig. 4. The proposed ER block is used to improve the representation ability of network by recalibrating elements of input tensor. "$\otimes$" and "$\odot$" denote matrix and element-wise multiplication, respectively. $\bar{\mathbf{X}} \in \mathbb{R}^{(T\times D)\times C}$ is the matrix, each column of which is constructed by reshaping the feature channel $\mathbf{X}_i \in \mathbb{R}^{T\times D},\ (i = 1, 2, \ldots, C)$. $\mathbf{I}^{\mathrm{G}} \in \mathbb{R}^{(T\times D)\times C}$ and $\mathbf{R}^{\mathrm{ER}} \in \mathbb{R}^{T\times D\times C}$ are calculated by means of Eq. (10) and Eq. (11), respectively.

$$\mathbf{I}^{\mathrm{G}} = \mathrm{Softmax}\big(\bar{\mathbf{X}}\mathbf{W}^{\mathrm{ER}}\bar{\mathbf{X}}^{\mathsf{T}}\big) \otimes \bar{\mathbf{X}} \in \mathbb{R}^{(T\times D)\times C}, \tag{10}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{(T\times D)\times C}$ is the matrix, each column of which is constructed by reshaping the channel $\mathbf{X}_i (i = 1, 2, \ldots, C)$, while $\mathbf{W}^{\mathrm{ER}} \in \mathbb{R}^{C\times C}$ is the transformation matrix. "$\otimes$" represents matrix multiplication

$$\mathbf{R}^{\mathrm{ER}} = \mathrm{Reshape}\big(\bar{\mathbf{X}} \odot \mathbf{I}^{\mathrm{G}}\big) \in \mathbb{R}^{T\times D\times C}, \tag{11}$$

where "$\odot$" denotes element-wise multiplication operation.

According to the principle of convolution layer, each convolution kernel corresponds to a filtering result. From the workflow of our proposed ER block (as shown in Fig. 4), we can see that this block is able to comprehensively examine all the filtering results and select the elements that are helpful to the task of depression level prediction. Thus, the ER block has advantages in obtaining a more discriminative representation of depression cues.

The Relationship of ER Block and SE Block. From a macro perspective, our proposed ER block and SE block both mainly include the global information extraction and recalibration processes. From a micro perspective, the global information and recalibration targets of these two blocks are different. The SE block generates the global information through GAP a layer and fully connected layers, as shown in Eq. (8), and takes the channels as the recalibration target. For our proposed ER block, we calculate the autocorrelation matrix among channels using Eq. (10) and take the result as the global information. In other words, the global information we extract captures the high-order statistics among channels. In addition, the ER block is designed to recalibrate each element of the tensor rather than the entire channels; that is to say, the ER block highlights the elements of the tensor that are related to depression level prediction task and suppresses the less useful ones. Thus, our proposed recalibration strategy is advantageous in capturing subtle information related to depression.

4 Experiments

In order to demonstrate the effectiveness of our proposed method, we construct experiments on two publicly available databases namely AVEC 2013 and AVEC 2014 depression databases. In this section, these two benchmark databases are described, after which the implementation details of our DAER network are presented. Ablation studies are then performed to verify the role of each block. Finally, we compare our method with some state-of-the-art works and analyze the reasons.

4.1 Databases and Evaluation Metrics

In the AVEC competitions held in 2013 and 2014, two databases (i.e., AVEC 2013 and AVEC 2014) for depression level prediction are published and provide the raw videos. Thus, in this paper, all experiments are conducted on these two databases to verify the effectiveness of our proposed method.

The AVEC 2013 depression database is derived from the Audio-Visual Depression Language Corpus (AViD-Corpus), which is recorded using a webcam and microphone. In concrete terms, each subject is asked to perform 14 different tasks according to the instructions displayed on the computer screen. Those 14 tasks include sustained vowel phonation, sustained loud vowel phonation, sustained smiling vowel phonation, speaking out loud while solving a task, counting from 1 to 10, etc. The age distribution of the subjects ranges from 18 to 63 years old and the average age is 31.5 years old with a standard deviation of 12.3 years. All subjects are German speakers. In this database, a total of 150 video clips from 82 subjects are included. And these recordings are divided into three parts by the publishers: training, development and test set. Each part contains 50 video clips. These video clips are set to 30 frames per second with the resolution of 640×480 pixels. Each subject is asked to fill in a BDI-II scale, and the corresponding score is the label.

The AVEC 2014 depression database is a subset of AVEC 2013 depression database. So their collection settings and age distribution are similar. Rather than the 14 tasks involved in AVEC 2013 database, there are only two tasks in AVEC 2014: namely, "Northwind" and "FreeForm". For the "Northwind" task, the subjects are asked to read an excerpt from the fable "Die Sonne underWind" (The Northwind and the Sun) in German. For the "FreeForm" task, the subjects need to respond to one of a number of questions (for example, "What is your favorite dish?") or describe a sad childhood memory in German. For these two tasks, 150 video clips from 82 subjects were recorded; the duration of these video clips ranges from 6 seconds to 4 minutes. In the same way, the publishers divide equally these recordings into training, development and test set. In our experiments, we combine the training, development and test sets of these two tasks to form a new database, which we still refer to as

the AVEC 2014 depression database. That is to say, there are 100 video clips in the training, development and test sets, respectively.

At present, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are two widely used indicators for evaluating the prediction accuracy. Eqs. (12) and (13) present the calculation formulas for RMSE and MAE, where N is the number of subjects. y_i and $\hat{y}_i$ are the ground truth and predicted BDI-II score of the ith subject, respectively. The smaller value of these two metrics, the better experimental performance is obtained

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{12}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{13}$$

4.2 Experimental Setup

From the description in Section 3, it is evident that the parameters in the experiment are mainly included in the process of generating the video segment representation sequence and the construction of DAER network. Thus, in the below, we will dpresent a detailed description of the parameter settings in these two processes.

4.2.1 Representation Sequence Generation

In [12], [37], the sample rate is reduced from 30 frames per second (fps) to 6fps, due to the fact that 6fps can be yield a better video analysis [42]. To this end, we also adopt 6fps to downsample the video clips in AVEC 2013 and AVEC 2014 depression databases. Moreover, we set the duration of the video segment to 3 seconds (i.e., 18 consecutive video frames). Note that the label applied to the video segment is the same as that of the corresponding long-term video. The overlapping of two adjacent video segments is 50%. For each frame in the video segments, we adjust the resolution to 480×480, then reshape the size of the frame to 224×224. In addition, we use the Euclidean loss as the objective function and replace "Softmax" with "ReLU" to train the ResNet50 using the ImageNet dataset like in [10], [12], [37]. Subsequently, we input each frame of the video segment into the trained ResNet50 and take the output of "avg_pool" layer as the frame representation. In this way, a video segment with 18 frames can be encoded into a representation sequence of size 18×2048, where each row is the representation of a frame.

4.2.2 The Network Structure of DAER

The architecture of our proposed DAER network is illustrated in Fig. 1. In this network, the "PCL" is used to reduce the dimension of the representation sequence and includes three Conv1D layers, which are set as (filters=1204, kernel_size=1), (filters=512, kernel_size=1) and (filters=256, kernel_size=1). "Covn2D_1" has 64 kernels with a size of 3×3, a stride size of 1 and the "same" for padding. The "FC" layer has 256 neurons.

The DA block contains two parts: namely, the self-attention part and the bilinear-attention part. For the self-attention part, Fig. 2 presents the corresponding structure. The three Conv2D layers have the same setting, namely 1 kernel with a size of 3×3, stride size of 1 and the "same" for padding. For the bilinear-attention part, Fig. 3 presents the corresponding structure, which includes three Conv1D layers with the "same" for padding. "Conv1D_1" and "Conv1D_2" have the same settings i.e., filters=16, kernel_size=3. The "Conv1D_3" is set to filters=1, kernel_size=3. The structure of ER block is presented in Fig. 4. As shown, the "Conv1D" has 64 filters with "kernel_size" of 1 and the padding is set to "same".

In all the above layers, "ReLU" is used as the activation function. The parameter of N in Fig. 1 is set to 2. The loss function used to train the DAER network is RMSE, as shown in Eq. (12). In this paper, we implement our DAER network in the Keras deep learning framework and use the Adam [43] optimizer with default momentum values (0.9, 0.999) for (β_1 and β_2). The weight decay is set to 0.0001. The learning rate is initialized to 0.0002 for AVEC 2013 and AVEC 2014 depression databases.

4.3 Ablation Analysis

As discussed above, we encode each frame of a video segment via the trained ResNet50 to obtain the corresponding representation and predict the BDI-II score using the DAER network, which mainly includes the DA and ER blocks. Thus, in this section, we conduct some ablation experiments to verify the role of the trained ResNet50 and those two blocks on the AVEC 2013 and AVEC 2014 depression databases.

4.3.1 The Role of the Trained ResNet50 for Encoding Frames

As described in Section 3.1, we use ImageNet to train ResNet50 to encode frames in the video segment for generating the corresponding representation sequence. However, one might consider whether it is more reasonable to use VGGFace descriptor for encoding facial images. To this end, we use these two models to encode video frames and employ our proposed DAER model to predict the BDI-II score. The corresponding experimental results are shown in Table 2.

From these comparison, we observe that "ResNet50" obtains better prediction accuracy than "VGGFace". This is because "VGGFace" model is aimed at the face recognition task. In other words, the encoding results of different frames in the same video segment relatively similar due to the inclusion of more identity information. Thus, the representation sequence generated by "VGGFace" is insensitive to facial dynamics and is not conductive to DAER network for capturing depression cues. However, the trained ResNet50 can characterize more diverse image patterns because ImagesNet contains images of multiple categories. Therefore, the representation sequence of video segment obtained using "ResNet50" has richer dynamic information, which is beneficial for the DAER network to extract the depression-related information from it.

4.3.2 The Role of the Dual Attention Block

In order to carefully explore the facial differences among individuals with different depression levels, we construct a DA block to extract the information related to depression in

TABLE 2
Performance of Depression Prediction on AVEC 2013 (AVEC 2014) Development and Test Sets Using Different Models to Encode Frames of the Video Segment

Databases	Encoding models	AVEC 2013		AVEC 2014	
		RMSE	MAE	RMSE	MAE
Dev	VGGFace[1]	8.18	6.21	8.09	6.17
	ResNet50	**7.97**	**5.86**	**7.79**	**5.82**
Test	VGGFace[1]	8.32	6.49	8.24	6.36
	ResNet50	**8.13**	**6.28**	**8.07**	**6.14**

[1]*https://www.robots.ox.ac.UK/vgg/software/vgg_face/. "VGGFace" refer to the use of VGGFace descriptor to encode frames of the video segment.*

facial videos from the perspective of temoral changes and feature components. To illustrate the effectiveness of DA block, we use the networks presented in Fig. 5a, 5b, and 5d to conduct experiments. The corresponding results on the development and test sets of AVEC 2013 and AVEC 2014 are presented in Tables 3 and 4.

From these tables, we can observe that the experimental performance of "CNN+self-attention" (as shown in Fig. 5b) is superior to that of "CNN" (as shown in Fig. 5a). This result reveals that self-attention is effective in capturing facial dynamic differences among individuals with different depression levels. From the comparison of bilinear-attention and self-attention, we can see that it is more helpful to investigate the role of feature components than temporal changes for improving the prediction accuracy. "CNN+DA block" (as shown in Fig. 5d) obtains the best experimental performance. This is mainly because the DA block can jointly extract depression-related information from the perspective of temporal changes and feature components.

TABLE 3
Performance of Depression Prediction on AVEC2013 and AVEC2014 **Development Sets** Using Different Networks

Network structures	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
CNN	8.97	7.03	8.85	6.61
CNN+self-attention	8.80	6.89	8.52	6.43
CNN+bilinear-attention	8.62	6.76	8.43	6.36
CNN+DA block	**8.34**	**6.33**	**8.39**	**6.16**
CNN+SE block	8.40	6.36	8.24	6.21
CNN+ER block	**8.16**	**6.29**	**8.04**	**6.00**
CNN+DA block+SE block	8.27	6.22	8.10	6.03
CNN+DA block+ER block	**7.97**	**5.86**	**7.79**	**5.82**

The networks of "CNN," "CNN+DA block" and "CNN+ER block" are shown in Fig. 5. The network of "CNN+DA block+ER block" is shown in Fig. 1.

TABLE 4
Performance of Depression Prediction on AVEC2013 and AVEC2014 **Test Sets** Using Different Networks

Network structures	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
CNN	9.35	7.58	8.97	7.11
CNN+self-attention	9.31	7.38	8.73	6.95
CNN+bilinear-attention	9.12	7.26	8.83	6.76
CNN+DA block	**9.03**	**7.21**	**8.79**	**6.88**
CNN+SE block	8.81	7.02	8.62	6.70
CNN+ER block	**8.69**	**6.54**	**8.32**	**6.42**
CNN+DA block+SE block	8.36	6.47	8.23	6.28
CNN+DA block+ER block	**8.13**	**6.28**	**8.07**	**6.14**

The networks of "CNN," "CNN+DA block" and "CNN+ER block" are shown in Fig. 5. The network of "CNN+DA block+ER block" is shown in Fig. 1.

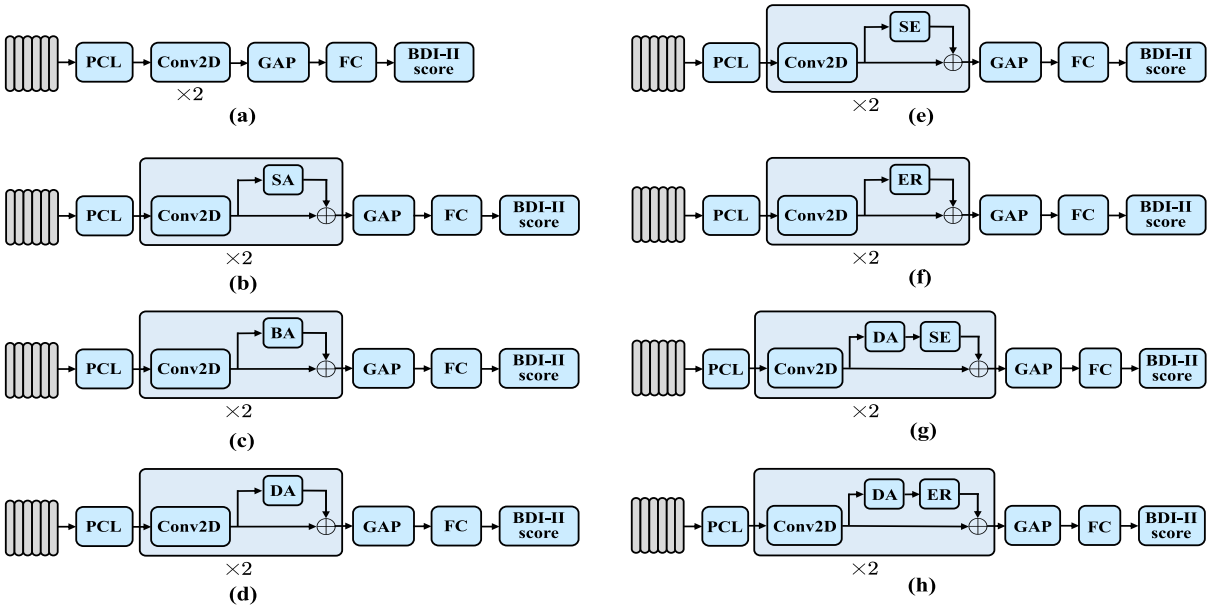


Fig. 5. Different network architectures are used to investigate the role of proposed DA and ER blocks for depression level prediction. In all three kinds of networks, "×2" indicates that the circled part is carried out twice. Note that the input is the representation sequence of a video segment through the trained ResNet50. The output is its corresponding BDI-II score. And the loss functions for three networks are all RMSE, as shown in Eq. (12). The "PCL" refers to Progressive Conv1D Layers to reduce the dimension of representation sequence. "SA," "BA" and "DA" refer to self-attention, bilinear-attention and dual-attention, respectively. "SE" and "ER" refer to Squeeze-Excitation block and Element Recalibration block, respectively. "GAP" and "FC" refer to the Global Average Pooling and Full Connection layer.

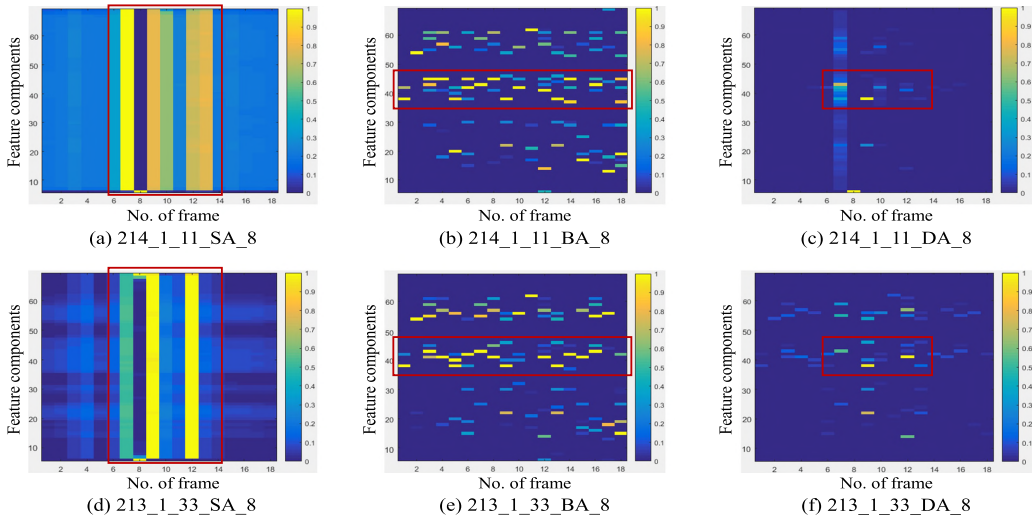


Fig. 6. Effects of different attention mechanisms on depression level prediction. "SA," "BA" and "DA" refer to self-attention, bilinear-attention and dual-attention, respectively. "214_1_11_SA_8" means that No. 214_1 subject with BDI-II score of 11 in the Northwind dataset of AVEC2014 is processed through the self-attention module. And the 8th channel of the multi-channel tensor is shown. The same is true for others in (b)-(f).

In order to further clarify the effects of self-attention and bilinear-attention on depression severity analysis, we present the corresponding visual results in Fig. 6. From this figure, we can see that self-attention tends to perceive the differences in representation sequences of individuals with different depression levels from the perspective of temporal changes, as shown in the part surrounded by the red box in Figs. 6a and 6d. The representation sequences after bilinear-attention processing are discriminative in terms of feature components, as shown in the part surrounded by the red box in Figs. 6b and 6e. Different from them, our proposed DA block can take into account the advantages of these two kinds of attention and jointly capture the differences of facial activities of individuals with different depression levels. Thus, as shown in the part surrounded by the red box in Figs. 6c and 6f, the processing results of the representation sequences of healthy and depressed individuals are obviously discriminative both from the perspective of temporal changes and feature components.

4.3.3 *The Role of the Element Recalibration Block*

The purpose of our proposed ER block is to recalibrate each element in the multichannel tensor. In other words, we intend to highlight the elements related to depression cues and suppress the less useful ones. Thus, in order to illustrate the effectiveness of ER block, we construct some comparative experiments with the architectures shown in Figs. 5e, 5f, 5g, and 5h. The results are presented in Tables 3 and 4.

As shown, "CNN+ER" (as shown in Fig. 5f) defeats "CNN+SE" (as shown in Fig. 5e) on prediction accuracy. The reason for this result is that the SE block treats all elements in the same channel as equal. While, our proposed ER block is able to examine all elements in the tensor one by one and pick out useful ones, so as to capture subtle depression cues. The same reason can also explain the results of the comparison between "CNN+DA+SE" (as shown in Fig. 5g) and "CNN+DA+ER" (as shown in Fig. 5h). In addition, we can see that the experimental performance is further improved when the SE block and DA block are combined. This finding also illustrates that the DA block is effective in capturing depression cues. It is easy to find that the network architecture of "CNN+DA block+ER block" (as shown in Fig. 1) achieves the best prediction accuracy. This is because the DA block is able to investigate the temporal changes and highlight the effective feature components. Meanwhile, the ER block is used to recalibrate each element of tensor in order to extract the subtle depression-related information.

Furthermore, we illustrate the advantages of ER block over SE block through the visual presentation in Fig. 7. As shown, different channels produce different visualization patterns (such as Figs. 7a, 7(b), 7c and 7d). These results also indicate that the impact of each channel on the task of depression level prediction is indeed not exactly the same. Moreover, as shown in the part surrounded bu the red box, the difference between Fig. 7a and 7c (or Figs. 7b and 7d) is weaker than that between Figs. 7e and 7g (or Figs. 7f and 7h). This is because the SE block assigns the same weight to all elements in each channel, but the proposed ER block can give corresponding weight to each element in the tensor to highlight the discriminative ones.

4.4 Comparisons With the State-of-The-Art Works

In this section, we compare our proposed method with some state-of-the-art works on the test sets of AVEC 2013 and AVEC 2014 to illustrate the effectiveness of our model. The corresponding experimental results are presented in Tables 5 and 6. From these comparisons, we can determine that our proposed approach achieves better experimental performance than those methods that use hand-crafted features. This fact also shows that deep neural network models have stronger representation ability in terms of depression cues extraction.

In addition, our proposed method outperforms those in [10], [12], [37] on the depression level prediction task. This is because the optical flow, C3D and VLDN used in the above three works are only able to extract the short-term facial changes, while the proposed DA block is able to

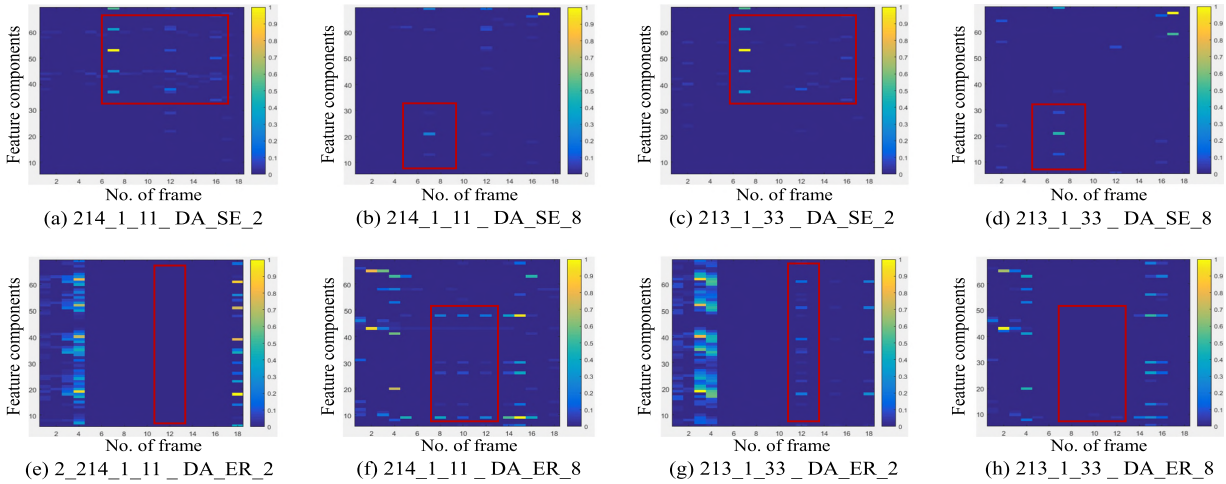


Fig. 7. Effects of SE and ER blocks on depression level prediction. "SE" and "ER" refer to Squeeze-Excitation block and Element Recalibration block, respectively. "214_1_11_DA_SE_8" means that No. 214_1 subject with BDI-II score of 11 in the Northwind dataset of AVEC2014 is processed through the DA and SE blocks. And the 8th channel of the multi-channel tensor is shown. The same is true for others in (b)-(f).

TABLE 5
Comparison of Depression Severity Analysis Performance With Some State-of-the-Art Methods on AVEC 2013 Test Sets

Category	Methods	RMSE	MAE
Hand-crafted features	Valstar *et al.* [28]	13.61	10.88
	Cummins *et al.* [44]	10.45	/
	Meng *et al.* [45]	11.19	9.14
	Wen *et al.* [31]	10.27	8.22
	He *et al.* [9]	9.20	7.55
	Niu *et al.* [8]	9.17	6.97
CNN-based methods	Zhu *et al.* [10]	9.82	7.58
	Jazaery *et al.* [12]	9.28	7.37
	Zhou *et al.* [7]	8.28	6.20
	Uddin *et al.* [37]	8.93	7.04
	Niu *et al.* [17]	8.97	7.32
	Zhou *et al.* [16]	8.37	6.63
	He *et al.* [46]	8.39	6.59
	He *et al.* [47]	8.46	6.83
	Shang *et al.* [15]	8.20	6.38
	Melo *et al.* [48]	**7.55**	**6.24**
	Ours	8.13	6.28

TABLE 6
Comparison of Depression Severity Analysis Performance With Some State-of-the-Art Methods on AVEC 2014 Test Sets

Category	Methods	RMSE	MAE
Hand-crafted features	Valstar *et al.* [29]	10.86	8.86
	Dhall *et al.* [30]	8.91	7.08
	Kaya *et al.* [49]	10.26	8.20
	Espinosa *et al.* [50]	9.84	8.46
	He *et al.* [9]	9.01	7.21
	Niu *et al.* [8]	9.10	7.19
CNN-based methods	Zhu *et al.* [10]	9.55	7.47
	Jazaery *et al.* [12]	9.20	7.22
	Zhou *et al.* [7]	8.39	6.21
	Uddin *et al.* [37]	8.78	6.86
	Niu *et al.* [17]	8.81	6.72
	Zhou *et al.* [16]	8.30	6.59
	He *et al.* [46]	8.30	6.51
	He *et al.* [47]	8.42	6.78
	Shang *et al.* [15]	7.84	6.08
	Melo *et al.* [48]	**7.65**	**6.06**
	Ours	8.07	6.14

capture the long-term facial dynamics in the video segment and compels the model to pay attention to the depression-related feature components in the representation sequence. Furthermore, each element of the tensor is recalibrated with the ER block to highlight the effective parts and suppress the less useful ones to boost the representation ability of network. Similarly, the advantages of these two blocks can also be used to explain why our method achieves better experimental performance than those proposed in [16], [17], [46]. However, on the test sets of AVEC 2013 and AVEC 2014, the prediction accuracy of [7], [15] and [36] is slightly superior to our method in terms of MAE and RMSE metrics respectively, which may be due to facial region division performed in [15] and [7] or the combination of hand-crafted and deep features in [15] and [36]. Besides, Melo *et al.* [48] obtains the best prediction accuracy because they examine the facial multi-scale temporal changes and encode the sudden facial variations. But our method is not enough in this aspect.

5 Conclusion

Physiological studies have revealed that depressed and healthy individuals present different types of facial changes. To this end, in this paper, we develop a DA and ER block to construct the DAER network, which is used to extract the facial representations of individuals with different depression levels. The DA block fuses the self-attention and bilinear-attention to capture the facial temporal changes and emphasize the effective feature components in the representation sequence, respectively. For its part, the ER block recalibrates each element of the tensor using global information in order to boost the network's representation ability. The experimental results on two publicly available databases (namely, AVEC 2013 and AVEC 2014 depression databases) demonstrate not only the effectiveness of each block, but also the superiority of our proposed method in the field of automatic depression level prediction. In the future, we will consider developing a depression level prediction method based on facial landmarks to mitigate personal privacy concerns caused by the use of face images.

REFERENCES

[1] C. Otte *et al.*, "Major depressive disorder," *Nature Rev. Dis. Primers*, vol. 2, no. 1, pp. 1–20, 2016.

[2] World Health Organization, "Depression and other common mental disorders: Global health estimates," World Health Organization, Tech. Rep. WHO/MSD/MER/2017.2, 2017.

[3] P. Philippot, R. S. Feldman, and E. J. Coats, *Nonverbal Behavior in Clinical Settings*, London, U.K.: Oxford Univ. Press, 2003.

[4] *Heiner Ellgring, Non-Verbal Communication in Depression*, New York, NY, USA: Cambridge Univ. Press, 2007.

[5] A. McPherson and C. Martin, "A narrative review of the beck depression inventory (BDI) and implications for its use in an alcohol-dependent population," *J. Psychiatr. Ment. Health Nurs.*, vol. 17, no. 1, pp. 19–30, 2010.

[6] Y. Kang, X. Jiang, Y. Yin, Y. Shang, and X. Zhou, "Deep transformation learning for depression diagnosis from facial images," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 13–22.

[7] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Jul.–Sep. 2020.

[8] M. Niu, J. Tao, and B. Liu, "Local second-order gradient cross pattern for automatic depression detection," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction Workshops Demos*, 2019, pp. 128–132.

[9] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1476–1486, Jun. 2019.

[10] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 578–584, Oct.–Dec. 2018.

[11] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3D networks for detecting depression from facial expressions," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–8.

[12] M. AlJazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 262–268, Jan.–Mar. 2021.

[13] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[14] Z. Du, W. Li, D. Huang, and Y. Wang, "Encoding visual behaviors with attentive temporal convolution for depression prediction," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–7.

[15] Y. Shang *et al.*, "LQGDNet: A local quaternion and global deep network for facial depression recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2021.3139651.

[16] X. Zhou, Z. Wei, M. Xu, S. Qu, and G. Guo, "Facial depression recognition by deep joint label distribution and metric learning," *IEEE Trans. Affect. Comput.*, 2020, to be published, doi: 10.1109/TAFFC.2020.3022732.

[17] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Trans. Affect. Comput.*, 2020, to be published, doi: 10.1109/TAFFC.2020.3031345.

[18] J. Salazar, K. Kirchhoff, and Z. Huang, "Self-attention networks for connectionist temporal classification in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7115–7119.

[19] K. J. Han, R. Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1D convolutions," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 54–61.

[20] M. Bilkhu, S. Wang, and T. Dobhal, "Attention is all you need for videos: Self-attention based video summarization using universal transformers," 2019, *arXiv: 1906.02792*.

[21] R. Zhang *et al.*, "Scan: Self-and-collaborative attention network for video person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4870–4882, Oct. 2019.

[22] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1571–1581.

[23] P. Fang, J. Zhou, S. Kumar Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8029–8038.

[24] T. Hu, J. Xu, C. Huang, H. Qi, Q. Huang, and Y. Lu, "Weakly supervised bilinear attention network for fine-grained visual classification," 2018, *arXiv: 1808.02152*.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," 2018, *arXiv: 1810.12348*.

[28] M. Valstar *et al.*, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 3–10.

[29] M. Valstar *et al.*, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 3–10.

[30] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 255–259.

[31] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[33] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4476–4484.

[34] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019.

[35] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu, and Q. Tian, "Pooling the convolutional layers in deep ConvNets for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1839–1849, Aug. 2018.

[36] A. Jan, H. Meng, Y.F. B.A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 3, pp. 668–680, Sep. 2018.

[37] M.A. Uddin, J.B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer Bi-LTSM," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2970418.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[39] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.

[40] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1971–1980.

[41] X. Zhong, O. Gong, W. Huang, L. Li, and H. Xia, "Squeeze-and-excitation wide residual networks in image classification," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 395–399.

[42] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4694–4702.

[43] P. DiederikKingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

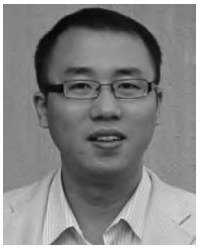
[44] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 11–20.

[45] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 21–30.

[46] L. He, J. Cheung-Wai Chan, and Z. Wang, "Automatic depression recognition using cnn with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.

[47] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *Int. J. Intell. Syst.*, pp. 1–18, 2021.
[48] W. Carneiro de Melo, E. Granger, and M. Bordallo Lopez, "MDN: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Trans. Affect. Comput.*, 2021, to be published, doi: 10.1109/TAFFC.2021.3072579.
[49] H. Kaya, F. Çilli, and A. Ali Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 19–26.
[50] H. Pérez Espinosa, H. Jair Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: INAOE-BUAP's participation at AVEC'14 challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 49–55.

Mingyue Niu received the PhD degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). He is currently a lecturer with the school of computer and information engineering, Tianjin Normal University. He has published the papers in *IEEE Transactions on Affective Computing*, *Neurocomputing*, ICASSP and INTERSPEECH. His research interests include affective computing and depression recognition and analysis. He is the reviewer of *IEEE Transactions on Systems, Man and Cybernetics*: *Systems and IEEE Transactions on Affective Computing*.



Ziping Zhao (Member, IEEE) received the PhD degree in automatic prediction of prosodic phrases from Nankai University, in 2008. He is currently a full professor of computer science, Tianjin Normal University, China. In 2018, he studied in the chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany as a visiting scholar. In 2016, he became the vice dean of the college of computer and information engineering at Tianjin Normal University. He has published 30+ publications in peer-reviewed books, journals, and conference proceedings, including ICASSP, INTERSPEECH, Neural Networks, and *IEEE Journal of Selected Topics in Signal Processing*. His research fields are affective computing and machine learning.

Jianhua Tao (Senior Member, IEEE) received the PhD degree from Tsinghua University, in 2001. He is winner of the National Science Fund for Distinguished Young Scholars and the deputy director in NLPR, CASIA. He has directed many national projects, including "863," National Natural Science Foundation of China. His research interests include speech synthesis, affective computing and pattern recognition. He has published more than eighty papers on journals and proceedings including *IEEE Transactions on Audio, Speech, and Language Processing*, and ICASSP, INTERSPEECH. He also serves as the steering committee member for *IEEE Transactions on Affective Computing* and the chair or program committee member for major conferences, including ICPR, Interspeech, etc.

Ya Li (Member, IEEE) received the bachelor's degree from the University of science and technology of China, in 2007 and the PhD degree from the Institute of automation, Chinese Academy of Sciences, in 2012. She is currently an associate professor with Beijing University of Posts and Telecommunications (BUPT). Her research interests include speech synthesis, affective computing, multimodal interaction. From 2012 to 2019, she has been an Assistant Professor and later an Associate Professor with the National Laboratory of Pattern Recognition (NLPR), Institute of automation, Chinese Academy of Sciences (CASIA). She has published more than 70 papers in journals and conferences, including Speech Communication, INTERSPEECH, ICASSP, ACII, etc. She won the first prize of science and technology progress award of China Electronics Society of 2018, and the second prize of Beijing Science and Technology award of 2014.

Björn W. Schuller (Fellow, IEEE) received the PhD degree from Technische Universität München, Germany. He is currently a professor of artificial intelligence with the Department of Computing, Imperial College London, U.K., and a full professor and head of the chair of embedded intelligence for health care and wellbeing with the University of Augsburg, Germany. He is the field chief editor of frontiers in Digital Health, president-emeritus of the AAAC, Golden Core awardee of the IEEE Computer Society, the fellow of ISCA, and senior member of the ACM. He (co-)authored five books and more than 1000 publications in peer-reviewed books, journals, and conference proceedings.