TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/23 pp2135–2148
DOI: 10.26599/TST.2024.9010246
Volume 31, Number 4, August 2026

Research Article

TREAT: Facial Depression Recognition by Learning Joint Depression Score and Level Distribution

Fan Zhang†, Liang Dong†, Byung-Gyu Kim, Jing Wang, Keqin Li, Saru Kumari, and Jianhui Lv*

Abstract: Automatic depression recognition is essential to depression diagnosis. In this paper, we investigate the problem of depression recognition from facial images, each of which is labeled with one Beck Depression Inventory (BDI-II) score. Because of the ambiguity between one facial image and the depression score, the annotators may not present the accurate score but tend to give those around the ground-truth one. To solve the problem, this paper adopts label distribution to annotate each image, in which each (score) label has a relevance degree. First, we apply the Gaussian distribution to generate the depression score distributions, in which the ground-truth score attains the highest degree, while the neighborhood scores also have degrees to some extent. Thus, each image can contribute to not only its ground-truth score but also neighborhood scores. Second, we generate the depression severity level distributions from the score distributions according to the mapping relationship between BDI-II score and severity level. Finally, we propose a novel method to learn joinT depression scoRE And level distribuTion, termed as TREAT. In the experiments, we compare TREAT with several state-of-the-art methods on three publicly released datasets AVEC 2013, AVEC 2014, and AVEC 2019, and the experimental results justified that TREAT achieves the best performance.

Key words: depression recognition; Beck Depression Inventory (BDI-II) score; depression severity; label distribution; deep learning

1 Introduction

Major depressive disorder[1], or known as depression, is the mental disorder of human beings, which may result in severe harm to one's health and life, such as persistent sadness, loss of interest or pleasure, self-denial, poor appetite, and even suicide. The diagnosis

• Fan Zhang is with The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China. E-mail: zhangf@jzmu.edu.cn.
• Liang Dong is with School of Electrical Engineering, Liaoning University of Technology, Jinzhou 121001, China. E-mail: dongliang53@cetc.com.cn.
• Byung-Gyu Kim is with School of AI Engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea. E-mail: bg.kim@sookmyung.ac.kr.
• Jing Wang is with School of Computer Science and Engineering, Southeast University, Nanjing 210018, China. E-mail: wangjing91@seu.edu.cn.
• Keqin Li is with School of Computer Science, State University of New York, New Paltz, NY 12561, USA. E-mail: lik@newpaltz.edu.
• Saru Kumari is with Department of Mathematics, Chaudhary Charan Singh University, Meerut 250001, India. E-mail: saryusiirohi@gmail.com.
• Jianhui Lv is with The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China, also with Peng Cheng Laboratory, Shenzhen 518057, China, and also with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China. E-mail: lvjh@pcl.ac.cn.
† Fan Zhang and Liang Dong contribute equally to this paper.
* To whom correspondence should be addressed.
Manuscript received: 2024-05-21; revised: 2024-10-11; accepted: 2024-12-04

of depression involves the subjective evaluations of psychologists, which makes this process labor-intensive and complicated[2]. As a result, automatic depression diagnosis has recently attracted widespread attention from the clinical research[3–5].

Among these automatic depression diagnosis methods, facial depression recognition[6] has emerged as one of the most promising methods[7–9]. In particular, human faces serve as a rich source of emotional cues, with facial expressions known to reflect the underlying affective states[10]. As a result, leveraging deep learning methods to analyze facial images holds the potential to uncover the subtle patterns of depressive symptoms, which helps enable more objective and efficient diagnosis of depression.

Overall, existing deep learning methods[8, 11–13] rough treat depression recognition as a single-label prediction task of the Beck Depression Inventory (BDI-II) score[14]. The BDI-II score describes the depression severity. It is assessed based on the BDI-II questionnaire[14], in which a total of 21 standardized questions are asked with each one scored from 0 to 3. Therefore, the BDI-II score ranges from 0 to 63. As demonstrated in Fig. 1, the facial image has a depression score of 43. Deep learning methods combine the representation learning ability of Deep Neural Network (DNN), such as Convolutional Neural Network (CNN)[15–17] and Transformer[18], in the process of learning BDI-II scores. For example, Zhu et al.[8] employed GoogleNet to learn the BDI-II scores from facial videosi, while Li et al.[13] applied Transformer to learn BDI-II scores.

However, as disclosed by Zhou et al.[12], these methods may face two challenges. The first is the ambiguity[19] between the BDI-II scores and facial images, which could lead to the fact that similar depression scores may undergo large variations in facial expression, while those with subtle differences in facial expression can imply significantly different depression scores[12]. The second one is lacking enough training data as facial depression data is expensive to collect, which could deteriorate the performance of deep learning methods.

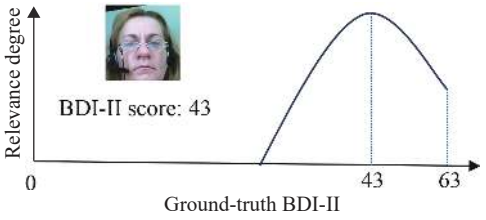

Fig. 1 An explanation of depression SD for a facial image. The ground-truth BDI-II score is 43. In the depression SD, the ground-truth score 43 has the highest relevance degree, and neighborhood score labels attains higher degrees than the distant ones.

To solve the above challenges, Zhou et al.[12] first proposed to introduce Label Distribution Learning (LDL)[20] to depression recognition from facial images. For each image, they generated a label distribution according to the ground-truth depression score. A label distribution spans the whole space of depression scores (i.e., 0–63), and each element indicates the relevance degree of the corresponding score label. For example, as demonstrated in Fig. 1, the ground-truth score has the highest relevance degree, while the neighborhood scores (i.e., those score values around 43) also have some lower description degrees. On one hand, the label distribution can efficiently capture the ambiguity between BDI-II scores and facial images. On the other hand, by learning from such label distributions, each image can contribute to not only the ground-truth score but also neighborhood scores, which helps enhance the training data and solve scarcity of enough training data.

Following Ref. [12], for each image, we first generate a label distribution from its ground-truth depression score, and then treat depression recognition as an LDL problem, that is, learning the depression Score Distributions (SDs) of the training instances and predicting unknown ones. In addition, we also consider depression severity level by considering the mapping relationship between depression scores and severity levels. The categories of severity regarding the cut-off scores[14] are shown in Table 1. For example, those with depression scores falling in the interval of [0, 13] are considered to have minimal depression severity, while those in the interval of [29, 63] have severe depression. Similarly, we believe depression severity also faces the same challenges as the depression score. First, there exists ambiguity between severity level and facial images. For example, a facial image with a depression score of 13 (should be considered with minimal depression according to Table 1) may also be considered as mildly depressed. Second, collecting enough training data with severity annotations are challenging.

Our key idea is to convert a depression SD into a severity Level Distribution (LD) according to the mapping relationship between BDI-II scores and

Table 1 Mapping relation between the BDI-II scores and depression severity levels[14].

BDI-II score	Depression severity level
0–13	Minimal
14–19	Mild
20–28	Moderate
29–63	Severe

severity levels. Then, we put forward a multi-task framework to learn joinT depression scoRE And level distribuTion, called TREAT. In summary, TREAT introduces two label distributions, i.e., depression score and severity distributions, which can efficiently leverage the ambiguity between depression score (and severity level) and facial images and solve the scarcity of enough training data. Since depression severity level is related to score, joint learning of these two distributions may help promote each other.

We conduct experiments on three public facial depression datasets, AVEC 2013[21], AVEC 2014[22], and AVEC 2019[23]. The comparison study justifies the advantages of TREAT against state-of-the-art deep learning methods. Furthermore, more analyses validate the superiority of learning joint distributions.

To summarize, our major contributions are as follows:

(1) We adopt label distribution to represent the depression annotations of each image. For depression score, we propose to apply Gaussian distribution to generate depression SD. For depression severity, we propose to generate depression severity LD via the mapping relationship between BDI-II score and severity.

(2) We put forward a novel method called TREAT to learn from these two distributions. TREAT also introduces two loss functions, i.e., absolute loss and margin loss to highlight the ground-truth score and severity level in the predictions.

(3) We conduct extensive evaluations of TREAT on two public datasets AVEC 2013 and AVEC 2014. The experimental results well justify the superiority of TREAT for depression recognition.

We organize the rest of this paper as follows. First, Section 2 briefly reviews the related works. Second, Section 3 introduces the details of the proposed TREAT method. Third, Section 4 reports the experimental results. Finally, Section 5 concludes this paper.

2 Related Work

In this section, we briefly introduce two lines of related studies, i.e., machine-learning-based depression recognition and LDL.

2.1 Depression recognition based on machine learning

Recently, automatic depression recognition by machine learning has attracted widespread attention within the research community. Among these works, facial depression recognition[6] is extremely popular because of the accessibility and richness of facial expression data. Overall, these methods can be categorized into hand-crafted methods and deep learning methods.

Earlier methods first applied hand-crafted features to represent facial images and then learned the depression scores[21]. To name just a few, Ojansivu and Heikkila[24] first used the Local Phase Quantization (LPQ) features as the representations of images and then learned the BDI-II scores. Cummins et al.[25] proposed to leverage the Pyramid of Histogram Of Gradient (PHOG) features[26] for depression score prediction. In addition, Valstar et al.[22] extracted the Local Gabor Binary Pattern (LGBP) descriptors[27] from facial videos for prediction of BDI-II scores. Dhall and Goecke[28] leveraged the Local Binary Pattern (LBP) features[29] to learn BDI-II scores.

Recently, deep learning methods can extract more expressive features and run in an end-to-end way, which have achieved significantly better performance than hand-crafted methods. To name just a few, Zhu et al.[8] fine-tuned pre-trained GoogleNet[15] models to learn the BDI-II scores from facial videos, which significantly advanced the state-of-the-art depression recognition results. Zhou et al.[9] borrowed the ResNet-50[16] and put forward DepressNet that blends different facial regions for predicting BDI-II scores. In addition, Li et al.[13] applied Transformer and put forward Hybrid Multi-Head cross attention Network (HMHN) to learn BDI-II scores. However, these works treated depression recognition simply as a problem of predicting the BDI-II scores, which ignore the ambiguity between facial images and BDI-II scores. To solve this challenge, Zhou et al.[12] first introduced label distribution to depression recognition. They converted BDI-II scores to depression distribution and employed DepressionNet to learning from such distribution. Our work further improves Zhou et al.[12]

by considering joint depression score and severity distributions, which leads to better performance. Our method is agnostic of the backbone networks and can be integrated into any DNN, such as CNN and Transformer.

2.2 LDL

Due to the smooth changes of facial appearance in the aging process, label distribution was first introduced by Geng et al.[30] to represent the age information of a facial image. Gaussian distribution was employed to generate a label distribution for one image by treating the chronological age label as the mean. Then, a novel method was designed to learn the age label distributions. Latter, Geng[20] formally defined LDL as a novel learning paradigm, which aims to learn the label distributions of the training instances and make prediction for unknown instances.

Researchers have proposed various LDL methods. To name just a few, Shen et al.[31] leveraged the differentiable random forests to learn label distributions. Wang et al.[32] considered the problem of classification in the setting of LDL. They introduced a novel re-weighting scheme and large margin to the process of LDL. In addition, Jia et al.[33] exploited the ranking information of labels in LDL. Jia et al.[34] put forward the description-degree percentile average, which can combine the ranking information and description degrees of labels. Huang et al.[35] employed the polynomial-based fuzzy broad learning systems to LDL, while Wen et al.[36] considered the ordinal LDL problem.

LDL has already seen extensive applications among various fields. In depression recognition, Zhou et al.[12] generated a label distribution via the Gaussian distribution with the ground-truth depression score as the mean. Then, they designed a novel LDL method based on metric learning to efficiently learn such depression distributions. In addition, in the application of emotion recognition, Li and Deng[37] leveraged a label distribution to denote the emotions of each image. Then, they proposed a deep model to learn the emotional label distributions. Shu et al.[38] formalized the problem of emotion recognition as emotion distribution learning. In age estimation, Shen et al.[31] generated label distributions from the chronological age and proposed a differential random forests to learn the age label distributions.

This paper applies LDL to depression recognition. Different from Ref. [12], we consider both depression scores distribution and depression severity LD and propose to learn joint distribution, which helps improve each other.

3 TREAT Method

Specifically, let $\mathcal{X}$ be the input space. Let $T = \{(\boldsymbol{x}_1, s_1, c_1), (\boldsymbol{x}_2, s_2, c_2), \ldots, (\boldsymbol{x}_n, s_n, c_n)\}$ denotes a training set, where $\boldsymbol{x}_i \in \mathcal{X}$ is the i-th instance (e.g., a facial image), $s_i \in [0, 63]$ is its corresponding real-value BDI-II score, and $c_i \in [1, 4]$ is the assigned depression severity level according to the mapping relationship between BDI-II scores and depression severity, as shown in Table 1.

In this paper, we put forward a novel method called TREAT for facial depression recognition. First, it generates depression score and severity distributions to represent the depression scores and severity levels, respectively. Next, it jointly learns the generated depression score and severity distributions. The whole framework of TREAT is illustrated in Fig. 2.

3.1 Generating depression score and severity distributions

Because of the ambiguity between facial images and depression scores, it is prohibitive to obtain the absolutely accurate annotations. For example, a doctor is highly likely to label an image with a ground-truth depression score of 12 to 13. On the other hand, due to the continuity of depression score, the neighborhood score labels around the ground-truth label may also contribute to one specific image to some extent. In light of the above two observations, in this paper, we employ the Gaussian function to generate a label distribution to represent the depression score for each image, in which each label attains a relevance degree to that image.

We denote by $\mathcal{Y} = \{y_1, y_2, \ldots, y_{64}\}$ the label space of the ordinal BDI-II scores ranging from 0 to 63. Concretely, for $\boldsymbol{x}_i$ with the ground-truth score s_i, the relevance degree of y_j to $\boldsymbol{x}_i$ is calculated by

$$d_{\boldsymbol{x}_i}^{y_j} = \frac{1}{\sqrt{2\pi}\sigma Z} \exp\left[-\frac{(y_j - s_i)^2}{2\sigma^2}\right] \tag{1}$$

where σ is the standard deviation parameter. Here, Z is normalization factor that equals

$$Z = \frac{1}{\sqrt{2\pi}\sigma} \sum_{j=1}^{64} \exp\left[-\frac{(y_j - s_i)^2}{2\sigma^2}\right] \tag{2}$$

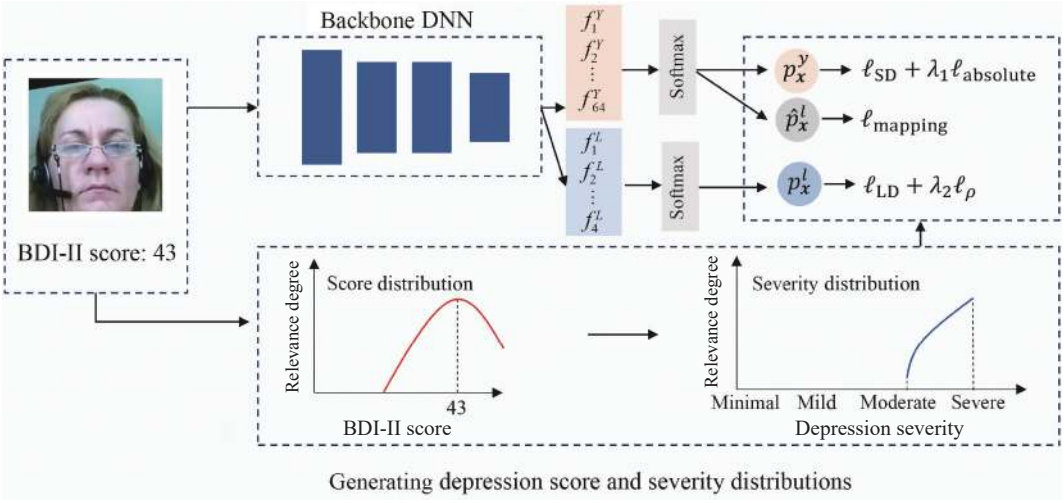

Fig. 2 Framework of the proposed method. First, we generate the depression SDs from the BDI-II scores via the Gaussian distribution. Then, we generate the depression severity distributions via the mapping relationship between BDI-II scores and depression severity. Finally, we train a neural network to jointly learn the depression score distribution and severity LDs.

The relevance degrees of all scores to $\boldsymbol{x}_i$ form a depression SD $\boldsymbol{d}_i^{\mathcal{Y}} = \left[d_{\boldsymbol{x}_i}^{y1}, d_{\boldsymbol{x}_i}^{y2}, \ldots, d_{\boldsymbol{x}_i}^{y64}\right]$ that satisfies $d_{\boldsymbol{x}_i}^{y_j} \geqslant 0$ and $\sum_j d_{\boldsymbol{x}_i}^{y_j} = 1$.

It is noteworthy that for each image, the depression SD can sufficiently describe its ground-truth depression score because that score has the highest degree in the SD. Compared with the original score label, the depression SD associates to each label a degree, with higher values for neighborhood labels and lower values for distant labels. As a result, it can efficiently represent the continuity of depression scores. On the other hand, since each score label has a degree, in the training process an image contributes to not only its ground-truth label but also neighborhood labels, which also help solves the scarcity of enough training samples.

The depression severity is categorized according to the BDI-II score, and as a result, it also has some continuity and ambiguity. For example, a facial image with a ground-truth depression score of 12 can be considered to be minimally depressed and mildly depressed as well. Likewise, we generate a label distribution to represent the depression severity of one specific image, in which each label (i.e., severity level) has a relevance degree to that image.

Next, we generate the depression severity LD according to the mapping relationship between BDI-II scores and severity levels. As illustrated in Fig. 3, the degree of a depression severity equals the sum of degrees for all score labels in the corresponding

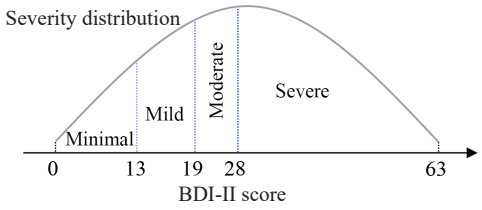

Fig. 3 An illustration of generating depression severity LD. The degree of one depression severity equals the sum of the degrees of scores belonging to it.

interval. For example, the degree of "minimal" depression equals the sum of degrees of all score labels ranging from 0 to 13. Specifically, let $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$ be the label space for depression severity. For $\boldsymbol{x}_i$ with a depression SD $\boldsymbol{d}_i$, the degree of l_j is calculated by

$$d_{\boldsymbol{x}_i}^{l_j} = \sum_{y \in \mathcal{I}(l_j)} d_{\boldsymbol{x}_i}^{y} \tag{3}$$

where $\mathcal{I}(l_j)$ contains all score labels in the interval of l_j. The degrees of all levels define a depression severity LD (LD) $\boldsymbol{d}_i^{\mathcal{L}} = \left[d_{\boldsymbol{x}_i}^{l_1}, d_{\boldsymbol{x}_i}^{l_2}, d_{\boldsymbol{x}_i}^{l_3}, d_{\boldsymbol{x}_i}^{l_4}\right]$.

In the single-label depression severity level, only the ground-truth label has a degree of one, while other labels have degrees of zeros, which ignores the ambiguity between depression severity level and facial images. In comparison, LD assigns a degree to each labels, which can effectively represent the ambiguity. Moreover, the depression level is also ordinal (i.e., "minimal" < "mild" < "moderate" < "severe"). Since the neighborhood score labels around the ground-truth label have higher degrees, the corresponding

neighborhood level labels around the ground-truth level label also have higher degrees. As a result, in the training process an image can contribute to not only the ground-truth depression severity label but also other labels, which can enlarge the training set.

After generating the depression SD and LD, we transform the depression recognition problem into learning these two distributions. We train a DNN with 68-dimension outputs to learn from these two distributions, where the first 64-dimension outputs learn the depression SD, and the remaining four-dimension outputs learn the depression LD.

Note that the computational cost to generate the depression SD for one image scales linearly regarding the depression score size (i.e., 64). As a result, the total computational cost to generate the depression SDs for all training images need complexity $\mathcal{O}(n)$. Similarly, the computational cost to generate the depression severity distribution for all training images need complexity $\mathcal{O}(n)$. As a result, compared with traditional methods, TREAT needs additional complexity $\mathcal{O}(n)$ to generate the depression score and severity distributions. However, these processes can be easily paralleled. The computational cost would reduce to $\mathcal{O}(n/n_0)$ if n_0 processes are available, which can significantly reduce the computational cost.

3.2 Learning depression SD

First, we add the softmax function to the first 64-dimension outputs of our DNN to learn the depression SD. Specifically, for $\boldsymbol{x}_i$, the predicted probability of y_j is calculated as

$$p_{\boldsymbol{x}_i}^{y_j} = \frac{\exp\left(f_j^{\mathcal{Y}}\right)}{\sum_{k=1}^{64}\exp\left(f_k^{\mathcal{Y}}\right)} \tag{4}$$

where $f_j^{\mathcal{Y}}$ is the j-th output of the first 64-dimension outputs of our DNN model. As widely used in the related LDL literature, Kullback-Leibler (KL) divergence is employed to minimize the distance between the ground-truth depression SD and the prediction of our DNN. We design the following loss to learn the depression SD,

$$\ell_{\mathrm{SD}} = -\sum_{j=1}^{64} d_{\boldsymbol{x}_i}^{y_j} \ln p_{\boldsymbol{x}_i}^{y_j} \tag{5}$$

3.3 Learning depression severity LD

In this subsection, we learn the depression LD from two perspectives. First, we can obtain the predicted depression LD from the predicted SD according to their mapping relationship. The predicted probability of $\boldsymbol{x}_i$ belonging to l_j equals

$$\hat{p}_{\boldsymbol{x}_i}^{l_j} = \sum_{y\in\mathcal{I}(l_j)} p_{\boldsymbol{x}_i}^{y} \tag{6}$$

where $p_{\boldsymbol{x}_i}^{y}$ is the predicted probability of $\boldsymbol{x}_i$ with the score y, as introduced in Eq. (4). We apply KL divergence and design the following mapping loss:

$$\ell_{\mathrm{mapping}} = -\sum_{j=1}^{4} d_{\boldsymbol{x}_i}^{l_j} \ln \hat{p}_{\boldsymbol{x}_i}^{l_j} \tag{7}$$

Besides, we also add the softmax function to the last four-dimension outputs of our DNN to learn the depression LD. The predicted probability of $\boldsymbol{x}_i$ belonging to l_j equals

$$p_{\boldsymbol{x}_i}^{l_j} = \frac{\exp\left(f_j^{\mathcal{L}}\right)}{\sum_{k=1}^{4}\exp\left(f_k^{\mathcal{L}}\right)} \tag{8}$$

where $f_j^{\mathcal{L}}$ is the j-th output of the last four-dimension outputs of our DNN model. Similarly, we use KL divergence to learn the depression LD and design the following loss:

$$\ell_{\mathrm{LD}} = -\sum_{j=1}^{4} d_{\boldsymbol{x}_i}^{l_j} \ln p_{\boldsymbol{x}_i}^{l_j} \tag{9}$$

Finally, we combine loss functions described by Eqs. (7) and (9) to learn the depression LD.

3.4 Multi-task learning model

Since depression severity level is highly correlated with depression score, as shown in Table 1, the learning task of these two may help each other. Therefore, in this paper, we use a unified multi-task learning model to jointly learn the depression SD and LD. To train our DNN, a combined multi-task loss is designed as

$$\ell_i = \ell_{\mathrm{SD}} + \ell_{\mathrm{mapping}} + \ell_{\mathrm{LD}} \tag{10}$$

However, Eq. (10) only learns the whole SD and LD but ignores the ground-truth depression score s_j and level l_j. As disclosed in Refs. [19, 32], it may produce a DNN model with a high accuracy to learn the SD and LD but fail to learn the scores and severity levels.

First, we encourage the predicted depression score to approach the ground-truth one. The predicted score can

be calculated as the expectation of the predicted SD, i.e.,

$$\hat{s}_j = \sum_{y \in \mathcal{Y}} y \cdot p_{\boldsymbol{x}_i}^{y} \tag{11}$$

We add an absolute loss to minimize their difference,

$$\ell_{\text{absolute}} = |s_i - \hat{s}_i| \tag{12}$$

Second, we encourage the predicted depression severity level to be the ground-truth one. To achieve that, we introduce a margin loss to the DNN, which highlights the ground-truth depression level in the predicted LD. Specifically, for $\boldsymbol{x}_i$, we introduce an extra margin $\rho > 0$ between the ground-truth depression level c_i and other level labels—the predicted probability of c_i is larger than that of other level labels by a margin ρ. The margin loss is designed as

$$\ell_{\rho} = \frac{1}{\rho} \sum_{j: l_j \neq c_i} \max\left(0,\ p_{\boldsymbol{x}_i}^{c_i} - p_{\boldsymbol{x}_i}^{l_j} + \rho\right) \tag{13}$$

Notice that the margin loss equals 0 if the predicted probability of c_i is larger than that of other labels by a margin ρ; otherwise, it results in a positive penalty.

Finally, we combine loss functions described by Eqs. (10), (12), and (13) to learn the SD, LD, and ground-truth depression scores and levels, and design the following loss to train a DNN:

$$\ell_i = \ell_{\text{SD}} + \ell_{\text{mapping}} + \ell_{\text{LD}} + \lambda_1 \ell_{\text{absolute}} + \lambda_2 \ell_{\rho} \tag{14}$$

where λ_1 and λ_2 are the trade-off parameters. The setting of these parameters is discussed in the experiments.

3.5 Explainability of learning depression SD

TREAT transforms the single-learning problem of the BDI-II scores into an LDL problem. Next, we analyze the explainability of doing that.

Concretely, for the i-th training instance $\boldsymbol{x}_i$, let $\boldsymbol{d}_i^{\mathcal{Y}} = \left\{d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \ldots, d_{\boldsymbol{x}_i}^{y_{64}}\right\}$ be the generated depression score and $\boldsymbol{p}_i^{\mathcal{Y}} = \left\{p_{\boldsymbol{x}_i}^{y_1}, p_{\boldsymbol{x}_i}^{y_2}, \ldots, p_{\boldsymbol{x}_i}^{y_{64}}\right\}$ the predicted depression SD. Let s_i be the ground-truth depression scores of $\boldsymbol{x}_i$, which has the highest relevance in $\boldsymbol{d}_i^{\mathcal{Y}}$. Define $\hat{s}_i$ as the score with the highest predicted relevance in $\boldsymbol{p}_i^{\mathcal{Y}}$, i.e.,

$$\hat{s}_i = \arg\max_{y \in \mathcal{Y}} p_{\boldsymbol{x}_i}^{y} \tag{15}$$

Then, we can prove the following theorem to show the explainability of LD SD.

Theorem 1 Suppose $\boldsymbol{d}_i^{\mathcal{Y}}$ is the conditional probability. Let y be the random variable. Then, the following inequality holds:

$$P(y \neq \hat{s}_i | \boldsymbol{x}_i) - P(y \neq s_i | \boldsymbol{x}_i) \leqslant \sum_j \left| p_{\boldsymbol{x}_i}^{y_j} - d_{\boldsymbol{x}_i}^{y_j} \right| \tag{16}$$

Proof First, we have

$$P(y \neq s_i \mid \boldsymbol{x}_i) = 1 - \mathcal{P}(y = s_i \mid \boldsymbol{x}_i) = 1 - d_{\boldsymbol{x}_i}^{s_i},$$

which yields

$$P(y \neq s_i \mid \boldsymbol{x}_i) - \mathcal{P}(y \neq \hat{s}_i \mid \boldsymbol{x}_i) = d_{\boldsymbol{x}_i}^{s_i} - d_{\boldsymbol{x}_i}^{\hat{s}_i}.$$

Then, according to Ref. [32], we can show

$$d_{\boldsymbol{x}_i}^{s_i} - d_{\boldsymbol{x}_i}^{\hat{s}_i} \leqslant \sum_j \left| p_{\boldsymbol{x}_i}^{y_j} - d_{\boldsymbol{x}_i}^{y_j} \right|,$$

which completes the proof the theorem. ■

In the above equation, the left-hand-side is the error probability, while the right-hand-side is the absolute loss between the two distributions. The error probability is bounded by the absolute loss between the distributions. According to Theorem 1, the predicted score is expected to converge to the ground-truth one as long as the learned SD converges to the ground-truth distribution. It explains why learning the depression SDs helps learn the depression scores.

4 Experiment

In this section, we conduct experiments to validate the performance of TREAT. We elaborate on the experimental setting, results and analysis, ablation study, and parameter analysis.

4.1 Methodology

4.1.1 Datasets

We evaluate the performance of our method on two publicly available datasets, AVEC 2013[21] and AVEC 2014[22], which are two most widely used depression datasets for facial depression recognition.

AVEC 2013 depression dataset[21]: AVEC 2013 was created to facilitate research in the automatic analysis of depression from multimodal data, including audio and video. There are totally 150 videos from 82 German-speaking subjects. Valstar et al.[21] used a webcam and a microphone over a period of two weeks to collect the data, whose ages range from 18 to 63 with an average age of 31.5 and a standard deviation of 12.3. They recorded the video for 30 frames per second (fps) with a resolution of 640 × 480 pixels. In Ref. [21], this dataset is evenly spitted into three subsets, i.e., the training set, development set, and test

set. For each subset, there are 50 videos, and each one is labeled with a depression severity level assessed based on the BDI-II questionnaire.

AVEC 2014 depression dataset[22]: AVEC 2014 is a subset of AVEC 2013. It has the same subjects as AVEC 2013 but introduces two additional tasks, FreeForm and Northwind. The FreeForm task asked the subjects to respond to several questions such as discussing a sad childhood memory; the Northwind task required the subjects to read an excerpt audibly from a fable[22]. For each task, there are 150 videos, which are partitioned evenly into three subsets, i.e., the training, development, and test sets, labeled with the ground-truth severity level.

The distribution of the depression scores of the samples from these two datasets is summarized in Fig. 4. For both datasets, we combine the training and development sets as the training set of our model and test its performance on the test set.

4.1.2 Implementation details

The proposed TREAT does not depend on any specific structures of neural networks, and it can be integrated into any deep models. Here, we implement TREAT with a state-of-the-art network, called HMHN[13] that includes two stages. The first stage consists of the grid-wise attention block and deep feature fusion block for the low-level visual depression feature learning. The second stage encodes high-order interactions among local features with multi-head cross attention block and attention fusion block[13]. For more details of HMHN, refer to Ref. [13].

Following Ref. [12], the standard deviation parameter σ is set to 1. Besides, we set λ_1 to 0.1 and λ_2 to 0.1.

4.1.3 Evaluation metrics

We apply two widely used metrics, i.e., Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to assess the performance of the comparing algorithms on AVEC 2013 and AVEC 2014. MAE is defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|s_i - \hat{s}_i| \tag{17}$$

and RMSE is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2} \tag{18}$$

where s_i and $\hat{s}_i$ are the ground-truth and predicted depression score of the i-th test instances, respectively.

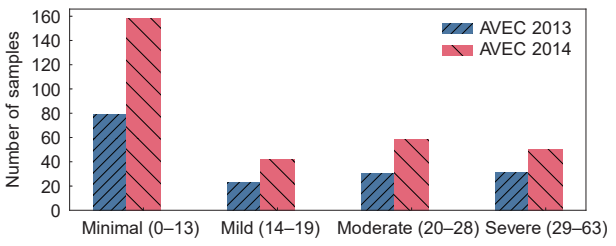

Fig. 4 Distribution of the depression scores of the samples from AVEC 2013 and AVEC 2014.

4.2 Results and analysis

First, we present two examples with their predicted depression SD by TREAT in Fig. 5. From Fig. 5, the predicted scores for these three typical examples are close to the ground-truth ones, which justifies the advantages of TREAT.

Second, to validate the performance of TREAT, we compare it with several state-of-the-art algorithms for facial depression recognition. First, for AVEC 2013, we compare TREAT against ten state-of-the-art proposals[7–9, 12, 13, 21, 39–41]. For AVEC 2014, we compare TREAT against nine state-of-the-art comparing algorithms[8, 9, 12, 13, 22, 40–43]. For TREAT, we set λ to 1, λ_1 to 0.1 and λ_2 to 0.001. For other comparing algorithms, we set their parameters as suggested in the literature. Tables 2 and 3 tabulate the results of these comparing methods on AVEC 2013 and AVEC 2014, respectively.

Among the comparing methods, DJ-LDML[12] also first transformed the BDI-II scores into label distributions and then learns such label distributions with a novel metric learning methods. Besides Deep Joint Lable Distribution and Metric Learning (DJ-LDML), the other comparing methods all directly learn the BDI-II scores, which are regarded as a regression problem.

According to Tables 2 and 3, TREAT achieves the best performance on both AVEC 2013 and AVEC 2014. The reasons are two-fold. First, compared with those that directly fit the BDI-II scores, TREAT leverages label distribution to represent the depression score, which not only captures ambiguity between facial images and BDI-II scores but also enhances the training data. Second, although DJ-LDML introduces an extra metric learning process to learn the label distributions, it fails to consider the depression severity level. Instead, TREAT takes into account the ambiguity between facial images and severity level and further

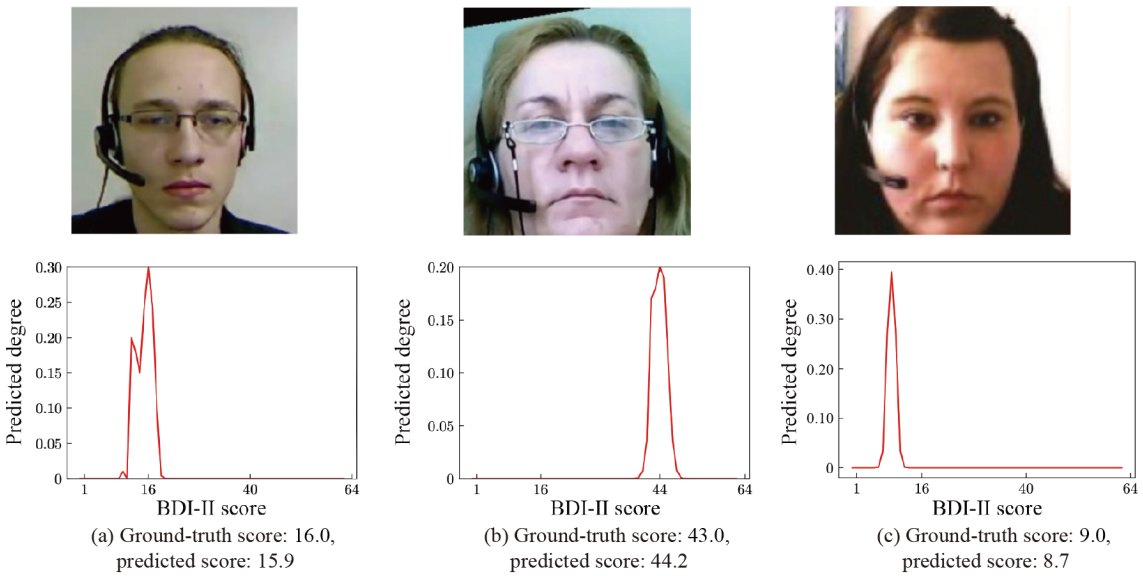

Fig. 5 Three example images and their predicted depression SDs by TREAT and the ground-truth BDI-II scores.

Table 2 Comparison results of TREAT with state-of-the-art methods on AVEC 2013. The best results are highlighted in bold.

Method	MAE	RMSE
Valstar et al.[21]	10.88	13.61
Wen et al.[7]	8.22	10.27
Kaya and Salah[39]	7.86	9.72
Zhu et al.[8]	7.58	9.82
Uddin et al.[40]	7.04	8.93
DepressionNet[9]	6.77	9.02
DJ-LDML[12]	6.63	8.37
He et al.[42]	6.83	8.46
Liu et al.[41]	6.08	7.59
HMHN[13]	6.05	7.38
TREAT	**6.01**	**7.22**

Table 3 Comparison results of TREAT with state-of-the-art methods on AVEC 2014. The best results are highlighted in bold.

Method	MAE	RMSE
Valstar et al.[21]	10.88	13.61
Zhu et al.[8]	7.47	9.55
Uddin et al.[40]	6.86	8.78
DepressionNet[9]	6.60	8.88
DJ-LDML[12]	6.59	8.30
He et al.[42]	6.78	8.42
Liu et al.[41]	6.04	7.98
De Melo et al.[43]	6.01	7.60
HMHN[13]	6.01	7.60
TREAT	**5.91**	**7.43**

transforms depression level into LD by the mapping relation between BDI-II score and severity level. TREAT jointly learns the depression SD and LD into a multi-task framework, which significantly boosts its performance.

To summarize, the experimental results well justify the superior performance of TREAT, which can be credited to introduction of joint depression SD and LD.

4.3 More results on AVEC 2019

Next, to further show the generalizability of TREAT, we evaluate it on AVEC 2019 dataset[23]. Different from AVEC 2013 and AVEC 2014, each image of AVEC 2019 is labeled with a Patient Health Questionnaire (PHQ-8) score[44], ranging from 0 to 24, with higher value indicating more severe depression. According to Ref. [44], the cut-off points for PHQ-8 scores are shown in Table 4. For this dataset, there are clips of 163 subjects for training, 56 subjects for validation, and 56 subjects for testing.

TREAT can simply adapt to PHQ-8 scores by changing the dimensions of depression scores and severity level to 24 and 5, respectively. As a result, we can also evaluate its performance on AVEC 2019. We compare TREAT with three state-of-the-art methods[12, 45, 46]. Figure 6 visualizes the comparison results on AVEC 2019. According to Fig. 6, TREAT performs best and outperforms the other three comparing methods by a margin. The results justify the superior performance and generalizability of TREAT.

Table 4 PHQ-8 cut-off points[44] to indicate differently severe depression.

BDI-II score	Depression severity
0–4	Minimal
5–9	Mild
10–14	Moderate
15–19	Moderately severe
20–24	Severe

4.4 Ablation study

Next, we conduct ablation study to investigate the advantages of several loss functions of TREAT, including LD loss (i.e., Eq. (9)), mapping loss (i.e., Eq. (7)), absolute loss (i.e., Eq. (12)), and margin loss (i.e., Eq. (13)). We run four degenerated versions of TREAT. Especially, TREAT without LD excludes learning the LD and margin loss.

We report the results of TREAT and its four degenerated versions in Tables 5 and 6. From Tables 5 and 6, TREAT all achieves higher performance than its degenerated versions. Especially, we find that excluding the absolute loss will significantly reduce the performance of TREAT. The reason may lie in that TREAT without the absolute loss focuses on the whole depression SD but ignores the ground-truth depression score.

4.5 Parameter sensitivity

TREAT introduces two hyper-parameters λ_1 and λ_2. In this subsection, we investigate the sensitivity of these two parameters.

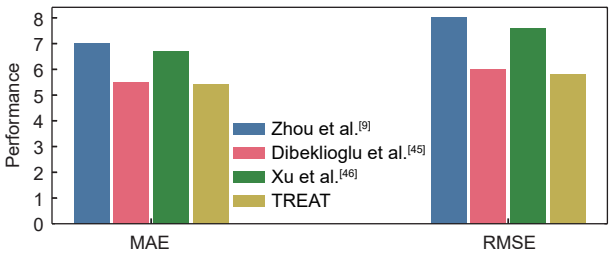

Fig. 6 Comparison results of TREAT with three state-of-the-art comparing methods on AVEC 2019.

Table 5 Results of ablation study on AVEC 2013 by comparing TREAT against the degenerated versions. W/o: without. The best results are highlighted in bold.

Method	MAE	RMSE
TREAT w/o LD	6.10	7.45
TREAT w/o mapping loss	6.08	7.30
TREAT w/o absolute loss	6.20	7.56
TREAT w/o margin loss	6.02	7.25
TREAT	**6.01**	**7.22**

Table 6 Results of ablation study on AVEC 2014 by comparing TREAT against the degenerated versions. The best results are highlighted in bold.

Method	MAE	RMSE
TREAT w/o LD	6.01	7.62
TREAT w/o mapping loss	6.00	7.55
TREAT w/o absolute loss	6.10	7.76
TREAT w/o margin loss	5.95	7.49
TREAT	**5.91**	**7.43**

First, we run TREAT with varying λ_1 from the candidate set {0.001, 0.01, 0.1, 0, 1, 10, 100, 1000} and report its achieved MAE and RMSE in Figs. 7 and 8, respectively. According to Figs. 7 and 8, TREAT is sensitive to λ_1 and achieves the best results with $\lambda_1 = 0.1$. In the experiments, the default value of λ_1 is 0.1.

Second, we run TREAT by varying λ_2 from the candidate set {0.001, 0.01, 0.1, 0, 1, 10, 100, 1000} and report its achieved MAE and RMSE in Figs. 9 and 10, respectively. From the results, we find that TREAT has robust performance when λ_2 is within [0.01, 10]. In the experiments, we set it to 0.1.

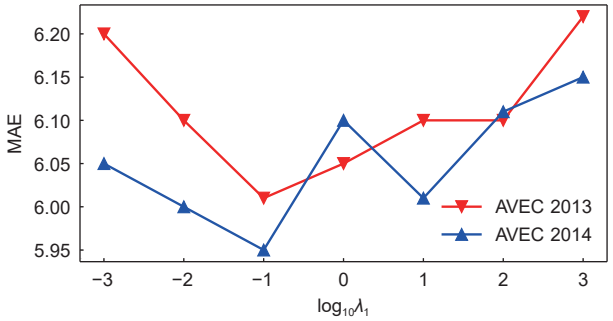

Fig. 7 Achieved MAE of TREAT with varying λ_1 on AVEC 2013 and AVEC 2014.

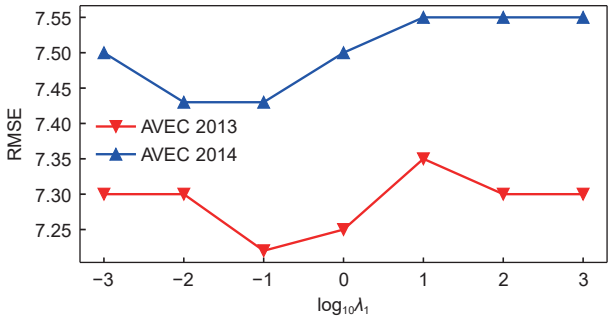

Fig. 8 Achieved RMSE of TREAT with varying λ_1 on AVEC 2013 and AVEC 2014.

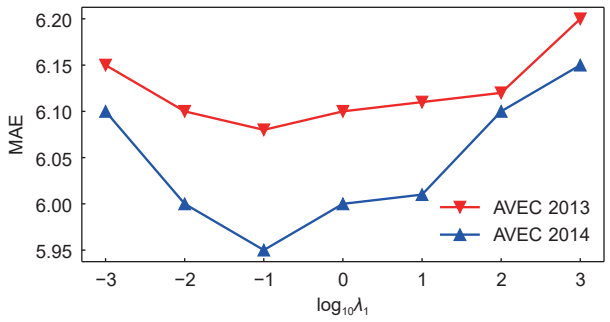

Fig. 9 Achieved MAE of TREAT with varying λ_2 on AVEC 2013 and AVEC 2014.

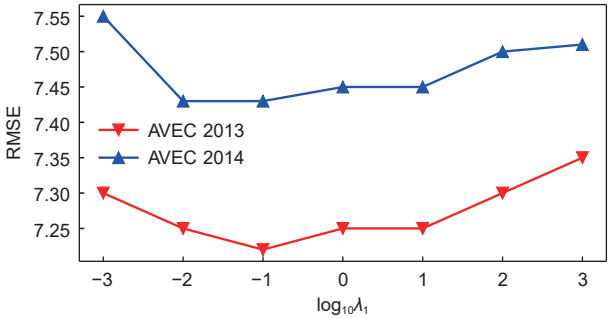

Fig. 10 Achieved RMSE of TREAT with varying λ_2 on AVEC 2013 and AVEC 2014.

5 Conclusion

In this paper, we propose a novel method called TREAT for facial depression recognition. Considering the ambiguity between facial images and BDI-II scores, we first transforms a score number into depression SD, in which the ground-truth score attains the highest degree and neighborhood scores also have degrees to some extent. A depression SD not only helps solve the ambiguity but also enhance the training set. Second, we transform a severity level into a depression severity distribution. Finally, we jointly learn these two distributions. The experimental results demonstrate that TREAT achieves the best performance on three commonly-used datasets, which can be credited to the introduction of joint depression score and severity distributions. Overall, TREAT achieves competitive performance for facial depression recognition.

However, there are still some limitations of our work. In Eq. (1), the Gaussian distribution applies the same standard deviation parameter σ for all training samples, which ignores the different distributions of ambiguity across the ordinal BDI-II scores. For example, the training samples falling in the interval of "mild" and "moderate" should have higher ambiguity than those falling into the other two intervals and as a result, may attain larger standard deviations. Besides, TREAT borrows an existing network structure called HMHN[13] without further considering the structures of depression score and severity distributions, which may deteriorate its performance. In the future, we will mainly investigate how to design varying standard deviations for different training samples (e.g., by considering different training samples falling in different severity intervals) and how to design an effective network structure to consider the characteristics of the depression score and severity distributions. Moreover, we will consider fusion multimodality[47] to further improve the performance in the future work.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2022ZD0115303), the National Natural Science Foundation of China (No. 62202247), and the Foundation of Yunnan Key Laboratory of Service Computing (No. YNSC24123).

References

[1] R. H. Belmaker and G. Agam, Major depressive disorder, *N. Engl. J. Med.*, vol. 358, no. 1, pp. 55–68, 2008.

[2] M. Maj, D. J. Stein, G. Parker, M. Zimmerman, G. A. Fava, M. De Hert, K. Demyttenaere, R. S. McIntyre, T. Widiger, and H. U. Wittchen, The clinical characterization of the adult patient with depression aimed at personalization of management, *World Psychiatry*, vol. 19, no. 3, pp. 269–293, 2020.

[3] K. Mao, Y. Wu, and J. Chen, A systematic review on automated clinical depression diagnosis, *npj Ment. Health Res.*, vol. 2, p. 20, 2023.

[4] K. Mao, D. Wang, T. Zheng, R. Jiao, Y. Zhu, B. Wu, L. Qian, W. Lyu, J. Chen, and M. Ye, Analysis of automated clinical depression diagnosis in a Chinese corpus, *IEEE Trans. Biomed. Circuits Syst.*, vol. 17, no. 5, pp. 1135–1152, 2023.

[5] A. K. Kim, E. H. Jang, S. H. Lee, K. Y. Choi, J. G. Park, and H. C. Shin, Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach, *J. Med. Internet Res.*, vol. 25, p. e34474, 2023.

[6] C. Bourke, K. Douglas, and R. Porter, Processing of facial emotion expression in major depression: A review, *Aust. N. Z. J. Psychiat.*, vol. 44, no. 8, pp. 681–696, 2010.

[7] 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, 2015.

[8] 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, 2018.

[9] 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, 2020.

[10] S. Li and W. Deng, Deep facial expression recognition: A survey, *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.

[11] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, and J. Yu, Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals, *Inf. Fusion*, vol. 104, p. 102161, 2024.

[12] 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.*, vol. 13, no. 3, pp. 1605–1618, 2022.

[13] Y. Li, Z. Liu, L. Zhou, X. Xuan, Z. Shangguan, X. Hu, and B. Hu, A facial depression recognition method based on hybrid multi-head cross attention network, *Front. Neurosci.*, vol. 17, p. 1188434, 2023.

[14] A. T. Beck, R. A. Steer, R. Ball, and W. Ranieri, Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients, *J. Pers. Assess.*, vol. 67, no. 3, pp. 588–597, 1996.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1–9.

[16] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proc. of Advances in Neural Information Processing Systems 25* (*NIPS 2012*), 2017. https://proceedings.neurips.cc/paper_files/paper/2012.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An Image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv: 2010.11929, 2021.

[19] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng, Deep label distribution learning with label ambiguity, *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, 2017.

[20] X. Geng, Label distribution learning, *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, 2016

[21] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, AVEC 2013: The continuous audio/visual emotion and depression recognition challenge, in *Proc. 3rd ACM Int. Workshop on Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 3–10.

[22] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, AVEC 2014: 3D dimensional affect and depression recognition challenge, in *Proc. 4th Int. Workshop on Audio/Visual Emotion Challenge*, Orlando, FL, USA, 2014, pp. 3–10.

[23] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. M. Messner, et al., AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition, in *Proc. 9th Int. Audio/Visual Emotion Challenge and Workshop*, Nice, France, 2019, pp. 3–12.

[24] V. Ojansivu and J. Heikkila, Blur insensitive texture classification using local phase quantization, in *Proc. 3rd Int. Conf. Image and Signal Processing*, Cherbourg-Octeville, France, 2008, pp. 236–243.

[25] 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 on Audio/Visual Emotion Challenge*, Barcelona, Spain, 2013, pp. 11–20.

[26] A. Bosch, A. Zisserman, and X. Munoz, Representing shape with a spatial pyramid kernel, in *Proc. 6th ACM Int. Conf. Image and Video Retrieval*, Amsterdam, The Netherlands, 2007, pp. 401–408.

[27] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition, in *Proc. 10th IEEE Int. Conf. Computer Vision*, Beijing, China, 2005, pp. 786–791.

[28] A. Dhall and R. Goecke, A temporally piece-wise fisher vector approach for depression analysis, in *Proc. 2015 Int. Conf. Affective Computing and Intelligent Interaction*, Xi'an, China, 2015, pp. 255–259.

[29] T. Ahonen, A. Hadid, and M. Pietikainen, Face description with local binary patterns: Application to face recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.

[30] X. Geng, C. Yin, and Z. H. Zhou, Facial age estimation by learning from label distributions, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, 2013.

[31] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, Deep differentiable random forests for age estimation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 404–419, 2021.

[32] J. Wang, X. Geng, and H. Xue, Re-weighting large margin label distribution learning for classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5445–5459, 2022.

[33] X. Jia, X. Shen, W. Li, Y. Lu, and J. Zhu, Label distribution learning by maintaining label ranking relation, *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1695–1707, 2023.

[34] X. Jia, T. Qin, Y. Lu, and W. Li, Adaptive weighted ranking-oriented label distribution learning, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp.

11302–11316, 2024.
[35] J. Huang, C. M. Vong, G. Wang, W. Qian, Y. Zhou, and C. L. P. Chen, Joint label enhancement and label distribution learning via stacked graph regularization-based polynomial fuzzy broad learning system, *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 9, pp. 3290–3304, 2023.
[36] C. Wen, X. Zhang, X. Yao, and J. Yang, Ordinal label distribution learning, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision*, Paris, France, 2023, pp. 23424–23434.
[37] S. Li and W. Deng, Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning, *Int. J. Comput. Vis.*, vol. 127, no. 6, pp. 884–906, 2019.
[38] Y. Shu, P. Yang, N. Liu, S. Zhang, G. Zhao, and Y. J. Liu, Emotion distribution learning based on peripheral physiological signals, *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2470–2483, 2023.
[39] H. Kaya and A. A. Salah, Eyes whisper depression: A CCA based multimodal approach, in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 961–964.
[40] 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.*, vol. 13, no. 2, pp. 864–870, 2022.
[41] Z. Liu, X. Yuan, Y. Li, Z. Shangguan, L. Zhou, and B. Hu, PRA-Net: Part-and-relation attention network for depression recognition from facial expression, *Comput. Biol. Med.*, vol. 157, p. 106589, 2023.
[42] 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.*, vol. 37, no. 12, pp. 10140–10156, 2022.
[43] W. C. De Melo, E. Granger, and M. B. López, MDN: A deep maximization-differentiation network for spatio-temporal depression detection, *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 578–590, 2023.
[44] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, The PHQ-8 as a measure of current depression in the general population, *J. Affect. Disord.*, vol. 114, nos. 1–3, pp. 163–173, 2009.
[45] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, Dynamic multimodal measurement of depression severity using deep autoencoding, *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 525–536, 2018.
[46] J. Xu, H. Gunes, K. Kusumam, M. Valstar, and S. Song, Two-stage temporal modelling framework for video-based depression recognition using graph representation, *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 161–178, 2025.
[47] X. Lyu, S. Rani, S. Manimurugan, and Y. Feng, A deep neuro-fuzzy method for ECG big data analysis via exploring multimodal feature fusion, *IEEE Trans. Fuzzy Syst.*, vol. 33, no. 1, pp. 444–456, 2025.

Fan Zhang received the MEng degree from Jinzhou Medical University, Jinzhou, China, in 2019, and has been studying for the PhD degree from China Medical University, Shenyang, China since 2021. She is currently an attending physician at Department of Neurology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China. Her research interests include intelligent processing, transmission, reconstruction of medical data, and intelligent analysis of neurodegenerative diseases, and depression-related diseases.

Byung-Gyu Kim received the BEng degree from Pusan National University, Busan, Republic of Korea, in 1996, the MEng and PhD degrees from the Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 1998 and 2004, respectively. In March 2016, he joined Sookmyung Women's University, Seoul, Republic of Korea, where he is currently a full professor. He has published over 250 international journal articles and conference papers. His research interests include image processing, video coding techniques, deep/reinforcement learning algorithm, embedded multimedia systems, and intelligent information system.

Liang Dong received the BEng degree in applied physics from Dalian University of Technology, Dalian, China, in 2004, and the MEng degree in optics from Dalian University of Technology, Dalian, China, in 2008. He is a professor at School of Electrical Engineering, Liaoning University of Technology, Jinzhou, China. His research interests include artificial intelligence, security, and image processing.

Jing Wang received the BEng degree in computer science from Suzhou University of Science and Technology, Suzhou, China, in 2013, the MEng degree in computer science from Northeastern University, Shenyang, China, in 2015, and the PhD degree in software engineering from Southeast University, Nanjing, China, in 2021. He is currently an assistant researcher at School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include pattern recognition and machine learning.

Keqin Li received the BEng degree in computer science from Tsinghua University, Beijing, China, in 1985, and the PhD degree in computer science from University of Houston, Houston, TX, USA, in 1990. He is currently a SUNY distinguished professor at State University of New York, New Paltz, NY, USA, and a national distinguished professor at Hunan University, Changsha, China. He has authored or co-authored more than 960 journal articles, book chapters, and refereed conference papers. He received several best paper awards from international conferences including PDPTA-1996, NAECON-1997, IPDPS-2000, ISPA-2016, NPC-2019, ISPA-2019, and CPSCom-2022. He holds nearly 70 patents announced or authorized by the Chinese National Intellectual Property Administration. He is among the world's top five most influential scientists in parallel and distributed computing in terms of single-year and career-long impacts based on a composite indicator of the Scopus citation database. He was a 2017 recipient of the Albert Nelson Marquis Lifetime Achievement Award for being listed in Marquis Who's Who in Science and Engineering, Who's Who in America, Who's Who in the World, and Who's Who in American Education for over twenty consecutive years. He received the Distinguished Alumnus Award from Department of Computer Science, University of Houston in 2018. He received the IEEE TCCLD Research Impact Award from the IEEE CS Technical Committee on Cloud Computing in 2022 and the IEEE TCSVC Research Innovation Award from the IEEE CS Technical Community on Services Computing in 2023. He was a winner of the IEEE Region 1 Technological Innovation Award (Academic) in 2023. He is a member of the SUNY Distinguished Academy. He is an AAAS fellow, an IEEE fellow, an AAIA fellow, and an ACIS Founding fellow. He is a member of Academia Europaea (Academician of the Academy of Europe). His research interests include cloud computing, distributed computing, high-performance computing, etc.

Saru Kumari received the PhD degree in mathematics from Chaudhary Charan Singh University, Meerut, India, in 2012. She is currently an assistant professor with Department of Mathematics, Chaudhary Charan Singh University, Meerut, India. Her research interests include information security and security of wireless sensor network.

Jianhui Lv received the BEng degree in mathematics and applied mathematics from the Jilin Institute of Chemical Technology, Jilin, China, in 2012, and the MS and PhD degrees in computer science from Northeastern University, Shenyang, China, in 2014 and 2017, respectively. He worked at the Network Technology Laboratory, Central Research Institute, Huawei Technologies Co. Ltd, Shenzhen, China, as a senior engineer from January 2018 to July 2019. He worked at Tsinghua University, Beijing, China, as an assistant professor from August 2019 to July 2021. He is currently a professor at The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China, an associate professor at both Peng Cheng Laboratory, Shenzhen, China, and Tsinghua Shenzhen International Graduate School, Shenzhen, China. His research interests include computer networks, artificial intelligence, information centric networks, Internet of Things, bio-inspired networking, evolutionary computation, cloud/edge computing, smart city, healthcare, etc. He has published more than 90 high-quality journal (such as *IEEE JSAC*, *IEEE TON*, *IEEE TFS*, *IEEE TCSS*, *IEEE TVT*, *IEEE TGCN*, *IEEE TCE*, *IEEE IOTJ*, *ACM TOMM*, *ACM TOIT*) and conference (such as IEEE INFOCOM, IEEE/ACM IWQoS, ACM WWW, AAAI, and IEEE ICPADS) papers. He has served as the leader guest editors in several international journals (such as *Applied Soft Computing*, *Digital Communications and Networks*, *Expert Systems*, *Wireless Networks*, *International Journal on Artificial Intelligence Tools*, *Mobile Information Systems*, *Internet Technology Letters*) and the guest editor of *IEEE TCE*. In addition, he is also an associate editor of *Internet Technology Letters*, *Journal of Multimedia Information System*, and *International Journal of Swarm Intelligence Research*.