frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 January 2023
DOI 10.3389/fnins.2022.1107284

OPEN ACCESS

EDITED BY
Xiaopeng Hong,
Harbin Institute of Technology, China

REVIEWED BY
Lang He,
Xi'an University of Posts and Telecommunications, China
Yong Li,
Nanjing University of Science and Technology, China

*CORRESPONDENCE
Hongsheng Lu
✉ luhs@tzc.edu.cn

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

RECEIVED 24 November 2022
ACCEPTED 13 December 2022
PUBLISHED 06 January 2023

CITATION
Zhao X, Liao Y, Tang Z, Xu Y, Tao X, Wang D, Wang G and Lu H (2023) Integrating audio and visual modalities for multimodal personality trait recognition *via* hybrid deep learning.
Front. Neurosci. 16:1107284.
doi: 10.3389/fnins.2022.1107284

Integrating audio and visual modalities for multimodal personality trait recognition *via* hybrid deep learning

Xiaoming Zhao[1], Yuehui Liao[1,2], Zhiwei Tang[1], Yicheng Xu[3], Xin Tao[1], Dandan Wang[1], Guoyu Wang[1] and Hongsheng Lu[1]*

[1]Taizhou Central Hospital (Taizhou University Hospital), Taizhou University, Taizhou, Zhejiang, China, [2]School of Computer Science, Hangzhou Dianzi University, Hangzhou, China, [3]School of Information Technology Engineering, Taizhou Vocational and Technical College, Taizhou, Zhejiang, China

Recently, personality trait recognition, which aims to identify people's first impression behavior data and analyze people's psychological characteristics, has been an interesting and active topic in psychology, affective neuroscience and artificial intelligence. To effectively take advantage of spatio-temporal cues in audio-visual modalities, this paper proposes a new method of multimodal personality trait recognition integrating audio-visual modalities based on a hybrid deep learning framework, which is comprised of convolutional neural networks (CNN), bi-directional long short-term memory network (Bi-LSTM), and the Transformer network. In particular, a pre-trained deep audio CNN model is used to learn high-level segment-level audio features. A pre-trained deep face CNN model is leveraged to separately learn high-level frame-level global scene features and local face features from each frame in dynamic video sequences. Then, these extracted deep audio-visual features are fed into a Bi-LSTM and a Transformer network to individually capture long-term temporal dependency, thereby producing the final global audio and visual features for downstream tasks. Finally, a linear regression method is employed to conduct the single audio-based and visual-based personality trait recognition tasks, followed by a decision-level fusion strategy used for producing the final Big-Five personality scores and interview scores. Experimental results on the public ChaLearn First Impression-V2 personality dataset show the effectiveness of our method, outperforming other used methods.

KEYWORDS

multimodal personality trait recognition, hybrid deep learning, convolutional neural networks, bi-directional long short-term memory network, Transformer, spatiotemporal

1. Introduction

In personality psychology, researchers believe that human personality is innate, and have developed various theoretical methods to understand and measure a person's personality. Costa and McCrae (1998) proposed a personality trait theory, in which personality characteristic were referred to as the main factors affecting the characteristics of individual behaviors, the critical factor in forming personality traits, and the basic unit for measuring personality traits. In Vinciarelli and Mohammadi (2014) personality is defined as: "personality is a psychological construct that can explain the diversity of human behaviors on the basis of a few, stable and measurable individual characteristics." At present, researchers have used psychological scales to establish various personality traits models, including Big-Five (McCrae and John, 1992), Cattell sixteen personality factor (16PF) (Karson and O'Dell, 1976), Myers-Briggs type indicators (MBTI) (Furnham, 1996), Minnesota multiple personality inventory (MMPI) (Bathurst et al., 1997), and so on. Among them, the Big-Five model has become the most fashionable measure model for automatic personality trait recognition. In particular, the Big-Five model, also known as the OCEAN model, aims to measure a person's personality through five dipolar scales: openness (O), conscientiousness (C), extroversion (E), agreeableness (A), and neuroticism (N). In affective neuroscience, the neural mechanisms of emotion expression are investigated by means of combining neuroscience with the psychological study of personality, emotion, and mood (Montag and Davis, 2018; Wang and Zhao, 2022; Zhang et al., 2022).

In recent years, researchers have employed computational techniques such as machine learning and deep learning methods (Gao et al., 2020; Liang et al., 2021; Wang and Deng, 2021; Yan et al., 2021; Ye et al., 2021) to model and measure human personality from the first impression behavior data, which is called personality computing (Junior et al., 2019). One of the most important research subject in personality computing is automatic personality trait recognition, which aims to identify people's first impression behavior data by computer and then analyze people's psychological characteristics (Zhao et al., 2022). Personality trait recognition has significant applications to human emotional behavior analysis, human-computer interaction, and interview recommendation. For example, Zhao et al. (2019) explored the influence of personality on emotional behavior by means of a hypergraph learning framework. When an enterprise recruits, human resource department can leverage personality trait recognition techniques to analyze personality characteristics of the job seekers by collecting their first-impression behavior data, and then select employees who can better meet the needs of the enterprise. To advance the development of personality trait recognition, the 2016 European Conference on Computer Vision (ECCV) released a publicly available personality dataset, i.e., ChaLearn-2016, and organized an academic competition of personality trait recognition (Ponce-López et al., 2016). Since 2016, personality trait recognition has become a hot research topic in psychology, affective neuroscience, and artificial intelligence.

In a basic personality trait recognition system, two important steps are involved: feature extraction and personality trait classification or prediction (Zhao et al., 2022). Feature extraction aims to derive appropriate feature parameters related to the expression of personality traits from the acquired first impression behavioral data. Personality trait classification or prediction aims to employ machine learning methods to conduct personality classification or prediction. The conventional classifiers or regressors such as support vector machines (SVM) and linear regressors can be adopted for personality trait classification or prediction. This paper will focus on feature extraction in a personality trait recognition system.

According to the types of extracted features characterizing personality traits, personality trait recognition techniques can be divided into hand-crafted based methods and deep learning based methods. Based on the extracted hand-crafted or deep learning features, previous works (Zhao et al., 2022) focus on performing personality trait recognition from single modality, such as audio-based personality trait recognition (Mohammadi and Vinciarelli, 2012), visual-based personality trait recognition (Gürpınar et al., 2016), etc. Although these works based on single modality have achieved good performance, there are still two limitations for them. First, the people's first impression behavior data in real-world scenery are often multimodal rather than single-modal for characterizing personality traits. For instance, both verbal and non-verbal information such as audio and visual modality are highly correlated with personality traits. In this case, it is thus necessary to adopt multiple input modalities for personality trait recognition. Second, although deep learning methods have been fashionable for personality trait recognition, each of them has its advantages and disadvantages. Therefore, integrating the advantages of different deep learning methods may further improve the performance of personality trait recognition, which will be investigated in this work.

To address these two issues above-mentioned, this paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework. As depicted in **Figure 1**, the proposed method combines three different deep models, including convolutional neural networks (CNN) (LeCun et al., 1998; Krizhevsky et al., 2012), bi-directional long short-term memory network (Bi-LSTM) (Schuster and Paliwal, 1997), recently emerged Transformer (Vaswani et al., 2017), to learn high-level audio-visual feature representations, followed by a decision-level fusion strategy for final personality trait recognition. In particular, for audio feature extraction, the pre-trained deep audio CNN model called VGGish (Hershey et al., 2017) is

used to learn high-level segment-level audio features. For visual feature extraction, the pre-trained deep face CNN model called VGG-Face (Parkhi et al., 2015) is leveraged to separately learn high-level frame-level global scene image features and local facial image features from each frame in dynamic video sequences. Then, these extracted deep audio-visual features are fed into a Bi-LSTM and a Transformer network (Vaswani et al., 2017) to individually capture long-term temporal dependency, thereby producing the final global audio and visual features for downstream tasks. Finally, a linear regression method is employed to conduct the single audio-based and visual-based personality trait recognition tasks, and yield six independent personality trait prediction scores. A decision-level fusion strategy is adopted to merge these personality trait prediction scores and output the final Big-Five personality scores and interview scores. Extensive experiments is conducted on the public ChaLearn First Impressions-V2 dataset (Escalante et al., 2017), and demonstrate the effectiveness of the proposed method on personality trait recognition tasks.

The main contributions of this paper are summarized as follows:

(1) This paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework, in which CNN, Bi-LSTM, and Transformer are combined to capture high-level audio-visual spatio-temporal feature representations for personality trait recognition.
(2) Extensive experiments are performed on the public ChaLearn First Impressions-V2 dataset and experimental results show that the proposed method outperforms other comparing methods on personality trait recognition tasks.

2. Related work

The majority of prior works for personality trait recognition concentrates on single modality such as audio or visual cues, as described below.

2.1. Audio-based personality trait recognition

In early works, the conventional extracted hand-crafted audio features are low-level descriptor (LLD) features including intensity, pitch, formants, Mel-Frequency Cepstrum Coefficients (MFCCs), and so on. Mohammadi and Vinciarelli (2012) derived the LLD features like intensity, pitch, and formants, and then employed a logistic regression to predict the Big-five personality traits in audio clips. An et al. (2016) extracted the typical Interspeech-2013 ComParE feature set (Schuller et al., 2013) and fed them into a SVM classifier to conduct the Big-Five personality trait recognition.

In recent years, researchers have tried to leverage deep learning (LeCun et al., 2015) models with a multilayer network structure to learn high-level audio feature representations for promoting the performance of personality trait recognition. Among them, the representative deep learning methods are CNN (LeCun et al., 1998; Krizhevsky et al., 2012), recurrent neural networks (RNN) (Elman, 1990) and its variants called long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), etc. Hayat et al. (2019) proposed an audio personality feature extraction method based on CNN. They fine-tuned the pre-trained CNN model called AudioSet in the first-impression behavior dataset and extracted high-level audio features for Big-Five personality prediction, demonstrating the advantages of CNN-based learned features compared with hand-crafted features. Zhu et al. (2018) presented a method of automatic perception of speakers' personality from speech in Mandarin. They developed a new skip-frame LSTM system to learn personality information from frame-level descriptor like MFCCs instead of hand-crafted prosodic features.

2.2. Visual-based personality trait recognition

In terms of the input type of visual data, visual-based personality trait recognition can be divided into two groups: static images-based and dynamic video sequences-based personality trait recognition.

For static images-based personality trait recognition, the extracted visual features mainly come from facial features, since facial morphology provides explicit cues for personality trait recognition. In early works, the commonly used hand-crafted facial features are color histograms, local binary patterns (LBP), global descriptor, aesthetic features, etc. Guntuku et al. (2015) extracted low-level hand-crafted features of facial images, including color histograms, LBP, global descriptor, and aesthetic features, and then employed the lasso regressor to predict the Big-five personality traits of users in self-portrait images. Recently, deep learning methods have been applied for static images-based personality trait recognition. Xu et al. (2021) explored the relationship between self-reported personality characteristics and static facial images. They investigated the performance of several deep learning models pre-trained on the ImageNet data, such as MobileNetv2, ResNeSt50, and the designed personality prediction neural network based on soft thresholding (S-NNPP) by means of fine-tuning them on the self-constructed dataset composed of facial images and personality characteristics.

For dynamic video sequences-based personality trait recognition, dynamic video sequences contain temporal information related to facial activity statistics, thereby providing

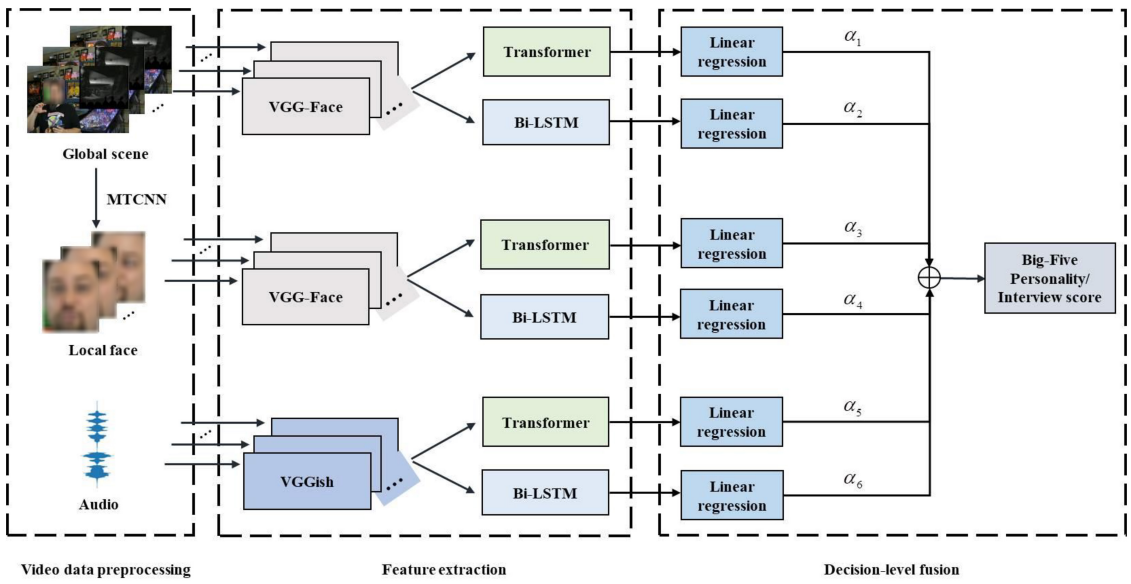

FIGURE 1
The flowchart of the proposed multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework.

useful and complementary cues for personality trait recognition (Junior et al., 2019). In early works, the hand-crafted video features related to facial activity statistics were usually adopted for personality trait recognition. Teijeiro-Mosquera et al. (2014) exploited the relationships between facial expressions in dynamic video sequences and personality impressions of the Big-Five traits. To characterize facial activity statistics, they extracted four kinds of behavioral cues for personality trait recognition, including statistic-based cues, Threshold (THR) cues, Hidden Markov Models (HMM) cues, and Winner Takes All (WTA) cues. Likewise, several recently developed deep learning methods have been employed for dynamic video sequences-based personality trait recognition. Gürpınar et al. (2016) extracted deep facial and scene feature representations in dynamic video sequences by fine-tuning a pre-trained VGG-19 model, and then input them into a kernel extreme learning machine to perform the prediction of Big-Five personality traits. Beyan et al. (2021) presented a classification method of perceived personality traits on the basis of novel deep visual activity (VA)-based features derived only from key-dynamic images in dynamic video sequences. They adopted a dynamic image construction, which aimed to learn long-term VA with CNN + LSTM, and detect spatiotemporal saliency to decide key-dynamic images.

3. The proposed method

To alleviate the problem of single modality based personality trait recognition, this paper proposes a multimodal personality trait recognition method integrating audio and visual modalities based on a hybrid deep learning framework. **Figure 1** depicts the flowchart of the proposed method. As depicted in **Figure 1**, the proposed method adopts two modalities as its input: one is the audio signals, the other is the visual signals including the global scene images and facial images. The used hybrid deep learning framework comprises of three different deep learning models like CNN, Bi-LSTM, and Transformer, which are used for high-level feature learning tasks. The proposed method consists of three key steps: video data preprocessing, audio-visual feature extraction, and decision-level fusion, as described below.

3.1. Video data preprocessing

For audio signals in the video data, we use the pre-trained VGGish model (Hershey et al., 2017) to extract high-level audio segment-level features. It is noted that the length of speech segments as input of VGGish is required to be 0.96 s. To this end, the original audio signals in the video data are divided into to a certain number of adjacent segments which last a time period of 0.96 s.

For visual signals in the video data, two preprocessing tasks are implemented. For global scene images in a video, 100 scene images are selected at equal intervals form each original video sample. Then, the resolution of each global scene image is resampled from the original 1280×720 pixels to 224×224 as inputs of VGG-Face model (Parkhi et al., 2015). For local face images in a video, we employ the popular Multi-Task Convolutional Neural Network (MTCNN) (Zhang et al., 2016)

to conduct face detection tasks. The resolution of face image detected in each frame is sampled to 224 × 224. Since some videos are affected by environmental factors such as illumination, MTCNN may detect face images with a low accuracy. As a tradeoff, 30 frames of detected face images are selected at equal intervals from the original video. For the video with less than 30 frames of detected face images, the first and last face images are repeatedly until the frame number of face video is 30.

3.2. Audio-visual feature extraction

Audio-visual feature extraction aims to learn the local and global feature representations from original audio and visual signals in a video for personality trait recognition, as described below.

3.2.1. Audio-visual local feature extraction

For the divided audio segment with 0.96 s, we leverage the VGGish model (Hershey et al., 2017) pre-trained on the AudioSet dataset (Gemmeke et al., 2017) to capture high-level segment-level deep audio features. The used VGGish model consists of 6 convolutional layers, 4 pooling layers, and 3 fully connected layers. The kernel size of convolutional layers and pooling layers is 3×3 and 2×2, respectively. Since the neuron number of the last fully connected layer in the VGGish network is 128, the learned audio features by the VGGish model are 128-dimension.

For each scene and face image in a video, we employ the VGG-Face model (Parkhi et al., 2015) pre-trained on the ImageNet dataset (Deng et al., 2009) to learn high-level frame-level deep visual feature representations for downstream scene and face global feature learning tasks, respectively. The VGG-Face model includes 13 convolution layers, 5 pooling layers, and 2 fully connected layers. Since the neuron number of the last full connection layer in the VGG-Face network is 4096, the dimension of visual frame-level features obtained by VGG-Face network is 4096.

Given i-th input video clip a_i $(i = 1, 2, \cdots N)$ and its corresponding Big-Five personality score y_i, we fine-tune the pre-trained VGGish network (Hershey et al., 2017) to obtain deep segment-level audio feature representations, as described below:

$$\min_{W^{VG}, \theta^{VG}} \sum_{i=1}^{N} L(\text{sigmoid}(W^{VG} \eta^{VG}(a_i; \theta^{VG})), y_i) \tag{1}$$

where $\eta^{VG}(a_i; \theta^{VG})$ represents the output of the last full connected layer in the VGGish network. θ^{VG} and W^{VG} separately denotes the network parameters of the VGGish network and the weights of the sigmoid layer. The cross-entropy loss function L is defined as:

$$L(VG, y) = -\sum_{j=1}^{N} y_j \log(y_j^p) \tag{2}$$

where y_j is the j-th ground-truth Big-Five personality score, and y_j^p is represented by the predicted Big-Five personality score.

For deep visual scene and face feature extraction on each frame of video, we fine-tune the pre-trained VGG-Face network (Parkhi et al., 2015) to learn high-level visual feature representations. The process of fine-tuning the pre-trained VGG-Face network is similar to the above-mentioned Eqs 1, 2.

3.2.2. Audio-visual global feature extraction

After completing the local audio and visual feature extraction tasks, it is necessary to individually learn the global audio features, visual scene features, and visual face features from the entire videos so as to conduct personality trait prediction tasks. To this end, we adopt the Bi-LSTM (Schuster and Paliwal, 1997) and recently emerged Transformer (Vaswani et al., 2017) to independently model long-term dependencies of temporal dynamics in video sequences, as described below.

Given an input sequence e_t, the learning process of the Bi-LSTM network is:

$$E = \text{Bi} - \text{LSTM}(W_{Bi-LSTM}, e_t) \tag{3}$$

where $E \in \mathbb{R}^{1 \times d}$ is the learned temporal features, and $W_{Bi-LSTM}$ is weight parameters of Bi-LSTM.

The original Transformer (Vaswani et al., 2017) is developed based on self-attention mechanisms like a Multi-Head attention without any recurrent structures and convolutions. A Multi-Head attention module consists of several Scaled Dot-Product Attention (SDPA) modules in parallel and then their outputs are concatenated as an input of a linear layer. Given the input query (Q), key (K), and value (V), the output of each SDPA module is defined as:

$$\text{Attention}\,(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where d_k is the feature dimension of the key matrix K.

3.3. Decision-level fusion

After obtaining audio-visual global features extracted by a Bi-LSTM model and a Transformer model, we adopt a linear regression layer to predict the Big-Five personality and interview scores. The linear regression layer is calculated as follows:

$$f_i(x) = x_i w_i + b \tag{5}$$

where x_i, w_i, and b represent the i-th input sample, the corresponding weight value, and bias, respectively. $f_i(x)$ is the i-th prediction score value.

As shown in Figure 1, when using the learned audio features, visual scene features, and visual face features as inputs of a linear regression layer, we can obtain six different recognition results. To effectively fuse these six different recognition results, a weighted decision-level fusion strategy is employed, as described below:

$$\tilde{f}(x) = \sum_{i=1}^{6} \alpha_i f_i(x) \tag{6}$$

where α_i is the weight value, $f_i(x)$ is the predicted value of each type of features, and $\sum_{i=1}^{6} \alpha_i = 1$. The mean squared error (MSE) loss is computed as follows:

$$\mathrm{MSE}(\tilde{f}(X)) = E[(\tilde{f}(X) - Y)^2] = E\left[\left(\sum_{i=1}^{6} \alpha_i (f_i(X) - Y)\right)^2\right] \tag{7}$$

where Y is the ground-truth score. Our goal is to minimize the MSE loss subject to $\sum_{i=1}^{6} \alpha_i = 1$. To this end, the Lagrangian expression of this problem is expressed as:

$$L(X, \lambda) = \mathrm{MSE}(\tilde{f}(X)) - \lambda\left(\sum_{i=1}^{6} \alpha_i - 1\right) \tag{8}$$

where λ is the Lagrange multiplier.

Then, we calculate the partial derivation of Eq. 8 based on α_m for $m = 1, 2, \cdots 6$, as defined as:

$$\frac{\partial L(X, \lambda)}{\partial \alpha_m} = E\left[2\sum_{i=1}^{6} \alpha_i (f_i(X) - Y)(f_m(X) - Y)\right] - \lambda \tag{9}$$

We set the gradient to be 0, and get:

$$2\sum_{i=1}^{6} \alpha_i E[(f_i(X) - Y)(f_m(X) - Y)] - \lambda = 0, m = 1, 2, \cdots 6 \tag{10}$$

Let $\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6]^T$, $\Omega = [w_{ij}] = E[(f_i(X) - Y)(f_j(X) - Y)]$, Eq. 10 can be transformed as:

$$\Omega\alpha = \frac{\lambda}{2} 1 \tag{11}$$

Then, the optimal weight vector $\boldsymbol{\alpha}$ can be obtained by:

$$\alpha = \frac{\Omega^{-1} 1}{1^T \Omega^{-1} 1} \tag{12}$$

4. Experiments

4.1. Dataset

To verify the effectiveness of the proposed method, the public ChaLearn First Impression-V2 (Escalante et al., 2017) is employed for personality and interview prediction. This dataset contains 10,000 video clips collected from more than 3,000 different YouTube videos. The language involved in video participants is English. The resolution of the video is 1280×720, and the duration of each video clip is about 15 s. This dataset annotates the "Interview" scene labels for interview analysis. The divided training set, testing set and validation set in this dataset contain 6,000, 2,000, 2,000 video clips, respectively. In this work, we use the training and validation sets for experiments because the testing set is only open to competitors. Each video in this dataset is labeled by using the Big-Five personality score [0,1]. Figure 2 shows several image samples from the ChaLearn First Impression-V2 dataset.

4.2. Implementation details

When training all used deep learning models, the batch size is set to 32, and the initial learning rate is $1 \times e^{-4}$. After each epoch, the learning rate will become a half of the original learning rate. The maximum epoch number of is 30, and the Adam optimizer is used. The MSE loss function is adopted. The experimental platform is NVIDIA GPU Quadro M6000 with 24 GB memory. In order to improve the generalization performance of trained deep learning models and avoid overfitting, the early stopping strategy (Prechelt, 1998) is used.

In this work, we choose a two-layer Bi-LSTM to capture temporal dynamics related to video sequences. The number of neurons in each layer of Bi-LSTM is 2048. The number of encoding layer in the Transformer model is 6 for its best performance, and its last layer output 1024-dimension features. To compare with these deep learning models, several classical regression models such as Support Vector Regression (SVR) with polynomial (poly), radial basis function (RBF), and linear kernel functions, Decision Tree Regression (DTR) are employed. In the SVR model, the degree of polynomial kernel function is 3, the penalty factor "C" of radial basis kernel function is 2, and the parameter "gamma" is 0.5. The DTR model is implemented for its default parameters, such as the splitting policy "split = best" at each node, "min _ samples _ split = 2" for splitting an internal node. For these classical regression models, a simple average-pooling strategy is conducted on these extracted audio-visual local features so as to produce the global features as their inputs.

The evaluation metric for evaluating the predicted personality trait or interview scores is defined as:

$$S = 1 - \sum_{j=1}^{N} \frac{\left|y_j^p - y_j\right|}{N} \tag{13}$$

where N is the number of samples, y_j^p is the predicted value, and y_j is the ground-truth value. The higher the value S is, the better the obtained performance on personality or interview prediction tasks is.

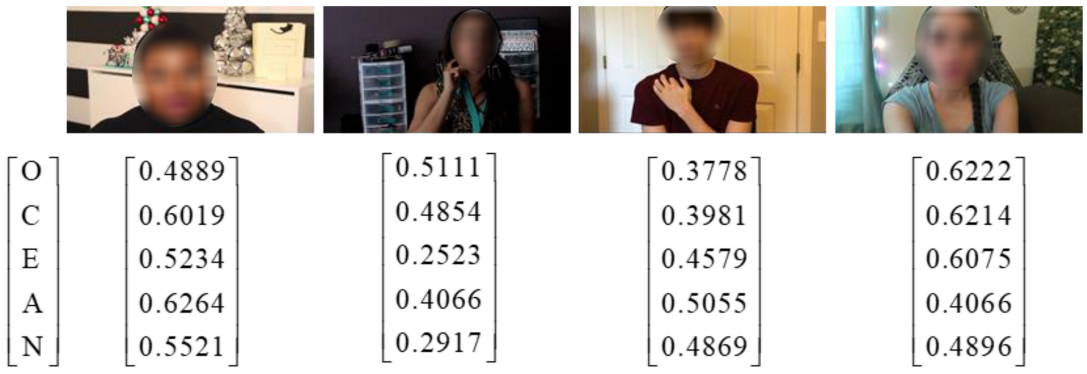

FIGURE 2
Image samples with the labeled Big-Five personality score from the ChaLearn First Impression-V2 dataset.

TABLE 1 Prediction results of deep audio features extracted by the pre-trained VGGish for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8540	0.8329	0.8624	0.8402	0.8744	0.8528	0.8319
SVR (rbf)	0.8967	0.8844	0.8932	0.9012	0.8906	0.8932	0.8920
SVR (linear)	0.8980	0.8846	0.8935	0.9025	0.8920	0.8941	0.8945
DTR	0.8541	0.8411	0.8542	0.8610	0.8453	0.8511	0.8511
Transformer	0.8972	0.8814	0.8920	0.9035	0.8907	0.8930	0.8915
Bi-LSTM	0.8986	0.8834	0.8932	0.9045	0.8928	0.8945	0.8947
Transformer + Bi-LSTM	**0.8989**	**0.8847**	**0.8938**	**0.9048**	**0.8935**	**0.8952**	**0.8953**

Bold values denote the highest performance.

4.3. Experimental results and analysis

In this section, two groups of experiments are carried out on the ChaLearn First Impression-V2 data set to verify the effectiveness of all used methods. One is the single-modal personality trait recognition, the other is multi-modal personality trait recognition.

4.3.1. Results of single-modal personality trait recognition

For single-modal personality recognition, we present the experiment results and analysis based on the single extracted audio features, visual scene features, and visual face features by using the pre-trained deep models.

Table 1 shows the prediction results of deep audio features extracted by the pre-trained VGGish for different methods. "Transformer + Bi-LSTM" denotes that the learned features with Transformer and Bi-LSTM are directly concatenated to form a whole feature vector as inputs of the latter linear regression layer for prediction. It can be seen from Table 1 that Transformer + Bi-LSTM performs best based on deep audio features. More specially, the average Big-Five personality prediction score is 0.8952 and the corresponding interview prediction score of 0.8953, thereby outperforming other used methods. The ranking order for other used methods is Bi-LSTM, SVR (linear), SVR (rbf), Transformer, SVR (poly), and DTR. This shows the advantages of Transformer + Bi-LSTM on audio personality trait recognition tasks. It is noted that Transformer + Bi-LSTM performs better than Transformer and Bi-LSTM, indicating that there is a certain complementarity between Transformer and Bi-LSTM.

Tables 2, 3 separately present personality prediction results of deep visual scene features and deep visual face features extracted by the pre-trained VGG-Face for different methods. It can be observed from Tables 2, 3 that Transformer + Bi-LSTM still obtains better performance other methods. In particular, Transformer + Bi-LSTM employs deep visual scene features and face features to produce the average Big-Five personality prediction scores of 0.9039 and 0.9124, respectively, and the interview prediction scores of 0.9057 and 0.9163, respectively. The ranking order for other used methods is Bi-LSTM, Transformer, SVR (poly), SVR (linear), SVR (rbf), and DTR. This shows the superiority of Transformer + Bi-LSTM on deep visual (scene and face) personality trait recognition tasks. The visual face images outperforms the visual scene images on personality trait recognition tasks. This may be because face images are more correlated with personality traits than scene images.

TABLE 2 Prediction results of deep visual scene features extracted by the pre-trained VGG-Face for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8921	0.8896	0.8896	0.8962	0.8850	0.8905	0.8890
SVR (rbf)	0.8841	0.8736	0.8804	0.8963	0.8780	0.8825	0.8818
SVR (linear)	0.8896	0.8872	0.8867	0.8922	0.8809	0.8873	0.8865
DTR	0.8636	0.8607	0.8627	0.8711	0.8586	0.8633	0.8639
Transformer	0.8941	0.8844	0.8909	0.9021	0.8884	0.8920	0.8920
Bi-LSTM	0.9042	0.9013	0.9012	0.9091	0.8993	0.9030	0.9050
Transformer + Bi-LSTM	**0.9043**	**0.9025**	**0.9035**	**0.9093**	**0.9000**	**0.9039**	**0.9057**

Bold values denote the highest performance.

TABLE 3 Prediction results of deep visual face features extracted by the pre-trained VGG-Face for different methods.

Models	O	C	E	A	N	Average score	Interview score
SVR (poly)	0.8871	0.8922	0.8923	0.8980	0.8855	0.8910	0.8963
SVR (rbf)	0.8841	0.8736	0.8804	0.8963	0.8780	0.8825	0.8818
SVR (linear)	0.8953	0.8922	0.8986	0.8974	0.8913	0.8950	0.8960
DTR	0.8714	0.8683	0.8702	0.8760	0.8674	0.8706	0.8721
Transformer	0.9023	0.9000	0.9029	0.9068	0.8968	0.9017	0.9017
Bi-LSTM	0.9103	**0.9155**	0.9129	0.9135	0.9085	0.9121	0.9161
Transformer + Bi-LSTM	**0.9110**	0.9148	**0.9130**	**0.9143**	**0.9087**	**0.9124**	**0.9163**

Bold values denote the highest performance.

TABLE 4 Comparisons of recognition results obtained by different methods.

Modality	O	C	E	A	N	Average score	Interview score
A	0.8989	0.8847	0.8938	0.9048	0.8935	0.8952	0.8953
S	0.9043	0.9025	0.9035	0.9093	0.9000	0.9039	0.9057
F	0.9110	0.9148	0.9130	0.9143	0.9087	0.9124	0.9153
A + S + F (EF)	0.9145	**0.9176**	0.9171	0.9158	0.9121	0.9154	0.9178
A + S + F (MF)	0.9151	0.9172	0.9156	0.9150	0.9123	0.9150	0.9180
A + S + F (LF)	**0.9167**	0.9163	**0.9176**	**0.9177**	**0.9150**	**0.9167**	**0.9200**

A, audio; S, scene; F, face; EF, early fusion; MF, model-level fusion; LF, late fusion. Bold values denote the highest performance.

In summary, the results in Tables 1–3 demonstrate that for single-modal personality recognition the visual face features perform best on personality trait and interview prediction tasks, followed by deep visual scene features and deep audio features. This shows that the facial images related to facial expression contain more discriminant information for personality trait recognition.

4.3.2. Results of multimodal personality trait recognition

For multimodal personality recognition tasks, we compare the performance of three typical multimodal information fusion methods, such as feature-level fusion, decision-level fusion, and model-level fusion. In feature-level fusion, the audio-visual global features learned by Bi-LSTM and Transformer networks, are concatenated into a whole feature vector as input of the linear regression layer for personality trait prediction. In this case, feature-level fusion is also called early fusion (EF). In model-level fusion (MF), the concatenated audio-visual global features are fed into a 4-layer full-collection layer network (1024-512-256-128) for personality trait prediction. In decision-level fusion, we adopt Eq. 12 to obtain the analytical solution of the optimal weight values in Eq. 6. In this case, decision-level fusion is also called late fusion (LF).

Table 4 presents the comparisons of recognition results obtained by different fusion methods such as EF, MF, and LF, as

TABLE 5 Comparisons with other existing methods.

References	Modality	Feature extraction	Fusion methods	Average score
Güçlütürk et al., 2016	Audio, visual	Audio:ResNet-17 Visual:ResNet-17	EF	0.9109
Güçlütürk et al., 2017	Audio, visual, text	Audio:ResNet-17 Visual:ResNet-17 Text:skip-thought vectors	EF	0.9118
Wei et al., 2017	Audio, visual	Audio:MFCCs Visual:DAN	LF	0.9130
Principi et al., 2021	Audio, visual	Audio:1D CNN Visual:ResNet-50	MF	0.9160
Escalante et al., 2022	Audio, visual, text	Audio:ResNet-18 Visual:ResNet-18 Text: skip-thought vectors	LF	0.9161
Ours	Audio, visual	Audio:VGGish Visual:VGG-Face	LF	**0.9167**

EF, early fusion; MF, model-level fusion; LF, late fusion. Bold values denote the highest performance.

well as the single modality methods. From the results in **Table 4**, we can see that: (1) among three used fusion methods, the used LF method combining audio, scene, and face obtains the best performance with an average score of 0.9167 on personality trait recognition tasks, and an average score of 0.9200 on interview prediction tasks. For personality trait recognition, the used EF method slightly outperforms the MF method, yielding an average score of 0.9154. By contrast, the used MF method slightly outperforms the EF method on interview prediction tasks. In particular, the MF method gives an average interview score of 0.9180. (2) All used fusion methods such as LF, MF, and EF provide superior performance to the single modality methods. This indicates the complementarity to some extent among audio, scene, and face modality on target recognition tasks.

4.3.3. Comparisons with other existing methods

To further verify the effectiveness of the proposed method, **Table 5** presents the comparisons of different used methods. **Table 5** shows that the proposed method obtains an average score of 0.9167, which is better than other reported results obtained by audio, visual, and text modalities. This demonstrates the advantage of our method on personality trait recognition tasks. These comparing works are described as follows.

Güçlütürk et al. (2016) provided an audio-visual personality trait recognition based on 17-layer deep residual networks (ResNet-17). They concatenated the learned features of audio-visual streams at feature-level as an input of a fully connected layer and reported an average score of 0.9109 for final personality trait prediction. In this case, the used network does not need any feature engineering or visual analysis like face detection, face landmark alignment. Similarly, they also presented an multimodal personality trait analysis integrating audio, visual, and text modalities by using the 17-layer deep residual networks (Güçlütürk et al., 2017). Here, they extracted skip-thought vectors as text features. They fused these modalities at feature-level and reported an average score of 0.9118. Wei et al. (2017) presented a deep bimodal regression method of personality traits on short video sequences. For audio modality, they extracted MFCCs and logfbank features. For visual modality, they employed a modified CNN model called Descriptor Aggregation Network (DAN) to extract visual features. Finally, they fused these predicted regression scores of audio-visual modalities at decision-level, and reported an average score of 0.9130. Principi et al. (2021) presented a multimodal deep learning method integrating the raw audio and visual modalities for personality trait prediction. For audio modality, a 14-layer 1D CNN was used for audio feature extraction. For visual modality, they employed a pre-trained ResNet-50 network for visual feature extraction. Finally, they employed a fully connected layer to jointly learn audio-visual feature representations at model-level for final personality trait recognition, and achieved an average score of 0.9160. Escalante et al. (2022) proposed a multimodal deep personality trait recognition method based on audio, visual, and text modalities. They adopted a ResNet-18 to extract audio and visual features, and skip-thought vectors as text features. Then, a late fusion strategy was utilized to fuse all three modalities, and yielded an average score of 0.9161.

5. Conclusion

This paper presents a multimodal personality trait recognition method based on CNN + Bi-LSTM + Transformer network. In this work, CNN, Bi-LSTM, and Transformer are combined to capture high-level audio-visual spatio-temporal feature representations for personality trait recognition. Finally, we compare multimodal personality prediction results based on three different fusion methods such as feature-level fusion, model-level fusion, and decision-level fusion. Experiments on the public ChaLearn First Impression-V2 dataset show that decision-level fusion achieves the best multimodal personality trait recognition results with an average score of 0.9167, outperforming other existing methods.

It is noted that this work only focuses on integrating audio and visual modalities for multimodal personality trait recognition. Considering the diversity of modal information related to the expression of personality traits, it is interesting to combine current audio-visual modalities with other modalities such as physiological signals, text cues, etc., to further improve the performance of personality trait recognition. In addition, exploring a more advanced deep learning model for personality trait recognition is also an important direction in our future work.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://chalearnlap.cvc.uab.cat/dataset/24/description/.

Author contributions

XZ contributed to the writing and drafted the article. YL, ZT, YX, XT, DW, and GW contributed to the data preprocessing and analysis, software and experiment simulation. HL contributed to the project administration and writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Natural Science Foundation of China (NSFC) and Zhejiang Provincial Natural Science Foundation of China under Grant Nos. 61976149, 62276180, LZ20F020002, and LQ21F020002.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

An, G., Levitan, S. I., Levitan, R., Rosenberg, A., Levine, M., and Hirschberg, J. (2016). "Automatically classifying self-rated personality scores from speech," in *Proceedings of the INTERSPEECH Conference 2016*, (Incheon: ISCA), 1412–1416. doi: 10.21437/Interspeech.2016-1328

Bathurst, K., Gottfried, A. W., and Gottfried, A. E. (1997). Normative data for the MMPI-2 in child custody litigation. *Psychol. Assess.* 9:205. doi: 10.1037/1040-3590.9.3.205

Beyan, C., Zunino, A., Shahid, M., and Murino, V. (2021). Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images. *IEEE Trans. Affect. Comput.* 12, 1084–1099. doi: 10.1109/TAFFC.2019.2944614

Costa, P. T., and McCrae, R. R. (1998). "Trait theories of personality," in *Advanced Personality*, eds D. F. Barone, M. Hersen, and V. B. Hasselt (Cham: Springer), 103–121. doi: 10.1007/978-1-4419-8580-4_5

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248–255. doi: 10.1109/CVPR.2009.5206848

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Escalante, H. J., Guyon, I., Escalera, S., Jacques, J., Madadi, M., Baró, X., et al. (2017). "Design of an explainable machine learning challenge for video interviews," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, (Piscataway, NJ: IEEE), 3688–3695. doi: 10.1109/IJCNN.2017.7966320

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Güçlütürk, Y., Güçlü, U., et al. (2022). Modeling, recognizing, and explaining apparent personality from videos. *IEEE Trans. Affect. Comput.* 13, 894–911. doi: 10.1109/TAFFC.2020.2973984

Furnham, A. (1996). The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Pers. Individ. Diff.* 21, 303–307. doi: 10.1016/0191-8869(96)00033-5

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Comput.* 32, 829–864. doi: 10.1162/neco_a_01273

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). "Audio set: an ontology and human-labeled dataset for audio events," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Piscataway, NJ: IEEE), 776–780. doi: 10.1109/ICASSP.2017.7952261

Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., et al. (2017). Multimodal first impression analysis with deep residual networks. *IEEE Trans. Affect. Comput.* 9, 316–329. doi: 10.1109/TAFFC.2017.2751469

Güçlütürk, Y., Güçlü, U., Van Gerven, M. A., and Van Lier, R. (2016). "Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proceedings of the European Conference on Computer Vision*, (Cham: Springer), 349–358. doi: 10.1007/978-3-319-49409-8_28

Guntuku, S. C., Qiu, L., Roy, S., Lin, W., and Jakhetiya, V. (2015). "Do others perceive you as you want them to? Modeling personality based on selfies," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, Association for Computing Machinery*, (New York, NY), 21–26. doi: 10.1145/2813524.2813528

Gürpınar, F., Kaya, H., and Salah, A. A. (2016). "Combining deep facial and ambient features for first impression estimation," in *Proceedings of the European Conference on Computer Vision*, (Berlin: Springer), 372–385. doi: 10.1007/978-3-319-49409-8_30

Hayat, H., Ventura, C., and Lapedriza, À (2019). On the use of interpretable CNN for personality trait recognition from audio. *CCIA* 319, 135–144.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Piscataway, NJ: IEEE), 131–135. doi: 10.1109/ICASSP.2017.7952132

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Junior, J. C. S. J., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., Baró, X., et al. (2019). First impressions: a survey on vision-based apparent personality trait analysis. *IEEE Trans. Affect. Comput.* 13, 75–95. doi: 10.1109/TAFFC.2019.2930058

Karson, S., and O'Dell, J. W. (1976). *A Guide to The Clinical Use of the 16 PF*. Chandigarh: Inst for Personality & Ability Test.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Cambridge, MA: MIT Press), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liang, Y., Li, S., Yan, C., Li, M., and Jiang, C. (2021). Explaining the black-box model: a survey of local interpretation methods for deep neural networks. *Neurocomputing* 419, 168–182. doi: 10.1016/j.neucom.2020.08.011

McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. *J. Personal.* 60, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

Mohammadi, G., and Vinciarelli, A. (2012). Automatic personality perception: prediction of trait attribution based on prosodic features. *IEEE Trans. Affect. Comput.* 3, 273–284. doi: 10.1109/T-AFFC.2012.5

Montag, C., and Davis, K. L. (2018). Affective neuroscience theory and personality: an update. *Personal. Neurosci.* 1:e12. doi: 10.1017/pen.2018.10

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, 411–412. doi: 10.5244/C.29.41

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., et al. (2016). "Chalearn lap 2016: first round challenge on first impressions-dataset and results," in *Proceedings of the European Conference on Computer Vision*, (Berlin: Springer), 400–418. doi: 10.1007/978-3-319-49409-8_32

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw.* 11, 761–767. doi: 10.1016/S0893-6080(98)00010-0

Principi, R. D. P., Palmero, C., Junior, J. C., and Escalera, S. (2021). On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Trans. Affect. Comput.* 12, 607–621. doi: 10.1109/TAFFC.2019.2956030

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon. doi: 10.21437/Interspeech.2013-56

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Teijeiro-Mosquera, L., Biel, J.-I., Alba-Castro, J. L., and Gatica-Perez, D. (2014). What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. *IEEE Trans. Affect. Comput.* 6, 193–205. doi: 10.1109/TAFFC.2014.2370044

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Cambridge, MA: MIT Press), 5998–6008.

Vinciarelli, A., and Mohammadi, G. (2014). A survey of personality computing. *IEEE Trans. Affect. Comput.* 5, 273–291. doi: 10.1109/TAFFC.2014.2330816

Wang, D., and Zhao, X. (2022). Affective video recommender systems: a survey. *Front. Neurosci.* 16:984404. doi: 10.3389/fnins.2022.984404

Wang, M., and Deng, W. (2021). Deep face recognition: a survey. *Neurocomputing* 429, 215–244. doi: 10.1016/j.neucom.2020.10.081

Wei, X. S., Zhang, C. L., Zhang, H., and Wu, J. X. (2017). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Trans. Affect. Comput.* 9, 303–315. doi: 10.1109/TAFFC.2017.2762299

Xu, J., Tian, W., Lv, G., Liu, S., and Fan, Y. (2021). Prediction of the big five personality traits using static facial images of college students with different academic backgrounds. *IEEE Access* 9, 76822–76832. doi: 10.1109/ACCESS.2021.3076989

Yan, A., Chen, Z., Zhang, H., Peng, L., Yan, Q., Hassan, M. U., et al. (2021). Effective detection of mobile malware behavior based on explainable deep neural network. *Neurocomputing* 453, 482–492. doi: 10.1016/j.neucom.2020.09.082

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2021). Deep learning for person re-identification: a survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intellig.* 44, 2872–2893. doi: 10.1109/TPAMI.2021.3054775

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* 23, 1499–1503. doi: 10.1109/LSP.2016.2603342

Zhang, S., Zhao, X., and Tian, Q. (2022). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Trans. Affect. Comput.* 13, 680–688. doi: 10.1109/TAFFC.2019.2947464

Zhao, S., Gholaminejad, A., Ding, G., Gao, Y., Han, J., and Keutzer, K. (2019). Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1–18. doi: 10.1145/3233184

Zhao, X., Tang, Z., and Zhang, S. (2022). Deep personality trait recognition: a survey. *Front. Psychol.* 13:839619. doi: 10.3389/fpsyg.2022.839619

Zhu, M., Xie, X., Zhang, L., and Wang, J. (2018). "Automatic personality perception from speech in mandarin," in *Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Taipei, 309–313. doi: 10.1109/ISCSLP.2018.8706692