International Journal of Computer Vision (2020) 128:2763–2780
https://doi.org/10.1007/s11263-020-01309-y

CR-Net: A Deep Classification-Regression Network for Multimodal Apparent Personality Analysis

Yunan Li[1,2] · Jun Wan[3,4] · Qiguang Miao[1,2] · Sergio Escalera[5] · Huijuan Fang[1,2] · Huizhou Chen[1,2] · Xiangda Qi[1,2] · Guodong Guo[6,7]

Received: 14 May 2019 / Accepted: 13 February 2020 / Published online: 17 March 2020

Abstract
First impressions strongly influence social interactions, having a high impact in the personal and professional life. In this paper, we present a deep Classification-Regression Network (CR-Net) for analyzing the Big Five personality problem and further assisting on job interview recommendation in a first impressions setup. The setup is based on the ChaLearn First Impressions dataset, including multimodal data with video, audio, and text converted from the corresponding audio data, where each person is talking in front of a camera. In order to give a comprehensive prediction, we analyze the videos from both the entire scene (including the person's motions and background) and the face of the person. Our CR-Net first performs personality trait classification and applies a regression later, which can obtain accurate predictions for both personality traits and interview recommendation. Furthermore, we present a new loss function called Bell Loss to address inaccurate predictions caused by the regression-to-the-mean problem. Extensive experiments on the First Impressions dataset show the effectiveness of our proposed network, outperforming the state-of-the-art.

Keywords Personality traits · Multimodal data · Convolutional neural networks · Classification-regression network · Bell Loss function

1 Introduction

The analysis of human affective behavior is an active research in computer vision nowadays, which can be widely used in a variety of applications, such as social relation analysis (Xia et al. 2017), analysis of depression (Klein et al. 2011) and job candidate screening (Naim et al. 2015; Ponce-López et al. 2016; Escalante et al. 2016), among others. Unconscious behaviors may produce facial expressions or words of a person that can reflect some traits of personality, influencing other people's impression about him/her. Evidence with psychological support has been shown in the case of job interviews (Barrick and Mount 1991). However, in real-world situations, estimating one's personality is still an open problem in psychology, linguistics and physiology (Wei et al. 2018). The advances in computer vision are providing support to advance the study of personality computing, ben-

Communicated by Wenjun Zeng.

✉ Jun Wan
jun.wan@ia.ac.cn

Yunan Li
yn_li@stu.xidian.edu.cn

Qiguang Miao
qgmiao@mail.xidian.edu.cn

Sergio Escalera
sergio@maia.ub.es

Guodong Guo
guoguodong01@baidu.com

[1] School of Computer Science and Technology, Xidian University, Xi'an, China

[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, China

[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[5] Universitat de Barcelona and Computer Vision Center, Barcelona, Spain

[6] Institute of Deep Learning, Baidu Research, Beijing, China

[7] National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China

efiting from the automatic analysis and recognition of facial expressions, audio, speech, scene, and so on (Zhang et al. 2016; Wei et al. 2018; Subramaniam et al. 2016; Güçlütürk et al. 2018; Kaya et al. 2017).

Apparent personality analysis is a key element in personality computing (Wei et al. 2018). Slightly different from real personality computing, apparent personality is that perceived by an observer regarding other people. In this paper, we focus on apparent personality analysis coming from first impressions scores on a large set of audio-visual recordings. Psychologists have proposed different models for describing personality traits. One of the most accepted models is the Big Five (Norman 1963). It involves five factors to provide a full picture of a person. The factors are *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. In this work, the Big Five is the personality trait model which we analyze on the ChaLearn First Impressions dataset (Escalante et al. 2018).

Psychologists have studied personality for decades (Corr and Matthews 2009; Pennebaker and King 1999; Mairesse and Walker 2007; Polzehl et al. 2010; Mohammadi and Vinciarelli 2015), with the questionnaire as the preferred choice in order to quantitatively estimate apparent personality traits scores (Corr and Matthews 2009). Further research has been done with the analysis of communication content (Pennebaker and King 1999; Mairesse and Walker 2007), audio (Polzehl et al. 2010; Mohammadi and Vinciarelli 2015) and biological signals (Zhao et al. 2018; Correa et al. 2018), among others. Owing to the advances of deep learning and the new Chalearn multimodal personality datasets (Escalante et al. 2018) released in computer vision, new insights have been presented in the area of personality computing (Xia et al. 2017; Basu et al. 2018; Wei et al. 2018; Subramaniam et al. 2016; Güçlütürk et al. 2016a; Kaya et al. 2017; Bekhouche et al. 2017; Ventura et al. 2017).

Predicting the apparent personality in the case of Big Five trait model is essentially a regression task. Either traditional machine learning-based methods or recent deep learn-ing-based methods can perform the regression task by mapping the set of features into real value scores. The mean square error (MSE) loss is usually used for this optimization. One problem with the MSE loss is the prediction of extreme values. When training with a batch of data that has ground truth scores covering a large range of values, it is common the optimization process produces predictions near the mean of ground truth scores in order to minimize the loss. This is even more pronounced in case where the ground truth scores follow a Gaussian distribution. This phenomenon is called "regression-to-the-mean" problem (Wang et al. 2018), and harms the proper regression of extreme values in predictions.

The recently released *First Impressions v2 dataset* (Escalante et al. 2018) contains apparent personality scores for the Big Five traits of people in video sequences. The annotation of recommendations on the invitation to job interview is also available in the dataset. It shows a high agreement among raters/observers regarding apparent personality (first impressions) annotations, as well as the interview recommendation variable (Escalante et al. 2018). We analyze the Big Five traits based on this dataset.[1]

In order to estimate personality traits and further provide assistance to job interview recommendation from short video sequences, we propose a deep Classification-Regression Network (CR-Net). The architecture of our network is shown in Fig. 1. In order to take the benefit of different modalities provided in the data, we first separate the visual and audio inputs from the video sequence. For the visual input, we consider both facial expressions and the motion of the person. We adopt a multi-focus scheme, which decomposes scene and face cues to pay attention to the global and local information about the person, respectively. For the audio input, we consider both input audio channel and the corresponding transcriptions, which are fused in early stages of the network. Compared to some existing combinations of classification and regression (Rothe et al. 2015; Huang and Ramanan 2017; Niu et al. 2016; Chen et al. 2018; Gao et al. 2018), Our CR-Net network applies classification features as a guidance to derive more discriminative features for regression. In this way, classification features are used to optimize the regression search space. Since the classification features are used as guidance rather than the input for regression, it can avoid the propagation of classification errors to the regression. Furthermore, in order to mitigate the regression-to-the-mean problem, we design a new loss function, called the Bell Loss, of which the shape is like an inverted Bell to avoid the minimum of the loss obtained at the mean value of the samples in a batch. Finally, with a weighted fusion, the multi-focus visual features and audio-text features are integrated and fed into the Extra Tree Regressor (ETR) for the final prediction.

The main contributions of this paper are threefold:

1. A network that takes features for classification as a guidance for regression. Unlike previous approaches that combine classification and regression predictions/outputs, we use the ResNet-34 as the backbone network for classification, and use classification features to guide the regression optimization.
2. A new loss function, namely the Bell Loss, is proposed. In order to mitigate the regression-to-the-mean problem (Wang et al. 2018) related to traditional MSE-oriented

[1] Note that our aim is to perform an analysis of our network and loss proposal in order to enhance first impressions recognition. We do not argue that interview recommendation variable has a direct application in real scenarios. Different jobs require different competences and studying automatic recommendation of job profiles is out of the scope of this work.

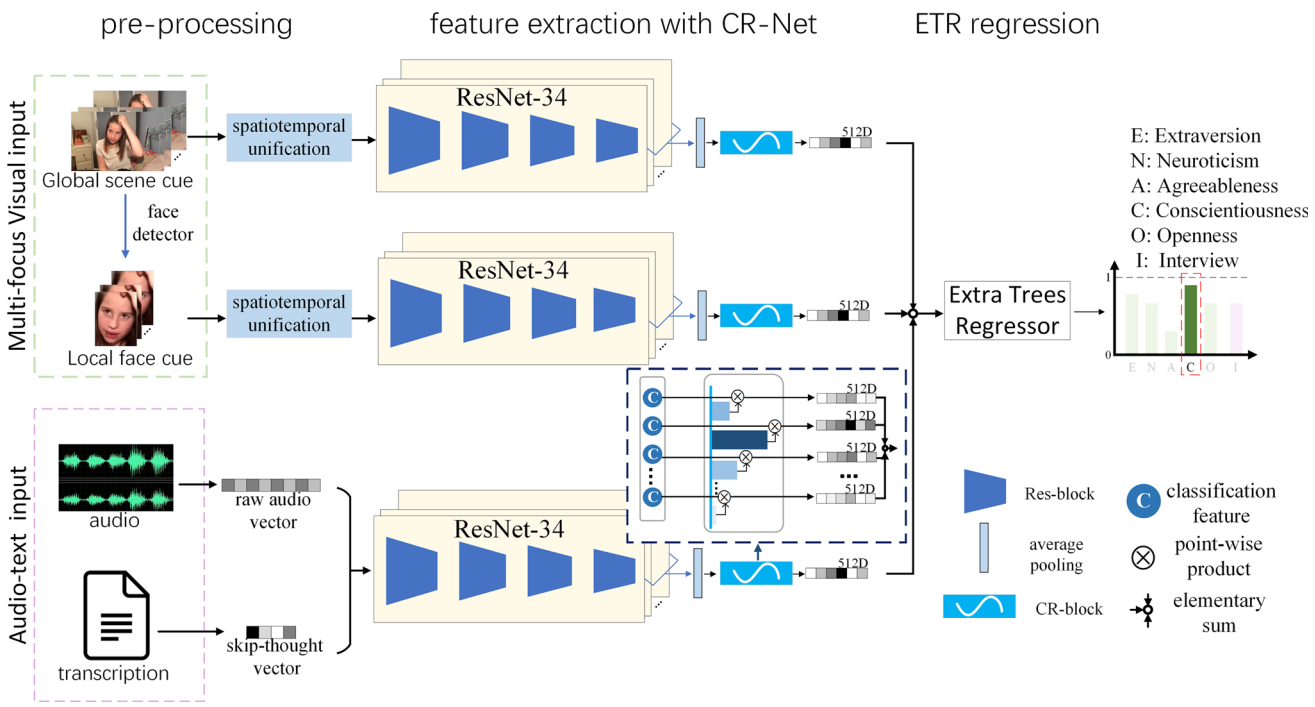

Fig. 1 Pipeline of the proposed method for personality trait prediction. The prediction process for trait *Conscientiousness* is taken as an example. We first split the input video sequence into the visual and audio-text input streams. For the visual stream, we employ a multi-focus scheme to capture the global scene cue and local face cue from the whole frame and the facial region (Zhang et al. 2016), respectively. The audio-text input is a fused feature of the audio vector and skip-thought vector of the transcription. Then different inputs are fed to the CR-Net. For each trait, we have one CR-block module in the network using the classification features as a guidance for regression. The features extracted by the CR-net are fused and sent to Extra Trees Regressor to produce final regression scores of the trait *Conscientiousness*. The other Big Five traits and the job interview recommendation variable can be obtained in the same way

loss functions, we design a new loss function inspired by the Gaussian curve. The Bell Loss has a high gradient even though the divergence between the prediction and label is small, which results in more robust regression results.

3. A comprehensive study on multimodal data for apparent personality analysis using the First Impressions dataset (Escalante et al. 2018). We exploit different modalities from the input audio-visual data. The video input is separated into global and local cues, whereas the audio and transcription are early fused in the network, considering their inner relations. The features of all modalities are fused for the final prediction.

The remainder of the paper is organized as follows. Related works are presented in Sect. 2, with the particular focus on apparent personality analysis, techniques combining classification and regression, and solutions to the regression-to-the-mean problem. Section 3 describes our CR-Net and the Bell Loss. Experiments are presented in Sect. 4. A discussion of the contributions of each module of our model is presented in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related Works

2.1 Learning personality traits from different modalities

Most of the research in the apparent personality analysis has been done from a linguistic data analysis perspective (Pennebaker and King 1999; Mairesse and Walker 2007). Audio data has also been widely considered (Polzehl et al. 2010; Mohammadi and Vinciarelli 2015). Physiologists also studied biological signals, like Electroencephalogram (EEG), Electrocardiogram (ECG) and Galvanic Skin Response (GSR) to relate them to affect or personality traits (Zhao et al. 2018; Correa et al. 2018). Facial expressions have played a very important role in relation to personality traits. Basu et al. (2018) used the RGB facial image together with the simultaneously obtained thermogram image to predict affective states. However, a single modality may not provide enough information about the personality. Words and audio features can also be influenced by environmental factors like background noise or sudden breaks, whereas the static image may only contain some instantaneous actions as a response to concrete stimulus rather than reflecting the apparent personality.

On the other hand, capturing signals such as EEG requires specific hardware, which may not be accessible for general applications. At the same time, the undesirable bias in capturing EEG signals can also affect the personality perception.

Recently, predicting apparent personality traits from social media has raised the attention of researchers. Several works to compute apparent personality from audio-visual data have been recently published thanks to the 2016 and 2017 ChaLearn Looking at People First Impressions Challenges (Zhang et al. 2016). Wei et al. (2018) employed a newly proposed Descriptor Aggregation Network (DAN) incorporating ResNet (He et al. 2016) to extract visual features and LSTM (Hochreiter and Schmidhuber 1997) to model handcrafted audio features. This work achieves a high performance, but relies on the combination of several models, which brings high memory and computational requirements. Ventura et al. (2017) employed a similar DAN+ network for inferring apparent personality traits from single facial images. However, it just proves that the facial image contains relevant information regarding personality. It lacks a comprehensive analysis of the effect of motion and additional information cues. Subramaniam et al. (2016) developed two bi-modal deep CNNs using audio and face images. One is based on 3D convolution networks (Ji et al. 2013) and the other is based on LSTM. This again suffers from high computational requirements. Güçlütürk et al. (2016a) employed an audiovisual deep residual network, which is based on the ResNet (He et al. 2016) and has two streams to process visual and audio data separately. The features of these two branches are fused for final regression. It is also extended to use audio transcriptions in Güçlütürk et al. (2018). Researchers further tried to predict interviewing recommendations score from the First Impressions dataset based on interviewees' personality (Escalante et al. 2018). Kaya et al. (2017) used pre-trained VGG-Face (Parkhi et al. 2015) and VGG-VD19 (Simonyan and Zisserman 2014) network to extract features of face and the whole frame, and used the openSMILE (Eyben et al. 2010) to extract audio features. Then they extracted features with Extreme Learning Machines (ELMs) (Huang et al. 2001) and obtained the final prediction with Random Forest (RF). Bekhouche et al. (2017) performed face alignment and extracted Pyramid Multi-Level features to train five Support Vector Regressors (SVRs) corresponding to Big Five traits. Most of the previous models rely on finetuning of existing pre-trained models. For example, the pre-trained VGG-Face model is commonly used (Zhang et al. 2016; Wei et al. 2018; Kaya et al. 2017).

2.2 Techniques Combining Classification and Regression

Some researchers combined classification schemes in order to support/guide regression problems. Apparent age estimation is an example that benefited from this combination, whe-re the simplest approach is to map the regression into a classification task. Rothe et al. (2015) defined a set of 101 possible output classes, which correspond to 101 discrete age values from 0 to 100, and use softmax value as the expectation to produce the final prediction. A similar idea, but grouping into age groups was performed by Tan et al. (2018). Niu et al. (2016) and Chen et al. (2018) dealt with apparent age estimation with an ordinal regression, which transforms the problem into several simple binary classifiers or CNNs. Based on the study of deep label distribution learning, Gao et al. (2018) used softmax to calculate the distribution of labels and used a L1 loss to predict the age. It can be understood as an extension of Rothe et al. (2015), which learns to regress with the expectation of all ages.

In the above examples, though age is continuous it can still be defined as a series of discrete integers, which is easy to fit into classification models. However, in the scenario of apparent personality, the score of personality traits is a real number that requires a precision up to four decimals according to recent publications. Therefore, directly treating apparent personality prediction as a classification task is not appropriate. Thus, the way which we combine the advantages of classification and regression is designed differently from the methods mentioned above. Unlike ordinal regression techniques, which aggregate ranking classification results for final prediction, we keep a regression loss to obtain accurate personality predictions. On the other hand, compared to Gao et al. (2018) which applied final regression based on classifier outputs, we sum up the weighted features used for classification as regression input. In this way, features for classification are also considered by the regressor, benefiting the final regression even in cases where the classification outputs may be wrong.

2.3 Dealing with Regression-to-the-Mean Problem

Regression-to-the-mean is a statistic phenomenon (Bland and Altman 1994a, b), whi-ch indicates a variable is extreme in its first measurement but closer to the mean in its second measurement.[2] In learning-based tasks, it occurs for regressions with MSE-oriented loss. An example of applications where it may happen is pixel-wise super-resolution task (Wang et al. 2018), where the extreme value of pixels (like 1 or 255) is always predicted towards the mean value (128). That is because if predicting towards the mean value, the average MSE loss can be lower when training with a large amount of data. The general solution in pixel-wise scenarios is to introduce some loss functions like perceptual loss (Johnson et al. 2016) and adversarial loss (Ledig et al. 2017), which take semantic information into consideration

[2] https://en.wikipedia.org/wiki/Regression_toward_the_mean.

to alleviate the regression-to-the-mean problem. However, the features for predicting apparent personality are highly abstract, with no clear definition of semantic information. Therefore, we designed a new loss function, namely Bell Loss, for the apparent personality regression problem. The shape of the Bell Loss is like an inverted Bell, and in this way, the decreasing loss ranges of samples with different ground truth values are not overlapped. Therefore, the Bell Loss avoids the minimum of the sum of the loss for samples in a mini-batch being obtained at the mean value of them, and then the regression-to-the-mean problem can be alleviated.

3 Methodology

In this section, we present the proposed CR-Net for apparent personality analysis and job interview recommendation. As depicted in Fig. 1, it has three main steps: data pre-processing, feature extraction with CR-Net and the ETR regression. For data pre-processing, we first split the input video sequence into the multi-focus visual input and the audio-text input. The multi-focus visual input involves two visual cues, which pay attention to both global scene and face, respectively. The audio-text input is a mixture of audio data and audio transcription encoded as skip-thought vectors (Kiros et al. 2015). The CR-Net takes the ResNet-34 as the backbone network, and uses the classification features as guidance for the regression process towards the end of the network. Finally, features of different modalities are fused and the final prediction is obtained via ETR.

3.1 Data Pre-processing

We perform a series of pre-processing steps for different data modalities, which help the inputs meet the requirement of CNNs for data.

3.1.1 Video Stream

To effectively extract features of video data and meet CNN requirements, it is better to sample the videos spatially and temporally. First, each video is sampled with 32 frames. This number is fixed to strike the balance between the effective feature extraction of motion relevant cues and computational requirements (Li et al. 2016, 2017). In order to avoid overfitting, when sampling the video sequences we cut the original video into 32 segments, and randomly select only one frame from each segment and resize it to 112 × 112 to fit the network input. Note that the random mechanism makes the selected frames vary through the entire training process. In other words, the frames of one video selected in each epoch can be different. In this way, we can augment the number of training samples online. After such a spatiotemporal pre-processing on the video sequence, we obtain the input stream for the network.

Fig. 2 An example of the complementarity of facial image and the whole scene image. **a** The facial expression. **b** The whole scene

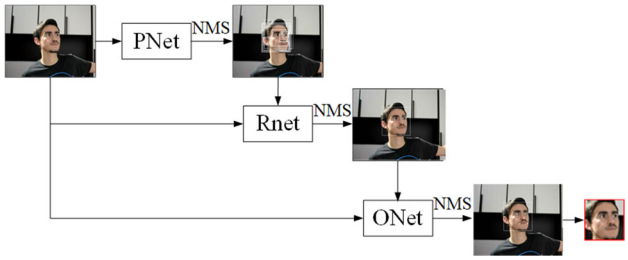

Fig. 3 The pipeline of face detection with MTCNN (Zhang et al. 2016). This network uses three subnets and Non-maximum suppression (NMS) to generate bounding boxes and merges them to output the facial region and keypoints

Previous methods for the first impressions task always focused on the face region, since it contains facial expressions relevant to personality. However, additional scene cues can contain complementary information. This may include actions, clothes, hair styles and even the background. Figure 2 is a good example.[3] If we only judge from the facial expression in Fig. 2a, we may think this man is irritable. However, with the whole scene in Fig. 2b, we understand that he is in a speech and is very impressive. Therefore, we divide the visual stream into global scene cue and local facial cue. The facial cue is obtained by detecting the face region (Zhang et al. 2016) with the pipeline as shown in Fig. 3. Both global and local visual cues are separate inputs of CR-Net, being each processed by an independent ResNet-34 in order to extract visual features.

3.1.2 Audio-Text Stream

Another input stream is audio and transcription. In this stream, the main data is the audio from the video sequence, which may reflect one's personality by the variations of speaking

[3] Images are from Lisa Feldman Barrett's Keynote speech "From Essences to Predictions: Understanding the Nature of Emotion" on European Society for Cognitive and Affective Neuroscience 2018.

manners and tones. There is no extra pre-processing but converting the acoustic wave to fixed-length vectors. To satisfy the requirement of input length of CNNs, we convert the acoustic wave to fixed-length vectors with a popular python audio library, *librosa*. With it we first load the audio as the floating series with the sampling rate of 16000. Then we record the maximum of length of the vectors. Most vectors are with the same length, under that sampling rate. Then for the vectors shorter than this, we extend their length by zero-padding. In this way, all the audio vectors are with the same length.

The audio transcription is easily accessible using speech-to-text techniques. In our case, we use the audio transcriptions already provided within the dataset as a complementary input to the audio channel. We adopt the skip-thought vector (Kiros et al. 2015) since it has shown to be effective and provide a compact representation of texts (Güçlütürk et al. 2018). Given the inherent correlation between audio and text, they are early-fused by a concatenation, as shown in Fig. 1, as the input of one ResNet-34 stream of CR-Net.

3.2 CR-Net Architecture

The CR-Net Network is based on ResNet-34 (He et al. 2016). The unique characteristic of the proposed CR-Net is the module of CR-block, as illustrated in Fig. 4. Unlike the previous networks that combine classification and regression, we do not directly use the expectancy of the classification as the result (Rothe et al. 2015), or the outputs of the classifier (Gao et al. 2018) as the input to regression. In this study, the ResNet-34 is used as the backbone for both classification and regression processes. At the first stage, we obtain the classification features with the cross-entropy loss. Then at the second stage, we generate weights from the classification features via the softmax function and yield a weighted sum of these features. We then use it for regression with the MSE, L1 and our proposed Bell Loss.

The optimizations for classification and regression are different. Using the cross-entropy loss, we just consider the probability of the samples belonging to the ground truth class during classification. The one-hot encoding strategy makes the optimization focus on the right distribution of the space of labels. However, regression is performed with the MSE loss, of which the response varies with the distance between prediction and ground truth. Having all samples contribute equally in the loss and making all of them within the same range of regression values lead the network to move all regression predictions to the ground truth mean score. If we want to take benefit of classification as a guidance for regression, the one-hot classification prediction can be used to estimate which sub-interval the sample belongs to. This will guide the regression through pruned ranges of values in order to achieve more accurate predictions.

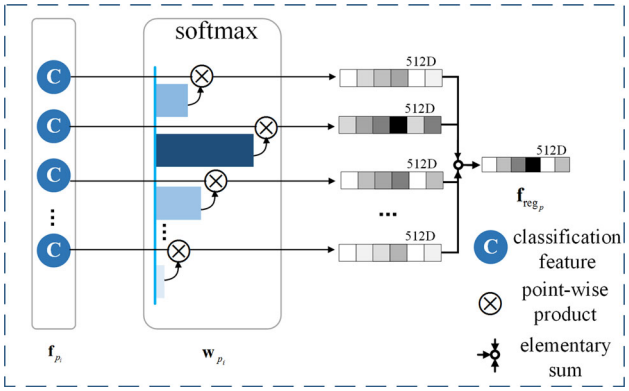

Fig. 4 The structure of CR-block. When predicting personality trait p, it first derives n weight values via a softmax layer for the corresponding n features $\{\mathbf{f}_{p_1}, \mathbf{f}_{p_2}, \ldots, \mathbf{f}_{p_n}\}$, which are obtained from the backbone ResNet-34. Here n is in accord with the number of classes, and the n weight values sum 1. Then These values are extended to 512-dimension vectors $\{\mathbf{w}_{p_1}, \mathbf{w}_{p_2}, \ldots, \mathbf{w}_{p_n}\}$, which serve as weights of their corresponding features. The weighted sum of the features is used as input for regression $\mathbf{f}_{p_{\text{reg}}}$

For each personality trait p, define n-class vector $\mathbf{C}_p = [C_{p_1}, C_{p_2}, \ldots, C_{p_n}]$, corresponding to n sub-intervals in the range of [0, 1]. In the CR-block, we obtain n features from the convolutional layer as:

$$\mathbf{f}_{p_i} = \mathcal{F}(\mathbf{a}, \theta_{p_i}), \tag{1}$$

where $\mathcal{F}$ is a convolutional process with parameter θ_{p_i} for the ith class and $\mathbf{a}$ is the output of the previous average pooling layer. Then the softmax function can be used to obtain the probability of the samples belonging to class i:

$$w_{p_i} = \frac{\exp(\mathbf{f}_{p_i})}{\sum_{i=1}^{n} \exp(\mathbf{f}_{p_i})}. \tag{2}$$

We take w_{p_i} as the weight for each feature $\mathbf{f}_{p_i}$ and sum them together:

$$\mathbf{f}_{p_{\text{reg}}} = \sum_{i=1}^{n} \mathbf{f}_{p_i} \odot \mathbf{w}_{p_i}, \tag{3}$$

where $\mathbf{w}_{p_i}$ is a vector extended by scalar w_{p_i} to match the dimension of $\mathbf{f}_{p_i}$ and $\odot$ indicates the Hadamard product. By summing the features, the regression can learn based on classes guidance. The personality prediction can be obtained with the following conditioned function:

$$\hat{y}_p = G(\mathbf{f}_{p_{\text{reg}}}, 2_p | \Psi_p), \tag{4}$$

where $\hat{y}_p$ is the personality prediction, G is the convolutional layer with parameter 2_p that maps $f_{p_{\text{reg}}}$ to $\hat{y}_p$, and Ψ_p defines the condition upon which the mapping function is

conditioned, and it can be expressed as:

$$\Psi_p = \mathbf{C}_p = [C_{p_1}, C_{p_2}, \ldots, C_{p_n}]. \quad (5)$$

3.3 Bell Loss

3.3.1 Limitations of Regression Loss Function

Image we have a batch of samples with the ground truth of 0.4, 0.5 and 0.6. When calculating the loss value, it always sums the differences between predictions and ground truth values according to the loss formulation. As shown in Fig. 5a, b, the shapes of the MSE and L1 losses in this batch obtain their minimum when the ground truth equals 0.5, being the mean of the range. Therefore, in the optimization with loss functions like MSE or L1 loss, the prediction tends to approach the mean value to assure a low loss. This is how the problem of regression-to-the-mean occurs in traditional regression tasks.

Another problem related to current regression loss functions is the accuracy of the predictions. As shown in Fig. 6, the gradient of MSE loss, which is marked in green dash line, decreases as the prediction approximates the ground truth. It may be insignificant if we do not need a high precision of the result. However, for a task like the first impressions requiring precision up to four decimals, this problem makes the optimization harder to give an accurate prediction. L1 loss marked in red dash-and-dotted line has an invariant gradient, which does not help either. Therefore, a loss function that keeps the gradient large enough when the predicted value is close to the ground truth value would be desirable to further help optimization and improve predictions.

3.3.2 Bell Loss Details

The design of the Bell Loss is inspired by the Gaussian curve, on which the gradient rises while the variable x approaches the expectation μ. To meet the demand of loss function, we define it as:

$$\mathcal{L}_{bell} = \gamma \left(1 - \mathrm{e}^{-\frac{(y-\hat{y})^2}{2\sigma^2}}\right), \quad (6)$$

where y and $\hat{y}$ are ground truth and prediction, respectively, σ is the derivation parameter that controls the amplitude of variation. A smaller σ leads to a higher gradient. γ is a scale parameter, which changes the loss value and makes it consistent with the other loss functions.

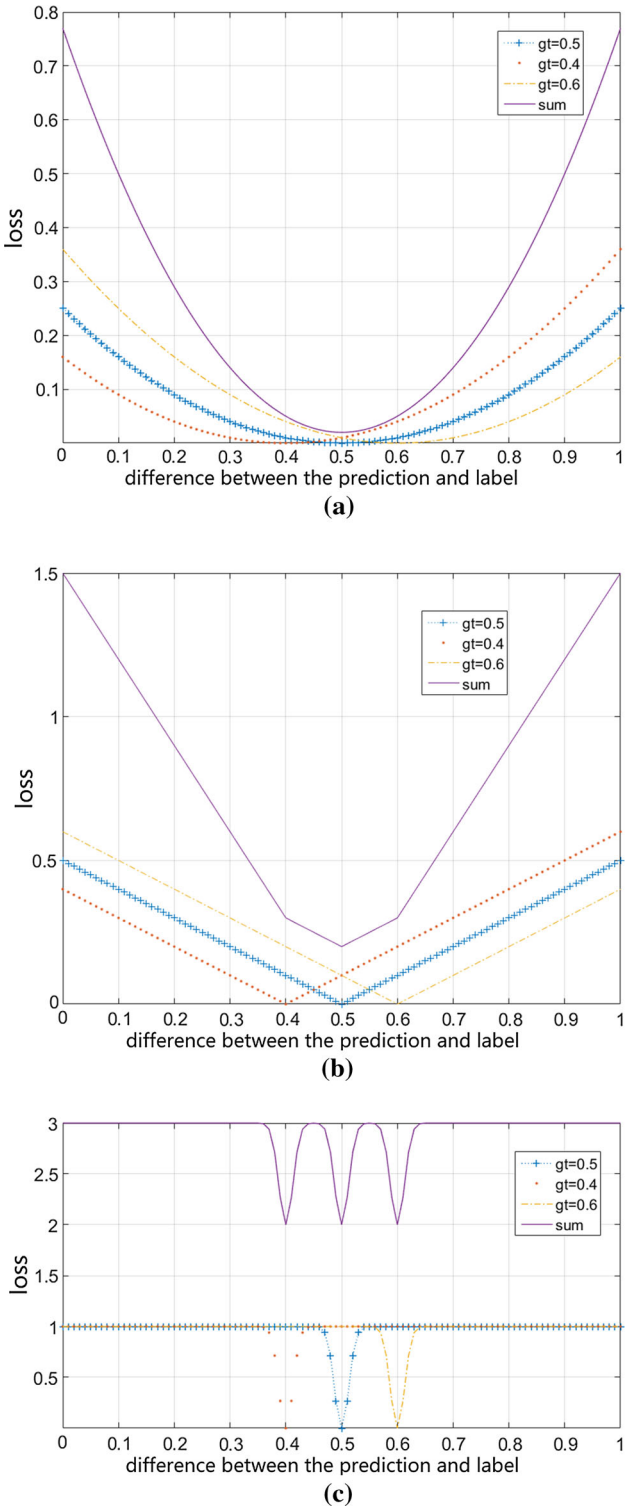

Fig. 5 An example of summing losses of a mini-batch with ground truth of 0.4, 0.5 and 0.6. **a** MSE loss. **b** L1 loss. **c** Bell loss

3.4 Regression with ETR

After learning with CR-Net, we extract the features for regression from each sub-network stream, namely global scene video stream, local face video stream and the audio-text stre-am. To obtain the final scores of personality and interview recommendation, the features of these streams are

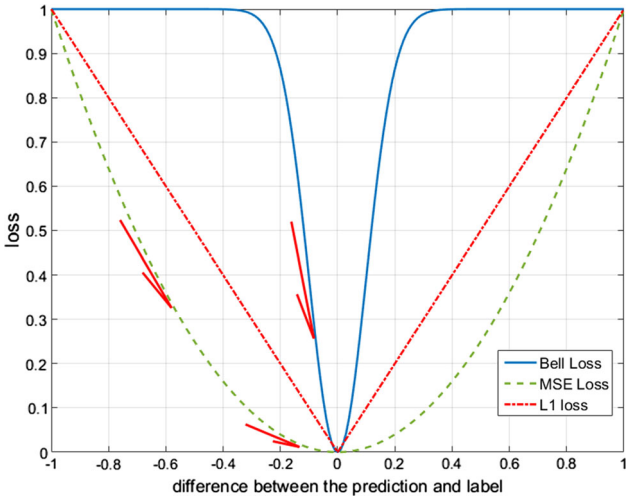

Fig. 6 A sketch of MSE loss, L1 loss and Bell Loss. The gradient of MSE loss moves rapidly when the prediction is far from the ground truth value but slowly when it becomes closer. It may hinder a precise prediction. On the contrary, the Bell Loss can produce a higher gradient even when the prediction is closer to ground truth score

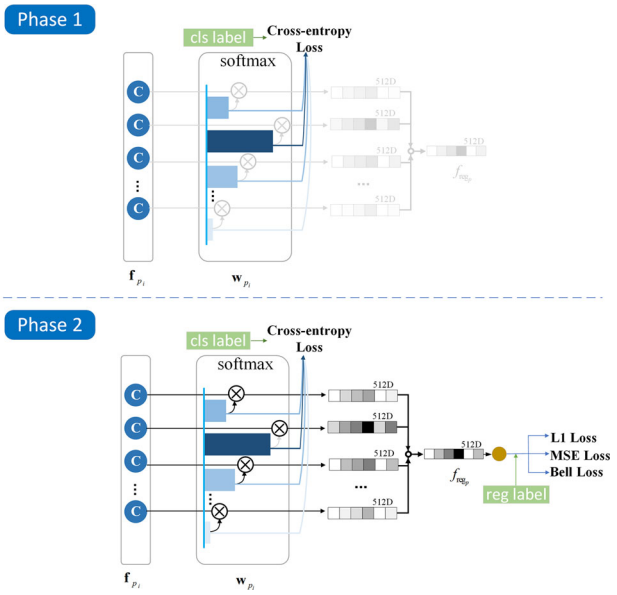

Fig. 7 Illustration of the two-stage training process. During the first stage, only the classification branch is trained using Cross-entropy Loss. At the second stage, the whole network is jointly trained with L1 loss, MSE loss, Bell Loss, and Cross-entropy Loss

integrated. In this paper, we use a late weighted fusion to obtain the final feature set. The weight for the fusion is empirically fixed as 7:5:3, emphasizing different levels of importance for the three streams.

The final regression is performed with Extra Trees Regressor (ETR). To the best of our knowledge, this is the first work to use ETR for apparent personality analysis. ETR is a regressor based on Extra-Trees algorithm (Geurts et al. 2006), which is a kind of tree-based ensemble learning method. ETR can be regarded a kind of extension of Random Forest with two main differences: 1) each tree in ETR is trained with the whole training data rather than applying bootstrapping, 2) instead of using the locally optimal cut-point for splitting the tree learner, ETR selects a random cut-point. From all the randomly generated splits, the split that yields the highest score is chosen to split the node.

For the regression process, ETR generates k decision trees, and randomly select m features for each training sample. At each node of the decision tree, it selects a cut-point at random. This random selection further benefits the network to be robust against overfitting to some extent. In our model, we empirically set $k = 1000$, and $m = 512$, as the control of the complete set of features.

3.5 Network Training

As illustrated in Fig. 7, the CR-Net is trained with a two-stage scheme. In the first stage, we train the classification branch with the objective function $\mathcal{L}_c$:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{P}\sum_{c=1}^{C}\rho_{p_c}^{i} log(\rho_{p_c}^{i}), \tag{7}$$

where N, P and C denote the number of samples, personality traits and classes, respectively, indexed by i, p and c.

In the second stage, we jointly optimize the entire network with a multi-task loss function:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{MSE} + \mathcal{L}_{bell} + \lambda\mathcal{L}_c, \tag{8}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{MSE}$ are the L1 and MSE losses, respectively, $\mathcal{L}_{bell}$ is the proposed Bell Loss. λ is the regularization parameter for $\mathcal{L}_c$, changing values related to the training iterations. We set it as:

$$\lambda = \frac{4 * E_{\max}}{(E+1)}, \tag{9}$$

where E indicates the current epoch and $E_{\max}$ is the maximum number of epochs. It decreases along with the training iterations, since at latter stages the classification branch becomes more stable.

4 Experimental Results

In this section, we first introduce the dataset used for the experiments. Then we describe the network implementation details and the protocols for evaluating the performance. Finally, we provide comparisons with the state-of-the-art.

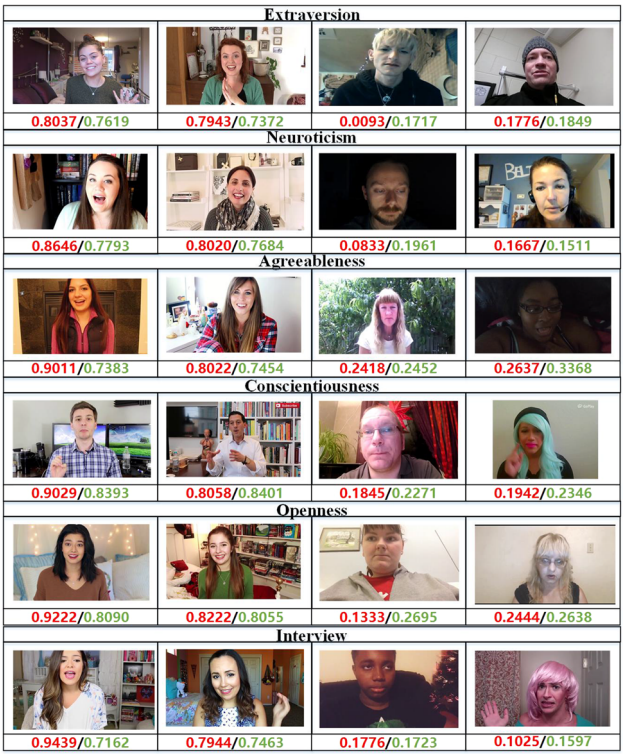

Fig. 8 Some examples of the First Impressions dataset. Each contains one person speaking in front of the camera. For each video, the apparent Big Five traits scores are provided. Traits scores are in the range [0, 1]. Each video also contains a scalar indicating the recommendation to invite him/her to a job interview based on the average of several raters' opinions. The *ground truth/our prediction* pairs are given under each sample video frame

4.1 Database

For training and evaluation, we choose the *First Impressions v2 dataset* (Escalante et al. 2018), which is an extension of the First Impressions v1 dataset (Ponce-López et al. 2016) and was used at the ChaLearn Explainable Computer Vision Multimedia and Job Candidate Screening Competition. As aforementioned, it is a representative dataset since it is the most relevant and largest public dataset on the topic of audio-visual personality perception. Some examples of this dataset together with the ground truth/our prediction are illustrated in Fig. 8. It comprises 10000 clips extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. This dataset can be divided into three parts: the training subset, the validation subset and the testing subset, of which the number of videos is split with the ratio 3:1:1. For each video, the Big Five personality scores are annotated with Amazon Mechanical Turk (AMT) workers. Meanwhile, in the First Impressions v2 dataset, the words of the video clips are transcribed by the professional transcription service Rev. The annotations of job interviewing completed by the AMT workers are also available, which are presented by real values in the range [0, 1] as well. The higher scores mean higher probabilities for the candidates being invited to the interviewing.

4.2 Experimental Setup

4.2.1 Training Parameters

Our experiments are conducted with PyTorch (Paszke et al. 2017) toolbox on a NVIDIA M6000 GPU. The training process is performed in two stages. For the first stage of optimization, we use the stochastic gradient descent (SGD) (Krizhevsky et al. 2012) with the initial learning rate, momentum and weight decay fixed to 0.002, 0.9 and 0.005, respectively. We feed the network with a mini-batch of 30 video sequences and audio-text. For the second stage, we use the Adam algorithm (Kingma and Ba 2014). The initial learning rate and weight decay are the same as the first stage, while β_1 and β_2 are set to 0.5 and 0.999, respectively. The batch size is set the same as the first stage. The learning rate decreases to its 1/10 after every 10 epochs, and the optimization is stopped after 50 epochs.

Note that the value of the traits is between 0 and 1. In this way, the MSE may be as small as 10^{-6} in some cases. In order to avoid a very small MSE to produce a vanishing gradient, we multiplied by 100 the value of both ground truth and prediction scores. And also, to balance the loss value of MSE and Bell Loss, we empirically set the parameters of Bell Loss $\sigma = 9$ and $\gamma = 300$.

4.2.2 Number of Classes

In the CR-Net, one important parameter is C_p, namely the number of classes. Using more classes means a fine division of the intervals. However, it does not always result in a better performance because the features are hard to reflect the subtle differences among classes when over-classified. A better solution is to consider C_p in terms of the distribution of the dataset. Based on this, we draw the histogram of the distribution of the Big Five traits, which is depicted in Fig. 9. We divide the interval stepped by 0.1, and derive 10 sub-intervals. As can be seen in Fig. 9, the data distribution is unbalanced. Most scores are within the intervals of 0.5–0.6 and 0.6–0.7, whatever the trait is. If we directly divide the scores into 10 classes, the uneven distribution could jeopardize the training process. Therefore we set C_p as 4, and scores with 0–0.5 fall into one class, and similar for 0.5–0.6, 0.6–0.7, and 0.7–1.0, respectively. In this way, the number of samples for each class is approximately balanced.

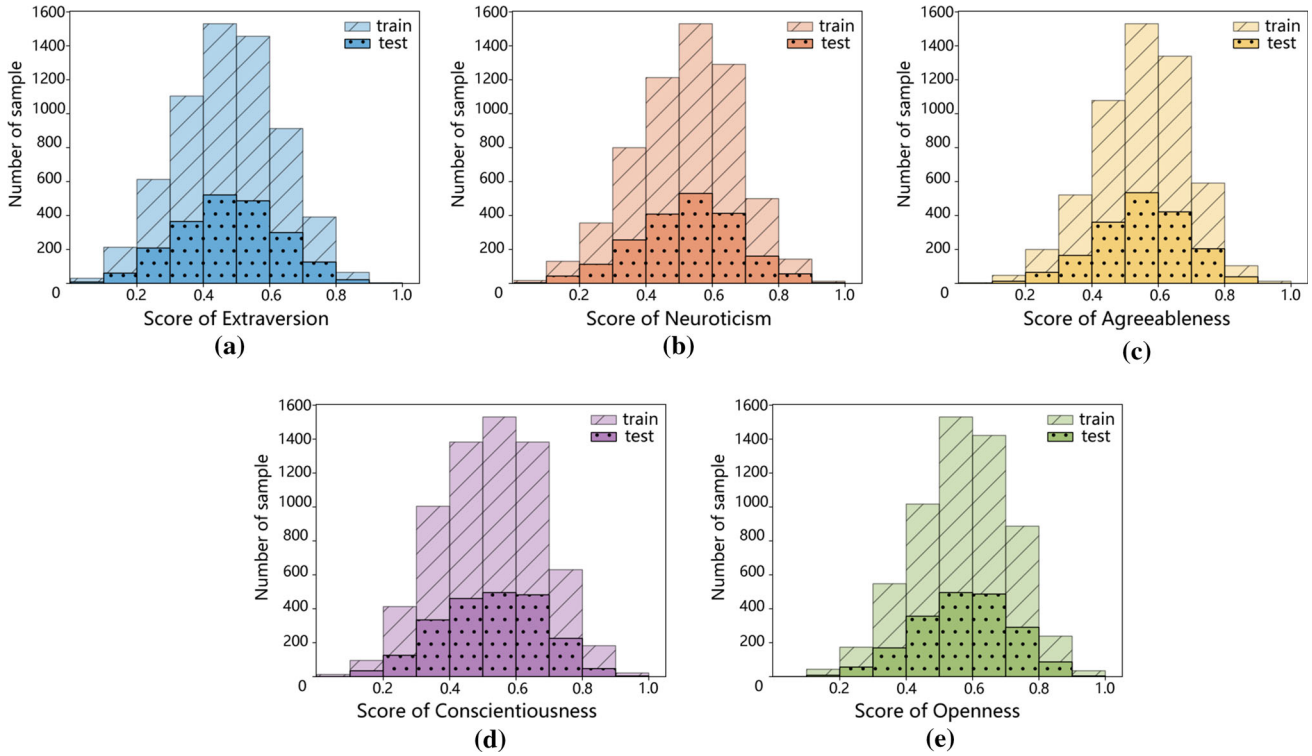

Fig. 9 Histogram of the score distribution of the five personality traits in the training/testing set of First Impressions v2 dataset

4.3 Evaluation Protocol

For a fair comparison, we employ the evaluation protocol used in most publications. The performance of each trait is scored in terms of Mean Absolute Error (MAE), which is formulated as:

$$E_p = 1 - \frac{1}{N}\sum_{i=1}^{N} |y_{p_i} - \hat{y}_{p_i}|, \tag{10}$$

where p indicates the personality trait, N indicates the number of samples, y_{p_i} and $\hat{y}_{p_i}$ denote the ground truth and the prediction of sample i, respectively.

4.4 Comparison with the State-of-the-Art

In this subsection, we show our results and compare them with state-of-the-art approaches, including the publications and top entries in both rounds of the First Impressions competitions.

As shown in Table 1, our method achieves 0.9188 score on average for the Big Five traits, and 0.9247 for the interview recommendation variable, showing better results than the methods in comparisons and any competition results. This happens even for methods trained with both the training and validation subsets, such as the one by Kaya et al. (2017). Note that some results in this table are reported for the First Impressions v1 dataset, on which the text data is not available. To have a fair comparison, we also provide the result of our method without using the text from audio transcriptions. Still, our result (the second row in Table 1) is better than the state-of-the-art.

In addition to the comparison of scores for Big Five traits and interview recommendation variable, we present a more comprehensive comparison in Table 2, including details about the training process in terms of data modalities and network details. As shown in Table 2, the visual cue is considered by all compared approaches. The reason behind it may be that the facial expression and movement always contribute to the determination of one's perceived personality when compared with his/her voice or words. Meanwhile, assembling more networks does not certainly imply the highest performance. The entry "ucas" combines 16 networks in their method, but the results are inferior to methods (Kaya et al. 2017; Zhang et al. 2016; Subramaniam et al. 2016), which have less than 5 networks on average.

5 Ablation Study

In this section, we discuss the contribution of each module of our method related to the final performance. We first evaluate

Table 1 Comparison with state-of-the-art methods on the first impressions dataset (test)

Rank	Method/entry	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
1	**Our**	0.9202	0.9146	**0.9177**	**0.9218**	**0.9195**	**0.9188**	**0.9247**
2	Our-w/o text	0.9200	**0.9150**	0.9176	**0.9218**	0.9191	0.9187	0.9244
3	Kaya et al. (2017)**	**0.9213**	0.9146	0.9137	0.9198	0.9170	0.9173	0.9209
4	Gürpinar et al. (2016)*	0.9180	0.9110	0.9070	0.9150	0.9140	0.9130	–
4	Zhang et al. (2016)*	0.9133	0.9100	0.9126	0.9166	0.9123	0.9130	–
6	Subramaniam et al. (2016)*	0.9150	0.9099	0.9119	0.9119	0.9117	0.9121	–
7	Güçlütürk et al. (2018)	0.9112	0.9104	0.9112	0.9152	0.9111	0.9118	0.9162
8	Bekhouche et al. (2017)	0.9155	0.9083	0.9103	0.9138	0.9101	0.9116	0.9157
8	Ventura et al. (2017)	0.9150	0.9070	0.9120	0.9140	0.9100	0.9116	–
10	Güçlütürk et al. (2016a)*	0.9107	0.9089	0.9102	0.9138	0.9111	0.9109	–
11	ucas (Ponce-López et al. 2016)*	0.9129	0.9064	0.9091	0.9107	0.9099	0.9098	–
12	Gürpınar et al. (2016b)*	0.9161	0.9021	0.9070	0.9133	0.9084	0.9094	–
13	ROCHCI (Escalante et al. 2018)	0.9026	0.9011	0.9032	0.8949	0.9047	0.9013	0.9018
14	Vo et al. (2018)***	0.8816	0.8814	0.8958	0.8772	0.8864	0.8845	–
15	FDMB (Escalante et al. 2018)	0.8788	0.8632	0.8910	0.8659	0.8747	0.8747	0.8721

The bold word "our" in the first column means ours achieves the best result on average, and the bold values in other columns mean the best results for each personality trait
*Trained and evaluated on the First Impressions v1 dataset, on which the text data and interview annotations are not available
**Ref. Kaya et al. (2017) is trained with both training set and validation set
***Ref. Vo et al. (2018) only MAE scores reported. For a unified comparison we calculate the scores according to Eq. (10)

Table 2 Training details of compared methods

Method/entry	Modality	Network	Number of networks	Mean trait*
Ours	Video & audio & text	CR-Net	3	0.9188
Kaya et al. (2017)	Video & audio	VGG-face	4	0.9173
		VGG-VD19		
		ELM		
		Random Forest		
Gürpinar et al. (2016)	Video & Audio	VGG-face	3	0.9130
		VGG-VD19		
		ELM		
Zhang et al. (2016)	Video and audio	VGG-face	5	0.9130
		ResNet		
		DAN/DAN+		
		Linear Regressor		
Subramaniam et al. (2016)	Video & audio	3D CNN	2	0.9121
		LSTM		
Güçlütürk et al. (2018)	Video & audio & text	ResNet	3	0.9118
		ridge regression		
Bekhouche et al. (2017)	Video	Pyramid multi-level	5	0.9116
		support vector regressor		
Ventura et al. (2017)	Video	DAN+	5	0.9116
Güçlütürk et al. (2016a)	Video & Audio	ResNet	2	0.9109
ucas (Ponce-López et al. 2016)	Video & Audio	lbptop	16	0.9098
		hog3d		
		VGG		
		AlextNet		
		ResNet		
Gürpınar et al. (2016b)	Video	CNN	2	0.9094
Vo et al. (2018)	Video & Audio & text	Mixture density nerual network	4	0.8845
		mixture of Gaussian distribution		
		dynamic cascade boosting network		

*Namely the *average* of Big Five traits marked in Table 1

the effect of different data modalities. After that, we analyze the two main contributions of our model, the CR-block and the Bell Loss. Then we verify the effect of different choices of the number of classes for CR-block. We further compare some commonly used regressors, including SVR and RF, against the ETR used in this paper. Finally, we provide a visual analysis of the learned features to visualize where our network focuses to regress for the apparent personality traits.

5.1 Performance of Different Data Modalities

Table 3 shows a comparison of performance using different data modalities. As we split two main cues, the results of scene and face data are also shown. The video data can yield a much better result than the other two modalities. The average score of video stream is about 2% higher than that of fusion of audio and text. It is also consistent with the result of current publications in Table 1. Meanwhile, one can see in general the scene data can achieve a slightly better result. The average score of scene cue is 0.9138, which is better than the face cue of 0.9133. However, not all traits follow this phenomenon. The result of face cue on "Extraversion" and "Neuroticism" is better than the scene cue. There may be two reasons for it. On one hand, features like whether the people are friendly or not (Extraversion) can be judged from their face directly, whereas the traits like sloppy (Conscientiousness) and imaginative (Openness) may also relate to their pose, hair style and body movements, which are beyond the face region. On the other hand, as we use the video stream rather than a single frame, the motion is more coherent in the scene data, which can leverage more temporal information than the facial alone. All in all, results show that the proposed combination of multiple cues enhances the result of the final recognition.

Table 3 Comparison on the performance of the network trained with different data modalities

Modality	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
Video							
Face	0.9167	0.9090	0.9130	0.9151	0.9125	0.9133	0.9177
Scene	0.9133	0.9087	0.9148	0.9183	0.9139	0.9138	0.9187
Multi-focus	0.9199	0.9123	0.9168	0.9204	0.9168	0.9172	0.9230
Audio-text							
Audio only	0.8942	0.8942	0.9005	0.8912	0.8996	0.8959	0.8974
Text only	0.8825	0.8818	0.8964	0.8800	0.8872	0.8856	0.8855
Audio + text	0.8953	0.8951	0.9010	0.8920	0.9002	0.8967	0.8981

Table 4 Comparison of network performance with CR-blocks and Bell Loss

Modality	Strategy		Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
	CR-block	Bell Loss							
Face	–	–	0.9041	0.8949	0.9025	0.9051	0.8927	0.8999	0.9049
	✓	-	0.9121	0.9019	0.9079	0.9045	0.9098	0.9072	0.9104
	-	✓	0.9134	0.9031	0.9060	0.9090	0.9065	0.9076	0.9094
	✓	✓	0.9167	0.9090	0.9130	0.9151	0.9125	0.9133	0.9177
Scene	-	-	0.9028	0.8972	0.9036	0.9063	0.9026	0.9025	0.9065
	✓	-	0.9057	0.9022	0.9087	0.9113	0.9086	0.9073	0.9098
	-	✓	0.9088	0.9054	0.9086	0.9142	0.9066	0.9087	0.9129
	✓	✓	0.9133	0.9087	0.9148	0.9183	0.9139	0.9138	0.9187
Audio + text	–	–	0.8891	0.8829	0.8937	0.8853	0.8914	0.8885	0.8909
	✓	–	0.8896	0.8901	0.8989	0.8877	0.8963	0.8925	0.8949
	-	✓	0.8919	0.8917	0.8996	0.8877	0.8976	0.8937	0.8952
	✓	✓	0.8953	0.8951	0.9010	0.8920	0.9002	0.8967	0.8981

5.2 Effectiveness of CR-Block and Bell Loss

In order to verify the two main contributions in this work, we evaluate the effectiveness of our CR-block and Bell Loss. As shown in Table 4, we use the checkmark to indicate which module we employ for learning. Meanwhile, the line without any checkmarks means the baseline network, i.e., ResNet-34 only.

We present the comparisons of the scene video stream, the face video stream and the audio-text stream, which are the three streams in the network. Comparing the result of ResNet-34 (the first line) and that of using CR-block (the second line), we find the CR-block improves the performance by about 0.0073, 0.0048 and 0.0040 for the face cue, scene cue, and the audio-text stream, respectively. The improvement of Bell Loss (the third line) is a little higher than CR-block, by about 0.0004, 0.0014 and 0.0012 for the three streams, respectively. It shows the benefit of the proposed loss at the final stage, being even more relevant than the gain provided by the guidance of classification at the first stage.

To better demonstrate how Bell Loss benefits the predictions for extreme samples, we also show the scattering plots of the absolute error with/without the Bell Loss against different ground truth values.

As can be seen from Fig. 10, without the Bell Loss, the number of samples with a high estimation error is apparently higher than using the Bell Loss. For those samples with extreme ground truth values, we can also see that it can lead to extremely wrong predictions without using the Bell Loss. Some of the divergence can even reach 0.8 under this condition.

We further perform a comparison when the networks are trained with different losses, including L2 loss, L1 loss, Bell loss, L2 + L1 losses, and all losses (namely L1 + L2 + Bell losses). We use the ResNet-34 as the backbone network and only change the loss functions for this comparison.

As shown in Table 5, when comparing the performances between L1 loss and L2 loss, we find that the ones with only L2 loss achieve in general a better result for all the three kind of input streams (Face, Scene, or Audio + text). The average improvements are 0.0077, 0.0035, and 0.0039 for "L2 loss only" versus "L1 loss only" for face, scene and audio-text stream, respectively. L1 and L2 losses, in average, provide performance improvements when used together compared to

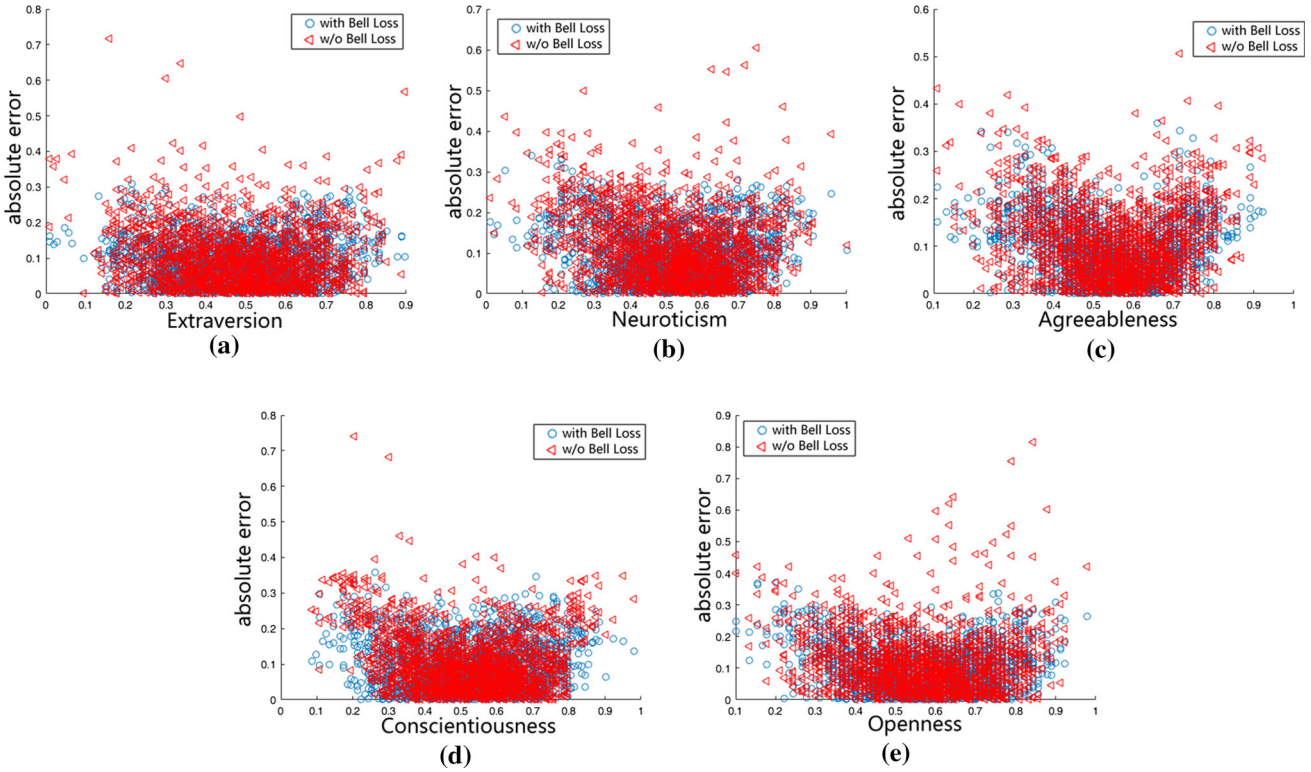

Fig. 10 The absolute error with or without the Bell Loss. The red triangles indicate the result learned without the Bell Loss whereas the blue circles refer to the result with Bell Loss. It is apparent the Bell Loss significantly reduces the prediction errors of samples with extreme ground truth (Color figure online)

Table 5 Comparison on the performance of the network trained with different losses

Modality	Strategy	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
Face	L1 loss only	0.9016	0.8861	0.8974	0.8801	0.8903	0.8911	0.9055
	L2 loss only	0.9044	0.8915	0.9024	0.9017	0.8941	0.8988	0.9036
	L1 + L2 losses	0.9041	0.8949	0.9025	0.9051	0.8927	0.8999	0.9049
	Bell Loss only	0.9123	0.9030	0.9052	0.9076	0.9044	0.9065	0.9126
	All losses	0.9134	0.9031	0.9060	0.9090	0.9065	0.9076	0.9094
Scene	L1 loss only	0.8999	0.8873	0.8924	0.9034	0.8986	0.8963	0.9036
	L2 loss only	0.8993	0.8960	0.9010	0.9027	0.9000	0.8998	0.9043
	L1 + L2 losses	0.9028	0.8972	0.9036	0.9063	0.9026	0.9025	0.9065
	Bell Loss only	0.9044	0.8987	0.9045	0.9078	0.9026	0.9036	0.9077
	All losses	0.9088	0.9054	0.9086	0.9142	0.9066	0.9087	0.9129
Audio + text	L1 loss only	0.8777	0.8802	0.8921	0.8824	0.8809	0.8827	0.8815
	L2 loss only	0.8834	0.8869	0.8965	0.8801	0.8859	0.8866	0.8838
	L1 + L2 losses	0.8891	0.8829	0.8937	0.8853	0.8914	0.8885	0.8909
	Bell Loss only	0.8897	0.8918	0.8955	0.8862	0.8967	0.8920	0.8920
	All losses	0.8919	0.8917	0.8996	0.8877	0.8976	0.8937	0.8952

their individual usage. Finally, our proposed Bell Loss outperforms "L2 + L1 losses" results (0.0066, 0.0011 and 0.0035 for face, scene and audio-text stream, respectively). When all three losses are used together, the performance is slightly improved as well, especially for the scene video stream, benefiting from the complementary nature of the three losses. One of the main reasons we hypothesize for this is that in those cases of having large differences between a random prediction and the corresponding ground truth, it may become hard for the Bell Loss to find a good initialization point. However,

Table 6 Comparison on the performance with different classification-regression techniques

Modality	Strategy	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
Face	Expectation	0.9019	0.8964	0.8991	0.9016	0.8978	0.8994	0.9042
	Expectation + reg	0.9061	0.9018	0.9039	0.9020	0.9034	0.9034	0.9066
	Our (CR-block)	0.9121	0.9019	0.9079	0.9045	0.9098	0.9072	0.9104
Scene	expectation	0.8882	0.8879	0.8953	0.8891	0.8909	0.8903	0.8930
	Expectation + reg	0.8933	0.8910	0.8979	0.8955	0.8959	0.8947	0.8982
	Our (CR-block)	0.9057	0.9022	0.9087	0.9113	0.9086	0.9073	0.9098
Audio + text	expectation	0.8856	0.8873	0.8947	0.8843	0.8917	0.8887	0.8903
	Expectation + reg	0.8877	0.8919	0.8983	0.8849	0.8940	0.8914	0.8920
	Our (CR-block)	0.8896	0.8901	0.8989	0.8877	0.8963	0.8925	0.8949

Table 7 Comparison on the performance with different classes for CR-block

	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness	Average	Interview
Face							
4-class	0.9167	0.9090	0.9130	0.9151	0.9125	0.9133	0.9177
7-class	0.9129	0.9071	0.9111	0.9111	0.9103	0.9105	0.9141
10-class	0.9110	0.9043	0.9101	0.9085	0.9081	0.9084	0.9130
Scene							
4-class	0.9133	0.9087	0.9148	0.9183	0.9139	0.9138	0.9187
7-class	0.9114	0.9084	0.9092	0.9169	0.9079	0.9108	0.9166
10-class	0.9074	0.9018	0.9075	0.9103	0.9084	0.9071	0.9105
Audio + text							
4-class	0.8953	0.8951	0.9010	0.8920	0.9002	0.8967	0.8981
7-class	0.8903	0.8913	0.8987	0.8987	0.8954	0.8949	0.8932
10-class	0.8862	0.8888	0.8946	0.8830	0.8915	0.8888	0.8891

according to Fig. 6, L2 and L1 losses have a larger gradient in this condition, and they can be beneficial at early optimization stages.

5.3 Comparison with Previous Classification-Regression Techniques

To further verify our CR-block in the apparent personality recognition task, we also compare our approach with two methods that involve a combination of classification and regression. One method in comparison uses the expectation of all the classes as the final prediction (Rothe et al. 2015), and we denote it as *expectation* in Table 6. The other method adds the regression loss on the expectation to learn the final result (Gao et al. 2018), and we denote it as *expectation + reg*. Similar to Sect. 5.2, The experiment is conducted on the face video stream, scene video stream and the audio-text stream. As it can be observed from Table 6, considering the value of personality as different classes and taking the expectation of them as the prediction is hard to get a good performance. The solution adding the regression loss achieves a better result when compared with the simple *expectation*. However, the information provided by the prediction of each class and its confidence is not accurate enough. Compared with these two strategies, ours takes the classification features as the guidance for the final regression, and achieves better recognition results.

5.4 Comparison of Different Classes for CR-Block

As mentioned in Sect. 4.2.2, we set the number of classes to 4, owing to the imbalance of data distribution. In this subsection, we evaluate the setting for a different number of classes.

As shown in Table 7, we defined the traits to be classified into 7 and 10 classes. To maintain a roughly equal number of samples in each class, we have a more sophisticated segment point. For the 7-class, we have the segment point of "0.3556, 0.4383, 0.5003, 0.5575, 0.6143, and 0.6849". For the 10-class, we have "0.3220, 0.3962, 0.4466, 0.4903, 0.5307, 0.5708, 0.6078, 0.6528 and 0.7102".

As shown in Table 7, with the increase of the number of classes, the ultimate performance is falling down. That is because although the problem of uneven distribution of data

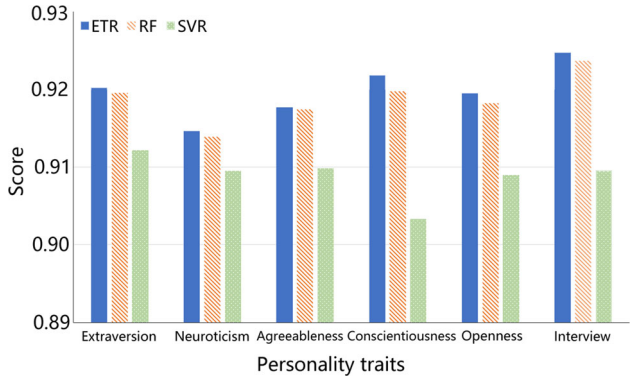

Fig. 11 Regression results of SVR, RF and ETR

is addressed, it is still hard for the classifier to discriminate the subtle differences between adjacent classes.

5.5 Comparison of Regressors

We compare the performance of different regressors. In this comparison, we select two popular regressors, the SVR and RF, which are also commonly used in current personality prediction works.

The comparison results are shown in Fig. 11. Compared with SVR, the results of RF and ETR are better. It demonstrates the effectiveness of ensemble learning. Moreover, the ETR outperforms RF, which shows that randomly selecting features can do better to avoid overfitting.

5.6 Feature Visualization

In this subsection, we visualize the feature maps, including both scene and face cues to illustrate what features are more important to recognize the apparent personality traits for the network. To achieve this, we first normalize the feature maps to the range [0, 1]. Then, we employ the python library - seaborn, and use the *heatmap* function with the parameter *colormap* set to "jet". The obtained feature maps are shown in Fig. 12.

One can see that for the face cue, what contribute the most to the results are relevant face keypoints such as eyes, nose and mouth. Those are more related to facial expressions of emotion. Regarding the scene cue, one can see the face has a significant contribution, and clothing and furnishings show some relevant activations as well.

We also quantitatively evaluate the relationship between highlighted features and face keypoints. For each visualized feature map, we select the 30% highest pixel values to obtain the highlighted features binary image. We find the face keypoints from the image of the face cue as in Zhang et al. (2016). For a precise evaluation, we employ the points of two eyes, the nose, and two corners of the mouth as in Zhang et al. (2016), with an additional point at the mid distance of the two mouth corners. Then we calculate the ratio of those points being inside the highlighted regions. From all 2000 testing videos with 32 frames from each, we have 73.96% of highlighted points, proving the relevance of face keypoints for recognizing apparent personality traits.

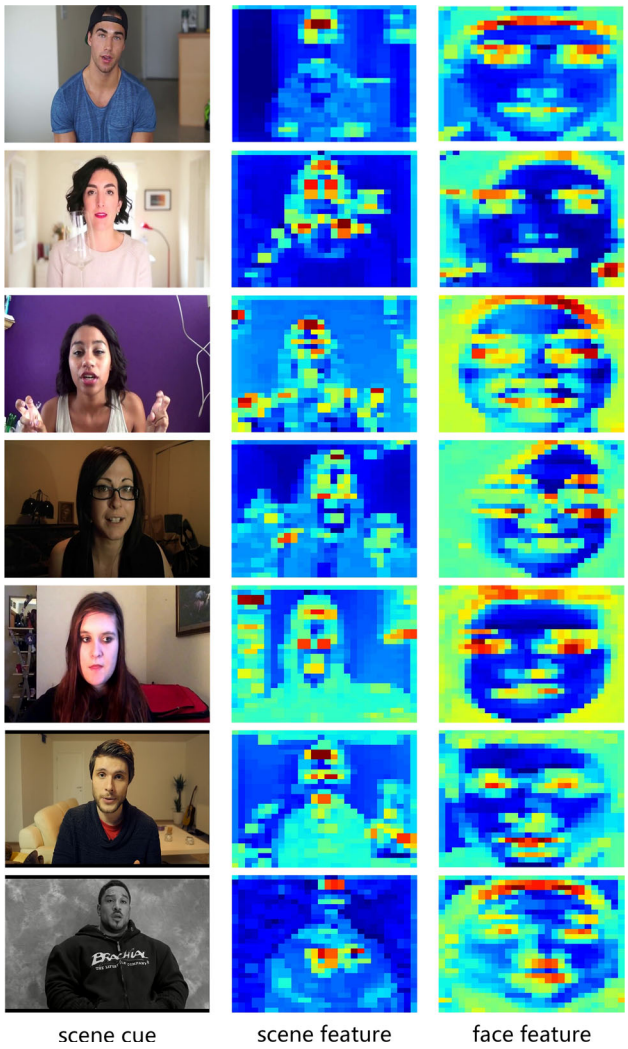

Fig. 12 Examples of feature visualization of scene and face cues

6 Conclusion

This work has presented a network scheme to deal with the problem of apparent personality computing and job interview recommendation using audio-visual recordings. We have proposed a deep Classification-Regression Network, which benefits from the learned classification features as a guidance to improve the regression performance. Furthermore, we have presented the Bell Loss function, which alleviates the regression-to-the-mean problem and promotes network optimization to reach more accurate predictions, by keeping high

gradient values when it approaches optimal solutions in optimization. Exhaustive evaluations of the proposed technique and loss, including the combination of multiple data modalities, the classification-regression module, and Bell Loss, have shown a higher recognition accuracy for personality traits and job interview recommendation on the First Impressions dataset when compared with the state-of-the-art.

Acknowledgements The work was supported by the National Key R&D Program of China under Grant #2018YFC0807500, the National Natural Science Foundations of China #61961160704, #61876179, #61772396, #61772392, #61902296, the Fundamental Research Funds for the Central Universities #JBF180301, Xi'an Key Laboratory of Big Data and Intelligent Vision #201805053ZD4CG37, the Science and Technology Development Fund of Macau (#0008/2018/A1, #0025/2019/A1, #0010/2019/AFJ, #0025/2019/AKP), Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/ Generalitat de Catalunya.

References

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26.

Basu, A., Dasgupta, A., Thyagharajan, A., Routray, A., Guha, R., & Mitra, P. (2018). A portable personality recognizer based on affective state classification using spectral fusion of features. *IEEE Transactions on Affective Computing*, *9*(3), 330–342.

Bekhouche, S. E., Dornaika, F., Ouafi, A., & Taleb-Ahmed, A. (2017). Personality traits and job candidate screening via analyzing facial videos. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1660–1663). IEEE.

Bland, J. M., & Altman, D. G. (1994a). Regression towards the mean. *BMJ: British Medical Journal*, *308*(6942), 1499.

Bland, J. M., & Altman, D. G. (1994b). Statistics notes: Some examples of regression towards the mean. *BMJ*, *309*(6957), 780.

Chen, S., Zhang, C., & Dong, M. (2018). Deep age estimation: From classification to ranking. *IEEE Transactions on Multimedia*, *20*(8), 2209–2222.

Corr, P. J., & Matthews, G. (2009). *The Cambridge handbook of personality psychology*, chap. MethodsofPersonalityAssessment (pp. 110–126). Cambridge: Cambridge University Press.

Correa, J. A. M., Abadi, M. K., Sebe, N., & Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2018.2884461.

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Gucluturk, Y., Guclu, U., et al. (2018). *Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos*. arXiv preprint arXiv:1802.00745.

Escalante, H. J., Ponce-López, V., Wan, J., Riegler, M. A., Chen, B., Clapés, A., et al. (2016). Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *ICPR* (pp. 67–73).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 1459–1462). ACM.

Gao, B. B., Zhou, H. Y., Wu, J., & Geng, X. (2018). Age estimation using expectation of label distribution learning. In *IJCAI* (pp. 712–718).

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.

Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., et al. (2018). Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, *9*(3), 316–329.

Güçlütürk, Y., Güçlü, U., van Gerven, M. A., & van Lier, R. (2016a). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European conference on computer vision* (pp. 349–358). Berlin: Springer.

Gürpınar, F., Kaya, H., & Salah, A. A. (2016b) Combining deep facial and ambient features for first impression estimation. In *European conference on computer vision* (pp. 372–385). Berlin: Springer.

Gürpinar, F., Kaya, H., & Salah, A. A. (2016) Multimodal fusion of audio, scene, and face features for first impression estimation. In *2016 23rd International conference on pattern recognition (ICPR)* (pp. 43–48). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of the 2004 IEEE international joint conference on neural networks* (vol. 2, pp. 985–990). IEEE.

Huang, S., & Ramanan, D. (2017). Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In *IEEE conference on computer vision and pattern recognition (CVPR)* (vol. 1).

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711). Berlin: Springer.

Kaya, H., Gürpinar, F., & Salah, A. A. (2017). Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVS. In *CVPR workshops* (pp. 1651–1659).

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., et al. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294–3302).

Klein, D. N., Kotov, R., & Bufferd, S. J. (2011). Personality and depression: Explanatory models and review of the evidence. *Annual Review of Clinical Psychology*, *7*, 269–295.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Li, R., et al. (2016). Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 25–30). IEEE.

Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Li, R., et al. (2017). Large-scale gesture recognition with a fusion of rgb-d data based on saliency theory and c3d model. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 2956–2964.

Mairesse, F., & Walker, M. (2007). Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 496–503).

Mohammadi, G., & Vinciarelli, A. (2015). Automatic personality perception: Prediction of trait attribution based on prosodic features

extended abstract. In *2015 International conference on affective computing and intelligent interaction (ACII)* (pp. 484–490). IEEE.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M.E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (vol. 1, pp. 1–6). IEEE.

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4920–4928).

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *British machine vision conference* (Vol. 1, p. 6).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). *Automatic differentiation in pytorch*.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296.

Polzehl, T., Moller, S., & Metze, F. (2010). Automatically assessing personality from speech. In *2010 IEEE fourth international conference on semantic computing (ICSC)* (pp. 134–140). IEEE.

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 *European conference on computer vision* (pp. 400–418). Berlin: Springer.

Rothe, R., Timofte, R., & Van Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 10–15).

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.

Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., & Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European conference on computer vision* (pp. 337–348). Berlin: Springer.

Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., & Li, S. Z. (2018). Efficient group-n encoding and decoding for facial age estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2610–2623.

Ventura, C., Masip, D., & Lapedriza, A. (2017). Interpreting CNN models for apparent personality trait regression. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1705–1713). IEEE.

Vo, N. N., Liu, S., He, X., & Xu, G. (2018). Multimodal mixture density boosting network for personality mining. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 644–655). Berlin: Springer.

Wang, X., Yu, K., Dong, C., & Change Loy, C. (2018). Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 606–615).

Wei, X. S., Zhang, C. L., Zhang, H., & Wu, J. (2018). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, *9*(3), 303–315.

Xia, F., Asabere, N. Y., Liu, H., Chen, Z., & Wang, W. (2017). Socially aware conference participant recommendation with personality traits. *IEEE Systems Journal*, *11*(4), 2255–2266.

Zhang, C. L., Zhang, H., Wei, X. S., & Wu, J. (2016). Deep bimodal regression for apparent personality analysis. In *European conference on computer vision* (pp. 311–324). Berlin: Springer.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

Zhao, G., Ge, Y., Shen, B., Wei, X., & Wang, H. (2018). Emotion analysis for personality inference from eeg signals. *IEEE Transactions on Affective Computing*, *9*(3), 362–371.