PersEmoN: A Deep Network for Joint Analysis of Apparent Personality, Emotion and Their Relationship

Le Zhang, Songyou Peng*, and Stefan Winkler, *Fellow, IEEE*

Abstract—Apparent personality and emotion analysis are both central to affective computing. Existing works solve them individually. In this paper we investigate if such high-level affect traits and their relationship can be jointly learned from face images in the wild. To this end, we introduce *PersEmoN*, an end-to-end trainable and deep Siamese-like network. It consists of two convolutional network branches, one for emotion and the other for apparent personality. Both networks share their bottom feature extraction module and are optimized within a multi-task learning framework. Emotion and personality networks are dedicated to their own annotated dataset. Furthermore, an adversarial-like loss function is employed to promote representation coherence among heterogeneous dataset sources. Based on this, we also explore the emotion-to-apparent-personality relationship. Extensive experiments demonstrate the effectiveness of *PersEmoN*.

Index Terms—Affective Computing, Emotion, Apparent Personality, Adversarial Learning, Multi-Task Learning, Deep Learning.

1 Introduction

PROLIFERATION of cameras, availability of cheap storage and rapid developments in high-performance computing have enabled exciting new developments in Human-Computer Interaction (HCI), in which affective computing plays an inevitable role. For instance, in video-based interviews, automatically computed personalities of candidates can serve as an important cue to assess their qualifications. However, affective computing remains a challenging problem in both computer vision and psychology despite many years of research.

Facial appearances strongly influence our judgement of the emotion and personality of other people. Such a judgement usually can be made after a very short time [1], although different studies have not yet reached a consensus about the accuracy of such appearance-based first impressions [2, 3]. As mentioned in a recent survey on personality computing [4], state-of-the-art studies consider either the actual personality traits measured from self or acquaintance reports, or the so-called apparent personality traits, which represent the impressions about someone's personality from an external observer's point of view.

In this paper, we are interested in the problem of analyzing apparent personality, emotion and their relationship. Apparent personality reflects the coherent patterning of behavior, cognition and desires (goals) over time and space, as perceived by an external observer. Emotion is an integration of feeling, action, appraisal, and wants at a particular time and location [5]. We can understand the emotion-to-apparent-personality relationship as weather to climate, i.e. what one expects is apparent personality while what one observes in a particular moment is emotion. Although they have distinct definitions, the relationship between personality and emotion has been revealed previously. Eysenck's personality model [6] showed that neurotics can be more sensitive to external stimulation and easily become upset or nervous due to minor stressors. Extraversion on the other hand has been linked to higher sensitivity to potentially rewarding stimuli, which in part explains the high levels of positive affect found in extraverts, since they will more intensely feel the excitement of a potential reward [7].

Apparent personality estimation has become an increasingly popular field of research; related challenges, e.g. ChaLearn 2016 on first impressions [8], together with publicly available datasets, have attracted wide attention. In this paper, we also focus on the task of apparent personality prediction, where we consider the Big Five personality traits (*Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness*) [8]. Our emotion analysis is based on Russell's circumplex model of affect [9], in which emotions are distributed in a two-dimensional circular space spanned by the dimensions of *arousal* and *valence*, instead of classifying pre-defined emotion categories. This is advantageous in the sense that it allows for a finer-grained representation of expressions and emotional states [10].

Deep convolutional neural networks (CNNs) reign undisputed as the new de-facto method for face based applications such as face recognition [11, 12], alignment [13], and so on. This motivates us to study the following fundamental problems:

- *Songyou Peng is the corresponding author.*
- *Le Zhang is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore.*
  E-mail: zhangl@i2r.a-star.edu.sg
- *Songyou Peng is with the Department of Computer Science, ETH Zurich, and Max Planck ETH Center for Learning Systems.*
  E-mail: songyou.peng@inf.ethz.ch
- *Stefan Winkler is with the National University of Singapore (NUS).*
  E-mail: winkler@comp.nus.edu.sg
- *All the authors are adjunct at the Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore, where this work was carried out.*

1) As both face recognition and affective computing can have faces as input, how transferable are deeply learned face representations for emotion and apparent personality analysis?
2) Is it beneficial to explore emotion, apparent personality and their relationship in a single deep CNN?

These tasks are non-trivial. Among the most significant challenges are:

- The scarceness of large-scale datasets which encompass both emotion and apparent personality annotations for learning such a rich representation for apparent personality, emotion and emotion-to-apparent-personality relationship. In particular, existing datasets only contain emotion attributes, while other datasets may only be annotated with apparent personality labels. Manually annotating data for both emotion and apparent personality may partly alleviate this. However, it is costly, time-consuming, and error-prone due to subjectivity.
- The discrepancy of existing datasets: Datasets are usually collected in different environments which may exhibit significant variations in illumination, scale, pose, etc. Each dataset may have vastly different statistical distributions.
- Annotations of emotion and/or apparent personality can be done at the image/frame level [14, 15] or at the video level [8]. How can we encapsulate both frame and video level understanding into a single network?

We address these challenges by proposing *PersEmoN*, an end-to-end trainable and deep Siamese-like network [16]. It consists of two CNN branches, which we call emotion network and personality network, respectively. Emotion network and personality network share their bottom feature extraction module and are optimized within a multi-task learning framework. An adversarial-like loss function is further employed to promote representation coherence between heterogeneous dataset sources. We show that *PersEmoN* works well for analyzing apparent personality, emotion and their relationship. A demo version of this paper has been presented in [17].

2 Related Work

The wealth of research in this area is such that we cannot give an exhaustive review. Instead, we focus on describing the most important threads of research on using deep learning for emotion and apparent personality analysis.

2.1 Deep Learning for Emotion Analysis

Emotion analysis has been investigated from different perspectives. [18] proposed a deep belief network for unsupervised audio-visual emotion recognition. However, its feasibility of large-scale supervised learning remains unclear. [19] investigated the usage of deep CNNs and Bayesian classifiers for group emotion recognition in the wild. Apart from visual inputs, the system also needs the scene context (such as background, clothes, etc.) which may not be available in many scenarios. [20] introduced convolutional deep belief networks to learn salient multi-modal features of emotions. Although workable as reported, their network structure is shallow, and it remains unclear how to transfer rich feature hierarchies from very deep networks for different modalities in their system.

Unlike popular classification approaches for discrete emotion categories, many recent works delve into different representations of human expressions and emotions. For instance, EmotioNet [21] provides an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. However, the performance gap between using EmotioNet and human-annotated datasets for training emotion networks is not well understood. Other approaches try to analyze emotion using continuous arousal-valence space [9]. For instance, Mollahosseini *et al.* [22] introduce a large-scale emotion dataset and show that their deep neural network outperforms conventional machine learning methods and off-the-shelf facial expression recognition systems. However, it requires a specially designed sampling strategy to alleviate data imbalance problems. An ensemble of memory networks [23], multiple datasets for cascade learning [24], and multiple LSTM layers [25] have been employed to predict the emotion scores. Due to the complexity of these networks, they are difficult to train however.

2.2 Deep Learning for Apparent Personality Analysis

Güçlütürk *et al.* [26] introduced a deep audio-visual residual network for multimodal apparent personality trait recognition. In their extended version [27], the authors analysed different cues (visual, acoustic and audiovisual). This may not be very practical as some of the cues could be missing during deployment of the system. [28] developed a volumetric convolution and Long-Short-Term-Memory (LSTM) based network to learn audio-visual temporal patterns. Although it outperforms many other approaches [8], its performance on pure visual inputs is not reported, which makes it difficult to understand the merits of their visual stream.

[29] identified apparent personality with a deep bimodal regression framework based on both video and audio inputs. However, the performance of all above-mentioned methods relies heavily on ensemble strategies, whereas we are able to achieve better results with a single visual stream with the proposed *PersEmoN*.

Gürpınar *et al.* [30] employed a pre-trained CNN to extract facial expressions as well as ambient information for apparent personality analysis. Although they achieve promising results, the system is not end-to-end trainable and needs a stand-alone regressor. A time-continuous prediction approach that learns the temporal relationships, rather than treating each time instant separately, was established in [31] for the prediction of Big Five traits, Attractiveness and Likeability. Nevertheless, a database annotated in a time-continuous manner is needed in their setting, which is difficult to obtain in practice.

Ventura *et al.* [32] investigated the reason of the success of CNN in apparent personality analysis. They showed that the face provides most of the discriminative information for this task. This motivated us to investigate the feasibility of using a very deep face verification network for this task. For more related work on apparent personality analysis, please refer to recent surveys [33, 34].

3 Methodology

In comparison to the aforementioned studies, our work aims to investigate whether emotion and apparent personality analysis can benefit from the face representations learned from a well-annotated face recognition dataset, *without having a dataset with both emotion and apparent personality annotations*. To this end, we show that state-of-the-art face recognition networks perform well for both emotion and apparent personality analysis. We also explore the feasibility of jointly training emotion and apparent personality analysis. More specifically, we propose *PersEmoN* within a multi-task learning framework to learn better representations for both emotion and apparent personality than those obtained by solving each task individually. On top of such representations, we demonstrate the feasibility of establishing a good emotion-to-apparent-personality relationship.

3.1 *PersEmoN* Overview

An overview of *PersEmoN* can be found in Fig. 1. We first detect and align faces for both apparent personality and emotion datasets with an open version of *MTCNN* [13]. For the apparent personality dataset, we employ a sparse sampling strategy. The personality network consists of a feature extraction module (*FEM*) and a personality analysis module (*PAM*) to predict the Big Five personality factors. A consensus aggregation function is employed to aggregate raw apparent personality scores before feeding them into the *PAM*. Similarly, the emotion network shares the *FEM* with the personality network and has its own emotion analysis module (*EAM*) targeted at predicting the arousal and valence dimensions [35] of emotion. An emotion-to-apparent-personality relationship analysis module (*RAM*) is also employed.

In the training phase, the system is aware of which dataset the image comes from and will automatically assign the image to its own branch. For instance, the images from the apparent personality set go through *FEM* and PAM. Meanwhile, they can also go through *FEM* and *EAM* and finally output the apparent personality traits through *RAM*. In the same way, the images from the emotion set can go through *FEM* and *EAM* to yield emotion outputs.

In the testing phase, the system can estimate the emotion and the apparent personality from *EAM* and *PAM* separately. During inference, we could use *FEM* and *PAM* to obtain the apparent personality traits. Similarly, we could use *FEM* and *RAM* to get the emotion outputs. As a side product, we could even use *RAM* to produce the apparent personality attributes from emotion (arousal and valence) inputs. Note that in the testing phase the proposed method also works with video-based emotion datasets by processing each video frame individually.

The detailed network structure of the various modules of *PersEmoN* is summarized in Table 1. In the following section, we elaborate on the different components mentioned above.

TABLE 1
Detailed architecture of *PersEmoN*. Conv denotes convolution units that may contain multiple convolution layers; residual units are shown in square brackets. For example, $[33, 64] \times 4$ denotes 4 cascaded convolution layers with 64 filters of size 33, and $S2$ denotes stride 2. FC is a fully connected layer, for which the number of output neurons are reported.

Module	Layer	Details
FEM	Conv 1	$[3 \times 3, 64] \times 1, S2$ $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 1$
	Conv 2	$[3 \times 3, 128] \times 1, S2$ $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
	Conv 3	$[3 \times 3, 256] \times 1, S2$ $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 4$
	Conv 4	$[3 \times 3, 512] \times 1, S2$ $\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 1$
	FC1	512
PAM	FC2	5
	Pooling	AVE
EAM	FC3	2
RAM	FC4	128
	FC5	2
Coherence	FC6	2

3.2 Personality and Emotion Networks

A shared *FEM*, embodied with a truncated *SphereFace* network [36] with its last two layers removed, is employed for both branches. Those two branches are dedicated to emotion and apparent personality-annotated datasets, respectively, and jointly optimized with the *FEM*.

To utilize rich information from each video frame for more effective network training, the personality network operates on a pool of sparsely sampled face frames from the entire video. Each face frame in this pool can produce its own preliminary prediction of the apparent personality score. We take inspiration from recent advances in video based human action recognition [37] and employ a consensus strategy among all the face frames from each video to give a video-level prediction on the apparent personality. The loss values of video-level predictions, other than those of face-level ones, are optimized by iteratively updating the model parameters. We use V and $\mathbf{Y}$ to represent a generic video input and its ground truth label. Given the i^{th} video $\{V_i^P, \mathbf{Y}_i^P\}(i \in \mathbb{N}^P)$, where $\mathbb{N}^P$ stands for the index set of apparent personality videos, and P denotes the data source, i.e. apparent personality dataset here, we divide it into K segments $\{S_{i1}^P, S_{i2}^P, \cdots, S_{iK}^P\}$ of equal duration. Now our personality network models a sequence of faces as follows:

$$\begin{aligned}\mathbf{P}(V_i, W^P) =& \mathbf{P}(I_{i1}^P, I_{i2}^P, \cdots, I_{iK}^P, W^P) \\ =& \mathcal{G}(\mathcal{F}(I_{i1}^P, W^P), \mathcal{F}(I_{i2}^P, W^P), \cdots, \mathcal{F}(I_{iK}^P, W^P))\end{aligned} \tag{1}$$

Here $(I_{i1}^P, I_{i2}^P, \cdots, I_{iK}^P)$ is a pool of face frames, where each face I_{ik}^P is randomly sampled from its corresponding segment S_{ik}^P. The function $\mathcal{F}(I_{ik}^P, W^P)$ represents the personality network with parameters W^P which operates on face I_{ik}^P and provides preliminary apparent personality scores. The segmental consensus function $\mathcal{G}$ aggregates the raw outputs from multiple face frames to obtain a final apparent personality score for each video. Although the

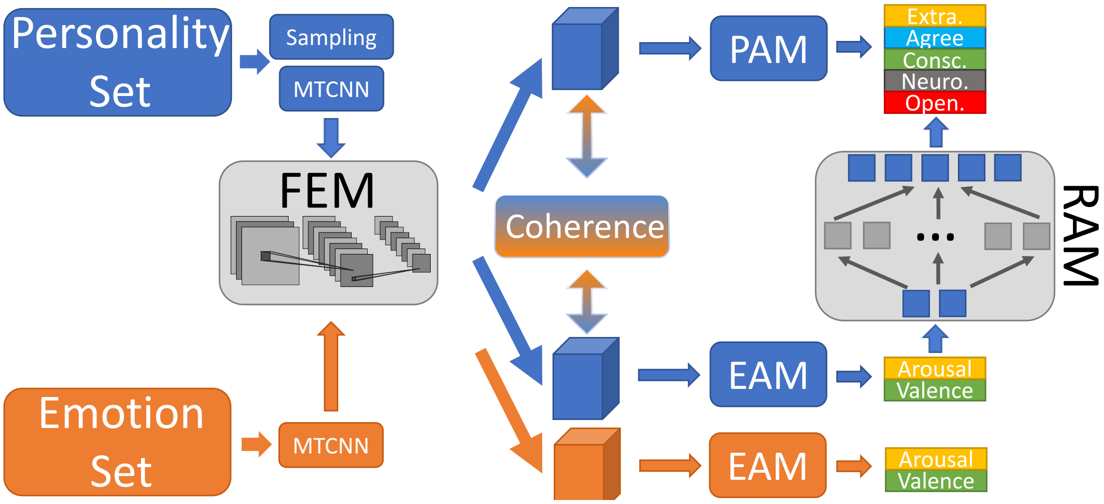


Fig. 1. Workflow of the proposed *PersEmoN*. The colours in the diagram represent the data flow of the system. In the training phase, the system is aware of which dataset the image comes from and will automatically assign the image to its own branch. During inference, we can use *FEM* and *PAM* to obtain the apparent personality traits. Similarly, we can use *FEM* and *RAM* to obtain the emotion outputs. The boxes preceding *PAM* and *EAM* are the feature representations. Please refer to the text for more details.

proposed method is generic and applicable for a wide range of functions such as *max*, *average*, *recurrent aggregation*, we use the *average* function similar to [37]. Based on this consensus, we optimize the personality network with the smooth ℓ_1 loss function [38] defined as:

$$\mathcal{L}_{per}(W^P) = \sum_{i \in \mathbb{N}^P} \text{smooth}_{\ell_1}(\mathbf{Y}_i^P - \mathbf{P}(V_i^P, W^P)) \quad (2)$$

The smooth ℓ_1 function is given below; m represents a margin parameter.

$$\text{smooth}_{\ell_1}(x) = \begin{cases} \frac{1}{2}(x)^2 & |x| < m, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \quad (3)$$

The emotion network works in a simpler manner by directly processing input faces, since frame level annotations are already available. More specifically, given a face image $\{I_i^E, \mathbf{Y}_i^E\}(i \in \mathbb{N}^E)$, the emotion network produces emotion scores as:

$$\mathbf{E}(I_i^E, W^E) = \mathcal{F}(I_i^E, W^E) \quad (4)$$

Similarly, the loss function for the emotion network is:

$$\mathcal{L}_{emo}(W^E) = \sum_{i \in \mathbb{N}^E} \text{smooth}_{\ell_1}(\mathbf{Y}_i^E - \mathbf{E}(I_i^E, W^E)) \quad (5)$$

3.3 Representation Coherence

People may appear in various scales and poses under different illumination conditions for different datasets. Besides, each dataset may exhibit different statistical distributions and annotation bias. Representations learned from each dataset individually without pursuing coherence between them may present significant discrepancy. A representation with good transferability should be dataset-invariant in the sense that the learned representations are coherent for different data samples from different datasets [39]. This is also beneficial to exploring the emotion-to-apparent-personality relationship in our case. To this end, a classifier is trained to classify which dataset the input image comes from. After convergence of the system, it cannot distinguish them because the final representation is dataset-invariant. This strategy reduces over-fitting by learning a generalizable representation, which is applicable not only to the tasks in question, but also to other tasks with significant commonalities. In our setting, since a shared network backbone is employed by two tasks, additional tasks act as a regularization which requires the system to perform well on a related task.

We take inspiration from [39] by training a dataset classifier, denoted as $\mathcal{D}$ with parameters $W^{\mathcal{D}}$, to perform binary classification to distinguish which dataset a particular datum comes from. For each feature representation from the *FEM*, we learn the dataset classifier with the following softmax loss. In each mini-batch, the loss is as follows when sampling from the personality dataset:

$$\mathcal{L}_{\mathcal{D}}^P(W^{\mathcal{D}}) = -\sum_{i \in \mathbb{N}^P} \sum_{k=1}^{K} \log q(I_{ik}^P, W^P, W^{\mathcal{D}}) \quad (6)$$

where $q(I, W, W^{\mathcal{D}}) = \text{softmax}(W^{\mathcal{D}} \cdot \mathcal{F}(I, W))$. Similarly, the loss for samples j from the emotion dataset is:

$$\mathcal{L}_{\mathcal{D}}^E(W^{\mathcal{D}}) = -\sum_{j \in \mathbb{N}^E} \log q(I_j^E, W^E, W^{\mathcal{D}}) \quad (7)$$

The overall loss for each mini-batch can be computed:

$$\mathcal{L}_{\mathcal{D}}(W^{\mathcal{D}}) = \mathcal{L}_{\mathcal{D}}^P(W^{\mathcal{D}}) + \mathcal{L}_{\mathcal{D}}^E(W^{\mathcal{D}}) \quad (8)$$

As in [39], an adversarial-like learning objective is introduced in the *FEM* which aims at "maximally confusing" the two datasets by computing the cross entropy between the output predicted dataset labels and a uniform distribution over dataset labels:

$$\begin{aligned} \mathcal{L}_{adv}(W^P, W^E) = &-\sum_{i \in \mathbb{N}^P} \sum_{k=1}^{K} \log q(I_{ik}^P, W^P, W^{\mathcal{D}}) \\ &-\sum_{i \in \mathbb{N}^E} \log q(I_i^E, W^E, W^{\mathcal{D}}) \\ &+ \log q(I_{ik}^P, W^E, W^{\mathcal{D}}) \\ &+ \log q(I_i^E, W^P, W^{\mathcal{D}}) \end{aligned} \quad (9)$$

Similar to the adversarial-learning, we perform iterative updates for both $\mathcal{L}_{\mathcal{D}}(W^{\mathcal{D}})$ and $\mathcal{L}_{adv}(W^P, W^E)$ given the fixed parameters from the previous iteration.

3.4 Emotion-to-Apparent-Personality Relationship Analysis

Here we investigate whether apparent personality can be inferred directly from emotion attributes. This is challenging due to the paucity of datasets which encompass both emotion and apparent personality annotations for us to learn such a relationship. We insert a relationship analysis module (*RAM*), which receives the emotion scores from the emotion analysis network and predicts apparent personality scores. More specifically, the input of *RAM* can be obtained by:

$$\begin{aligned}\mathbf{E}(V_i^P, W^E) =& \mathbf{E}_i = \mathbf{E}(I_{i1}^P, I_{i2}^P, \cdots, I_{iK}^P, W^E)\\ =&(\mathcal{F}(I_{i1}^P, W^E), \mathcal{F}(I_{i2}^P, W^E), \cdots, \mathcal{F}^E(I_{iK}^P, W^E))\end{aligned} \tag{10}$$

As we already defined, $(I_{i1}^P, I_{i2}^P, \cdots, I_{iK}^P)$ is a pool of faces from the apparent personality dataset where each face I_{ik}^P is randomly sampled from its corresponding segment S_{ik}^P. $\mathcal{F}(I_{ik}^P, W^E)$ represents the emotion network with parameters W^E which operates on face I_{ik}^P to give preliminary results on the emotion scores. *RAM* employs the same consensus strategy among all the faces from the video to output the aggregated apparent personality score $\mathbf{R}$ of video V_i^P:

$$\mathbf{R}(\mathbf{E_i}, W^{\mathbf{R}}) = \mathbf{R}(\mathcal{G}(\mathbf{E}_i), W^{\mathbf{R}}), \tag{11}$$

where $W^{\mathbf{R}}$ represents the weights of *RAM*. *RAM* is trained by optimizing the following objective function:

$$\mathcal{L}_{RAM}(W^{\mathbf{R}}) = \sum_{i \in \mathbb{N}^P} \text{smooth}_{\ell_1}(\mathbf{Y}_i^P - \mathbf{R}(\mathbf{E}_i, W^{\mathbf{R}})) \tag{12}$$

3.5 Overall Loss Functions

Every module of *PersEmoN* is differentiable, allowing end-to-end optimization of the whole system. The learning process of *PersEmoN* aims to minimize the following loss:

$$\begin{aligned}\mathcal{L} = \lambda_1 \mathcal{L}_{per}(W^P) + \lambda_2 \mathcal{L}_{emo}(W^E) + \lambda_3 \mathcal{L}_{\mathcal{D}}(W^{\mathcal{D}})\\ + \lambda_4 \mathcal{L}_{adv}(W^P, W^E) + \lambda_5 \mathcal{L}_{RAM}(W^{\mathbf{R}})\end{aligned} \tag{13}$$

4 Experiments

4.1 Dataset and Evaluation Protocol

We choose two large challenging datasets to evaluate *PersEmoN*. The Aff-Wild emotion dataset [35] consists of 298 YouTube videos (252 for training and 46 for testing) with a total length of about 30 hours (over 1M frames). The videos show the reaction of individuals to various clips from movies, TV series, trailers, etc. Each video is labeled by 6-8 annotators with frame-wise valence and arousal values, with a total of 200 annotators. Both valence and arousal values range from -1 to 1. The representation of emotions via arousal/valence values is illustrated in Fig. 2. For apparent personality, we use the ChaLearn personality dataset [8], which consists of $10k$ short video clips with 41.6 hours (4.5M frames) in total. In this dataset, people face and speak to the camera. Each video is annotated with apparent personality attributes as the Big Five personality traits in $[0,\ 1]$. The annotation was done via Amazon Mechanical Turk.

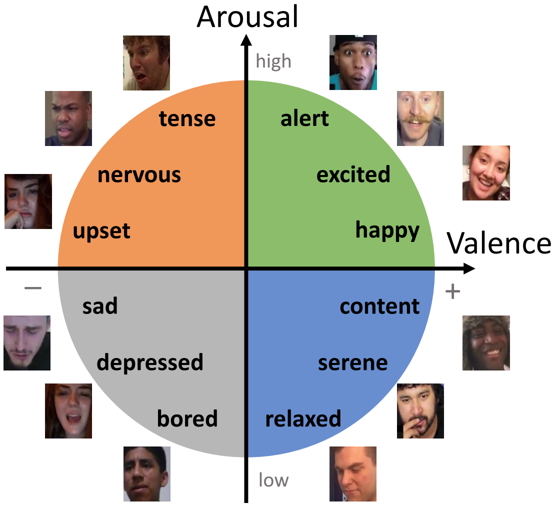


Fig. 2. Emotion wheel showing the connection between emotion categories and arousal-valence space.

To assess the quality of emotion predictions from *PersEmoN*, we calculate the mean square error (MSE) between the predicted values of emotion scores and the ground truth. For the evaluation of the apparent personality recognition, we apply two metrics used in the ECCV 2016 ChaLearn First Impressions Challenge [8], namely mean accuracy A and coefficient of determination R^2, which are defined as follows:

$$A = 1 - \frac{1}{N^t} \sum_i^{N^t} |\mathbf{Y}_i^P - \mathbf{P}_i|, \tag{14}$$

$$R^2 = 1 - \sum_i^{N^t} (\mathbf{Y}_i^P - \mathbf{P}_i)^2 / \sum_i^{N^t} (\bar{\mathbf{Y}}^P - \mathbf{P}_i)^2 \tag{15}$$

where N^t denotes the total number of testing samples, $\mathbf{Y}^P$ the ground truth, $\mathbf{P}_i$ the prediction, and $\bar{\mathbf{Y}}^P$ the average value of the ground truth.

4.2 Implementation

We initialize *FEM* with a truncated 20 layer version of the *SphereFace* model [36]. *PAM* is embodied with a fully connected (FC) layer with 5 outputs, while *EAM* has only 2 output neurons in the FC layer. We use *sigmoid* and *tanh* to squash the outputs for *PAM* and *EAM* respectively. We use a single-hidden-layer feed-forward network to analyze the emotion-to-apparent-personality relationship. More specifically, *RAM* is implemented with two FC layers where the first one receives 2 emotion scores as input and output 128 features with ReLU nonlinearity. The same consensus function and *sigmoid* nonlinearity are used to obtain the apparent personality traits for RAM. Additional architecture details are provided in Table 1.

PersEmoN is implemented in Caffe [40]. We train the whole network with an initial learning rate of 0.01. For each mini-batch, we randomly select 100 images from the Aff-Wild dataset and 10 videos from Chalearn. For each video, 10 frames are further sparsely sampled in a randomized

TABLE 2
Results (MSE) of emotion task on Aff-Wild.

Method	Arousal	Valence
CNN-M [35]	0.140	0.130
MM-Net [23]	**0.088**	0.134
FATAUVA [24]	0.095	**0.123**
DRC-Net [25]	0.094	0.161
PersEmoN (ours)	0.108	0.125

manner, i.e. $K = 10$. Hence, the overall batch size is equal to 200. We train the network for $56k$ iterations and decrease the learning rate by a factor of 10 in the $32k^{th}$ and $48k^{th}$ iteration. As the main goal of the system is to estimate emotion and apparent personality attributes, $\mathcal{L}_{emo}$ and $\mathcal{L}_{per}$ should be the main objective functions, and hence their weights are set to $\lambda_1 = \lambda_2 = 1$. Other loss functions serve as regularizers to further improve the results, and hence their weights are set to relatively smaller values, $\lambda_3 = \lambda_4 = \lambda_5 = 0.1$. The margin parameter in all the smooth ℓ_1 loss (Eq. (3)) is set to $m = 0.05$.

4.3 Evaluation of Emotion

We first report the results of emotion prediction on the Aff-Wild dataset. *PersEmoN* is compared with a strong baseline method CNN-M and 3 benchmark methods from the Aff-Wild challenge [35]. More specifically, MM-Net [23] consists of a carefully designed deep face feature learner to learn discriminative features for affective levels and then employs multiple memory networks for feature aggregation. FATAUVA [24] first learns the facial part-based response through attribute recognition CNNs, which is further used to supervise the learning of action unit (AU) detection. Finally, it employs AUs as a mid-level representation to estimate the intensity of valence and arousal. DRC-Net [25] is based on Inception-ResNet modules redesigned specifically for the task of facial affect estimation. It consists of a shallow Inception-ResNet, a deep Inception-ResNet and an inceptionResNet with LSTMs. These networks extract facial features at different scales and simultaneously estimate both valence and arousal in each frame.

Since annotations of the test data are not public, our results in Table 2 were evaluated by the official organizer. A total of 9 evaluations were obtained from the organizers. As demonstrated in Table 2, our method achieves competitive accuracy to these state-of-the-art methods on the test data.

Simplicity is central to our design; the strategies adopted in *PersEmoN* are complementary to those more complicated approaches, such as ensemble of memory networks used in MM-Net, multiple datasets used for cascade learning employed in FATAUVA-Net and multi-scale inputs adopted in DRC-Net. Furthermore, all these other methods are much more difficult to train than ours. Multiple LSTM layers are used in MM-Net and DRC-Net, while FATAUVA-Net cannot perform end-to-end but cascade training. Although *PersEmoN* was not optimized for emotion recognition like the other methods, it still yields competitive results for the emotion task.

4.4 Evaluation of Apparent Personality

Recognition of Big Five personality traits appears more interesting to us because apparent personality is a higher-level attribute compared to emotion. Table 3 compares some of the latest apparent personality recognition methods. In contrast to other approaches, ours can be trained end-to-end using only one pre-trained model. Moreover, unlike most methods which fuse both acoustic and visual cues, *PersEmoN* uses only video as input.

TABLE 3
Comparison of the deep personality network properties of *PersEmoN* vs. the top teams in the 2016 ChaLearn First Impressions Challenge.

	Fusion	**Modality**		**End-to-End**
		Audio	**Video**	
PersEmoN	late	✗	✓	✓
NJU-LAMDA [29][1]	late	✓	✓	✓
evolgen [28]	early	✓	✓	✓
DCC [26]	late	✓	✓	✓
ucas [8]	late	✓	✓	✗
BU-NKU-v1 [30]	early	✗	✓	✗
BU-NKU-v2 [41][2]	early	✓	✓	✗

[1] winner, 1^{st} ChaLearn First Impressions Challenge (ECCV 2016).
[2] winner, 2^{nd} ChaLearn First Impressions Challenge (ICPR 2016)

The quantitative comparison between *PersEmoN* and state-of-the-art works on apparent personality recognition is shown in Table 4. The teams from NJU-LAMDA to BU-NKU-v1 are the top five participants in the 1^{st} ChaLearn Challenge on First Impressions [8]. Note that BU-NKU was the only team not using audio in the challenge, and their predictions were rather poor comparatively. After adding the acoustic cues, the same team won the 2^{nd} ChaLearn Challenge on First Impressions [42]. Importantly, *PersEmoN* only considers visual streams. Yet as is evident in Table 4, even when only taking into account *PAM*, *PersEmoN* already achieves superior performance over others, not only on the average A and R^2 scores, but both scores for all traits.

Since *RAM* can also predict the apparent personality attributes from the output of *EAM*, as shown in Fig. 1, it can provide our personality network with complementary information. To demonstrate this, we fuse the predicted attributes of both *RAM* and *PAM*; we use late fusion by a weighted average which assigns a weight of 6 to the personality network and 1 to the *RAM* (weights are obtained by performing a grid search on a separate validation set). The results are presented in Table 4 as "*PAM+RAM*". In this case, we observe another performance boost and the highest overall accuracy.

4.5 Emotion-to-Apparent-Personality Relationship

Here we show the possibility of determining apparent personality traits from 2-dimensional affective components. As can be noticed in Table 4 under "Ours (*RAM*)", we achieve satisfactory apparent personality predictions with only 2-dimensional arousal-valence inputs.

An illustration of the emotion-to-apparent-personality relationship is shown in Fig. 3, where each "disk" represents a certain apparent personality trait with respect to the corresponding values of arousal and valence. The discoveries are consistent with [43]: Agreeableness and Conscientiousness are fairly near each other (the two traits share similar emotions), while Neuroticism is located far away from Openness. The "disk" for Extraversion (not shown in

TABLE 4
Apparent personality prediction benchmarking using mean accuracy A and coefficient of determination R^2. Note that there are no R^2 scores reported for BU-NKU-v2.

	Average		*Extraversion*		*Agreeableness*		*Conscientiousness*		*Neuroticism*		*Openness*	
	A	R^2	A	R^2	A	R^2	A	R^2	A	R^2	A	R^2
PAM+RAM	**0.917**	**0.485**	**0.920**	**0.552**	**0.914**	**0.349**	**0.921**	0.570	**0.914**	**0.500**	**0.915**	**0.457**
Ours (*PAM*)	0.916	0.478	**0.920**	0.544	0.913	0.338	**0.921**	**0.571**	0.913	0.489	0.914	0.448
Ours (*RAM*)	0.903	0.373	0.911	0.449	0.908	0.264	0.902	0.349	0.908	0.442	0.907	0.364
NJU-LAMDA [29]	0.913	0.455	0.913	0.481	0.913	0.338	0.917	0.544	0.910	0.475	0.912	0.437
evolgen [28]	0.912	0.440	0.915	0.515	0.912	0.329	0.912	0.488	0.910	0.455	0.912	0.414
DCC [26]	0.911	0.411	0.911	0.431	0.910	0.296	0.914	0.478	0.909	0.448	0.911	0.402
ucas [8]	0.910	0.439	0.913	0.489	0.909	0.292	0.911	0.520	0.906	0.457	0.910	0.439
BU-NKU-v1 [30]	0.909	0.394	0.916	0.514	0.907	0.234	0.913	0.487	0.902	0.363	0.908	0.372
BU-NKU-v2 [41]	0.913	-	0.918	-	0.907	-	0.915	-	0.911	-	0.914	-

the Figure) is close to Agreeableness. This demonstrates that our RAM network indeed has the ability of learning the emotion-to-apparent-personality relationship. Based on this, we believe that *PersEmoN* can serve as a strong practical baseline for automatically annotating apparent personality based on arousal and valence.

4.6 Ablation Study

4.6.1 Effectiveness of Joint Training

Our novel multi-task learning approach aims to learn a generalizable representation, which is applicable not only to the task in question, but also to other tasks with significant commonalities. In *PersEmoN*, since a shared *FEM* is employed by all tasks, additional tasks act as regularization, which requires the system to perform well on a related task. The backpropagation training from different tasks will directly impact the representation learning of shared parameters. It prevents overfitting by solving all tasks jointly and allowing for the exploitation of additional training data.

Table 5 illustrates the effectiveness of this strategy. As the annotations for the test set of Aff-Wild are not released, we divide the original training set into training and validation set with a ratio of $10 : 1$ and evaluate all models on the validation set for the emotion task using MSE. We use the symbols ✓ and ✗ to represent cases where the corresponding functionality is enabled or disabled, respectively. The 2^{nd} and 3^{rd} row in Table 5 shows the case where we train *EAM* and *PAM* on top of *FEM* individually. The 4^{th} row shows the results of both tasks when we disable *RAM*. The last row is our final results when all modules are activated. When we compare the 3^{rd} with the 1^{st} row, we observe an improvement in emotion MSE, which indicates the superiority of jointly training emotion with apparent personality. Similarly, an improvement in apparent personality MSE (3^{rd} vs. 2^{nd} row) verifies that such a strategy is reciprocal. Finally, incorporating the joint training of *RAM* improves the results in both tasks. We believe these improvements originate from the back-propagation training of CNN, during which the shared parameters within the *FEM* directly impact the generalization ability of the whole system.

4.6.2 Consensus Function $\mathcal{G}$

Average temporal pooling has been reported to work well in modeling long-term temporal dependencies for deeply learned representations by [37]. This is also in line with our empirical results on apparent personality recognition. To demonstrate this, we compare average pooling with two other alternatives. One is max pooling, which helps to select the most salient information in its receptive field and has been heavily encoded in popular network structure such as ResNet, VGG and so on. The other is recurrent aggregation, for which we choose the popular LSTM [44]. LSTM has been shown to work better than conventional recurrent networks due to its learnable memory gate to avoid gradient vanishing or explosion.

TABLE 5
Effectiveness of jointly training *PersEmoN*.

Modules in Training			MSE in Prediction	
Emotion	Apparent Personality	Relationship	Emotion	Apparent Personality
✓	✗	✗	0.096	-
✗	✓	✗	-	0.057
✓	✓	✗	0.080	0.033
✓	✓	✓	0.071	0.027

In our implementation, both feature representations from *FEM* as well as LSTM are jointly optimized. More specifically, we train LSTM to aggregate all the feature maps from *FEM* with 10 input frames from the apparent personality dataset. The hidden neurons of LSTM are set to 128. After that, we use average pooling to integrate all temporal information, as done in [45]. Finally, a FC layer is employed to directly regress the apparent personality scores. We achieve an average accuracy of 91.4%, 90.6% and 90.1% for average pooling, max pooling and LSTM, respectively. Max pooling performs worse than average pooling and better than LSTM. This indicates that selecting the most salient information from a video frame does not necessarily capture its overall statistics better. The reason for the weakness of LSTM could be that apparent personality is an orderless concept where temporal dependencies may not be so relevant.

4.6.3 Number of Segments K

In our implementation, $K = 10$. We empirically find that the apparent personality results are not sensitive when K is within $[5, 20]$. However, when both emotion and personality network are jointly optimized, we observe that a balanced input is always beneficial in both tasks. We use a batch size of 100 for both emotion and apparent personality datasets. In this way, 10 input videos for apparent personality are

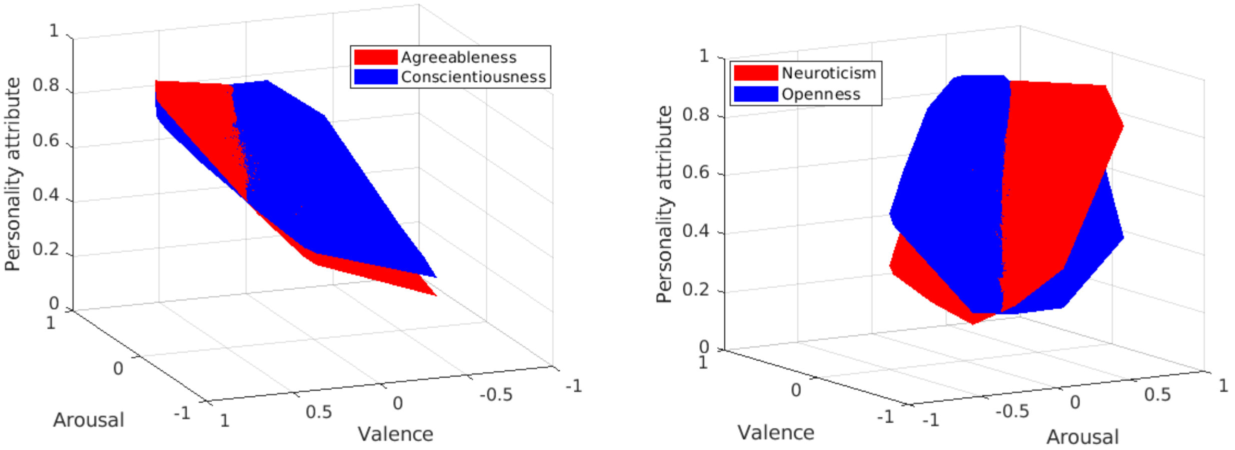
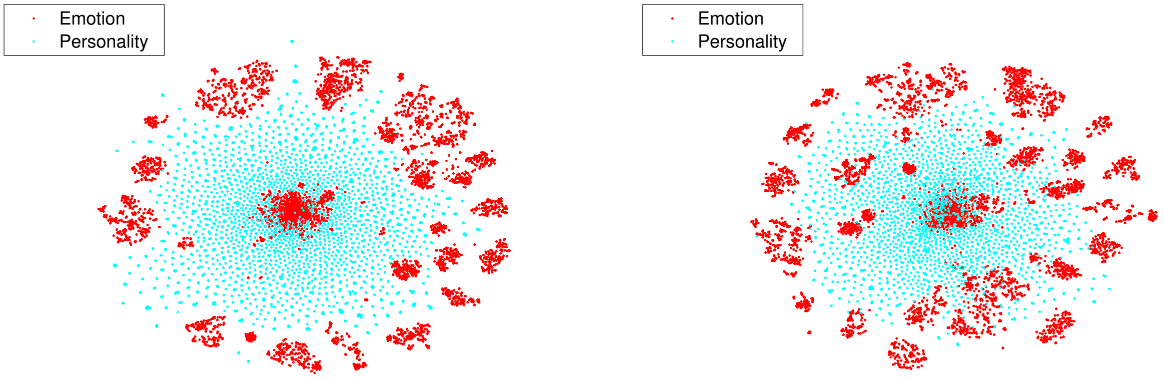


Fig. 3. Illustration of the relationship between various apparent personality traits and the arousal-valence emotion space, acquired from the input and output of *RAM*. Best viewed in color.



(a) Without coherence

(b) With coherence

Fig. 4. Visualization of the distribution of learned features in *PersEmoN* for both emotion and apparent personality datasets with and without coherence strategy. Using the coherence strategy, a large number of features from the emotion dataset are pulled inside the ring, making the two distributions more similar and increasing the overlap between the distributions significantly. Zoom in for more details. Best viewed in color.

used in each batch. Setting K to a larger value, for example 100, will lead to a lower number of either input videos for apparent personality or emotion frames. This further reduces the final performance in both tasks.

4.6.4 Coherence Strategy

As reported by [39], a representation with good transferability should be dataset invariant. To verify this, we also report the MSE scores of two tasks when we remove the coherence strategy from *PersEmoN* in Table 6. We observe that this strategy leads to about 15% improvement in terms of MSE for emotion (Aff-Wild) and apparent personality (Chalearn).

We visualize the distribution of the deeply learned features from *FEM* (the *fc*5 layer of *SphereFace*) in Fig. 4. More specifically, we project the 512-dimensional features on both emotion and apparent personality datasets into 2 dimensional space and visualize their distributions using t-SNE [46]. t-SNE visualizes high-dimensional data by giving each data point a location in a two- or three-dimensional map. The visualizations produced by t-SNE are often significantly better than other advanced techniques. Without a coherence strategy, distributions of those deep features on different datasets can be well classified, i.e. features from emotion dataset are mainly distributed in the outer ring of the x/y plane. Using the coherence strategy, a large number of features from the emotion dataset are pulled inside the ring, making the two distributions more similar and their overlap significantly larger.

TABLE 6
Effectiveness of coherence strategy in *PersEmoN* in terms of MSE.

MSE	With Coherence	Without Coherence
Emotion	0.071	0.082
Apparent Personality	0.027	0.032

5 Conclusions

For the first time, we investigate the feasibility of jointly analyzing apparent personality, emotion, and their relationship within a single deep neural network. This is challenging due to the scarceness of datasets which encompass both emotion and apparent personality annotations. To tackle this issue we propose *PersEmoN*, an end-to-end trainable deep network with two CNN branches called emotion and apparent personality network. With shared bottom feature extraction layers, these two networks regularize each other within a multi-task learning framework, where each one is dedicated to their own annotated dataset. We further employ an adversarial-like loss function to promote representation coherence between heterogeneous dataset sources, which leads to further performance boosts. We demonstrate the effectiveness of *PersEmoN* on two apparent personality and emotion datasets. We find that the proposed joint training of both emotion and apparent personality networks can lead to a more generalizable representation for both tasks.

Acknowledgement

This research is supported by the SERC Strategic Fund from the Science and Engineering Research Council (SERC), A*STAR (project no. a1718g0046).

References

[1] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychological Science*, vol. 17, no. 7, pp. 592–598, 2006.

[2] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Personality and Social Psychology Bulletin*, vol. 35, no. 12, pp. 1661–1671, 2009.

[3] C. Y. Olivola and A. Todorov, "Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences," *Journal of Experimental Social Psychology*, vol. 46, no. 2, pp. 315–324, 2010.

[4] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[5] W. Revelle and K. R. Scherer, "Personality and emotion," in *Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press, 2009, pp. 304–305.

[6] H. J. Eysenck, *Dimensions of Personality*. Transaction Publishers, 1950, vol. 5.

[7] R. A. Depue and P. F. Collins, "Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion," *Behavioral and Brain Sciences*, vol. 22, no. 3, pp. 491–517, 1999.

[8] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn LAP 2016: First round challenge on first impressions – dataset and results," in *Proc. ECCV*, 2016.

[9] J. A. Russell, "A circumplex model of affect," *J. Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, "A deep network for arousal-valence emotion prediction with acoustic-visual cues," *arXiv preprint arXiv:1805.00638*, 2018.

[11] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014.

[12] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[14] S. C. Guntuku, L. Qiu, S. Roy, W. Lin, and V. Jakhetiya, "Do others perceive you as you want them to? Modeling personality based on selfies," in *Proc. 1st International Workshop on Affect & Sentiment in Multimedia*. ACM, 2015, pp. 21–26.

[15] A. Dhall and J. Hoey, "First impressions – predicting user personality from Twitter profile images," in *Proc. International Workshop on Human Behavior Understanding*. Springer, 2016, pp. 148–158.

[16] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. NIPS*, 1994.

[17] S. Peng, L. Zhang, S. Winkler, and M. Winslett, "Give me one portrait image, I will tell you your emotion and personality," in *Proc. ACM Multimedia*, 2018.

[18] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*, 2013.

[19] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi, "Emotion recognition in the wild using deep neural networks and bayesian classifiers," in *Proc. ICMI*, 2017.

[20] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. WACV*, 2016.

[21] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. CVPR*, 2016.

[22] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Computing*, 2017.

[23] J. Li, Y. Chen, S. Xiao, J. Zhao, S. Roy, J. Feng, S. Yan, and T. Sim, "Estimation of affective level in the wild with multiple memory networks," in *Proc. CVPRW*, 2017.

[24] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien, "FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation," in *Proc. CVPRW*, 2017.

[25] B. Hasani and M. H. Mahoor, "Facial affect estimation in the wild using deep residual and convolutional networks," in *Proc. CVPRW*, 2017.

[26] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proc. ECCV*, 2016.

[27] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Trans. Affective Computing*, vol. 9, no. 3, pp. 316–329, 2017.

[28] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *Proc. ECCV*, 2016.

[29] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep

bimodal regression for apparent personality analysis," in *Proc. ECCV*, 2016.

[30] F. Gürpınar, H. Kaya, and A. A. Salah, "Combining deep facial and ambient features for first impression estimation," in *Proc. ECCVW*, 2016.

[31] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE Trans. Affective Computing*, vol. 8, no. 1, pp. 29–42, 2017.

[32] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting CNN models for apparent personality trait regression," in *Proc. CVPRW*, 2017.

[33] J. C. S. Jacques Junior *et al.*, "First impressions: A survey on computer vision-based apparent personality trait analysis," *arXiv preprint arXiv:1804.08046*, 2018.

[34] H. J. Escalante *et al.*, "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," *arXiv preprint arXiv:1802.00745*, 2018.

[35] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal in-the-wild challenge," in *Proc. CVPRW*, 2017.

[36] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017.

[37] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*, 2016.

[38] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015.

[39] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. ICCV*, 2015.

[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ICMI*, 2014.

[41] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *Proc. ICPR*, 2016.

[42] H. J. Escalante, V. Ponce-López *et al.*, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *Proc. ICPR*, 2016.

[43] M. S. Yik and J. A. Russell, "Predicting the big two of affect from the big five of personality," *J. Research in Personality*, vol. 35, no. 3, pp. 247–277, 2001.

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

[45] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. CVPR*, 2016, pp. 1325–1334.

[46] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Machine Learning Research*, vol. 9, pp. 2579–2605, Nov 2008.

Le Zhang received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC) in 2011. He received his M.Sc. and Ph.D. degrees from Nanyang Technological University (NTU) in 2012 and 2016, respectively. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Prior to that, he was a postdoctoral researcher at the University of Illinois' Advanced Digital Sciences Center (ADSC), Singapore. His current research interests include deep learning and computer vision.

Songyou Peng is currently a PhD student in Max Planck ETH Center for Learning Systems, a joint program between ETH Zurich and the Max Planck Institute for Intelligent Systems. He received the Erasmus Mundus M.Sc. in Computer Vision and Robotics in 2017 and received B.Eng degree from Xi'an Jiaotong University in 2015. He was a research engineer at Institute for Infocomm Research, A*STAR, Singapore. Prior to that, he was a research engineer at Advanced Digital Sciences Center, a research center of the University of Illinois at Urbana-Champaign. His research interests are computer vision and machine learning.

Stefan Winkler is Deputy Director at AI Singapore and Associate Professor at the National University of Singapore (NUS). Prior to that, he was Distinguished Scientist and Program Director at the University of Illinois' Advanced Digital Sciences Center (ADSC) in Singapore. He also co-founded two start-ups and worked for a Silicon Valley company. He has a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, and a Dipl.-Ing. (M.Eng./B.Eng.) degree from the University of Technology Vienna, Austria. He is an IEEE Fellow and has published over 150 papers. His research interests include video processing, computer vision, machine learning, perception, and human-computer interaction.