Learning Person-specific Cognition from Facial Reactions for Automatic Personality Recognition

Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar and Hatice Gunes

Abstract—This paper proposes to recognise the true (self-reported) personality traits from the target subject's cognition simulated from facial reactions. This approach builds on the following two findings in cognitive science: (i) human cognition partially determines expressed behaviour and is directly linked to true personality traits; and (ii) in dyadic interactions, individuals' nonverbal behaviours are influenced by their conversational partner's behaviours. In this context, we hypothesise that during a dyadic interaction, a target subject's facial reactions are driven by two main factors: their internal (person-specific) cognitive process, and the externalised nonverbal behaviours of their conversational partner. Consequently, we propose to represent the target subject's (defined as the listener) person-specific cognition in the form of a person-specific CNN architecture that has unique architectural parameters and depth, which takes audio-visual non-verbal cues displayed by the conversational partner (defined as the speaker) as input, and is able to reproduce the target subject's facial reactions. Each person-specific CNN is explored by the Neural Architecture Search (NAS) and a novel adaptive loss function, which is then represented as a graph representation for recognising the target subject's true personality. Experimental results not only show that the produced graph representations are well associated with target subjects' personality traits in both human-human and human-machine interaction scenarios, and outperform the existing approaches with significant advantages, but also demonstrate that the proposed novel strategies help in learning more reliable personality representations.

Index Terms—True personality recognition, Dyadic interaction, Person-specific cognition simulation, Facial reaction generation, End-to-end graph representation learning, Multi-dimensional edge feature

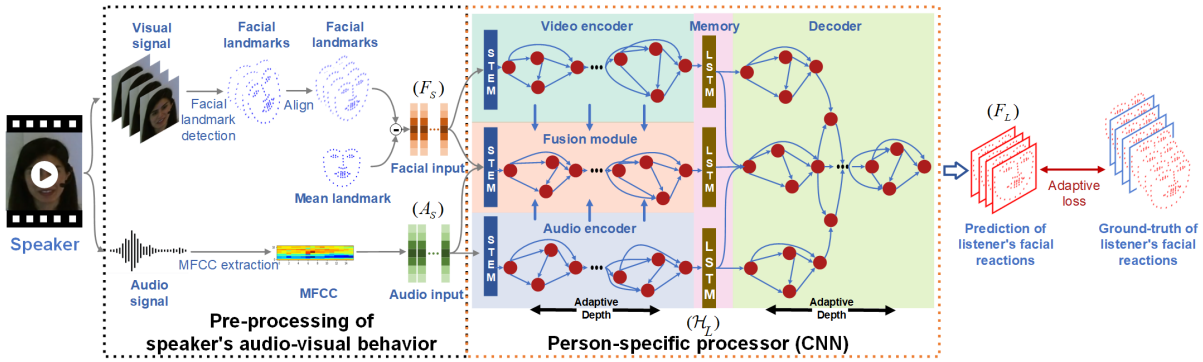
1 Introduction

Understanding human personality can benefit a wide range of applications such as (mental) health condition analysis [1], [2], candidate screening for recruitment [3], as well as personalised, adaptive human-agent interactions (e.g., [4]). Recent advances in machine learning (ML) have enabled the development of non-invasive automatic personality trait analysers that recognise subjects' personality traits from their audio-visual non-verbal behaviours [5], [6], [7], [8], [9], [10] as there is solid psychological and biological evidence [11], [12], [13], [14] claiming that nonverbal behaviours are reliable predictors of personality. In most of these approaches, ML models are trained with the personality labels provided by the external observers (annotators). Therefore, these ML models play the role of an external artificial observer that observes the target subjects' nonverbal distal cues, i.e., audio signals (e.g., delta-mel-cepstral, speech duration, pitch, and pause rate, etc.) [8], [15], [16], [17], visual cues (e.g., facial actions and gestures) [18], [19], [20], observable inter-personal cues [8], [21], [22] etc., and output the external observer's perception of the target subjects' personality. However, people externalize their personality through distal cues (e.g., energy), which undergo a perception bias based on what the observer actually perceives, and become proximal cues (e.g., loudness). As a result, the aforementioned approaches can be treated as Automatic Personality Perception (APP) solutions (inference from proximal cues) [23].

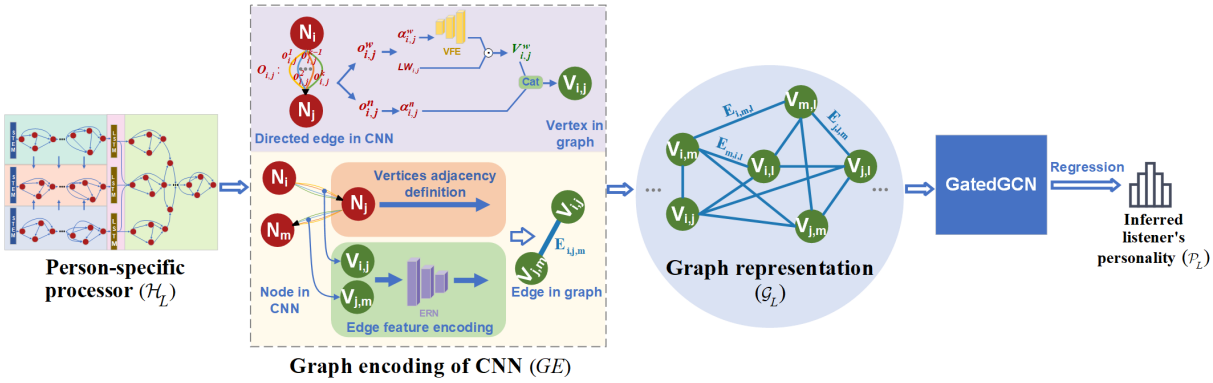
In some scenarios, the goal is to infer true personality from machine detectable distal cues, i.e., Automatic Personality Recognition (APR) [23]. While APP approaches predict apparent personality (perception) based on proximal behavioural cues, APR aims to recognise the true personality that impacts the generation of distal behavioural cues. Thus, APP models that were trained as external observers to provide personality perceptions may not be reliable for recognising true personality traits (**Problem 1**). Moreover, the majority of these APP solutions [8], [19], [20], [24], [25] recognise personality traits from single frames or thin slices of behaviour, independently, by re-using clip-level personality labels as the frame/thin slice-level labels to train ML models that can provide a personality prediction for each frame/thin slice. This is problematic as people with different personality traits may express very similar nonverbal audio-visual behaviours in a single frame or a thin slice. As a result, such strategies may lead to the same input pattern being paired with multiple labels during the training, making them theoretically impossible to learn a good hypothesis **(Problem 2)**. Although recent approaches [5], [6], [26] address this issue by modelling personality using an entire clip, (i.e., recognising personality traits at the

- *Siyang Song and Hatice Gunes are with the AFAR Lab, Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FT, United Kingdom. E-mail: ss2796@cam.ac.uk, Hatice.Gunes@cl.cam.ac.uk*
- *Zilong Shao and Linlin Shen are with Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. Linlin Shen is also with Shenzhen Institute of Artificial Intelligence and Robotics for Society, PR China and the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China. E-mail: shaozilong2019@email.szu.edu.cn, llshen@szu.edu.cn. (Corresponding Author: Prof. Linlin Shen. E-mail: llshen@szu.edu.cn)*
- *Shashank Jaiswal is with the BlueSkeye AI, Nottingham, United Kingdom. E-mail: shashank@blueskeye.com*
- *Michel Valstar is with the Computer Vision Lab, School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, United Kingdom. E-mail: michel.valstar@nottingham.ac.uk*

Manuscript received April 29, 2022



(a) Neural Architecture Searching (NAS) for a person-specific CNN to simulate the target subject's (listener's) cognition.



(b) Recognition of the target subject's (listener) personality based on the graph representation that summarises the architecture and parameters of the target subject's person-specific CNN model.

Fig. 1. The pipeline of the proposed approach. (a) Our approach starts with searching for a person-specific processor (multi-modal CNN) architecture (unique topology, weights and depth) that can reproduce the target listener's facial reactions according to the speaker's audio-visual non-verbal signals (Sec. 3.1); (b) Then, we parameterize the person-specific processor as a graph representation to represent the listener's cognition and feed it to a graph neural network for the target listener's personality recognition (Sec. 3.2). It should be noted that we individually **search for a person-specific processor with a unique architecture and parameters for each subject** as the person-specific cognition/personality representation.

clip-level), they only select a set of key frames to represent an entire clip. This may ignore the short-term behaviours displayed by the discarded frames (**Problem 3**), which may contain crucial cues for personality recognition.

In this paper, we propose a novel audio-visual automatic true personality recognition framework that addresses the problems highlighted above. It is built on the definition that true personality influences the cognitive process of individual's distal cues externalization [23] (e.g., facial reactions). In particular, recent works [27], [28] show that in dyadic and group interactions, subjects' nonverbal behaviours (e.g., facial reactions) are influenced by, and therefore can be predicted from, the behaviours of their conversational partner(s). Therefore, this paper assumes that during a dyadic interaction, the target subject's (listener) facial reactions are driven by two main factors: (i) the target subject's internal (person-specific) cognition, and (ii) the externalised nonverbal behaviours of the conversational partner (the speaker). Therefore, we propose to learn a person-specific CNN for each subject, which reproduces the subject's facial reaction in response to the conversational partner. Consequently, the explored person-specific CNN can represent the target subject's cognitive process during the facial reaction generation, which is well associated with the subject's true personality (**addressing Problem 1**). More importantly, each person-specific CNN is explored using the behaviours contained in all available frames of the target video (**addressing Problem 3**) and thus its architecture and parameters contain the clip-level information, which is then encoded as a graph representing the target subject's personality. This allows the training of the GNN-based personality model to be implemented by pairing the clip-level representation with the clip-level personality labels (**avoiding Problem 2**). The pipeline of the proposed approach is illustrated in Fig. 1. The main contributions of this paper are summarised as follows:

- We propose to use the simulated person-specific cognition of the target subject as the source descriptor to recognise the subject's true personality traits. To the best of our knowledge, this is the first audio-visual approach that uses person-specific CNN architecture and weights to represent the target subject's cognition, and recognizes the true (self-reported) personality traits from the simulated cognition.
- We propose a novel audio-visual non-invasive human person-specific cognition simulation strategy which automatically searches for an optimised multi-modal person-specific CNN for each subject to reproduce the subject's facial reactions. The explored person-specific CNN has a unique combination of layers (operations), weights and depth, and plays the role of the target subject's person-specific cognitive process for generating the unique and person-

specific facial reactions.
- We propose a novel graph encoding strategy to parameterize the unique architecture and parameters of an explored CNN into an graph representation, where each CNN edge that contains a set of operations (convolution, pooling, etc.) is treated as a vertex, while the existence of edges between vertices in the graph are decided by the CNN's architecture.
- We propose a novel transformer-based feature learning strategy that deep learns a task-specific multi-dimensional edge feature for each pair of adjacent vertices in the graph representation. To the best of our knowledge, this is the first approach that deep learns multi-dimensional edge features for graphs to represent convolution neural network architectures.
- We conduct a set of experiments under both human-human and human-machine dyadic interaction settings, which not only validate the superior performance of the proposed approach in recognising true personality traits but also systematically demonstrate the influence of the various internal (methodological) and external (subject demographic) factors on the proposed approach.

Compared with our earlier conference version [29], the extended journal version has following additional contributions and novelties:

Methodologies: Firstly, we introduce a depth searching strategy, allowing each person-specific CNN to not only have unique architectural parameters but also a unique depth. In addition to aligning all person-specific CNNs as graph representations with the same topology as the conference version, we further propose to encode CNNs of variable depths as heterogeneous graph representations that have different typologies. **Secondly**, we propose a novel end-to-end vertex feature learning strategy to encode task-specific vertex features from corresponding OPs and LWs, replacing the hand-crafted vertex feature encoding strategy introduced in the conference version. **Finally**, we propose a novel transformer-based multi-dimensional edge feature learning strategy which employs attention operations to learn salient task-specific relationship cues between vertices.

Experiments: Firstly, we have conducted additional ablation studies for different demographic groups. **Secondly**, we have added an experiment that compares the proposed approach (i.e., using the weights and architectural parameters of the explored person-specific CNN as the person-specific cognition representation), with the system that uses the personalized weights of the standard CNN architecture (ResNet) as the person-specific cognition representation. **Thirdly**, we have conducted additional ablation studies for evaluating the new methodological contributions described above. **Finally**, we have conducted all experiments on an additional self-reported personality dataset that was collected under human-machine interaction scenarios.

Presentations: Firstly, we have added a detailed overview with a set of formulations to explain the full pipeline of the proposed approach at the beginning of the Sec. 3. **Secondly**, we have added texts and figures to explain the new methodological contributions and new experimental results described above. **Thirdly**, we additionally provide the pseudocode of the VFE and EFE in the supplementary material. **Finally**, we provide the detailed settings of all reproduced baselines in the supplementary material.

2 Related Work

This section first reviews previous audio-visual automatic personality analysis approaches in Sec.2.1. Then, it summarizes biological and psychological studies which found that personality can be reflected by human cognition, providing the theoretical basis for our work, i.e., recognising true personality traits from the simulated human cognitive processes (Sec. 2.2).

2.1 Audio-visual automatic personality analysis

Early audio-visual automatic personality analysis approaches usually extract hand-crafted features to describe audio-visual human non-verbal behaviours or interpersonal relationship between subjects, including low-level features such as histogram of oriented gradients (HOG) [30], Local Phase Quantization (LPQ) [31] and mid-level cues such as statistics of mid-level behaviour attributes [8], [32], facial attributes (gazes, head motions, etc.) [33], human posture and gesture cues while speaking [34], co-occurrent patterns of behaviours [35], body skeleton activity [21], Quantised Local Zernike Moments (QLZM) [9], visual focus of attention [21], etc. These hand-crafted features are then fed to traditional machine learning models such as support vector machine regressor (SVR) or logistic regression to generate apparent personality predictions.

Due to recent advances in deep learning, most existing approaches employ Convolution Neural Networks (CNNs) to learn task-specific deep features from each frame or a thin video slice. For example, Ventura et al. [7] propose a Descriptor Aggregation Network (DAN) to extract a frame-level feature at multiple spatial resolutions, and use such multi-level visual features to infer personality at the frame-level. To learn both personality-related audio and visual cues, two-stream bi-modal networks are proposed in [20] and [19], which firstly learn frame-level audio-visual features and then combine them at the the fully connected layer to provide frame-level personality prediction. The video-level prediction is then obtained by averaging predictions of all frames. Principi et al. [18] propose a multi-modal CNN to jointly learn audio and visual information from every image sequence (thin video slice) and audio segment. The extracted features are combined with attribute-specific models to predict personality traits.

Since personality trait models focus on evaluating the aspects of personality that are relatively stable over a long period of time for the target subject [36] (usually much longer than the duration of a single audio-visual clip), the frame/thin slice-level behaviours may not be reliable in reflecting personality traits [36]. Consequently, approaches that model personality traits based on clip-level/long-term behaviours also have been investigated. One popular solution is to summarise frame/thin slice-level features of an entire audio-visual clip into a global statistical descriptor to infer personality [34], [37], e.g., averaging all

frame/thin slice-level vectors or using a histogram to represent frame/thin slice-level feature distributions. To consider important dynamic cues, Li et al. [5] first divide the video into 32 slices, and then randomly select a face image and a face-background image from each of them, which are then stacked as the clip-level stream. Subsequently, the video-level prediction is made by these selected frames. Beyan et al. [26] propose to generate multiple dynamic facial images [38], [39], [40] to represent each video segment and then choose a set of dynamic facial images that have the highest spatio-temporal saliency as the key frames to construct the video-level representation.

In addition, multiple audio-visual personality computing datasets [1], [41], [42], [43], [44], [45], [46], [47], [48], [49] have been proposed and recorded in the past decades, where a large part of them [41], [42], [43], [44] are labelled with apparent personality traits (i.e., external human observers annotate personality traits for each clip according to their impressions). These datasets are either recorded under lab conditions (i.e., participants are asked to conduct a set of tasks) [43], [44] or video blogs collected from public websites [41], [42]. Several datasets provided self-reported big-five personality traits annotations [1], [45], [46], [47], [48], where the NoXI and UDIVA datasets [46], [47] were recorded under human-human dyadic interaction scenarios, while the VHQ dataset [1] was recorded under three human-robot dyadic interaction scenarios. In addition, the MHHRI dataset [48] was collected for personality computing under both human-human and human-robot interaction scenarios.

In summary, while modelling personality traits at the frame/segment-level is problematic, the recent clip-level representations usually failed to utilise the full scale of the available information in the data, as they select a subset or key frames to represent an entire video. To avoid these problems, Song et al. [10] propose a domain adaption approach to learn a set of intermediate convolution layers from all available data as the person-specific representation for the target subject, which achieved a comparable performance to the state-of-the-art method [5]. However, similar to the approaches described above, it still directly infers apparent personality based on the subjects' observable behaviours. In other words, all the aforementioned studies focused on automatic personality perception analysis.

2.2 The relationship between personality and human cognition

According to previous biological studies [11], [12], personality traits (e.g., Extraversion, Conscientiousness, and Neuroticism) are well associated with human brain structures [50] and activities such as brain local volumes [51] and gray and white matter [52], which are key factors in deciding and controlling human cognitive processes. For example, Kumari et al. [14] investigated brain fMRI activity based on the "n-back" task, and found that brain responses during cognitive activities are related to Extraversion and Neuroticism traits. Previous psychological studies also frequently claimed that people's personality is well associated with their cognitive processes in various daily activities such as risk taking [53], creativity [13], [54], and music learning [13]. An exploratory factor analysis was conducted by [13],

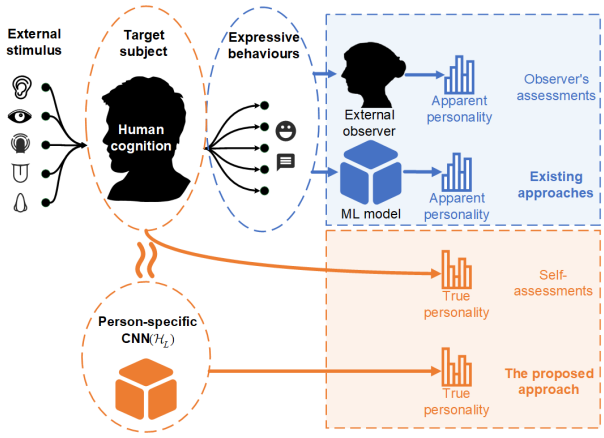


Fig. 2. The difference between the proposed approach (depicted in orange) and existing approaches (depicted in blue). While existing approaches attempt to directly using the target subject's external non-verbal behavioural data to predict personality perception, our approach learns to recognise true personality by modelling the target subject's internal person-specific cognition.

whose results show that creativity and primary cognitive processes are correlated with the Extraversion and psychoticism (Neuroticism) traits. Importantly, the relationship between human cognition and personality are relatively stable, as a longitudinal study conducted by Schaie et al. [55] showed that some of the personality-cognition relations could last for over 35 years. This finding gives us the inspiration that human cognition can be a reliable and stable source for recognising personality.

As reviewed in Sec. 2.1, the main difference between our approach and the existing approaches (illustrated in Fig. 2), is the fact that the existing approaches attempt to achieve automatic personality perception directly from observable non-verbal behaviours of the target subjects, where the ML model acts as an external observer. Instead, our approach draws inspiration from the aforementioned works on the interrelationship between personality and human cognition, and learns to recognise true personality by simulating and modelling target subjects' person-specific cognitive processes.

3 Methodology

The proposed approach recognises each subject's true personality traits based on three steps: (i) person-specific cognition (CNN) simulation (Sec. 3.1); (ii) person-specific graph representation generation (Sec. 3.2); and (iii) personality recognition based on the produced person-specific graph representation (Sec. 3.3).

Person-specific cognition (CNN) simulation: our approach starts with simulating and modelling each target subject's (listener) cognition by individually searching for an optimal person-specific multi-modal CNN $\mathcal{H}_L$. The explored person-specific CNN is expected to accurately reproduce the listener's facial reactions F_L in response to the conversational partner's (the speaker) audio A_S and facial behaviours F_S, i.e., the signals that the subject received (ex-

plained in Fig. 1(a)) in a dyadic interaction. Mathematically speaking, the $\mathcal{H}_L$ is achieved by:

$$\mathcal{H}_L = \text{NAS}(F_L, A_S, F_S) \tag{1}$$

where NAS denotes the DARTs-based neural architecture searching algorithm. Here, the $\mathcal{H}_L$ is defined by its depth $\mathcal{D}_L^{\text{CNN}}$, operation parameters (OPs) $\mathcal{O}_L$ and layers' weights (LWs) $\mathcal{W}_L$ of operations:

$$\mathcal{H}_L = \{\mathcal{D}_L^{\text{CNN}}, \mathcal{O}_L, \mathcal{W}_L\} \tag{2}$$

In summary, we individually search for a person-specific CNN for each listener by considering the listener's facial reaction as well as the corresponding speaker's audio and facial behaviours. Specifically, in this stage, our goal is to adjust the person-specific CNN to fit the provided listener-speaker dyadic interaction data, where we search and validate the person-specific CNN on the same data. Representative loss curves for the search process are illustrated in Fig. 9.

Person-specific graph representation generation: In this paper, we hypothesize that the well explored CNN $\mathcal{H}_L$ represents the person-specific cognition of the target listener, and thus $\mathcal{H}_L$ is well associated with the listener's true personality. However, it is not possible to directly feed the explored CNN to a ML predictor for personality recognition, as it can not be directly processed by any existing ML model. Since each CNN network can be well described by a graph, where a set of layers that contain parameters can be treated as vertices and their connection relationship can be treated as edges, we parameterize the $\mathcal{H}_L$ into a learnable graph representation $\mathcal{G}_L(V, E)$ as the corresponding listener's person-specific cognition representation for personality recognition:

$$\begin{aligned} \mathcal{G}_L(V, E) =& \text{GE}(\mathcal{H}_L) \\ =& \text{GE}(\mathcal{D}_L^{\text{CNN}}, \mathcal{O}_L, \mathcal{W}_L) \end{aligned} \tag{3}$$

where GE denotes the proposed graph encoding strategy (explained in Fig. 1(b)); V and E represent the nodes and edges of the graph representation $\mathcal{G}_L$.

Personality recognition: Finally, the produced graph representation $\mathcal{G}_L$ is fed to a GNN model to recognise the target listener's true personality as:

$$\mathcal{P}_L = \text{GNN}(\mathcal{G}_L) \tag{4}$$

where $\mathcal{P}_L$ represents the predicted five personality traits of the target listener.

3.1 Simulating person-specific cognition

This section explains how we search for a person-specific multi-modal CNN that represents the target listener's cognition. Specifically, we introduce the input and target of the CNN (Sec. 3.1.1), the CNN settings that allows each person-specific CNN to accurately simulate the cognition of the target listener (Sec. 3.1.2), the loss function for searching and training person-specific CNNs (Sec. 3.1.3), and the architectural parameters' optimization strategy (Sec. 3.1.4). The complexity analysis of the person-specific searching is provided in the supplementary material.

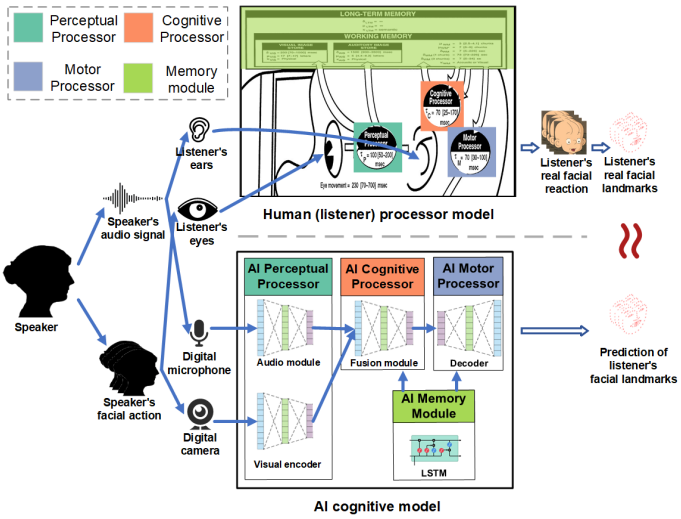


Fig. 3. Using person-specific CNN to simulate human cognition.

3.1.1 Input and target

Previous findings [27], [28] suggest that during a dyadic interaction, the listener's facial reactions are driven by two main factors: (i) listener's **person-specific cognition**, and (ii) the externalised nonverbal behaviours of the conversational partner (the **speaker**). Based on this, the person-specific CNN model $\mathcal{H}_L$ that represents the cognition of the listener is explored to output facial reactions F_L of the listener when given audio signal A_S and facial behaviours F_S of the speaker as the input. This can be formulated as:

$$F_L = \mathcal{H}_L(A_S, F_S) \tag{5}$$

Once $\mathcal{H}_L$ is obtained, it takes on the role of the corresponding listener's cognitive processor in generating facial reactions during the provided dyadic interaction. Consequently, the learnt $\mathcal{H}_L$ is *sufficiently* informative for modeling the listener's true personality traits not only because the true personality relates to the listener's cognition but also because true personality is a key factor in governing how non-verbal behaviours are generated and displayed by humans [23]. In this paper, we use the speaker and listener's facial landmark sequences to represent the input and target facial movements, respectively. In this paper, we empirically set the sequence length as 80 frames (around 3 seconds). This is because that this duration is not only enough to contain a complete facial behaviour/reaction that consists of multiple facial expressions [56], but also not too long to contain several reactions in response to multiple stimulus. The aligned facial landmarks are obtained for each frame using OpenFace 2.0 [57], which are then transformed based on a pre-defined mean face shape in order to keep only facial behaviours without the identity information (as suggested by [58]). Also, we use 64 bin log-mel spectra as the audio representation, where each audio frame is computed by a 40 ms hanning window with stride size of 40 ms. This way, the number of audio frames for each video is the same as the number of video frames.

3.1.2 Multi-modal cognitive processor model

Basic topology: The basic topology of each person-specific CNN is inspired by the Model Human Processor (MHP) [59] (visualized in Fig. 3), which is set to have a visual encoder and an audio encoder that simulate the *human perceptual processor*, an audio-visual decoder that simulates the *human motor processor* and a fusion module that partially simulates the *human cognitive processor*, i.e., jointly processing audio-visual cues at multiple levels. Since during each person-specific CNN's optimisation, it takes audio-visual sequences of the speaker, and outputs facial reaction sequence of the listener, we also employ a Long-Short-Term-Memory network (LSTM) to process the latent feature sequence generated from the two encoders, which aims to simulate the *human working memory module* of the MHP. The basic topology of each person-specific CNN is also illustrated in Fig. 1(a)).

Model settings for person-specific cognition simulation: We follow the similar settings of the DARTs [60] to represent each module as a directed acyclic sub-graph that is made up of several cells. Each cell contains a set of nodes that represent latent features as well as a set of CNN edges, where each edge contains a set of operations defined by OPs and LWs. Specifically, in the proposed person-specific CNNs, a node N_j represents a set of feature maps generated from its adjacent parent nodes $N_i, N_{i+1}, \cdots N_{j-1}$ $(i < j)$, where a pair of adjacent nodes (N_i and N_j) are connected by a CNN edge $O_{i,j}$ which consists of a set of pre-defined operations $o^k_{i,j}$ (e.g., convolution, pooling, etc.):

$$O_{i,j} = \{o^k_{i,j} | k = 1, 2, \cdots, K\} \tag{6}$$

where each $o^k_{i,j}$ contains a set of layer weights (LWs) $w^k_{i,j}$ (e.g., kernel weights of a convolution layer). In particular, for those operations that do not have learnable LWs (e.g., pooling, identity mapping, etc), we define their LWs as $o^k_{i,j} = \varnothing$. As a result, the Eqa. 6 can be re-written as:

$$O_{i,j} = \{o^k_{i,j}(w^k_{i,j}) | k = 1, 2, \cdots, K\} \tag{7}$$

During the propagation, the feature maps in the node N_j are produced from all of its adjacent parent nodes N_i $(i < j$ and $textadj(i, j) = 1)$ via all operations of corresponding CNN edges $O_{i,j}$, which can be formulated as:

$$N_j = \sum^{i<j} (O_{i,j}(N_i) \times \text{adj}(i, j)) \tag{8}$$

where $\text{adj}(i, j)$ denotes the connectivity between N_i and N_j (i.e., 0 denote N_i and N_j are not connected while 1 denoting they are connected). Here, each operation $o^k_{i,j}$ in $O_{i,j}$ is assigned to have a operation parameter (OP) $\alpha^k_{i,j}$ to represent its importance. This way, when feeding feature maps contained in the node N_i to a CNN edge $O_{i,j}$, the output can be represented as:

$$O_{i,j}(N_i) = \sum_{k=1}^{K} (\alpha^k_{i,j} \times o^k_{i,j}(w^k_{i,j}(N_i))) \tag{9}$$

This process is also illustrated in Fig. 4(c). To simulate uncertain and complex human cognitive processes of facial reactions, we set the n_{th} $(n > 2)$ fusion cell to take four inputs: the outputs of the n_{th} visual cell C^{Visual}_n and n_{th} audio cell C^{Audio}_n, the output of the $(n-1)_{th}$ and $(n-2)_{th}$ fusion cells (C^{Fusion}_{n-1}, C^{Fusion}_{n-2}), which can be formulated as:

$$C^{\text{Fusion}}_n = \|\{C^{\text{Fusion}}_{n-1}, C^{\text{Fusion}}_{n-2}, C^{\text{Audio}}_n, C^{\text{Visual}}_n\} \tag{10}$$

where $\|$ is the concatenation operator. Consequently, the input audio and visual signals can be combined and jointly processed at multiple levels (illustrated in Fig. 1(a)). Secondly, in each cell, we set each node to connect to all of its previous nodes to represent all possible information flow, allowing the extracted features (nodes) to be potentially influenced by the information of multiple previous states (parent nodes) during the CNN propagation (illustrated in Fig. 4(b)). Thirdly, we set each CNN edge to have a set of unique OPs and LWs rather than setting all cells to share the same set of OPs [60], [61]. Finally, since depth is also a key factor that impacts a CNN's cognitive process, we also search for a unique number of cells for the person-specific CNN of each target listener (illustrated in Fig. 1(a)). This way, the person-specific CNN $\mathcal{H}_L$ that represents the target listener's cognition can be defined as:

$$\begin{aligned} \mathcal{H}_L =& \{\mathcal{D}^{\text{CNN}}_L, \mathcal{O}_L, \mathcal{W}_L\} \\ =& \{(\mathcal{O}^{\text{Audio}}_L, \mathcal{O}^{\text{Video}}_L, \mathcal{O}^{\text{Fusion}}_L, \mathcal{O}^{\text{Decoder}}_L), \\ & (\mathcal{W}^{\text{Audio}}_L, \mathcal{W}^{\text{Video}}_L, \mathcal{W}^{\text{Fusion}}_L, \mathcal{W}^{\text{Decoder}}_L), \\ & (\mathcal{D}^{\text{AVF}}_L, \mathcal{D}^{\text{Decoder}}_L)\} \end{aligned} \tag{11}$$

where $\mathcal{D}^{\text{AVF}}_L$ denotes the depth of the audio, visual and fusion modules, i.e., these three modules are set to have the same number of cells. In summary, compared with training a person-specific CNN with a fixed architecture for each subject [10], which only represents the person-specific cognition of the subject using a set of unique LWs, the person-specific CNN explored by our approach allows the subject's cognition to be represented by not only a set of unique LWs but also unique OPs and depth (i.e., the architecture of the CNN). In other words, the complex human cognition would theoretically be better represented by the CNN explored by our approach (evaluated in Sec. 4.4).

Employed operations (search space): In this paper, we pre-define $\upsilon = 5$ operations that have layer weights (LWs), and $\kappa = 5$ operations that do not have LWs for each edge $O_{i,j}$. Here, all OPs and LWs of the target listener's person-specific CNN are defined as α_L and $\mathcal{W}_L$, respectively. The details of the person-specific CNN settings (e.g., cell, nodes and operations) are provided in Table. 1. As the first work that searches for a person-specific CNNs to represent each listener's person-specific cognition as their personality representation, there is no previous study suggesting the optimal operations and search space. Since the main goal of this paper is to validate the concept that the personalized architecture and weights of the explored person-specific CNN can reflect the corresponding listener's self-reported personality traits, we found that the standard deep learning operations (convolution, pooling, etc) that have been frequently used in previous deep learning models (e.g., classification [62], segmentation [63], etc.) can already allow most explored person-specific CNNs to accurately reproduce their target listeners' facial reactions. Consequently, we decided to define the search space based on these standard deep learning operations. Although more operations can be

Operation Name	Size	Num of LWs
Max Pooling	1×3	0
Average Pooling	1×3	0
Separable Convolution	1×3	$3 \times C_{\text{in}} \times C_{\text{out}}$
Separable Convolution	1×5	$5 \times C_{\text{in}} \times C_{\text{out}}$
Dilated Convolution	1×3	$3 \times C_{\text{in}} \times C_{\text{out}}$
Dilated Convolution	1×5	$5 \times C_{\text{in}} \times C_{\text{out}}$
Transposed Convolution	1×3	$3 \times C_{\text{in}} \times C_{\text{out}}$
Up-Sampling (Linear)	N.A.	0
Up-Sampling (Nearest)	N.A.	0
Identity Mapping	N.A.	0

TABLE 1
The operations used in this paper. C_{in} and C_{out} denote the numbers of input and output feature maps, respectively.

employed in the search space, this would further increase the computational cost of the searching process. In future work, we will propose a more efficient person-specific CNN searching strategy, and specifically investigate the optimal searching space for representing personality-related person-specific cognition.

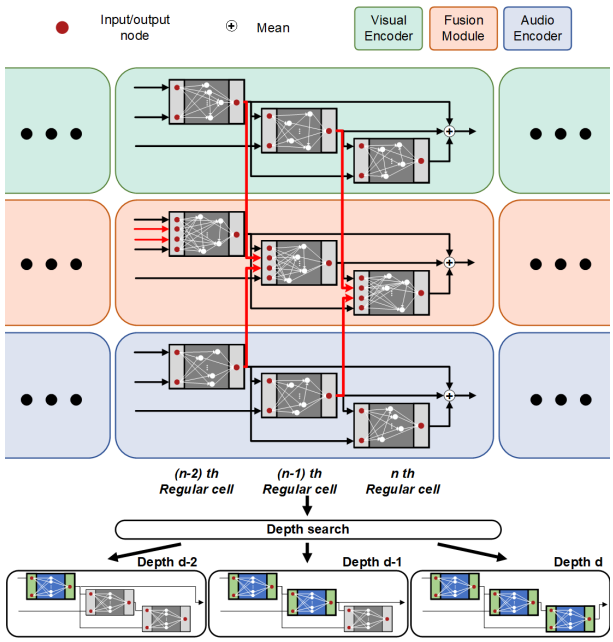
3.1.3 Adaptive loss function

To supervise the searching process of each person-specific CNN (as described by Eq. 5) we first highlight some important aspects of human psychology and behaviour pattern. Firstly, facial reactions of similar emotions or intentions can be displayed by different facial spatio-temporal patterns, which is partially caused by the differences in listeners' facial identities, responding times and personalities. While differences in facial identities can be partially addressed by projecting faces of different subjects to a mean face, we consider that there is always a time delay for a listener to generate a facial reaction in response to speaker's behaviours. This is because the execution of the corresponding cognitive processes takes some time [59]. Importantly, the duration of the time delay may vary not only for different listeners but also for the same listener depending upon other external factors.

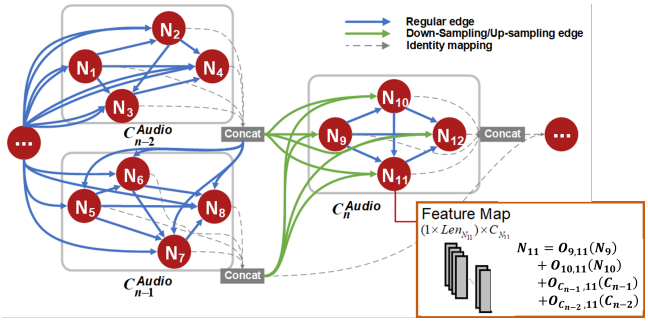
In light of this, we introduce an adaptive factor τ to model this uncertainty. Let us define an audio-facial input $A_S(t_1, t_2)$ and $F_S(t_1, t_2)$ that represent the speaker's audio-facial non-verbal behaviours expressed from time t_1 to t_2. We propose the following adaptive loss (A-loss) function to measure the similarity between the predicted listener's facial reaction and the ground-truth:

$$\begin{aligned} &L_{A-loss}(t_1, t_2, \tau) \\ &= L_{A-loss}(F_L^p(t_1, t_2), F_L^g(t_1 + \tau, t_2 + \tau)) \\ &= \sum_{i=t_1}^{t_2} \sum_{j=1}^{68} \min(L_\star(x_{i,j}^p, x_{i+\tau,j}^g) + L_\star(y_{i,j}^p, y_{i+\tau,j}^g), \varepsilon) \end{aligned} \tag{12}$$

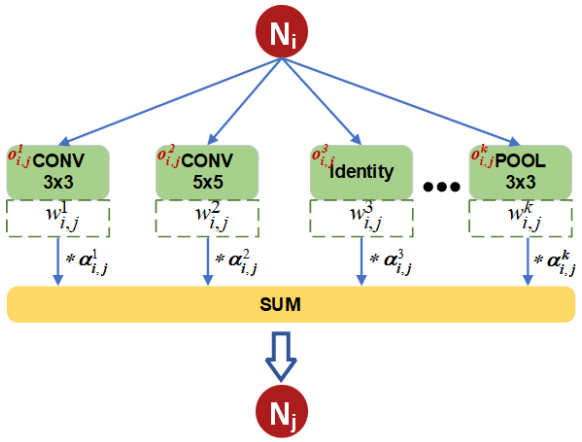
where $F_L^p(t_1, t_2)$ denotes the predicted landmarks of the listener's facial reaction corresponding to the input $A_S(t_1, t_2)$ and $F_S(t_1, t_2)$; $F_L^g(t_1 + \tau, t_2 + \tau)$ are the listener's real facial reaction landmarks induced by $A_S(t_1, t_2)$ and $F_S(t_1, t_2)$, where τ represents the time delay; $(x_{i,j}^p, y_{i,j}^p)$ denotes the predicted coordinates of the jth facial landmark of the ith frame and $(x_{i+\tau,j}^g, y_{i+\tau,j}^g)$ is the corresponding ground-truth



(a) Illustration of the fusion module and the depth search. We initially set each module to contain M_{reg} linear stacked cells ($M_{\text{reg}} = 3$ in the figure), and then gradually mask out each cell from the end of the module. If masking out the $m+1_{th}$ cell leads to the best performance in facial reaction generation, the final optimal depth for the corresponding module is m.



(b) The details of nodes' connections. Each node in a cell is influenced by the information coming from all its parent nodes, and each cell takes the outputs from previous two cells.



(c) A CNN edge example in explored person-specific CNNs.

Fig. 4. The details of the explored person-specific multi-modal CNN.

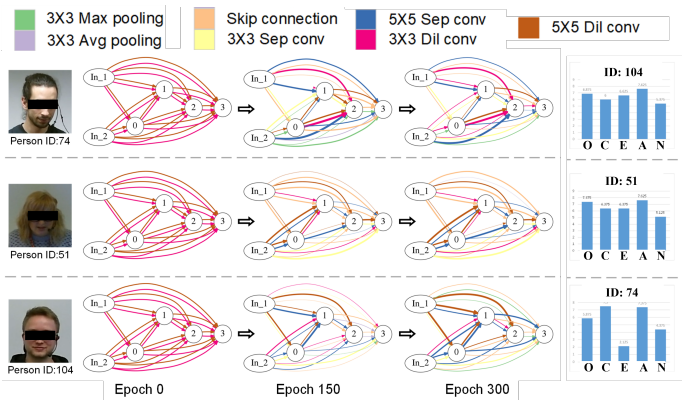


Fig. 5. Visualization of person-specific CNNs explored on the NoXI dataset, where the initial CNN architectures for all subjects are the same (Epoch 0). After the searching, we can see that the explored CNN for each subject is unique and person-specific (Epoch 300).

coordinate. Specifically, the ε is a constant value employed to avoid extremely large loss values caused by outliers (e.g. incorrectly detected face regions) which can lead to a misguided CNN search. $L_\star$ represents the similarity measurement between the prediction and ground-truth. In this paper, $L_\star$ is defined as the Mean Square Error (MSE).

To achieve the proposed adaptive loss (i.e., computing a τ at each time), in practice we use a sliding time-window to compare the prediction of listener's facial reactions with a set of ground-truth candidates (the duration of the ground-truth candidates is longer than the time-window, as illustrated in the last section of Fig. 1(a)). Specifically, we set R ground-truth candidates, i.e., $F_L^g(t_1+r, t_2+r), r = 1, 2, \cdots R$, and only choose $r = \tau$ that allows the loss $L_{A-loss}(t_1, t_2, \tau)$ to have the lowest value:

$$\tau = \text{argmin} L_{A-loss}(t_1, t_2, \tau) \tag{13}$$

As a result, the delay period can be automatically adapted for each listener at each training iteration.

3.1.4 Person-specific CNN optimization

To search for an optimal multi-modal CNN (described in Sec. 3.1.2) for each listener, we conduct a single-level optimization based on the continuous relaxation algorithm [60]. It adjusts all OPs, LWs as well as depths of the person-specific CNN at the same time during the optimization. In comparison to the widely-used bi-level optimization strategy [60] which separately optimizes OPs in the validation set and LWs in the training set, i.e., freezing one of them while optimizing the other, the proposed single-level optimization strategy allows the OPs, LWs and depths to be simultaneously optimized. This aims to replicate how the human cognition operates with all cognitive processes jointly activated during reaction generation - there is no evidence suggesting that some parts of the human cognitive processors are frozen during the reaction generation. In addition, this strategy allows the OPs, LWs and depths to be optimized using the full audio-facial frames instead of a sub-segment of it, i.e., the explored CNN is a clip-level representation without ignoring any frames. The pseudocode of the proposed single-level OPs, LWs and depths optimization is provided in Algorithm 1. Representative examples of person-specific CNNs' optimization processes are visualized in Fig. 5.

Algorithm 1 Single-level optimization

Require: A multi-modal CNN that is parametrized by OPs, LWs and depths of an audio encoder, a visual encoder, a fusion module and a decoder, which are denoted as $\mathcal{A}_{\text{V,A,F,D}}^{t=0}$, $\mathcal{W}_{\text{V,A,F,D}}^{t=0}$ and $\mathcal{D}_{\text{AVF,D}}^{t=0}$.

Ensure: An optimal person-specific multi-modal CNN that can reproduce the target subject's facial reactions, which is parametrized by OPs $\mathcal{A}_{\text{V,A,F,D}}^{\text{Optimal}}$, LWs $\mathcal{W}_{\text{V,A,F,D}}^{\text{Optimal}}$ and depths $\mathcal{D}_{\text{AVF,D}}^{\text{Optimal}}$.

1: **repeat**

2: Updating OPs $\mathcal{A}_{\text{V,A,F,D}}^{t}$ $(t > 1)$ on the training set by descending $\nabla_{\mathcal{A}} L_{A-loss}(\mathcal{W}_{\text{V,A,F,D}}^{t-1} - \eta_{\mathcal{A}} \nabla_{\mathcal{W}} L_{A-loss}(\mathcal{W}_{\text{V,A,F,D}}^{t-1}, \mathcal{A}_{\text{V,A,F,D}}^{t-1}, \mathcal{D}_{\text{AVF,D}}^{t-1}), \mathcal{A}_{\text{V,A,F,D}}^{t-1}, \mathcal{D}_{\text{AVF,D}}^{t-1})$.

3: Updating LWs $\mathcal{W}_{\text{V,A,F,D}}^{t}$ on the training set by descending $\nabla_{\mathcal{W}} L_{A-loss}(\mathcal{A}_{\text{V,A,F,D}}^{t}, \mathcal{W}_{\text{V,A,F,D}}^{t-1}, \mathcal{D}_{\text{AVF,D}}^{t-1})$.

4: Choosing the optimal depths $\mathcal{D}_{\text{AVF,D}}^{t}$ to achieve the best training loss $L_{A-loss}(\mathcal{A}_{\text{V,A,F,D}}^{t}, \mathcal{W}_{\text{V,A,F,D}}^{t}, \mathcal{D}_{\text{AVF,D}}^{t})$

5: **until** Convergence

6: $\mathcal{A}_{\text{V,A,F,D}}^{\text{Optimal}} = \mathcal{A}_{\text{V,A,F,D}}^{\text{Convergence}}$; $\mathcal{W}_{\text{V,A,F,D}}^{\text{Optimal}} = \mathcal{W}_{\text{V,A,F,D}}^{\text{Convergence}}$ and $\mathcal{D}_{\text{AVF,D}}^{\text{Optimal}} = \mathcal{D}_{\text{AVF,D}}^{\text{Convergence}}$

3.2 Graph representation of the person-specific CNN

Let's recall that the main hypothesis of this paper is that if a CNN can reproduce the target subject's facial reactions, it represents the person-specific cognition of the subject, which is well associated with the subject's true personality. Consequently, we search for a person-specific CNN for each subject. Since the explored CNN is a directed acyclic graph, we encode each explored CNN as a graph $\mathcal{G}(V, E)$, and treated it as the personality representation for the corresponding subject. This process is formulated in Eqa. 3, where each graph representation $\mathcal{G}$ is made up of a set of vertices V and edges E. Specifically, we represent each CNN edge $O_{i,j}$ as a vertex $V_{i,j}$ in the corresponding graph representation $\mathcal{G}(V, E)$. Meanwhile, the edge presence $A_{i,j,m}$ between $V_{i,j}$ and $V_{j,m}$ in $\mathcal{G}(V, E)$ is decided by the relationship between their corresponding CNN edges $O_{i,j}$ and $O_{j,m}$. If $A_{i,j,m} = 1$, the edge feature $E_{i,j,m}$ in $\mathcal{G}(V, E)$ is obtained by considering both vertex features $V_{i,j}$ and $V_{j,m}$. These can be formulated as:

$$\begin{aligned} V_{i,j} &= \text{VFE}(O_{i,j}) \\ E_{i,j,m}, A_{i,j,m} &= \text{EFE}(V_{i,j}, V_{j,m}, O_{i,j}, O_{j,m}) \end{aligned} \tag{14}$$

where VFE denotes the proposed vertex feature encoding strategy described in Sec. 3.2.1, and EFE denotes the proposed edge feature encoding strategy described in Sec. 3.2.2. We additional provide the pseudocode of the VFE and EFE in the supplementary material.

3.2.1 Vertex feature encoding

Given a CNN edge $O_{i,j}$, we categorize all its operations into two parts: υ operations $o_{i,j}^{w}$ that have LWs (e.g., convolution) and κ operations $o_{i,j}^{n}$ that do not have LWs (e.g., pooling). Then, we propose a vertex feature encoding (VFE) strategy

to achieve its corresponding vertex $V_{i,j}$ for the graph representation as:

- **Step 1 LWs alignment:** we first notice that the number of LWs are different (ranging from hundreds to tens of thousands in our study) in each CNN edge because they have different number of input and output feature maps. Consequently, we follow the idea of [64] to select a fixed number of most representative weights from each operation $o_{i,j}^k$, which are denoted as $S\omega_{i,j}^k$. For examples we choose weights of five kernels with the top-5 highest L1 values (sum of absolute weights) from each convolution operation. This way, the LWs representation $\text{LW}_{i,j}$ of the CNN edge $O_{i,j}$ (having K operations) is denoted as:
$$\text{LW}_{i,j} = [S\omega_{i,j}^1, S\omega_{i,j}^2, \cdots, S\omega_{i,j}^K] \tag{15}$$
where the $\text{LW}_{i,j}$ would have a fixed dimension for all CNN edges.
- **Step 2 Fusion of OPs and LWs:** Since each OP $\alpha_{i,j}^k$ reflects the importance of the operation $o_{i,j}^k$ (as well as its LWs $\omega_{i,j}^k$), we use OPs to weight corresponding LWs. For operations that have LWs, their original OPs $\alpha_{i,j}^w$ are projected to a OP-LW weighting vector $\text{OP-LW}_{i,j}$ that has the same dimension as LWs representation $\text{LW}_{i,j}$:
$$\begin{aligned}\text{OP-LW}_{i,j} &= \text{VEN}(\alpha_{i,j}^w) \\ &= \text{VEN}([\alpha_{i,j}^{w_1}, \alpha_{i,j}^{w_2}, \cdots, \alpha_{i,j}^{w_v}])\end{aligned} \tag{16}$$
where VEN is a Multi-Layer Perceptron (MLP) that has two hidden layers. Then, the OPs and LWs for operations that have LWs are combined by computing the dot product between the weighting vector $\text{OP-LW}_{i,j}$ and the LWs representation $\text{LW}_{i,j}$, which can be denoted as:
$$V_{i,j}^w = \langle \text{OP-LW}_{i,j}, \text{LW}_{i,j} \rangle \tag{17}$$
- **Step 3 Vertex feature generation:** Finally, we concatenate the obtained $V_{i,j}^w$ with OPs $\alpha_{i,j}^n$ of κ operations that do not have LWs as the final vertex feature:
$$V_{i,j} = [\alpha_{i,j}^n, V_{i,j}^w] \tag{18}$$
Since the dimension of both $\alpha_{i,j}^n$ and $V_{i,j}^w$ are fixed, all vertex features would have the same dimension.

This process is illustrated in Fig. 1(b) and depicted in purple.

3.2.2 Edge feature encoding

For a graph representation $\mathcal{G}(V, E)$, we define a pair of vertices $V_{i,j}$ and $V_{j,m}$ are connected (i.e., the edge $E_{i,j,m}$ exists ($A_{i,j,m} = 1$)) if their corresponding CNN edges $O_{i,j}$ and $O_{j,m}$ are connected to the same node N_j in the CNN (illustrated in Fig.1(b)). While most existing approaches only use a single binary value (0 or 1) to define the relationship between a pair of vertices in graphs, this single-value binary edge feature usually fail to describe all task-related relationship cues, as sometimes the relationship between vertices can be described by multiple attributes. On the contrary, we aim to produce a person-specific graph representation that not only encodes parameters (contained in vertex features) and the architecture (encoded as the graph topology) of the person-specific CNN, but also the underlying relationship between CNN edges, which may provide additional personality-related cues.

To this end, we propose a novel multi-dimensional edge feature encoding strategy that represents the relationship (edge) between each pair of connected vertices as a multi-dimensional vector. In particular, we propose to produce the edge feature $E_{i,j,m}$ directly from the obtained vertices $V_{i,j}$ and $V_{j,m}$:
$$E_{i,j,m} = \text{ERN}(V_{i,j}, V_{j,m}) \tag{19}$$
where ERN is an attention-based edge relationship network. It takes a pair of vertices' feature $V_{i,j}$ and $V_{j,m}$ as the input, and outputs an edge feature $E_{i,j,m}$ that contains the task-specific relationship feature that are related to both vertex features $V_{i,j}$ and $V_{j,m}$. The detailed process of the ERN is illustrated in Fig. 6.

It should be noted that both ERN and VEN are jointly trained with the personality recognition model in an end-to-end manner. This way, they learn to generate personality-related vertex and edge features (personality-related features) from the explored person-specific CNN's parameters and architecture.

3.3 Personality recognition model

In this paper, each subject's personality traits are recognised from the graph representation of the subject's person-specific CNN. We formulate the personality recognition as a multi-task graph regression problem (jointly recognizing 5 traits). Particularly, we employ the state-of-the-art residual gated graph convolution neural network (residual GatedGCN) [65] provided by [66] as the personality recognition model to process the produced graph representations, as it is the state-of-the-art GNN model which can process heterogeneous graphs and graphs containing multi-dimensional edge features. We empirically employ a network that consists of six GatedGCN layers (the detailed settings are provided in supplementary material). Then, two fully connected (FC) layers are attached to the last GatedGCN layer to concatenate all produced vertices features, where a ReLU activation and a dropout (0.3) are followed by each FC layer. The size of the output layer is set to 5 to jointly recognise the *five* personality traits of **Extraversion** (**Ext**), **Agreeableness** (**Agr**), **Openness** (**Ope**), **Conscientiousness** (**Con**), and **Neuroticism** (**Neu**).

4 EXPERIMENTS

This paper evaluates the proposed approach on a human-human dyadic interaction dataset and a human-machine dyadic interaction dataset, which are described in Sec. 4.1. We then present the implementation details in Sec. 4.2 followed by evaluation metrics in Sec. 4.3. We compare the personality recognition performance achieved by the proposed approach to other existing personality recognition approaches that directly infer personality from subjects' external behaviours in Sec.4.4, showing the advantages of the proposed novel strategy which recognises personality traits from the simulated human cognition. Finally, we experiment the influence of different CNN searching and graph representation encoding settings on personality recognition

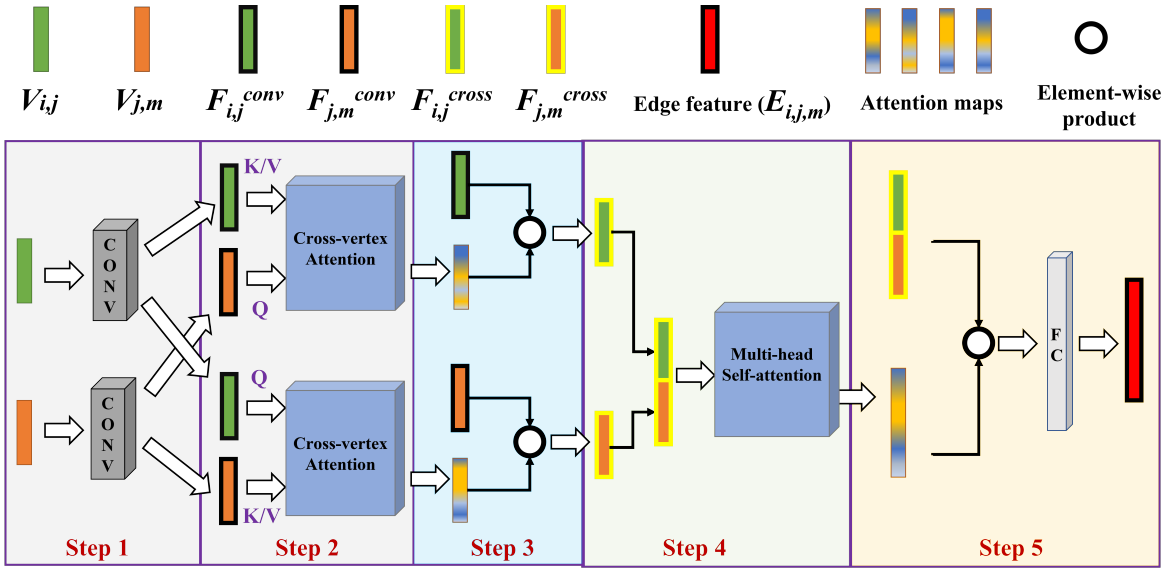


Fig. 6. Illustration of the ERN. **Step 1:** Learning a pair of 1D representations $F_{i,j}^{\text{conv}}$ and $F_{j,m}^{\text{conv}}$ from a pair of connected vertex features $V_{i,j}$ and $V_{j,m}$; **Step 2:** generating cross-vertex attention maps $A_{i,j}^{\text{cross}}$ and $A_{j,m}^{\text{cross}}$, where $A_{i,j}^{\text{cross}}$ emphasizes a part of $F_{i,j}^{\text{conv}}$'s information that is correlated with $F_{j,m}^{\text{conv}}$ while $A_{j,m}^{\text{cross}}$ highlighting a part of $F_{j,m}^{\text{conv}}$'s information that is correlated with $F_{i,j}^{\text{conv}}$; **Step 3:** Generating weighted features $F_{i,j}^{\text{cross}}$ and $F_{j,m}^{\text{cross}}$; **Step 4:** Concatenating $F_{i,j}^{\text{cross}}$ and $F_{j,m}^{\text{cross}}$, and producing self-attention maps; and **Step 5:** Generating the final edge feature $E(i,j,m)$. In this figure, the **K**, **Q**, **V** depicted in purple represent 'key', 'query' and 'value' of the attention operation.

in Sec. 4.5, where we also specifically compare the person-specific CNNs that are explored using NAS with person-specific CNNs that have a fixed architecture. Additionally, we conduct a set of experiments to systematically evaluate the sensitivity of our approach for different demographic groups (provided in supplementary material.

4.1 Datasets

In this paper, we evaluate our approach in both human-human and human-machine dyadic interaction scenarios. While many existing datasets [41], [42], [43], [44] are built for personality perception prediction study, some publicly available datasets [1], [46], [47], [48], [67] also can be used for audio-visual true personality recognition studies. However, most of these available datasets are not suitable for our study as our models assume dyadic interaction.

Human-human interaction: The NoXi dataset [46] is a multi-lingual human-human dyadic interaction dataset that was designed to generate spontaneous interactions with emphasis on adaptive behaviours in unexpected situations. It consists of 84 sessions in which one participant acts as an Expert and the other acts as a Novice interacting on a chosen topic of expertise via video conferences. The participants were allowed to continue the conversation until it reached a natural end. During the interaction, participants can interrupt each other for either changing the topic or inducing a mild debate whenever possible. This dataset contains 84 pairs of audio-visual clips (168 clips in total from 89 participants) with participants' ages ranging from 21 to 50 years old. The average and standard deviation of clips' duration are 18m6s and 6m28s, respectively. All participants provided self-assessments of their Big-Five Personality Traits using the Saucier's Mini-Markers [68].

Human-machine interaction: We conducted the human-machine experiments on the Virtual Human Questionnaire (VHQ) database [1]. The VHQ database consists of 165 videos collected from 55 participants, where each participant completed 3 questionnaire interview sessions. During each session, participants were asked to answer a set of questions verbally based on one of three questionnaires: BFI-10 [69], PHQ-9 [70] or GAD-7 [71]. In this database, 55 videos (corresponding to 55 subjects) were recorded under the human-machine dyadic interaction mode. More specifically, a virtual human agent interviewer (Fig. 7) was projected directly in front of the participant and ask questions, which was implemented using the ARIA-VALUSPA Platform [72]. The self-reported labels of the Big-Five personality traits were obtained by asking participants to fill the BFI-44 questionnaire online.

(a) The virtual human (b) Detected virtual facial displays

Fig. 7. Examples of a virtual human display (Fig. (a)) and automatically detected (aligned) faces (Fig. (b)) in VHQ dataset.

4.2 Implementation details

Persons-specific CNN settings: For subjects in the NoXI dataset, their multi-modal person-specific CNNs have 6 cells (a pre-defined convolution cells, 3 down-sampling cells and 2 regular cells) for each encoder and 5 cells (2 regular cells and 3 up-sampling cells) for each decoder. The employed LSTMs have 3 hidden layers. During the neural architecture searching, the input speaker's audio-visual signal lasts for 80 frames and the listener's candidate ground-truth consists of 105 frames and the delay factor r ranges from 0 to 25 frames, i.e., selecting 80 consecutive frames as the final reaction. Since the audio data in the VHQ dataset is very noisy, we only search for a single-modal person-specific CNN for each subject, which takes the virtual human's facial landmarks as the input and aims to reproduce the target subject's facial reactions. We noticed that the virtual human only spoke a set of pre-defined sentences during the interaction. Thus, we also categorized all sentences into 4 classes: depression-related questions, anxiety-related questions, personality-related questions, and other sentences (e.g., virtual human asks the subject to repeat the answer.), and then encoded each as a one-hot vector (e.g., 1000, 0100, 0010, and 0001). Consequently, the multi-modal CNNs explored in the VHQ dataset are implemented by concatenating the deep-learned virtual human's face feature with the proposed sentence categorical feature at the last FC layer of the encoder.

Neural Architecture Search: In this paper, all person-specific CNNs for each dataset have the same initial architecture and parameters, where OPs and LWs are initialized with the Xavier strategy [73]. Meanwhile, we used the same training strategies to obtain all person-specific CNNs in each dataset. In particular, during each person-specific CNN's searching, we fed facial landmark sequences to each CNN based on their time stamps in the corresponding video, i.e. from the beginning of the video to the video's end. This not only ensures that the OPs and LWs of the CNN always converge (during searching) to the same set of values for a particular video, but also ensures that the difference between individually explored person-specific CNNs is only influenced by person-specific reactions rather than the initialization of weights or the order in which the frames are used for searching. During the searching, the batch size was set to 60 audio-visual clips, while 2 Adam optimizers were independently used to jointly adjust OPs and LWs, with the learning rate of 0.05 and 0.001, respectively.

Personality model training details: In this paper, we conduct a 12-fold subject-independent cross-validation on the NoXI dataset. For each fold, 154 videos were used for training and hyperparameter optimisation and 14 videos were used for testing (each subject appeared in either training or test set, not both). Due to the limited number of data, we conduct a leave-one-subject-out cross-validation on the VHQ dataset. For each fold, 154 videos were used for training and hyperparameter optimisation, and the remaining video was used for testing. For both NoXI and VHQ datasets, we report the accuracy on the test sets averaged over all folds. In this paper, all experiments were conducted on the PyTorch platform using Nvidia V100 GPUs.

4.3 Evaluation metrics

Two common metrics are used to evaluate the personality recognition performance: the Pearson Correlation Coefficient (PCC) and the mean accuracy measurement (ACC), which has been adopted in relevant challenge events (e.g., the ChaLearn challenge [41].

4.4 Comparison to existing approaches

To compare the proposed approach with other video-based automatic personality analysis solutions, we reproduced four existing personality computing approaches that have been reported on ChaLearn dataset [41], which are DCC [20], NJU-LAMDA [19], CR-Net [5], and PALs [10] as well as spectral representation [74], [75]. The detailed reproduction settings are provided in the supplementary material.

Table 2 and Table 3 compare the variations of the proposed models to the existing state-of-the-art audio-visual personality analysis approaches (automatic personality perception (APP) solutions) on the NoXi and the VHQ datasets. Table 4 compares the results achieved by our best systems (the graph representations of multi-modal CNNs that are learned using adaptive loss, independent parameter settings and the graph representations are constructed using end-to-end vertex and edge feature learning strategy) and the results achieved by other methods under both interaction scenarios with highlighted statistical significance. It can be observed that for both datasets, the predictions produced by the graph representations of the explored multi-modal CNNs are positively correlated with all self-reported personality traits. Specifically, these graph representations achieved PCC>0.37 for *Con*, *Ext*, and *Neu* traits on the NoXI dataset, which shows significant advantages over the other listed methods. Meanwhile, the graph representations of the explored multi-modal model (A-MModal (M)) also achieved the best average ACC result and the second best PCC result in recognizing the self-reported personality traits under human-computer interaction scenarios, i.e., it generated the best PCC results between predictions and ground-truth of the *Neu* trait with PCC of 0.363, showing more than 8% relative improvements over the second best method [10]. In addition, we also train a CRNet-based baseline (M-CRNet) that takes the audio-visual signal of both the listener and speaker, to predict the listener's personality. While the proposed approaches and the M-CRNet both using the listener and speaker's data to predict the listener's true personality traits, the results demonstrate that using the representation of the simulated person-specific cognition still provide significant advantages over directly extracting features from external behaviours, in recognizing all five true personality traits.

It also can be observed from Table 5 that the explored person-specific CNNs can accurately predict the corresponding target subjects' facial reactions, which are evidenced by the promising PCC results (more than 0.75 for all systems) and RMSE results of the facial reaction (facial landmarks) generation. This indicates that the simulated person-specific cognition can reproduce similar facial reactions for the majority frames of the target subject's video. In other words, the explored person-specific CNN can accurately represent the target subject's cognition over

time (a video's duration). On the contrary, these person-specific CNNs have poor performance in reproducing other subjects' facial reactions with the PCC between the reproduced reaction landmarks and the ground-truth being less than 0.1, which means each person-specific CNN can not reflect other subjects' cognition for facial reactions (i.e., the simulated cognition of the target subject is different from others' cognition). As a result, we assume the proposed approach can partially encode the target subject's person-specific cognitive process that is stable over time but different from other subjects.

Discussion: In summary, the results presented above indicate that despite CNNs and humans having different cognitive mechanisms, if a CNN can simulate a subject's cognitive process for generating facial reactions, this CNN's architectural parameters are positively associated with the subject's self-reported personality traits. Compared to existing solutions that directly predict personality traits from the listener's non-verbal behaviours, the proposed approach that recognises self-reported personality traits from the simulated cognition seems a more reliable solution. It is clear that the performance in recognising *Neu* and *Ext* traits are better than the other traits, which is consistent with what has been frequently claimed by previous studies [14], [50], i.e., *Ext* and *Neu* traits are well associated with human cognition. We also observed that the approaches which predict personality using video-level features (e.g., CR-Net, PALs, Spectral and the proposed approach) have clear advantages over the approaches that infer personality from a single frame or a thin slice (DCC and NJU-LAMDA), demonstrating that long-term information is more reliable for modelling self-reported personality traits. It should be noted that the advantage of our approach in human-machine interaction scenarios is not as clear as the human-human interaction scenarios. This can be explained by the fact that the virtual human used in the VHQ dataset only has limited non-verbal facial behaviours, they are not as rich as real human speakers in the NoXI dataset. Thus, the listeners' facial reactions in human-machine interaction scenarios may be less correlated with the non-verbal behaviours expressed by the virtual human.

	Methods	Ope	Con	Ext	Agr	Neu	Avg.
ACC	DCC [20]	0.755	0.787	0.772	0.736	0.791	0.768
	NJU-LAMDA [19]	0.741	0.826	0.827	0.753	0.789	0.787
	Spectral [75]	0.868	0.909	0.903	0.898	0.910	0.898
	PALs [10]	0.845	0.819	0.916	0.837	0.911	0.866
	CR-Net [5]	0.892	0.916	0.924	0.888	0.913	0.907
	M-CRNet	0.898	0.905	0.913	0.902	0.907	0.905
	Ours (A-MModal (S))	0.871	0.911	0.915	0.903	0.916	0.903
	Ours (MModal (M))	**0.902**	0.919	0.927	**0.923**	0.926	0.919
	Ours (A-MModal (M))	0.895	**0.925**	**0.928**	0.920	**0.931**	**0.920**
PCC	DCC [20]	-0.153	-0.078	0.037	-0.024	0.121	0.008
	NJU-LAMDA [19]	-0.110	0.118	0.115	-0.067	0.032	0.017
	Spectral [75]	0.135	0.246	0.265	0.192	0.277	0.223
	PALs [10]	0.129	0.091	0.270	0.106	0.264	0.172
	CR-Net [5]	0.181	0.271	0.301	0.177	0.325	0.251
	M-CRNet	0.176	0.273	0.275	0.201	0.318	0.249
	Ours (A-MModal (S))	0.161	0.322	0.333	0.239	0.358	0.283
	Ours (MModal (M))	**0.196**	0.354	0.403	0.281	0.450	0.337
	Ours (A-MModal (M))	0.189	**0.376**	**0.420**	**0.289**	**0.481**	**0.351**

TABLE 2
Personality recognition results on the NoXi dataset. *MModal* denotes the graph representations of the explored multi-modal (audio-visual) CNNs. *(M)* and *(S)* represent the multi-level and single-level fusion, respectively. *A-* represents that the CNNs were trained with adaptive loss. *M-CRNet* consists of two CR-Net models that input audio, face frames, original frames of the listener and speaker, respectively, and aims to predict the listener's personality. For all our systems, the graph representations were obtained using OP-LW (VEN) vertices features and end-to-end learned multi-dimensional edge features while GatedGCN was employed as the personality recognition model.

	Methods	Ope	Con	Ext	Agr	Neu	Avg.
ACC	DCC [20]	0.835	0.837	0.840	0.838	0.840	0.838
	NJU-LAMDA [19]	0.842	0.838	0.839	0.841	0.841	0.840
	Spectral [75]	0.838	0.842	0.843	0.846	0.846	0.843
	PALs [10]	0.843	**0.843**	0.843	0.846	0.845	**0.844**
	CR-Net [5]	0.845	0.842	0.840	0.844	0.845	0.843
	M-CRNet	**0.851**	0.830	0.833	0.841	0.828	0.837
	Ours (A-SModal (S))	0.839	0.839	**0.844**	**0.848**	0.846	0.843
	Ours (MModal (M))	0.842	0.838	0.841	0.847	0.845	0.843
	Ours (A-MModal (M))	0.840	0.842	0.842	0.847	**0.848**	**0.844**
PCC	DCC [20]	-0.020	0.098	0.133	-0.072	0.185	0.065
	NJU-LAMDA [19]	0.077	0.155	0.118	0.112	0.234	0.139
	Spectral [75]	-0.039	0.167	0.221	0.158	0.256	0.153
	PALs [10]	0.135	0.187	**0.279**	0.132	0.336	**0.214**
	CR-Net [5]	0.138	**0.191**	0.166	0.143	0.280	0.184
	M-CRNet	**0.141**	0.183	0.150	0.145	0.239	0.172
	Ours (A-SModal)	-0.021	0.127	0.267	**0.193**	0.319	0.177
	Ours (MModal)	0.088	0.136	0.192	0.191	0.358	0.193
	Ours (A-MModal)	0.063	0.172	0.211	0.189	**0.363**	0.200

TABLE 3
Personality recognition results on the VHQ dataset. *MModal* denotes the graph representations built on the face and sentence categorical features extracted from the speaker (please check Sec. 4.2). The rest of the settings are the same as in Table 2.

4.5 Ablation studies

In this section, we explicitly investigate the sensitivity of the proposed approach to different NAS settings (i.e., modality, parameter sharing strategy, loss function and topology alignment) and graph representation learning settings. The statistical significance testing results achieved by the best system and the second best system in terms of each ablation study, as well as the detailed settings for each 'second best system' are provided in supplementary material.

We first demonstrate the importance of: (i) applying NAS to obtain unique architecture and parameters for each person-specific CNN; and (ii) encoding person-specific CNNs as graph representations in Fig. 8. Specifically, the system 'NAS+MLP' is obtained by simply concatenating all OPs and LWs of the person-specific CNN as a vector, whose dimension is reduced by CFS [76]. Meanwhile, the system 'Unet+MLP' is achieved using the same strategy as the 'NAS+MLP' system without NAS, i.e., all person-specific Unets of the system 'Unet+MLP' have the same and fixed architecture. Each person-specific Unet consists of a audio encoder, a visual encoder, a fusion module and decoder, and each module is made up of a set of ResNet blocks. Firstly, predictions of all three NAS-based models achieved positive correlations across all traits under both interaction scenarios, demonstrating that the explored person-specific CNNs are indeed positively associated with target subjects' personality traits. Then, it can be observed from the figure that it is superior to encode each person-specific CNN to a graph representation than simply concatenating all OPs and LWs of the person-specific CNN as a vector.

	Traits	Ope	Con	Ext	Agr	Neu	Avg.
NoXI	Spectrum (BP) [75]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	DCC [19]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	NJU-LAMDA [20]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	PALs [10]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	CR-Net [5]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	M-CRNet	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
VHQ	Spectrum (BP) [75]	+(***)	-	-	+(**)	+(***)	+(***)
	DCC [19]	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)
	NJU-LAMDA [20]	+(***)	+(*)	+(***)	+(***)	+(***)	+(***)
	PALs [75]	+(***)	-	+(***)	+(***)	+(*)	-
	CR-Net [5]	+(***)	+(*)	+(***)	+(***)	+(***)	+(***)
	M-CRNet	+(***)	+(***)	+(***)	+(***)	+(***)	+(***)

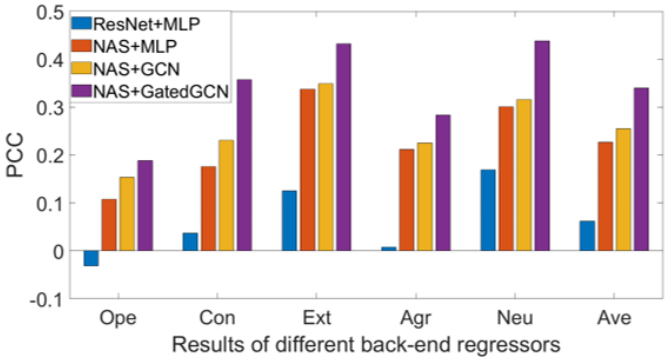
TABLE 4
Statistical significance testing results in terms of PCC achieved by our best system and the five reproduced systems on the NoXI and the VHQ dataset, where + / - denotes that there is / there is no statistically significant difference between our approach and the other approach (The significance level of $* \ P < 0.05$, $** \ P < 0.01$, $*** \ P < 0.001$). To conduct the T-Test, we used the 12-fold results on the NoXI dataset. For VHQ dataset, we conducted 10 times leave-one-subject-out cross-validation for all approaches and the used these 10 results to compute the P-values.

Cognitive model	PCC	MSE ($\times 10^{-5}$)
Audio-to-face	0.769	**2.882**
Face-to-face	**0.770**	2.894
PS-Multi-to-face (S)	**0.781**	2.390
IP-Multi-to-face (S)	0.775	2.503
A-IP-Multi-to-face (S)	**0.781**	**2.260**
PS-Multi-to-face (M)	0.794	2.362
IP-Multi-to-face (M)	0.798	**2.245**
A-IP-Multi-to-face (M)	**0.802**	2.331
A-IP-Dep-Multi-to-face (M)	0.649	7.190

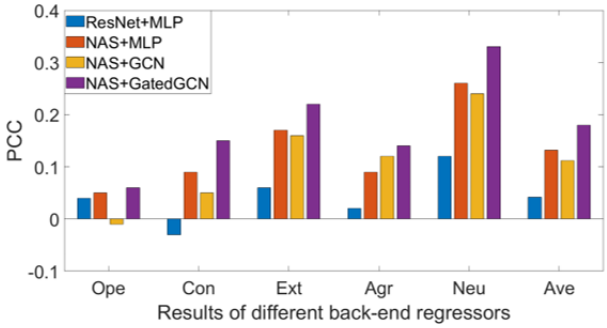
TABLE 5
Facial reactions prediction results on the NoXI dataset. *PS-* and *IP-* denote the parameter sharing and independent parameter strategy, respectively; *Multi-* refers to the multi-modal audio and face features of the speaker were used as the input; *A-* represents that the CNNs were explored with the adaptive loss; *Dep-* denotes that the depth is considered as a variable during the CNN search.

Cognitive model	PCC	MSE ($\times 10^{-5}$)
Face-to-face	0.596	6.632
PS-Multi-to-face	0.588	6.550
IP-Multi-to-face	0.602	6.279
A-IP-Multi-to-face	0.612	**6.098**
A-IP-Dep-Multi-to-face	**0.619**	6.177

TABLE 6
Facial reactions prediction results on the VHQ dataset. The *Multi-* in this table refers to the face and sentence categorical features of the speaker were used as the input.



(a) Results on the NoXI dataset.



(b) Results on the VHQ dataset.

Fig. 8. The results achieved by our best system and several baselines.

This validates that the proposed graph representation is a superior way to summarise architectures and parameters of the CNN. Finally, person-specific CNNs explored by NAS (NAS+MLP) generated better results than these of Unet-based person-specific CNNs (Unet+MLP), which shows that the CNNs explored by NAS can better simulate personality-related cognition for each subject. This is also evidenced by the better facial reaction generation performance displayed in Table 5 and Table 6. We conclude these results as the person-specific CNNs explored by NAS have not only unique weights but also unique architectures, which would theoretically have better capability to fit complex human cognition. In addition, we provide the example neural architecture searching loss curves in Fig. 9, showing that despite the limited number of frames in each pair of speaker and listener's videos, the person-specific CNN can still be well explored to fit to the person-specific facial reactions of the given video, i.e., training losses are well converged.

4.5.1 Person-specific CNN settings

We first show the personality recognition performance achieved by different person-specific CNN settings in Fig. 10(a) and Fig. 10(b). Our best settings for CNN topology, parameter sharing strategy, loss function and topology alignment are: multi-modal, independent parameters (IP), adaptive loss, and block distillation, respectively.

Modalities: Firstly, it can be observed from the results on the NoXI dataset that predictions generated by all settings are positively correlated with the self-reported values across all five traits. In general, the multi-modal system achieved better results than single-modal systems for both personality recognition and facial reaction generation tasks. This demonstrates that both non-verbal audio and facial behaviours of the speaker contribute to the listener's facial reactions, where each of them contain some unique aspects in forming reactions. In other words, the person-specific cognition triggered by each modality provides unique and useful clues for personality recognition. Meanwhile, as we can see from the results on the VHQ dataset, even simply adding the encoded virtual human sentence categorical feature (explained in Sec. 4.2) can improve the recognition performance of Ope, Con and Neu traits. In comparison to the real human speaker, the facial behaviours of the virtual human may not be the key factor to trigger the

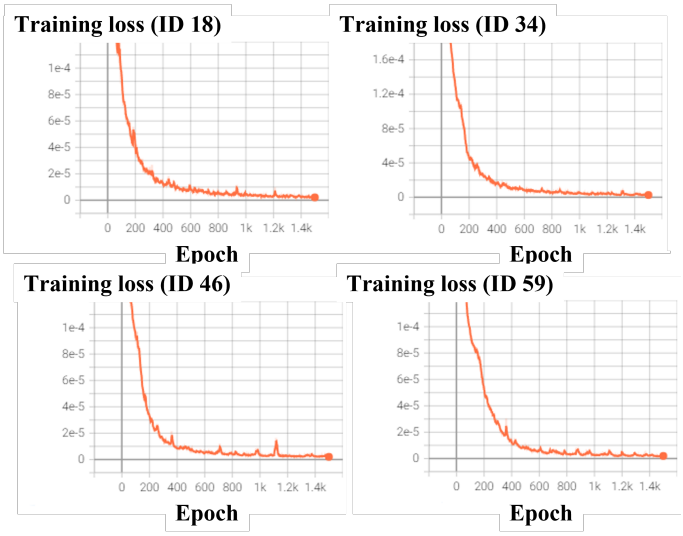
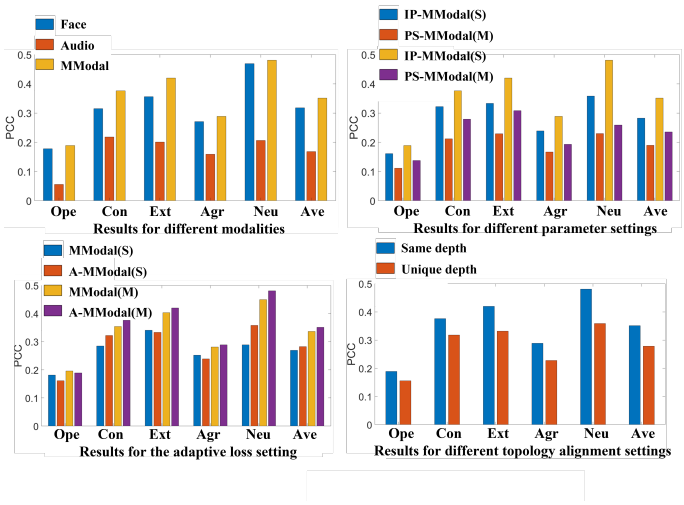


Fig. 9. The training loss curves of four subjects' person-specific CNNs.

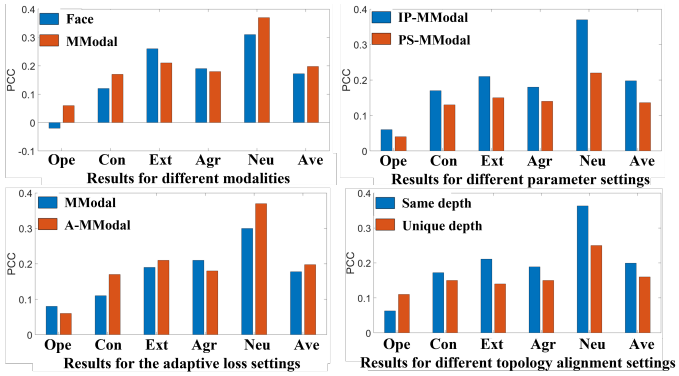
listener's facial reactions, and thus the explored person-specific CNNs may not learn good hypotheses of the listeners' cognitive processes. However, the sentence categorical feature provides the context information which provides a controlled condition for a subject's reaction as well as strong supervision for the search of person-specific CNNs. Thus, adding sentence categorical feature as the extra modality improves the recognition of most traits.

Parameter sharing strategies: For the results achieved on both datasets, it is clear that graph representations of persons-specific CNNs explored by the independent parameter (IP) strategy have clear advantages over the results achieved by the widely-used parameter sharing (PS) strategy [60], [61] over all five traits. These results validate our assumption that human cognition consists of a set of cognitive processes, each of which undertakes a unique function and can be different from others. Therefore, each part of the explored CNN should also have its own weights to better simulate a unique cognitive process/function.

Loss function settings: As we can see from the results on the NoXI dataset, despite most of our systems trained with standard MSE loss already achieved good performance in recognising Con, Ext and Neu traits, models trained using the proposed adaptive loss provided further improvements. Meanwhile, the system that used the adaptive loss achieved the similar results in recognising Ope and Agr traits with no significant differences. Specifically, the use of adaptive loss still brought more than 5.8% average improvement for Con, Ext and Neu traits. It can be observed from the results of human-machine interactions, the graph representations of person-specific CNNs trained with adaptive loss better recognised all five traits. Meanwhile, the systems that used the adaptive loss generated better facial reaction results. Since Neu and Ext traits can be better reflected by human cognition [14], [50], we hypothesize that the proposed adaptive loss can partially address the uncertainty of subjects' responding time, allowing the explored CNNs to better simulate target subjects' facial reaction-related cognition, which are well associated with Con, Ext and Neu traits.



(a) Personality recognition results on the NoXI dataset.



(b) Personality recognition results on the VHQ dataset.

Fig. 10. The results of different person-specific CNN settings. The definition of *MModal*, *S*, *M*, *PS-*, *IP-*, *A-* can be found in the captions of Table. 2, Table. 3, Table. 5 and Table. 6.

Depth settings: We also evaluate the influence of depth settings on person-specific cognition simulation and personality recognition. It can be found from Table 5 and Table 6 that the person-specific CNNs with their unique depths do not show clear advantages in reproducing listeners' facial reactions. These results suggest that even CNNs that were searched using the same depth have a comparable or even better capability to represent the target subjects' cognition. Moreover, as we can see from Fig. 10(a) and Fig. 10(b), the personality recognition results achieved by the heterogeneous graph representations of person-specific CNNs that have various depths are not as good as the isomorphic graph representations of person-specific CNNs that have the same depth. This may indicate that the differences in typologies can not reflect the differences of personality. In addition, the typologies of heterogeneous graph representations are varied a lot, which leads the training process of the corresponding GCNs to become more difficult.

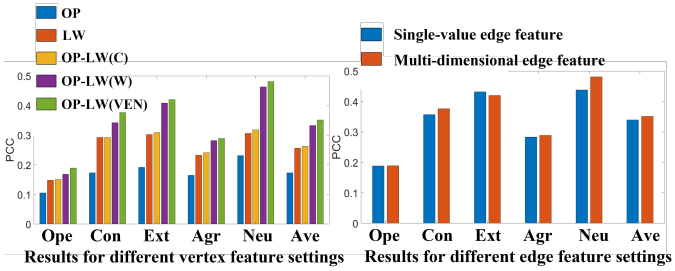
4.5.2 Graph representation

In this section, we demonstrate the advantages of the proposed end-to-end vertex feature and edge feature learning strategy for constructing graph representations in Fig. 11. Our best settings for vertex feature, LWs representation

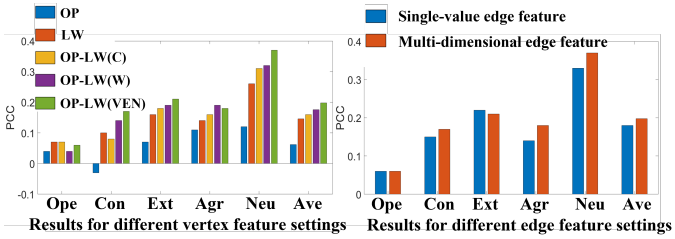
and edge feature are: vertex feature learned by the strategy proposed in Sec. 3.2.1 (denoted as OP-LW (VEN)) and edge features learned by ERNs.

Vertex feature settings: We first compare the proposed deep-learned vertex feature (OP-LW (VEN)) to four hand-crafted vertex features: 1. the OP feature: a vector that concatenates all OPs of a CNN edge; 2. the LW representation; 3. OP-LW (C) feature: a vector that concatenates all OPs and the LW representation of a CNN edge; 4. OP-LW (W) feature: a vector obtained by concatenating OPs that do not have LWs and a weighted vector that produced by multiplying LWs with their corresponding OPs. It can be seen that graph representations of the LW and OP-LW (C) vertex features have a similar capability for recognising true personality traits, both of which outperformed the graph representations that only use OPs as the vertex feature. This can be explained by the fact that the OP vertex feature ignores all LWs which are crucial in deciding CNNs' generalization capabilities. Then, we can conclude that the personality-related cues reside in both OPs and their LWs. Meanwhile, the OP-LW (C) feature does not show a clear advantage over the LWs feature, whose performance is also not comparable to the OP-LW (W) and OP-LW (VEN) features, demonstrating that simply concatenating OPs and LWs is not a proper way to combine their clues. In other words, the best recognition results of all five traits are achieved either by OP-LW (VEN) feature or OP-LW (W) vertex feature. As a result, we concluded that using each OP to weight corresponding LWs is a more superior way to combine OPs and LWs. Moreover, the OP-LW (VEN) setting shows significant advantages over the OP-LW (W) on Ope and Con traits under human-human interaction and Ope, Con. Ext, and Neu traits under human-machine interaction. Thus, we assume that the OP-LW (VEN) setting allows a better weighting vector to be learned to construct the each vertex, which not only considers the original OPs but also task-specific information.

Edge feature settings: We also compare the proposed end-to-end learned multi-dimensional edge features to the widely-used binary adjacency edge feature (0 or 1). It can be seen that the graph representations equipped with the proposed deep-learned multi-dimensional edge features outperformed the graph representations that only use a binary adjacency matrix to define the connectivity between vertices, with more than 3.4% and 10% average improvements under human-human and human-machine interaction scenarios, respectively. More importantly, the improvements brought by these edge features are significant for some traits (Con, Neu in the human-human interaction setting and Con, Agr, Neu in the human-machine interaction setting) as well as the average performance (please check the supplementary material). Such results validate the usefulness of the proposed end-to-end multi-dimensional edge feature learning strategy, which can better describe the relationship between adjacent vertices with multiple task-specific relationship clues, particularly for the Con and Neu trait. In other words, the task-specific multi-dimensional edge features lead the produced graph representations to have superior message passing mechanism when they are processed by GNNs, resulting in more discrminative latent personality representations.



(a) Personality recognition results on the NoXI dataset.



(b) Personality recognition results on the VHQ dataset.

Fig. 11. The results of different vertex and edge feature learning settings. The definition of settings can be found in Sec. 4.5.2.

5 Conclusions and Future Work

This paper proposes the first work which recognises true personality traits from the graph representation of an automatically explored person-specific CNN's architecture and parameters, where each CNN simulates the cognition of each target subject in terms of person-specific facial reactions. Our approach is evaluated on datasets of different nature (i.e., they are recorded under human-human vs. human-machine dyadic interaction scenarios), and the achieved results suggest the following conclusions: (i). the graph representations of person-specific CNNs are positively associated with the target subjects' self-reported personality traits, showing that the CNNs explored by our approach may have their own personalities, which are similar to their corresponding subjects; (ii). the proposed approach has clear advantages over most existing APP approaches which predict personality directly from non-verbal behaviours the target subject, demonstrating that it is reliable to recognise self-reported (true) personality from the simulated cognition of subjects; (iii). we found that the graph representations learned by the proposed approach are particularly informative for recognising Ext and Neu traits under both interaction scenarios; (iv). the proposed approach performed better personality recognition and facial reaction prediction under the human-human interaction scenario than the human-machine scenario, indicating that nonverbal behaviours expressed by human speakers are more powerful to trigger the listeners' personality-related facial reactions; (v). many human demographic attributes (e.g., age, gender, education level, and interpersonal relationship) can influence the performance of the proposed approach, where the gender and age are the most influential factors. This is caused by the fact that the facial reactions of a similar intention or emotion can be varied due to the these factors; and (vi). among several technical settings, the proposed adaptive loss function, independent parameters

strategy and end-to-end vertices/edges feature learning strategies have largely enhanced the personality recognition performance.

The main limitation of this work is that searching for a unique CNN architecture for each subject takes a relatively long time, i.e., the training and inference duration of the proposed approach are expected to be longer than most existing approaches. Therefore, it may not be suitable for *fast* personality assessment requirements. Another limitation is that we only used audio-visual modalities but ignored other human signals such as psychological signals (EEG, heart rates, skin temperature, etc.) and verbal information, which contribute important information to one's communication and reactions. As a result, a potential future direction is to accelerate the person-specific cognition simulation algorithm so that it does not require searching for a person-specific CNN for each person from scratch. Then, additional modalities (e.g., verbal signal, psychological signals, etc.) might enable the CNNs to be more similar to the target subjects' cognition in a dyadic interaction, where dialogue response generation can be utilised to predict listeners' verbal responses. All these modalities can in principle be combined via the proposed fusion module, i.e., combining them at multiple levels, as each is influenced by the others. There remain of course some modality-specific issues to resolve, so while it's definitely possible there is also substantial future research to be done in this area. Meanwhile, from the application perspective, this work opens up a new avenue of research for predicting and recognizing socio-emotional phenomena (personality, affect, engagement, etc.) from the simulations of person-specific cognitive processes that will have further implications for relevant fields including neuroscience, and cognitive, behavioural and emotion sciences. Another future work will focus on extending and evaluating our approach to analyze mental health or other human internal states with domain-specific loss functions under clinical settings, i.e. representing them with CNN parameters, or creating data-driven robot coaches that can express personalized behaviours during dyadic interactions [77], [78].

ACKNOWLEDGMENTS

The work of Siyang Song was funded partially by the EPSRC/UKRI under grant ref. EP/R030782/1, and partially by the European Union's Horizon 2020 research and innovation programme project WorkingAge, under grant agreement No. 82623. Hatice Gunes is supported by the EPSRC/UKRI under grant ref. EP/R030782/1. Linlin Shen and Zilong Shao are supported by the National Natural Science Foundation of China under Grant 91959108.

REFERENCES

[1] S. Jaiswal, S. Song, and M. Valstar, "Automatic prediction of depression and anxiety from behaviour and personality attributes," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[2] M.-T. Lo, D. A. Hinds, J. Y. Tung, C. Franz, C.-C. Fan, Y. Wang, O. B. Smeland, A. Schork, D. Holland, K. Kauppi *et al.*, "Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders," *Nature genetics*, vol. 49, no. 1, p. 152, 2017.

[3] 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 *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds., 2016, pp. 400–418.

[4] H. Gunes, O. Çeliktutan, and E. Sariyanidi, "Live human-robot interactive public demonstrations with automatic emotion and personality prediction," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, pp. 1–8, 2019.

[5] Y. Li, J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi, and G. Guo, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," *International Journal of Computer Vision*, pp. 1–18, 2020.

[6] L. Zhang, S. Peng, and S. Winkler, "Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship," *IEEE Transactions on Affective Computing*, 2019.

[7] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting cnn models for apparent personality trait regression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1705–1713.

[8] S. Fang, C. Achard, and S. Dubuisson, "Personality classification and behaviour interpretation: An approach based on feature categories," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 225–232.

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

[10] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2021.

[11] M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student's handbook*. Taylor & Francis, 2005.

[12] L. Willerman, R. Schultz, J. N. A. Rutledge, and E. Bigler, *Brain Structure and Cognitive Function*. Soc Neuroscience, 1994.

[13] K. A. Corrigall, E. G. Schellenberg, and N. M. Misura, "Music training, cognition, and personality," *Frontiers in psychology*, vol. 4, p. 222, 2013.

[14] V. Kumari, S. C. Williams, J. A. Gray *et al.*, "Personality predicts brain responses to cognitive demands," *Journal of Neuroscience*, vol. 24, no. 47, pp. 10 636–10 641, 2004.

[15] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE transactions on affective computing*, vol. 8, no. 1, pp. 29–42, 2015.

[16] Z.-T. Liu, A. Rehman, M. Wu, W. Cao, and M. Hao, "Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features," *IEEE Transactions on Multimedia*, 2020.

[17] G. An and R. Levitan, "Lexical and acoustic deep learning model for personality recognition." in *INTERSPEECH*, 2018, pp. 1761–1765.

[18] R. D. P. Principi, C. Palmero, J. C. Junior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Transactions on Affective Computing*, 2019.

[19] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2018.

[20] 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 *European Conference on Computer Vision*. Springer, 2016, pp. 349–358.

[21] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. 5, pp. 705–721, 2016.

[22] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2019.

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

[24] H. Kaya, F. Gurpinar, and A. Ali Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–9.

[25] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. Junior, C. Jacques, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund *et al.*, "Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2177–2188.

[26] C. Beyan, A. Zunino, M. Shahid, and V. Murino, "Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images," *IEEE Transactions on Affective Computing*, 2019.

[27] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, 2013.

[28] W. Mou, H. Gunes, and I. Patras, "Your fellows matter: Affect analysis across subjects in group videos," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–5.

[29] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Personality recognition by modelling person-specific cognitive processes using graph representation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 357–366.

[30] J. Joshi, H. Gunes, and R. Goecke, "Automatic prediction of perceived traits using visual cues under varied situational context," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 2855–2860.

[31] S. Eddine Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 10–13.

[32] L. Teijeiro-Mosquera, J.-I. Biel, J. L. Alba-Castro, and D. Gatica-Perez, "What your face vlogs about: expressions of emotion and big-five traits impressions in youtube," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 193–205, 2015.

[33] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion—a systematic study," *IEEE Transactions on Affecive Computing*, vol. 3, no. 4, pp. 443–455, 2012.

[34] L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez, "Multimodal analysis of body communication cues in employment interviews," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 437–444.

[35] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 15–22.

[36] S. M. Kassin, *Essentials of psychology*. Prentice Hall, 2003.

[37] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1660–1663.

[38] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042.

[39] S. Song, E. Sánchez-Lozano, M. Kumar Tellamekala, L. Shen, A. Johnston, and M. Valstar, "Dynamic facial models for video-based dimensional affect estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[40] S. Song, E. Sanchez, L. Shen, and M. Valstar, "Self-supervised learning of dynamic representations for static images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1619–1626.

[41] 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 *European Conference on Computer Vision*. Springer, 2016, pp. 400–418.

[42] J.-I. Biel and D. Gatica-Perez, "Voices of vlogging," in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[43] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 816–832, 2011.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[45] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.

[46] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 350–359.

[47] C. Palmero, J. Selva, S. Smeureanu, J. C. J. Junior, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera, D. Leiva *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." in *WACV (Workshops)*, 2021, pp. 1–12.

[48] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, 2017.

[49] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[50] O. F. Kernberg, "What is personality?" *Journal of Personality Disorders*, vol. 30, no. 2, pp. 145–156, 2016.

[51] C. G. DeYoung, J. B. Hirsh, M. S. Shane, X. Papademetris, N. Rajeevan, and J. R. Gray, "Testing predictions from personality neuroscience: Brain structure and the big five," *Psychological science*, vol. 21, no. 6, pp. 820–828, 2010.

[52] J. Jackson, D. A. Balota, and D. Head, "Exploring the relationship between personality and regional brain volume in healthy aging," *Neurobiology of aging*, vol. 32, no. 12, pp. 2162–2171, 2011.

[53] N. Kogan and M. A. Wallach, "Risk taking: A study in cognition and personality." 1964.

[54] R. R. McCrae, "Openness to experience as a basic dimension of personality," *Imagination, Cognition and Personality*, vol. 13, no. 1, pp. 39–55, 1993.

[55] K. W. Schaie, S. L. Willis, and G. I. Caskie, "The seattle longitudinal study: Relationship between personality and cognition," *Aging Neuropsychology and Cognition*, vol. 11, no. 2-3, pp. 304–324, 2004.

[56] J. Haberman, T. Harp, and D. Whitney, "Averaging facial expression over time," *Journal of vision*, vol. 9, no. 11, pp. 1–1, 2009.

[57] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[58] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 372–381.

[59] S. Card, T. MORAN, and A. Newell, "The model human processor-an engineering model of human performance," *Handbook of perception and human performance.*, vol. 2, no. 45–1, 1986.

[60] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018.

[61] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4095–4104.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[63] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[64] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[65] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[66] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," *arXiv preprint arXiv:2003.00982*, 2020.

[67] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.
[68] G. Saucier, "Mini-markers: A brief version of goldberg's unipolar big-five markers," *Journal of personality assessment*, vol. 63, no. 3, pp. 506–516, 1994.
[69] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.
[70] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.
[71] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.
[72] "Aria-valuspa platform," https://github.com/ARIA-VALUSPA/AVP, 2019.
[73] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[74] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 158–165.
[75] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020.
[76] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.
[77] N. Churamani, M. Axelsson, A. Caldir, and H. Gunes, "Continual learning for affective robotics: A proof of concept for wellbeing," *Proc. ACII 2022 Demos and Workshops*, 2022.
[78] M. Axelsson, M. Spitale, and H. Gunes, "Robots as mental well-being coaches: Design and ethical recommendations," *arXiv preprint arXiv:2208.14874*, 2022.

Siyang Song is a Research Associate with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K.. He received his PhD in the Computer Vision Lab and Horizon Center for Doctoral Training at the University of Nottingham, UK. His research interests include automatic emotion, personality and depression analysis by developing various self-supervised learning, Neural Architecture Search and graph modelling techniques.

Zilong Shao is a master student in Shenzhen University majoring in computer technology. He received the B.S. degree of computer science and technology from Shenzhen University, Shenzhen, China in 2019. His current research interests include model compression, neural architecture search, personality analysis and deep learning.

Shashank Jaiswal is a post-doctoral Research Fellow at the School of Computer Science, University of Nottingham. He received his PhD in computer science at the University of Nottingham in 2018. His research interests include automatic facial expression recognition and its applications in the diagnosis of mental health conditions.

Linlin Shen is currently a professor at Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also a Honorary professor at School of Computer Science, University of Nottingham, UK. He serves as the director of Computer Vision Institute and China-UK joint research lab for visual information processing. He received the BSc and MEng degrees from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K. He was a Research Fellow with the University of Nottingham, working on MRI brain image processing. His research interests include deep learning, facial recognition, analysis/synthesis and medical image processing. Prof. Shen is listed as the Most Cited Chinese Researcher by Elsevier. He received the Most Cited Paper Award from the journal of Image and Vision Computing. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP 2013 and ICPR 2016.

Michel Valstar is a Professor in the Computer Vision and Mixed Reality Labs at the University of Nottingham. He received his masters degree in Electrical Engineering at Delft University of Technology in 2005 and his PhD in computer science with the intelligent Behaviour Understanding Group (iBUG) at Imperial College London in 2008. His main interest is in automatic recognition of human behaviour. In 2011 he was the main organiser of the first facial expression recognition challenge, FERA 2011. In 2007 he won the BCS British Machine Intelligence Prize for part of his PhD work. He has published technical papers at authoritative conferences including CVPR, ICCV and SMC-B and his work has received popular press coverage in New Scientist and on BBC Radio.

Hatice Gunes is a Professor with the Department of Computer Science and Technology, University of Cambridge, U.K., leading the Affective Intelligence and Robotics Lab. Her expertise is in the areas of affective computing and social signal processing cross-fertilizing research in human behavior understanding, computer vision, machine learning, and human–robot interaction. She has published over 125 papers in the above areas, and her research highlights include RSJ/KROS Distinguished Interdisciplinary Research Award Finalist at IEEE RO-MAN'21, Distinguished PC Award at IJCAI'21, Best Paper Award Finalist at IEEE RO-MAN'20, Finalist for the 2018 Frontiers Spotlight Award, Outstanding Paper Award at IEEE FG'11, and Best Demo Award at IEEE ACII'09. Prof Gunes is the former President of the Association for the Advancement of Affective Computing (AAAC), and is a member of the Human-Robot Interaction Steering Committee. In 2019 she was awarded the prestigious EPSRC Fellowship as a personal grant and was named a Faculty Fellow of the Alan Turing Institute.