A Media-Guided Attentive Graphical Network for Personality Recognition Using Physiology

Hao-Chun Yang, *Student Member, IEEE,* and Chi-Chun Lee, *Senior Member, IEEE*

Abstract—Physiological automatic personality recognition has been largely developed to model an individual's personality trait from a variety of signals. However, few studies have tackled the problems of integration methodology from multiple observations into a single personality prediction. In this study, we focus on finding a novel learning architecture to model the personality trait under a *Many-to-One* scenario. We propose to integrate not only the information on the user but also consider the effect of the affective multimedia stimulus. Specifically, we present a novel Acoustic-Visual Guided Attentive Graph Convolutional Network for enhanced personality recognition. The emotional multimedia content guides the formation of the physiological responses into a graph-like structure to integrate latent inter-correlation among all responses toward affective multimedia. Then these graphs would be further processed by the Graph Convolutional Network (GCN) to jointly model instances and inter-correlation levels of the subject's responses. We show that our model outperforms the current state of the art on two large public corpora for personality recognition. Further analysis reveals that there indeed exists a multimedia preference for inferring personality from physiology, and several frequency-domain descriptors in ECG and the tonic component in EDA are shown to be robust for automatic personality recognition.

1 Introduction

Personality is an important psychological construct that can be characterized by a few stable and measurable attributes. It has long been regarded as a key internal construct due to its role in influencing an individual's emotion, modulating behaviors, and triggering decisions, i.e., knowing one's personality would effectively provide us a sneak peek of an individual's behavioral patterns. Developing computational methods that enable automatic personality recognition (APR) [1] has drawn tremendous interest because of its wide application across different domains. For example, in *Human-Computer Interaction, (HCI)*, researches have shown that personalized adaptation based on individual personality traits can improve user experiences [2], [3], [4]; personality-driven recommendation systems have also enabled precision marketing for different media/product consumption, such as music [5], [6], [7], movie [8], [9], and e-commerce [10], [11], [12]. Lastly, personality traits have also been shown to be correlated to lifespan health. Higher levels of *Neuroticism* and *Conscientiousness* will lead to higher phishing (fraud) vulnerability [13], [14], while people scoring higher in honesty and humility are more likely to become fraud victims. All of these demonstrate that an APR-integrated system would benefit the delivering of personalized media content with impact [15], and continuously advancing a robust APR system is a critical technical endeavor.

Most of the prior research on APR has focused on modeling different signal modalities as measured by users. For example, a major effort of APR development has used lexical information [16] that enables personalized profiling on social media platform [17]. Recently, the proliferation of miniaturized sensors has enabled low-cost and precise monitoring of different human internal physiological signals. This property of continuous sensing and seamless sensor deployment possess a great advantage to ubiquitous computing by harnessing these properties that traditional modalities do not afford [18]. In contrast to expressive cues (such as written language), these bio-signals provide a scientifically grounded indicator to model personality traits directly from neurophysiological evidence. These physiological signals, such as *electroencephalogram (EEG)* and *Electrodermal activity (EDA)*, represent the reaction of the central and peripheral nervous systems (CNS and ANS) that is deferentially activated when an individual encounters emotional stimulation [19]. Most if not all of these physiological-based recognition works share a common experimental setting, that is, by using emotion-rich audio-visual data as stimuli to elicit the subject's internal physiological responses, one can then build models on these physiological measurements for automatic recognition of personality [20], [21], [22].

Affective multimedia stimuli play an important role in all of these studies. The intriguing connection between personality and media stimulation has been well documented in various studies. For example, there is a significant preference bias for Extroverts on the choice of TV programs and music genre [23]; individuals with higher *Openness* often favor reflective/complex music (such as jazz), while people with higher *Neuroticism* prefer more emotional music [24]. Cristani et al. 's study [25] also demonstrates that visual patterns extracted from "favorite" Flickr images can be used to predict user traits. In Colombo et al. 's study [26], *Creativity*, i.e., one of the personality dimensions indexing one's curiosity and tolerance for ambiguity, could be distinguished from pre-selected commercials; their psychophysiological test reveals that less creative individuals would consistently be activated by all stimuli videos, while highly creative individuals would get less activated during plain stimuli. Hence, a key presumption is that one's traits are considered as a stable state over time; when an emotional-rich media content is exposed to users to trigger the latent arousal, this physiological changes as conditioned on the media content would differ between people with different

personalities. Developing a computational framework in modeling physiology under affective media stimulation has been at the core of realizing an APR system.

However, most of the current physiological APR systems neglect the information about these multimedia cues in their modeling frameworks. That is, no computational work has yet addressed the issue of personality recognition from physiology with joint consideration of the exposed acoustic-visual stimuli. They ignore that an individual's bodily signals are triggered through these multimedia stimuli, which serve as a latent conditional control toward physiological responses. We argue that to develop a robust and enhanced APR recognition model from physiology, these media content signals should be integrated to properly model the intricate dependencies of personality as a function of affective media stimuli and physiological responses.

Hence in this research, we propose a novel learning network that would aggregate ***Multiple-Observations*** into ***Single-Personality***. All the physiological cues of a single subject while simultaneously considering the original multimedia content for personality trait recognition. According to our hypotheses, incorporating the stimuli's acoustic-visual information as pre-regularization into learning would further achieve a more accurate personality prediction. In summary, the contributions of this paper are three-fold:

1): To our best knowledge, this is the first work that addresses the problem of multiple instance integration for personality recognition. While all previous studies [27], [28] aggregate multiple observations to a single individual personal prediction by either early/late fusion method, we present a data-driven voting method which would weigh an uncertain number of instances for enhancing modeling.
2): We propose a novel content-graph personality recognition framework that is evaluated on two publicly available large physiological datasets, Amigos [21] and Ascertain [20]. The original video elicitations are preprocessed into *Visual-Semantic* and *Acoustic-Affective* embeddings to be used as representation to inject the multimedia content information into the learning framework. Furthermore, we propose a graphical structure that is designed to model both instances and inter-correlation levels of physiology.
3): We analyze the learned attention distribution to uncover dominant videos for personality elicitation. Additional statistical testings are performed to uncover the hidden correlation between subjective emotional feelings and personality traits. Finally, several physiological descriptors are highlighted as key indicators in revealing personality traits.

We first introduce the idea of integrating multimedia for personality recognition using physiology in our previous work [29]. We have substantially extended the preliminary conference version in the followings: (1) we advance our algorithm in both graphical construction and prediction step, which alleviate the impractical constraint in requiring additional emotional tagging on video stimuli as done in our previous work, (2) we evaluate our framework on an additional and larger corpus and includes acoustic modality of multimedia stimuli verifying the robustness of our system. The recent STOA multiple-instance learning methods were also jointly leveraged to verify our contribution to media-guided learning strategy. (3) we conduct comprehensive analysis under both intra- and inter- corpus scenarios to identify the key physiological indicators.

The rest of this paper is organized as follows. Section 2 discusses related works. Section 3 introducs the datasets. Section 4 5 details our proposed framework and the experiment settings respectively. Section 6 summarizes the personality recognition results where section 7 shows the analysis. Finally, section 8 concludes the paper.

2 Related Works

In this section, we present a review of the existing works that are closely related to our study in terms of automatic personality recognition and affective graphical modeling.

2.1 Automatic Personality Recognition

Recently, there is a growing number of studies in developing robust APR [30]. Depending on the triggering type and modality characteristic, APR could be further divided into either *Spontaneous* or *Triggered* personality recognition. Textual data is one of the major modalities that has been widely studied in assessing one's personality and is considered a spontaneous type. Specifically, there has been a systematic effort in constructing a dictionary with psychological evidence with the indication of an individual's personality trait [31], [32]. Recently, a lot more research has focused on analyzing profiles in social networks [33], [34], [35] due to its inclusion of multimodal data. That is, behavior collected using audio-video data on an individual has also been shown to reveal personality traits. For example, studies in [36], [37] have shown that self-reported personality traits could be inferred from spontaneous talks using both acoustic-prosodic cues. Furthermore, the profile picture of a virtual avatar put on the social network is known to be indicative of one's trait as well [38], [39], [40]. Finally, an individual's facial expressions as recorded in the video also reveal one's personality-related characteristic [41], [42]. A major computational effort has largely been concentrated on modeling expressive and spontaneous behavior data.

In contrast, fewer computational works have investigated the development of APR using physiology. The internal nature of this modality as it is reactive to the *triggered* external stimuli makes the measurement feasible, and this provides us a chance to compute personality traits from a different perspective compared to *spontaneous* setting. Abadi et al. [43] fuse an individual's reactive facial expression with the ECG and GSR data for improved trait recognition; they apply a linear regression model to obtain Fl-scores of 70% and 69% for predicting extroversion and openness. In [44], emotional level (Arousal / Valence) is additionally examined for personality recognition, and they conclude that physiological responses using similar emotional clips could reveal more personality differences. Correa et al. [45] propose a multi-task cascaded network using EEG data, which firstly predicts the emotion status then finally aggregates all the responses from a single person for personality recognition; their approach achieves an improvement of 2.7% mean f1-scores on average.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing

From all these literature reviews, we could conclude that all these studies focused on either feature developing or multimodal fusion, yet there have not been any computational works specifically tackling the problem of the ***Many-to-One*** personality aggregation algorithm development. An efficient and reliable method to aggregate multiple observations from a single subject for personality modeling remains unstudied. Hence in this study, comprehensive works on multiple instance aggregation methods were covered, and we further propose a media-guided attentive graphical model to help the multiple observations aggregation.

2.2 Affective Graphical Modeling

Integrating graphical modeling into deep learning approaches has been growing in the field of affective computing. Graph Neural Networks (GNNs) is an effective representation learning framework in modeling non-Euclidean space of complex structural relationships between interdependency objects [46], [47]. Several research works have applied GCNs on brain images due to their graphical nature. For example, Zhong et al. [48] propose a robustly regularized GCN model that achieves state-of-the-art accuracy on EEG-based emotion prediction. In [49], a dynamic GCN model is designed to automatically learn the connection among brain regions for an enhanced EEG emotion recognition. Graphical learning is not only limited to brain signals. Ghosal et al. [50] propose to model group interaction as a temporal graphical process, in which the dialogue interactions could be jointly modeled for better speech emotion recognition. Furthermore, a visual-based GCN model is proposed by Bhattacharya et al. [51] to learn an emotion assessment with gait data.

One of the assumptions in applying GCN models is that the data should be graph-like, which is composed of nodes and edges. According to the formation of the edges, we further categorize the GCN modeling: 1. The edges are linked by node attributes (mostly relative distances). These edges are formed through known domain rules or characteristics, such as brain structures [52], [53], [54] or skeleton keypoints [55], [56], [57]. Usually, these edges are initialized at the beginning and would remain constant throughout the whole learning progress; 2. The edges are automatically learned from the data. These types of works mostly targeting on developing algorithms that could infer the latent graphical structure through data itself without prior knowledge [58]. Temporally varying data would also require a dynamic graph learning strategy to adapt the structural variation over time [59]. Although these methods automatically infer the graphical structure, it often has a high requirement on both the quality and quantity of data for edge learning.

Motivated by these studies, this work proposes a multimodal graphical construction technique that embeds auditory or visual information into physiological graph modeling. With the additional constraint from these original physiological responses' inducers, the proposed model can help to better mine the input signals, learning a more discriminative and robust personality modeling.

TABLE 1: Big-Five personality and associated adjectives. [60]

Personality Trait	Adjectives
Agreeableness (Agr)	Appreciative, Forgiving, Generous, Kind, Sympathetic
Conscientiousness (Con)	Efficient, Organized, Planful, Reliable, Responsible, Thorough
Creativeness (Cre)	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests
Emotion Stability (Emo)	Unenvious, Relaxed, Unexcitable, Patient, Undemanding, Imperturbable
Extraversion (Ext)	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative

TABLE 2: Personality label distribution (Low/High) for two separate datasets.

	Agr	Con	Cre	Emo	Ext
Amigos	16/22	16/22	24/14	19/19	24/14
Ascertain	30/28	31/27	33/25	29/29	29/29

TABLE 3: The list of the repetitive video stimuli used in both Amigos and Ascertain databases.

Video ID		Source Movie
Amigos	Ascertain	
1	10	August Rush
2	13	Love Actually
4	18	House of Flying Daggers
6	20	My Girl
7	23	My Bodyguard
9	31	Prestige
10	34	Pink Flamingos
11	36	Black Swan
12	4	Airplane
13	5	When Harry Met Sally
16	9	Hot Shots

3 Datasets

In this study, we use two large physiological datasets collected under a similar scenario for our algorithm development and evaluation. In each dataset, a series of emotional videos with intended affective stimuli (annotated with high/low arousal or valence, *-Int*) were delivered as multimedia elicitation to arouse the participants' affective responses. The participants were asked to self-disclose their subjective feelings (*-Sb*) at the end of each video, while their physiological responses (ECG, EDA) were recorded with bio-sensors throughout the time. Meanwhile, personality measures for the big-five dimensions were also compiled using a Big-Five marker scale (BFMS) questionnaire [61]. The Big-Five framework has become one of the widely used personality trait measurements [62] and could be easily interpreted by referring to their associated personality dimensions as presented in Table 1. Specifically, we carry out the personality recognition experiments as a binary classification problem, i.e., for each dataset, the labels for personality are divided into high and low classes using the median value of each dimension as the threshold as demonstrated in Table 2. Several details of the datasets are listed below:

- **Amigos (Am)** [21]: A total of 16 short emotional videos (duration<250s) were carefully chosen from previous re-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing

TABLE 4: An overview of physiological low-level descriptors extracted from [67]. "F*" indicates 15 statistical functions[1].

Modality	Low-Level Descriptors
ECG(51)	number_of_artifacts, RMSSD, meanNN, sdNN, cvNN, CVSD, medianNN, madNN, mcvNN, pNN50, pNN20, Triang, Shannon_h, ULF, VLF, LF, HF, VHF, Total_Power, LFn,HFn, LF/HF, LF/P, HF/P, DFA_1, DFA_2,Shannon, FD_Higushi, Average_Signal_Quality, F* Cardiac_Cycles_Signal_Quality
EDA(68)	F*SCR_Onsets, F*SCR_Peaks_Amplitudes, F*EDA_Phasic, F*EDA_Tonic

search as elicitation. 40 participants aged between 21 and 40 (mean age 28.3) were recruited in a laboratory environment. Two participants were excluded due to missing value of original personality scores, hence in total, there are 38 subjects left for experiments.

- **Ascertain (As)** [20]: Ascertain is one of the largest datasets aiming for studying physiological responses under emotional content stimuli. There are 36 short movie clips (duration 51˜127s) used for affective elicitation with 58 university students (mean age 30) collected in this dataset. The whole data collection was conducted in the laboratory environment using a commercial physiological sensor. Note that there are 11 video stimuli listed in Table 3 which are the same across both Amigos and Ascertain.

4 METHODOLOGY

4.1 Physiological Low-Level Descriptors (LLDs)

We first apply a low-pass filter cut-off at 60Hz on both ECG and EDA signals. For ECG features, we first calculate the RR-intervals for each video elicitation. Then, several standard Heart Rate Variabilities (HRVs) like Standard Deviation of the time interval between successive normal heartbeats (SDNN) in the time domain or Low Frequency to High-Frequency ratio (LF/HF) in the frequency domain are calculated, which is known to be an important marker of autonomic nervous system (ANS) modulation [63]. As for EDA data, we compute the tonic and phasic component [64], which has previously been shown as an important measure linking the physiological status toward affective responses [65]. Besides, we also extract the Skin Conductance Responses (SCR) onsets, peaks, and amplitudes, which are commonly used for revealing the event-related alteration during psychophysiological tests [66]. All of these LLDs would act as key indicators of how the participants react toward the multimedia stimulation, such that we could use it to infer the participant's personality traits. The exact features and dimensions are listed in Table 4, and we use the open-source toolkit [67] for feature extraction. A standard z-normalization is then performed subject-wise on each feature dimension to mitigate the issue of individual differences.

4.2 Multi-Media Graph Building

In this research, our goal is to perform personality recognition in a multi-instances setting, i.e., given uncertain number of video elicited physiological responses, we aim to find a mapping between these multiple physiological data points into single personality score. Specifically, considering a set of subject i's d-dimensional LLDs $\boldsymbol{x_i} = \{x_i^1, \ldots, x_i^{N_v}\} \subset \mathbb{R}^d$ while N_v denotes the number of the video stimuli during the experiment, our objective is to find a *Many-To-One* mapping $\boldsymbol{\mathcal{F}}$, which maps $\boldsymbol{x}$ to personality label $\boldsymbol{y} = \{y_1, \ldots, y_{N_s}\}$, where N_s is number of subjects. To handle this multi-instances problem, we utilize the idea of graphical signal processing.

We first transform each subject's LLDs into subject-wise graph-like structural representation $\boldsymbol{\mathcal{G}} = \{\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}\}$. For each subject i, the node-set $\mathcal{V}_i$ is comprised of all his observed LLDs $\boldsymbol{x_i}$. As for the edges $\mathcal{E}_i$, there are two types of connectivity mechanism for graph building:

- ***Visual-Semantic*** ($\mathcal{E}^{Video}$): The first type focuses on describing the visual semantic cues of the source video stimuli. For each video j, we utilize the 3D spatial-temporal convolutional network [68], which is pre-trained on a large video understanding corpus Kinetics [69] to extract the frame-level video semantic vectors $\boldsymbol{v_j^{frame}} = \{v_j^1, \ldots, v_j^{t_k}\} \subset \mathbb{R}^{d_v}$, where t_k is number of time steps of video while d_v is the embedding dimension. Then, an average pooling over time is applied to obtain the video-level embedding. A further dimensionality reduction method UMAP [70] was performed to prevent the curse of dimensionality. The UMAP embedding method has been verified as an improved non-supervise dimension reduction tool that shown better discriminative power particularly on retain both the local and global structure on the data distribution [71], [72]. More Specifically, the traditional method such as PCA [73] or t-SNE [74] could not meet our need to embed the relative distance in multimedia semantic space, which the former only focus on finding the principal components (eigenvectors) while the latter only preserve the local instances distance but omit the multimedia-space global structure. We regard this dimensionally-reduced embedding v_j^{video} as representing semantic information of the video content.

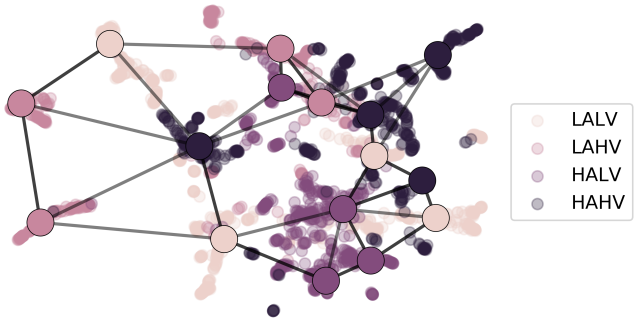


Fig. 1: The 2D visualization of the Visual-Semantic embeddings with $K = 3$ in Amigos dataset. Each small circle in the graph refers to the frame-level embedding of the video, while the larger circles are the final aggregated video-level vectors. The colors are according to the original intended emotional level (*Int*). L: Low, H: High, A: Arousal, V: Valence.

1. max, min, mean, median, std, skewness, kurtosis, min position, max position, 25_percentile, 75_percentile, 75_percentile-25_percentile, 1_percentile, 99_percentile, 99_percentile-1_percentile

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing

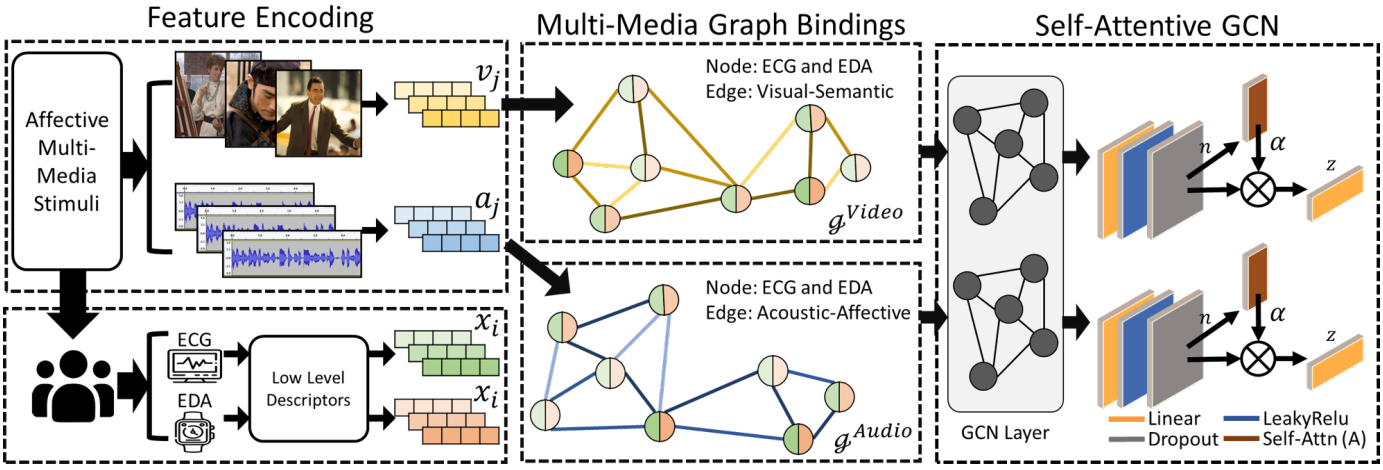


Fig. 2: Our proposed Acoustic-Visual Guided Attentive Graph Convolution Network.

Moreover, since the physiological responses are induced through these video content, better modeling of these videos means to include a prior description of latent control of the physiology from these video content. Hence, to properly incorporate these visual semantic into our modeling, we initially retrieve the K-nearest neighbor of all extracted v_j^{video} as depicted in Fig.1. Then, any of the two nodes (physiological responses) in which their source stimuli v_j^{video} are in their K-nearest neighbor would be considered linked in $\mathcal{E}^{Video}$.

- ***Acoustic-Affective*** **($\mathcal{E}^{Audio}$)**: In addition to visual elicitation, acoustic cues are also another important source for affective arousal [75]. Here we utilize the Open-Source acoustic tool-box OpenSMILE [76] to characterize the acoustic component from the videos. We extract the frame-level ComParE16 features sets $\boldsymbol{a_j^{frame}} = \{a_j^1, \ldots, a_j^{t_k}\} \subset \mathbb{R}^{d_a}$ for each video j, followed by averaging over time with dimension reduction method applied to generate a single video-level acoustic embedding a_j^{video}. Then, we use the same technique as in visual-semantic graph to retrieve the K-nearest neighbors for the edge binding $\mathcal{E}^{Audio}$.

To this end, for each subject i's physiological responses, we have bound them into two different graphical representations $\mathcal{G}_i^{Video}$ and $\mathcal{G}_i^{Audio}$ which take into account of the auditory and visual component of the elicitation, respectively. These graphical representations will be further processed through graph convolutional operators in the following section.

4.3 Self-Attentive GCN

Our model is primarily motivated as a variation of Graph Convolutional Network (GCN) [77] which performs a spectral convolution for modeling structural data. The power of capturing non-linear inter-relationship among instances (nodes) makes it a great fit for our Multiple-to-Single personality recognition. The core GCN layer can be interpreted as a special case of a first-order differentiable message-passing framework:

$$H^{(l+1)} = \sigma(\tilde{\mathrm{D}}^{-\frac{1}{2}}\tilde{\mathrm{A}}\tilde{\mathrm{D}}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

Here, H^l denotes the l^{th} layer in the network, and D,A refers to the degree and adjacency matrix decomposed from the above semantic graphs $\boldsymbol{\mathcal{E}}$. The $\sim$ is a re-normalization trick in which that the self-connection is added to each node of the graph. The model input H^0 is equivalent to the node matrix $\boldsymbol{\mathcal{V}}$ of the graph with shape $N_v \times d$. During the forward pass, each node would perform message sharing among the linked nodes, then multiplied by a learnable weight matrix W of shape d^lxd^{l+1}, and finally activated by a non-linearity function σ. Through this process, each node's (a subject's single observed LLD) would first cross-refer to its neighboring node (other linked LLDs which were visually/auditorially closed in the source of the stimuli as depicted in Section 4.2), then non-linearly integrated through trainable GCN parameters as personality-refined representation. Through multiple stacked layers of the GCN blocks, eventually, we would get a set of representations that contains maximally personality information. The representation would be a $N_v \times d^l$ matrix, where each row n_j acts as the d^l-dimensional latent states for each video elicitation.

Since our goal is to learn an automatic mapping from multiple physiological data points of an individual into a single individual-level personality trait, we employ an automatic soft weighting mechanism to aggregate the above GCN-generated representation. Specifically, we integrate a self-attention technique [78] into our model:

$$\alpha_i^j = \frac{exp(\boldsymbol{A}(n_i^j))}{\sum_{j=1}^{N_v} exp(\boldsymbol{A}(n_i^j))} \tag{2}$$

where $\boldsymbol{A}$ is a trainable network for outputting the attention weights. Note that during the calculation of the $\boldsymbol{\alpha_i}$, the output of the GCN n_j is fed separately to obtain the attention weight for each video j. We consider this step as a data-driven regularized learning of a graph representation for personality classification. Finally, a single graph-level output would be obtained for each subject i as:

$$z_i = \boldsymbol{n_i^\intercal \alpha_i} \tag{3}$$

which is then fed into the standard deep neural network $\boldsymbol{P}$ for final binary classification. The whole network $\boldsymbol{\mathcal{F}}$ outputs

a single personality prediction for each subject, and we update the model through standard cross-entropy loss:

$$\min_{\mathcal{F}} -\frac{1}{n}\sum[y\log\mathcal{F}(\mathcal{G}) + (1-y)(1-\log\mathcal{F}(\mathcal{G}))] \quad (4)$$

The overall network is shown in Fig.2 and is implemented with back-propagation using the sparse matrix multiplication kernel [79].

5 Experimental Settings

5.1 Hyperparamters and Evaluation Metric

The exact architecture of our acoustic-visual guided attentive GCN includes three blocks of networks: GCN block $\boldsymbol{G}$ of a standard GCN layer with dimension $[d - d/10]$; attention block $\boldsymbol{A}$ that is composed of a single trainable matrix with dimension $[d/10 - 1]$ to output an attention weight for each node; the final prediction layer $\boldsymbol{P}$ is constructed using a dense layer with dimension $[d/10 - 2]$. Several hyperparameters are grid-searched: dropout rate between $[0.2, 0.5]$, learning rate among $[0.01, 0.005, 0.001]$, number of connected edges K is set between $[10\% \sim 90\%]$ of the number of original video stimuli. Batch size is fixed as 8, the max epoch is 150 with early stopping patience 10, and the optimizer is Adam. To prevent overfitting, we carry out all experiments under subject-independent 10-fold cross-validation(CV). For each set of hyperparameter, we repeat the experiments 10 times with 10 different randomly initialized model parameters and average the calculated metrics. Finally, the highest results from our hyperparameter grid-searched space were reported. The final evaluation metric used is the unweighted average recall (UAR) and weighted F1-score(F1).

5.2 Comparison Models

To comprehensively evaluate our proposed personality aggregation network, we first conduct our experiments utilizing linear SVM and vanilla DNN with the commonly performed early/late(concatenation/majority vote) aggregation methods. Since to our best knowledge that no other researches are specifically tackling the similar *Many-to-One* aggregation methods, we also re-implement some state-of-the-art algorithms that were studied in other fields to verify our idea of using acoustic/visual information to help personality aggregation. Model details are listed below:

- **SVM and DNN**: The naive classification approach without using the deep self-attention soft-voting scheme. Here, we use the SVM with linear kernel and grid-searched regularization parameter C among $0.1, 1, 10$ [83]. There are two variations to integrate multiple physiological responses into a single personality prediction; i.e., through feature concatenation (-*C*) and majority vote (-*V*). The concatenation method is originally used in the Amigos [21] dataset paper, while majority voting was broadly applied for personality aggregation [28], [80].
- **Attention Multiple Instance Learning (AMIL)** [81]: Multiple instance learning can be seen as a variation of supervised learning that is designed to predict a single label while a bag of instances is given. In this scenario, the bag could be viewed as the multiple stimulated physiological responses aroused from various video elicitation, while the single label refers to a subject's personality trait. The improved AMIL adopts the use of self-attention mechanism 2 as a soft-voting scheme during the bag prediction, which has been regarded as one of the state-of-the-art methods for multiple instances prediction tasks. However, comparing with graphical models, this method focuses on an instance-level integration but ignores the potential structural information between instances. We regard this method as the very naive data-driven personality aggregation baseline. Hyperparameters like learning rate and dropout were also grid-searched for a fair comparison.
- **Set Transformer (SET)** [82]: Inspired from the transformer model [84] which has achieved great success in various tasks, Lee et, al proposed a novel *Many-to-One* network called Set-Transformer. By taking advantage of the characteristic of the Multi-Head Attention (MHA), this model explicitly forces the architecture to learn the interactions among data from the same set. In our scenario, each set is composed of all Physiologies from a single subject, while the model would return a single output as the subject's personality trait. Due to its superior performance on various instance aggregation tasks, we would regard this model as the most closely state-of-the-art personality aggregation method. We re-implement the network with the same encoder-decoder architecture, in which the encoder is a Set Attention Block (SAB) while the decoder is composed of Pooling by Multihead Attention (PMA) block followed by a dense for final output[2]. Hyperparameters like learning rate and dropout were also grid-searched, while the number of attention heads was searched among $[1, 2, 3]$.
- **Vanilla GCN (GCN) [29]**: The vanilla GCN models without consideration of multi-media information. Instead of utilizing the latent semantic correlation among the video elicitations depicted in 4.2 to build the graph, here we directly calculate the Pearson correlation between any two nodes of a subject's physiological responses. Those samples larger than zero would be thought to as likely correlated physiological responses and are connected in the edge matrix $\mathcal{E}$. This method is a GCN baseline without the integration of multi-media content.
- **Media-Guided Attentive GCN (GCN-[A/V/AV])**: The multi-media content-aware GCN illustrated in section 4.2. Note that the essential difference when comparing with our previously proposed method [29] is that in the previous approach, there is a strict requirement to obtain the intended (-*int*) emotional annotation beforehand to build the graphs for better personality recognition. While in this paper, we relax this requirement in our newly designed model, i.e., the graph formation is fully data-driven and could lead to more explainable results when analyzing the graphical binding for each individual. Several variations of our model will be evaluated: $\boldsymbol{A}$: using the acoustic information $\mathcal{E}^{Aideo}$ for graph building; $\boldsymbol{V}$: utilizing the semantic visual cues $\mathcal{E}^{Video}$ for graph building; $\boldsymbol{AV}$: For each subject, both acoustic and visual graph would be jointly integrated into the network for joint modeling. Specifically, each graph would pass through separated GCN blocks $\boldsymbol{G}$ and attention block $\boldsymbol{A}$. Then the final graph-level embedding would be concatenated into a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing

TABLE 5: The personality recognition results using the metric of UAR and F1. The † mark is the highest results which statistically improved(Student's t-test, $p < 0.05$) against the **DNN-C** and **DNN-V** baseline within a modality, while the boldfaces are the results that both statistically improved and also the cross-modality global highest.

Amigos

		SVM-C [21]		**SVM-V** [80]		**DNN-C**		**DNN-V**		**AMIL** [81]		**SET** [82]		**GCN** [29]		**GCN-A**		**GCN-V**		**GCN-AV**	
		UAR	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*
ECG	Agr	0.500	0.431	0.500	0.479	0.573	0.566	0.552	0.560	0.639	0.648	0.653†	0.659	0.625	0.628	0.587	0.598	0.650	0.661†	0.643	0.652
	Con	0.476	0.398	0.500	0.484	0.659	0.635	0.560	0.543	0.606	0.618	0.566	0.559	0.590	0.602	0.610	0.622	**0.679**†	0.685†	0.629	0.637
	Cre	0.500	0.486	0.500	0.489	0.586	0.602	0.586	0.608	0.666†	0.686†	0.576	0.602	0.597	0.628	0.647	0.682	0.636	0.667	0.623	0.653
	Emo	0.345	0.315	0.447	0.472	0.544	0.540	0.618	0.618	0.626	0.623	0.587	0.583	0.653	0.651	0.634	0.633	**0.734**†	**0.734**†	0.689	0.686
	Ext	0.578	0.613	0.509	0.543	0.611	0.613	0.549	0.577	0.622	0.645	0.568	0.600	0.622	0.648	0.632	0.656	0.699†	0.731†	0.623	0.650
EDA	Agr	0.455	0.399	0.500	0.425	0.609	0.621	0.590	0.575	0.647	0.647	0.651	0.658	0.597	0.647	**0.681**†	**0.696**†	0.628	0.634	0.634	0.640
	Con	0.580	0.582	0.500	0.412	0.578	0.578	0.585	0.594	0.612	0.612	0.564	0.568	0.659†	**0.706**†	0.652	0.656	0.631	0.642	0.634	0.646
	Cre	0.586	0.624	0.500	0.489	0.638	0.667	0.552	0.576	0.632	0.632	0.631	0.658	0.642	0.700	0.662	0.694	**0.742**†	**0.776**†	0.688	0.720
	Emo	0.526	0.525	0.395	0.428	0.613	0.611	0.595	0.593	0.642†	0.642	0.587	0.582	0.615	0.652†	0.626	0.623	0.632	0.627	0.637	0.633
	Ext	0.500	0.498	0.500	0.489	0.657	0.671	0.546	0.568	0.660	0.660	0.566	0.592	0.553	0.626	0.691	0.719	0.728	0.744	**0.743**†	**0.760**†
ECG+EDA	Agr	0.500	0.431	0.500	0.425	0.636	0.634	0.523	0.553	0.634	0.636	0.584	0.597	0.645	0.654	0.641	0.645	0.647	0.649	0.657†	0.658†
	Con	0.563	0.561	0.500	0.425	0.539	0.553	0.528	0.568	0.624	0.629	0.564	0.572	0.578	0.582	0.648	0.655	0.629	0.640	0.652†	0.660†
	Cre	0.550	0.569	0.500	0.489	0.605	0.613	0.524	0.564	0.657	0.679	0.626	0.644	0.640	0.691	0.680†	0.707†	0.656	0.683	0.658	0.682
	Emo	0.346	0.525	0.395	0.394	0.546	0.549	0.524	0.606	0.655	0.651	0.611	0.607	0.605	0.608	0.624	0.622	0.703†	0.701†	0.676	0.673
	Ext	0.500	0.503	0.500	0.489	0.603	0.608	0.529	0.570	0.652	0.675	0.616	0.647	0.607	0.667	0.612	0.645	0.696†	0.723†	0.673	0.688

Ascertain

		SVM-C [21]		**SVM-V** [80]		**DNN-C**		**DNN-V**		**AMIL** [81]		**SET** [82]		**GCN** [29]		**GCN-A**		**GCN-V**		**GCN-AV**	
		UAR	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*
ECG	Agr	0.500	0.416	0.500	0.310	0.543	0.547	0.534	0.525	0.613	0.611	0.528	0.526	0.630	0.629	0.619	0.618	0.645	0.640	0.646†	0.644†
	Con	0.428	0.427	0.498	0.477	0.579	0.579	0.537	0.507	0.605	0.603	0.552	0.547	0.596	0.594	0.627†	0.628†	0.614	0.613	0.626	0.625
	Cre	0.500	0.465	0.525	0.475	0.557	0.583	0.535	0.498	0.612	0.614	0.619	0.627	0.591	0.597	0.627	0.635	0.634	0.643	0.636†	**0.645**†
	Emo	0.488	0.487	0.343	0.574	0.613	0.611	0.622	0.590	0.595	0.594	0.553	0.551	0.678	0.677	0.662	0.660	0.691†	0.690†	0.669	0.667
	Ext	0.386	0.383	0.414	0.413	0.531	0.527	0.491	0.484	0.598	0.596	0.607	0.605	0.610	0.608	0.634	0.631	0.617	0.615	0.666†	0.665†
EDA	Agr	0.481	0.482	0.465	0.466	0.561	0.561	0.528	0.526	0.618	0.617	0.650	0.650	0.596	0.592	0.751	0.749	0.613	0.612	**0.774**†	**0.774**†
	Con	0.542	0.543	0.544	0.503	0.622	0.624	0.576	0.540	0.623	0.624	0.568	0.566	0.588	0.586	0.639	0.638	0.642†	**0.642**†	0.634	0.629
	Cre	0.500	0.413	0.500	0.413	0.592	0.597	0.524	0.471	0.621	0.624	0.602	0.610	0.567	0.577	**0.642**†	**0.645**†	0.615	0.622	0.637	0.644
	Emo	0.534	0.534	0.431	0.431	0.612	0.610	0.519	0.514	0.636	0.634	0.572	0.570	0.674	0.670	0.667	0.664	**0.716**†	**0.712**†	0.709	0.707
	Ext	0.345	0.344	0.517	0.517	0.547	0.545	0.534	0.507	0.629	0.628	0.547	0.540	0.603	0.600	0.643	0.641	0.648†	0.645†	0.629	0.626
ECG+EDA	Agr	0.548	0.501	0.412	0.412	0.615	0.625	0.496	0.513	0.617	0.616	0.625	0.625	0.637	0.635	0.722†	0.717†	0.626	0.635	0.713	0.711
	Con	0.508	0.506	0.505	0.425	0.614	0.614	0.509	0.540	0.612	0.613	0.577	0.576	0.595	0.594	**0.643**†	**0.642**†	0.612	0.619	0.621	0.622
	Cre	0.500	0.465	0.500	0.413	0.600	0.614	0.508	0.504	0.610	0.616	0.651†	0.659†	0.589	0.594	0.630	0.636	0.604	0.625	0.607	0.614
	Emo	0.405	0.394	0.431	0.438	0.514	0.514	0.504	0.547	0.614	0.611	0.574	0.566	0.667	0.666	0.624	0.619	0.707†	0.705†	0.652	0.648
	Ext	0.368	0.366	0.414	0.411	0.579	0.579	0.497	0.482	0.593	0.591	0.557	0.554	0.600	0.598	0.669	0.666	0.617	0.644	**0.681**†	**0.680**†

single vector to serve as the fusion from both acoustic and visual perspective before feeding into $\boldsymbol{P}$ for personality prediction.

6 Personality Recognition Results

Table 5 summarizes our personality recognition results. Our proposed Media-Guided Attentive GCN reaches the best UAR results under different settings for all personality traits recognition on both datasets. More precisely, our best model reaches a relative maximum improvement of 4.7%, 0.2%, 10.4%, 12.1%, and 8.6% on **Am** and 15.9%, 2.1%, 4.2%, 10.3%, 10.2% on **As** in *Agr*, *Con*, *Cre*, *Emo* and *Ext* respectively over the naive *DNN-C* approach. Several notable observations are summarized below.

6.1 Baseline VS GCN

Firstly, we notice that the simple SVM method hardly predicts the personality traits either through direct concatenation of all feature dimensions (*SVM-C*) or by using the majority voting scheme (*SVM-V*). When comparing the results obtained using the naive DNN approaches, we could conclude that the non-linear neural learning method is much more effective in modeling the latent complexity of the personality patterns as reflected in the physiological signals. However, the recognition rates obtained using *DNN* models are not satisfying either. We hypothesize that directly concatenating physiological features (*DNN-C*) of a subject across all emotion stimuli would lead to an extremely large feature dimension creating issues of overfitting, while the naive majority voting method (*DNN-V*) underestimate the variable importances that different video simulations have on physiological responses for personality recognition tasks.

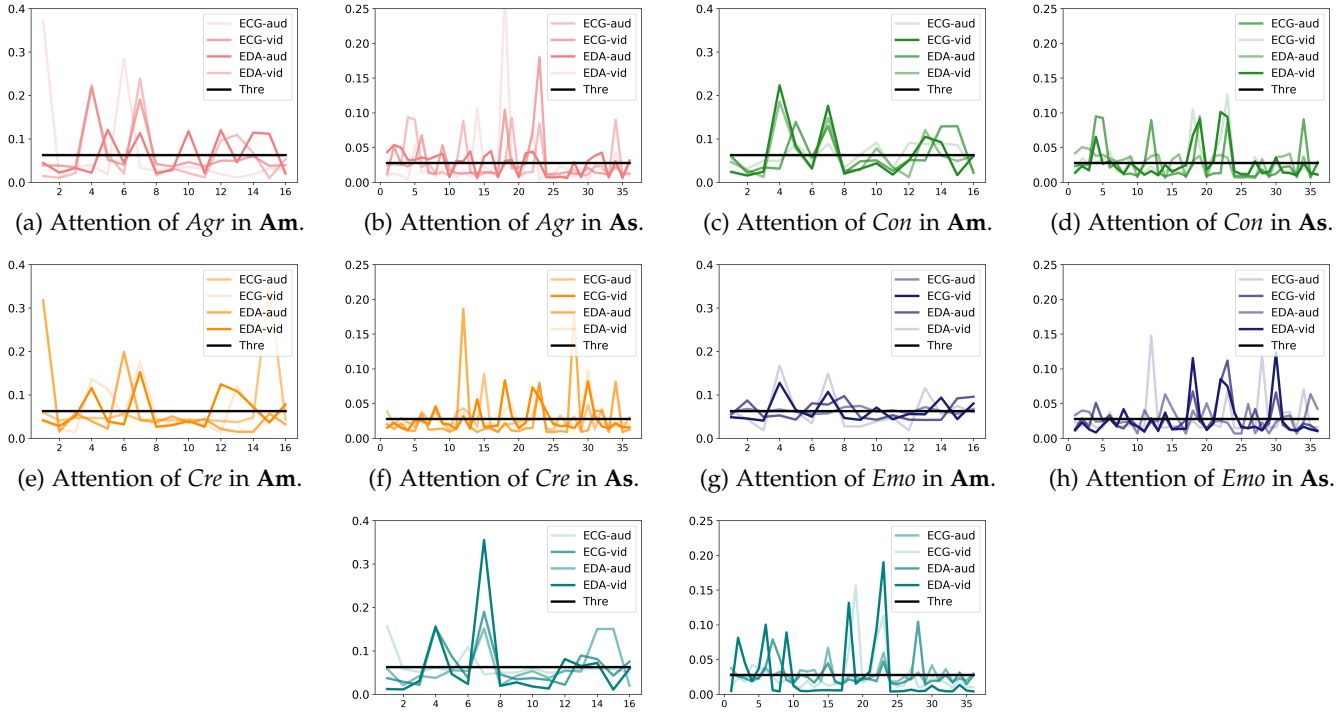
To handle these issues, we integrate the self-attention mechanism into the model. More precisely, we observe significant improvement on both datasets using either *AMIL*, *SET*, or *GCN* architecture. Furthermore, we notice that there is a larger gain for *AMIL* in the dataset **Am**, while a better recognition is obtained using *GCN* in dataset **As**. We think this result may come from the difference in sample sizes between the two corpora. Since the *GCN* method additionally consider the structural information (edges of graphs), this explicit modeling of inter-responses often leads to better performance when learning with larger datasets in contrast to the relatively simple *AMIL* method.

On the other hand, surprisingly the *SET* model, which we regard as the STOA model on multiple instances aggregation, result in unsatisfying recognition in most of the dimensions. Among all the non-multi-media-regularized models, it only achieves the best recognition on *Agr* with specific modalities (ECG,EDA in **Am** while EDA in **As**). We hypothesize that this result is due to the overly-powered architecture of the *SET* model (complex blocks of the MHA models). In other words, the extremely high degree of freedom could somehow let the model learn an enhanced recognition, but more often it would lead to trivial recognition especially on the small sample tasks in our scenario.

Finally, our proposed content attentive GCN method

2. https://github.com/juho-lee/set_transformer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing



(a) Attention of *Agr* in **Am**. (b) Attention of *Agr* in **As**. (c) Attention of *Con* in **Am**. (d) Attention of *Con* in **As**.

(e) Attention of *Cre* in **Am**. (f) Attention of *Cre* in **As**. (g) Attention of *Emo* in **Am**. (h) Attention of *Emo* in **As**.

(i) Attention of *Ext* in **Am**. (j) Attention of *Ext* in **As**.

Fig. 3: Line plot of the attention of all models. Note that the heavier the color of the line refers to higher UAR in Table 5, while the horizontal black line (Thre) is defined as $1/N_v$ which state as the average-level of attention received if no audio/video inducer is explicitly important arousing the personality enriched physiologies.

reaches the best recognition rates under different stimulation settings for each personality dimension. During the graph binding step, we manually force each node to only bind edges among nodes from a similar intended emotional stimulation. We believe that this acts as a hard constraint as if we force our model to focus on learning subtle structural information of physiology under similar affective stimulation, and this fine-grained representation improves our personality recognition results. More detailed comparisons would be elaborated in the following section.

6.2 Personality and Modality Specific Results

We are also curious about under *WHICH* modality and under *WHAT* multi-media stimuli would most effectively reveal the personality traits by examining an individual's physiology. In this section, we mainly focus on the comparison of GCN results under different graph binding methods using either Audio or Video in ECG or EDA modality. To begin with, we first focus on the *Agr*. We could immediately notice that the *Audio* component of the emotional stimuli is a key stimulus while EDA is an important indicator in recognizing *Agreeableness*. This setting consistently reaches the highest performance of 68.1% and 75.1% on both datasets. Besides, the fusion of audio and video further improves the recognition in **As**, where multi-modality integration benefits the prediction with a larger sample size dataset. Second, for the dimension of *Conscientiousness*, we regard this as a more challenging personality dimension to recognize from physiology in which there seems to be no specific modality or preferential elicitation. We also observe that we reach the lowest improvement in this dimension. In *Creativeness*, we see from the results that EDA acts as an important measure in revealing how eager a person learns new things and enjoys new experiences. It is interesting to observe that *Visual* cues in **Am** helps in revealing more personality variability while *Audio* elicitation triggered more differences in **As**.

In *Emotion Stability*, generally, the *Visually*-aroused physiology consistently achieves higher recognition in contrast to the *Audio*-based elicitation. Besides, it is interesting to observe that under visual stimuli, although EDA reaches the highest recognition 71.6% in **As**, there is a drop to 63.2% in **Am**. In contrast, the ECG consistently reaches competitive results of 73.4% and 69.1% UAR, which may infer that the cardiovascular-related signal serves as a more robust indicator of evaluating a subject's emotional stableness. Finally, in *Extraversion*, we observe that the EDA outperforms ECG. Besides, similar to *Cre*, *Visual*-component of the emotional stimuli is a more important elicitation in triggering this trait, though fusing both acoustic and visual information could further boost the performances.

In summary, our proposed Acoustic-Visual Guided Attentive GCN reaches state-of-the-art recognition on most of the personality recognition tasks in both datasets. In *Agr*, the combination of *Audio* elicitation and EDA signal could expose one's trait polarity more effectively. As for the *Cre* and *Ext*, although no specific types of stimuli are dominated, the EDA signal consistently serves as a good indicator revealing personality traits. Finally, we also show that an individual's mood stability can be robustly recognized by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3090040, IEEE Transactions on Affective Computing

TABLE 6: The most important video elicitations for each setting.

	Amigos					**Ascertain**				
	ECG		**EDA**		**>=3**	**ECG**		**EDA**		**>=3**
	Aud	Vid	Aud	Vid		Aud	Vid	Aud	Vid	
Agr	1, 6	4, 7	5, 7, 10, 12, 14	4, 7, 12, 13, 14	7	4, 5, 12, 19, 34	6, 9, 18, 22, 23	2, 3, 13, 16, 22	5, 14, 18, 22, 34	22
Con	5, 7, 10, 12, 14	4, 5, 7, 13, 14	5, 7, 14, 15	4, 5, 7, 10, 13	5, 7, 14	4, 5, 12, 19, 34	6, 18, 19, 22, 23	1, 2, 3, 16, 32	4, 18, 19, 22, 23	19
Cre	15	4, 5, 7, 13, 16	1, 6	4, 7, 12, 13, 16		12, 15, 23, 30, 32	14, 18, 22, 23, 30	12, 23, 28, 31, 34	12, 18, 19, 23, 30	23, 12, 30
Emo	2, 8, 9, 13, 16	4, 7, 10, 14, 16	2, 6, 8, 15, 16	1, 4, 7, 13, 15	16	12, 23, 28, 32, 34	4, 18, 22, 23, 30	14, 15, 20, 26, 35	8, 18, 22, 23, 30	23
Ext	1, 6	4, 5, 7, 13, 14	7, 14, 15	4, 7, 12, 13, 14	7, 14	1, 6, 15, 23, 30	4, 18, 19, 22, 23	7, 8, 15, 23, 28	6, 9, 18, 22, 23	23
>=3		4, 5, 7, 13, 14	7, 14, 15	4, 7, 12, 13		12, 23, 34	18, 22, 23		18, 22, 23	

measuring their ECG signal when exposing affective visual stimuli.

7 Analysis and Discussion

In this section, we provide analyses in understanding the potential modulation that multimedia content has on physiological personality recognition. We gather our model's supervised-learned self-attention weights $\boldsymbol{\alpha}$ for each subject then average them into video-level statistics shown in Fig.3. This attention (range from 0~1) would indicate that the important videos which could largely arouse the individual's physiology that reveal his/her personality trait. Here we focus on every single modality aroused by either the *Audio* or the *Visual* component of the emotion elicitations. Several statistical analyses would be conducted in the following.

7.1 Key Elicitations

Firstly, we are curious about whether there exist specific videos that consistently act as the most important (key) elicitation for different personality dimensions. Hence, we analyze the attention distribution in Fig.3 and select the most important video IDs using these two criteria:

- The attention weight that is larger than the threshold $1/N_v$
- The attention weight that is in the top 5 of all video elicitations

The results are shown in Table 6. We could immediately notice that the video "*House of Flying Daggers*" (ID is 4 in **Am** and 18 in **As** respectively) are key visual elicitation, which is consecutively selected as important stimuli using the attention mechanism in both datasets. Moreover, this video's visual content is considered important in predicting personality traits across *ALL* dimensions. Another important video is "*My Bodyguard*" (ID is 7 in **Am** and 23 in **As**). Again, the network learns to select this video as key visual elicitation in almost all personality dimensions (the only exception is in dataset **As**'s *Agr* dimension). Moreover, if we take a closer look into the dimension of *Ext*, we can conclude that not only visual but audio cues are also crucial for predicting this trait in both datasets. It is interesting to observe that there indeed exist interrelationships between personality traits and affective multimedia, and all these samples show support of the idea that individuals' personalities would be reflected in their physiology with a variable degree that is conditioned on the auditorial/visual elicitation they receive.

7.2 Emotional Multimedia

Since all the multi-media elicitations in both datasets have been carefully chosen from previous literature as affective elicitations, we would also like to investigate whether there exists a latent modulation between the emotional content, subjective emotion ratings, and personality traits. We first average all the emotion ratings (*-SB*) to serve as the indication of the level of emotional content for each video. Then we perform the Pearson correlation to examine the correlation between the attention importances $\boldsymbol{\alpha}$ and the level of emotional content in video elicitations (Note that to ensure the fidelity of this analysis, we only target the modalities under *Audio* or *Video* elicitation which reaches the highest UAR in Table 5). Interestingly, we find no significant correlation between traits of *Agr, Con, Cre, Emo* toward emotion labels for either ECG or EDA signals. However, we find that in both **Am** and **As** dataset, the attention weights for modeling *Extraversion*, i.e., learned from ECG signals using visual semantic guided graph, are statistically correlated with the feeling of "Disgust" ($r = 0.64909, p = 0.0065153$) and "Familiarity"($r = -0.3599, p = 0.031084$) respectively. In other words, it implies that the more "**Unfamiliar**" or "**Uncomfortable**" the visual elicitations are delivered, the more discriminative cardiovascular variance would be aroused, in which this variance would reflect an individual's degree of outgoing and sociative.

7.3 Important Physiological Indicators

We also investigate the key physiological descriptors that are indicative of an individual's personality trait. We first participant's physiological features according to the videos that are learned with high attention (both auditorial and visual perspective) in Table 6. Then these physiological descriptors are gathered from all individuals in each dataset, then we perform the Students t-test to search for the important physiological features. Finally, we only retain those descriptors that are consistently selected in both **Am** and **As** datasets to indicate their robust cross-corpus evidence. The final selected features are listed in Table 7

We first notice that no feature is selected for ECG modality in *Aggreableness*. This is consistent with our previous conclusion that ECG is not an informative modality for modeling an individual's social harmony characteristic. Then we observe that "Shannon entropy" related descriptors are consistently picked among the rest of the personality dimensions. Shannon entropy is commonly calculated in the domain of information theory, and it has been verified for its discriminative power on modeling human emotion

TABLE 7: The important physiological descriptors that are stimulated under the *key* videos are shown in Table 6. Note that all the descriptors listed are statistically correlated ($p - value \leq 0.05$) with a particular personality dimension in both Am and As datasets.

	ECG	EDA
Agr		EDA_Tonic_99_per, EDA_Tonic_min_pos
Con	CVSD, Entropy_SVD, Fisher_Inf, Shannon_h, VHF, madNN, mcvNN	EDA_Phasic_LF, EDA_Tonic_99minus1_per, EDA_Tonic_HF, EDA_Tonic_LF, EDA_Tonic_ULF, EDA_Tonic_VLF, EDA_Tonic_kurtosis, EDA_Tonic_max
Cre	CCSQ_min_pos, Shannon_h, VHF	EDA_Tonic_VLF SCR_Peaks_Amp_kurtosis, SCR_Peaks_Amp_skewneww
Emo	CCSQ_std, DFA_1, HF/P, HFn, LFn, Shannon, Triang	EDA_Phasic_LF, EDA_Phasic_mean, EDA_Tonic_HF, EDA_Tonic_LF, EDA_Tonic_VLF, EDA_Tonic_mean, EDA_Tonic_up_quar, SCR_Peaks_Amp_99minus1_per, SCR_Peaks_Amp_min_pos
Ext	CCSQ_low_quar, Correlation_Dimension, Shannon, pNN50	EDA_Tonic_ULF, SCR_Peaks_Amp_99_per

reactions [85], [86]. We also notice that more frequency-domain Heart Rate Variabilities (HRVs) are linked toward *Emotion Stabliltiy*, which is consistent with previous studies in showing the relationship of this feature with emotional intelligence/stress-related indices [87], [88]. Similar to Zohar's study [89], systematic and significant associations between personality traits and HRV's frequency components were found. Yet unlike our study, they conclude that the LF/HF(low frequency to high-frequency ratio) was related to the personal traits, which could originate from the substantially different experiment design in which they didn't have any types of emotional stimuli in their study.

As for the EDA modality, we immediately see that in comparison to ECG, more descriptors are selected across all personalities, which is consistent with our recognition results that EDA is a relatively more robust descriptor. We then further observe that in general, "tonic" related descriptors are more important than "phasic" features. According to the definition, the tonic component reflects the *background* characteristics, which models the underlying slowly changing baseline that is caused by the drifting skin conductance level (SCL) and other unconscious activities. This usually is considered as static emotion changes that indicate a similarity in the static nature of personality construct [90]. Our findings also show consistent with previous studies. Cumulated researches have associated the reliability (the phasic component) between EDA and *Agreeableness-Antagonism (Agr)*. For example, in Crider's study [91], specific phasic components were assessed over two sessions of iterated auditory stimulation in a sample of 22 male undergraduates. The two reliability measures have shown positively correlated with Minnesota *California Psychological Inventor (CPI)* and inversely correlated with *Multiphasic Personality Inventory(MMPI)* scales, where the higher CPI indicates the higher responsibility and self-control while higher MMPI generally linked to a poorly socialized, impulsive disposition indicative of psychopathy. The phasic component has also been connected to *Conscientiousness* as over-controlled and under-controlled personality types in Blocks study [92]. We also see that the Skin Conductance Reactivity(SCR) has also been indicated as a key descriptor distinguishing personality, which was also reported in Norris's study [93], that they found that individuals lower in Emotion Stability exhibited greater skin conductance reactivity to the emotional picture stimuli.

All these studies have suggested that individual differences could be elicited by external stimuli and monitored through measured physiological responses. In summary, through additional multimedia constraints, our model automatically learns that the combination of latent slowly changing plus certain event-related measures could most properly model an individual's personality traits.

8 Conclusion and Future Work

In this work, we aim to find a data aggregation strategy to aggregate multiple induced physiological observations into a single subject's personality. To the best of our knowledge, this is also one of the first frameworks with comprehensive analyses that learn to embed the source multimedia stimuli into a graph structural network for automatic personality recognition. To extract the multimedia information, we propose two distinct strategies based on either the data-driven *Visual-Semantic* cues or the expert-defined *Acoustic-Affetive* features. These two types of multimedia information provide the network with a prior view on the source stimuli in guiding the learning of the physiological responses with an aid of a graphical structure. A Graph Convolutional Network is applied followed by a Self-Attention mechanism to complete our media-guided attentive graphical network for personality modeling. We evaluate the proposed methods on two large benchmark databases Ascertain [20] and Amigos [21]. Our experimental results demonstrate that the method achieves state-of-the-art on these two datasets.

Three further analyses are conducted to understand the role of affective media in inducing physiological responses reflecting personality traits. We first examine the attention weights automatically learned from the model to identify the potential video preference for personality elicitation. The cross-database comparison shows that both "*House of Flying Daggers*" and "*My Bodyguard*" are two of the key personality inducers among all personalities. We further perform statistical analysis between learned attention weights and subjective emotional ratings of the media stimuli. Correlational analysis result suggests that more "*unfamiliar*" nor "*uncomfortable*" the visual elicitations are delivered, the more reflective ECG signal would be aroused on assessing the level of *Extraversion*. Finally, through examining each physiological descriptor, we conclude that "Shannon Entropy" related features in ECG, as well as "tonic" component which modeling slowly baseline changes in EDA are robust descriptors for personality identification.

There are multiple future directions. Firstly, we will investigate additional technical approaches for joint modeling the temporal variation of physiology during media stimuli. Although we already extracted statistical descriptors, which contain certain temporal information, employing time series-based modeling could potentially further improve the recognition. Second, larger and diverse databases

should be collected. Through this research, we have demonstrated that multimedia content could act as a personality trigger, yet most of the current researches merely focusing on a short and limited set of emotional videos as stimuli. An immediate next step is to gather and inspect a wider and diverse range of multimedia sources of different domains (such as movies, social media, and gaming), which can be used as emotion elicitation triggers. Automatically video tagging algorithms should also be incorporated to analyze the delivered multimedia contents(characters, actions, environment context, etc...) that would help us understand the neuropsychological working function of the emotional videos to the induced physiologies. Better understand exactly what components of multimedia content would trigger physiological responses while providing evidence of a subject's personality would help in advancing a variety of human-centered multimedia applications [94], [95].

REFERENCES

[1] J. Chen, E. Haber, R. Kang, G. Hsieh, and J. Mahmud, "Making use of derived personality: The case of social media ad targeting," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[2] A. Marcus, *Design, User Experience, and Usability: User Experience Design for Diverse Interaction Platforms and Environments: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings*. Springer, 2014, vol. 8518.

[3] K. L. Norman, *Cyberpsychology: An introduction to human-computer interaction*. Cambridge university press, 2017.

[4] S. M. Sarsam and H. Al-Samarraie, "Towards incorporating personality into the design of an interface: a method for facilitating users interaction with the display," *User Modeling and User-Adapted Interaction*, vol. 28, no. 1, pp. 75–96, 2018.

[5] B. Ferwerda and M. Schedl, "Personality-based user modeling for music recommender systems," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 254–257.

[6] M. Onori, A. Micarelli, and G. Sansonetti, "A comparative analysis of personality-based music recommender systems." in *Empire@ RecSys*, 2016, pp. 55–59.

[7] A. Paudel, B. R. Bajracharya, M. Ghimire, N. Bhattarai, and D. S. Baral, "Using personality traits information from social media for music recommendation," in *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*. IEEE, 2018, pp. 116–121.

[8] J. Golbeck and E. Norris, "Personality, movie preferences, and recommendations," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 1414–1415.

[9] I. Cantador, I. Fernández-Tobías, and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," in *CEUR workshop proceedings*. Shlomo Berkovsky, 2013.

[10] C. Bologna, A. C. De Rosa, A. De Vivo, M. Gaeta, G. Sansonetti, and V. Viserta, "Personality-based recommendation in e-commerce." in *UMAP Workshops*. Citeseer, 2013.

[11] M. Tkalcic and L. Chen, "Personality and recommender systems," in *Recommender systems handbook*. Springer, 2015, pp. 715–739.

[12] I. Fernández-Tobías, M. Braunhofer, M. Elahi, F. Ricci, and I. Cantador, "Alleviating the new user problem in collaborative filtering by exploiting personality information," *User Modeling and User-Adapted Interaction*, vol. 26, no. 2-3, pp. 221–255, 2016.

[13] T. Halevi, J. Lewis, and N. D. Memon, "Phishing, personality traits and facebook," *CoRR*, vol. abs/1301.7643, 2013. [Online]. Available: http://arxiv.org/abs/1301.7643

[14] T. Halevi, N. Memon, and O. Nov, "Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks," *Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks (January 2, 2015)*, 2015.

[15] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen, "Personalized persuasion: Tailoring persuasive appeals to recipients personality traits," *Psychological science*, vol. 23, no. 6, pp. 578–581, 2012.

[16] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[17] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," *User modeling and user-adapted interaction*, vol. 26, no. 2-3, pp. 109–142, 2016.

[18] E. Kanjo, L. Al-Husain, and A. Chamberlain, "Emotions in context: examining pervasive affective sensing systems, applications, and analyses," *Personal and Ubiquitous Computing*, vol. 19, no. 7, pp. 1197–1212, 2015.

[19] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[20] 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*, no. 2, pp. 147–160, 2018.

[21] 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.

[22] J. Pinto, A. Fred, and H. P. da Silva, "Biosignal-based multimodal emotion recognition in a valence-arousal affective framework applied to immersive video visualization," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3577–3583.

[23] A. Hall, "Audience personality and the selection of media and media genres," *Media Psychology*, vol. 7, no. 4, pp. 377–398, 2005.

[24] T. Chamorro-Premuzic and A. Furnham, "Personality and music: Can traits explain how people use music in everyday life?" *British Journal of Psychology*, vol. 98, no. 2, pp. 175–185, 2007.

[25] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 213–222.

[26] B. Colombo, L. Grati, and C. Di Nuzzo, "The role of individual creativity levels in the cognitive and emotive evaluation of complex multimedia stimuli. a study on behavioral data and psychophysiological indexes," 2013.

[27] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, pp. 1–27, 2019.

[28] I. M. A. Agastya, D. O. D. Handayani, and T. Mantoro, "A systematic literature review of deep learning algorithms for personality trait recognition," in *2019 5th International Conference on Computing Engineering and Design (ICCED)*. IEEE, 2019, pp. 1–6.

[29] H.-C. Yang and C.-C. Lee, "A siamese content-attentive graph convolutional network for personality recognition using physiology," in *ICASSP 2020-200 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

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

[31] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[32] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, and N. Howard, "Common sense knowledge based personality recognition from text," in *Mexican International Conference on Artificial Intelligence*. Springer, 2013, pp. 484–496.

[33] A. C. E. Lima and L. N. De Castro, "A multi-label, semi-supervised classification approach applied to personality prediction in social media," *Neural Networks*, vol. 58, pp. 122–130, 2014.

[34] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2015, pp. 170–174.

[35] S. Kleanthous, C. Herodotou, G. Samaras, and P. Germanakos, "Detecting personality traces in users social activity," in *International conference on social computing and social media*. Springer, 2016, pp. 287–297.

[36] G. An, S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine, and J. Hirschberg, "Automatically classifying self-rated personality scores from speech." in *INTERSPEECH*, 2016, pp. 1412–1416.

[37] G. An, S. I. Levitan, J. Hirschberg, and R. Levitan, "Deep personality recognition for deception detection." in *Interspeech*, 2018, pp. 421–425.

[38] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1101–1104.

[39] T. Fernando *et al.*, "Persons personality traits recognition using machine learning algorithms and image processing techniques," *Advances in Computer Science: an International Journal*, vol. 5, no. 1, pp. 40–44, 2016.

[40] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. E. Moghaddam, and L. Ungar, "Analyzing personality through social media profile picture choice," in *Tenth international AAAI conference on web and social media*, 2016.

[41] 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.

[42] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "Tensorflow-based automatic personality recognition used in asynchronous video interviews," *IEEE Access*, vol. 7, pp. 61 018–61 023, 2019.

[43] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," vol. 1, pp. 1–8, 2015.

[44] J. Wache, R. Subramanian, M. K. Abadi, R.-L. Vieriu, N. Sebe, and S. Winkler, "Implicit user-centric personality recognition based on physiological responses to emotional videos," pp. 239–246, 2015.

[45] J. A. Miranda-Correa and I. Patras, "A multi-task cascaded network for prediction of affect, personality, mood and social context using eeg signals," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 373–380.

[46] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km

[47] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019. [Online]. Available: http://arxiv.org/abs/1901.00596

[48] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *CoRR*, vol. abs/1907.07835, 2019. [Online]. Available: http://arxiv.org/abs/1907.07835

[49] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, 2018.

[50] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. F. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 154–164. [Online]. Available: https://doi.org/10.18653/v1/D19-1015

[51] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "STEP: spatial temporal graph convolutional networks for emotion perception from gaits," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 1342–1350. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/5490

[52] J. Kim, Y. Hong, G. Chen, W. Lin, P.-T. Yap, and D. Shen, "Graph-based deep learning for prediction of longitudinal infant diffusion mri data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 133–141.

[53] D. Yao, M. Liu, M. Wang, C. Lian, J. Wei, L. Sun, J. Sui, and D. Shen, "Triplet graph convolutional network for multi-scale analysis of functional connectivity using functional mri," in *International Workshop on Graph Learning in Medical Imaging*. Springer, 2019, pp. 70–78.

[54] Y. Hong, G. Chen, P.-T. Yap, and D. Shen, "Reconstructing high-quality diffusion mri data from orthogonal slice-undersampled data using graph convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 529–537.

[55] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, "Spatio-temporal graph convolution for skeleton based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[56] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[57] R. Liu, C. Xu, T. Zhang, W. Zhao, Z. Cui, and J. Yang, "Si-gcn: Structure-induced graph convolution network for skeleton-based action recognition," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[58] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 5, pp. 1–12, 2019.

[59] F. Manessi, A. Rozza, and M. Manzo, "Dynamic graph convolutional networks," *Pattern Recognition*, vol. 97, p. 107000, 2020.

[60] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[61] M. Perugini and L. DI BLAS, "The big five marker scales (bfms) and the italian ab5c taxonomy: Analyses from an etic–emic perspective," 2002.

[62] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[63] R. D. Lane, K. McRae, E. M. Reiman, K. Chen, G. L. Ahern, and J. F. Thayer, "Neural correlates of heart rate variability during emotion," *Neuroimage*, vol. 44, no. 1, pp. 213–222, 2009.

[64] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[65] G. Berna, L. Ott, and J.-L. Nandrino, "Effects of emotion regulation difficulties on the tonic and phasic cardiac autonomic response," *PloS one*, vol. 9, no. 7, p. e102971, 2014.

[66] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[67] D. Makowski, "Neurokit: A python toolbox for statistics and neurophysiological signal processing (eeg, eda, ecg, emg...)," *Memory and Cognition Lab'Day*, vol. 1, 2016.

[68] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[69] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: http://arxiv.org/abs/1705.06950

[70] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[71] Y. Jiale and Z. Ying, "Visualization method of sound effect retrieval based on umap," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1. IEEE, 2020, pp. 2216–2220.

[72] M. Bahri, B. Pfahringer, A. Bifet, and S. Maniu, "Efficient batch-incremental classification using umap for evolving data streams," in *International Symposium on Intelligent Data Analysis*. Springer, 2020, pp. 40–53.

[73] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[74] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[75] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, 2000.

[76] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[77] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[78] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, J. Su, X. Carreras, and K. Duh, Eds. The Association for Computational Linguistics, 2016, pp. 551–561. [Online]. Available: https://doi.org/10.18653/v1/d16-1053

[79] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Zhao, J. Li, A. J. Smola, and Z. Zhang, "Deep graph library: Towards efficient and scalable deep learning on graphs," *CoRR*, vol. abs/1909.01315, 2019. [Online]. Available: http://arxiv.org/abs/1909.01315

[80] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling, "Eye movements during everyday behavior predict personality traits," *Frontiers in human neuroscience*, vol. 12, p. 105, 2018.

[81] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 2132–2141. [Online]. Available: http://proceedings.mlr.press/v80/ilse18a.html

[82] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3744–3753.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[85] Y. Xia, L. Yang, H. Shi, Y. Zhuang, and C. Liu, "Changes of permutation pattern entropy and ordinal pattern entropy during three emotion states: Natural, happiness and sadness," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[86] S.-H. Wang, H.-T. Li, E.-J. Chang, and A.-Y. A. Wu, "Entropy-assisted emotion recognition of valence and arousal using xgboost classifier," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2018, pp. 249–260.

[87] S. Laborde, A. Brüll, J. Weber, and L. S. Anders, "Trait emotional intelligence in sports: A protective role against stress through heart rate variability?" *Personality and Individual Differences*, vol. 51, no. 1, pp. 23–27, 2011.

[88] A. Arza, J. Garzón, A. Hemando, J. Aguiló, and R. Bailón, "Towards an objective measurement of emotional stress: Preliminary analysis based on heart rate variability," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 3331–3334.

[89] A. H. Zohar, C. R. Cloninger, R. McCraty *et al.*, "Personality and heart rate variability: exploring pathways from personality to cardiac coherence and health," *Open Journal of Social Sciences*, vol. 1, no. 06, p. 32, 2013.

[90] D. Yu and S. Sun, "A systematic exploration of deep neural networks for eda-based emotion recognition," *Information*, vol. 11, no. 4, p. 212, 2020.

[91] A. Crider and R. Lunn, "Electrodermal lability as a personality dimension." *Journal of Experimental Research in Personality*, 1971.

[92] J. Block, *Personality as an affect-processing system: Toward an integrative theory*. Psychology Press, 2002.

[93] C. J. Norris, J. T. Larsen, and J. T. Cacioppo, "Neuroticism is associated with larger and more prolonged electrodermal responses to emotionally evocative pictures," *Psychophysiology*, vol. 44, no. 5, pp. 823–826, 2007.

[94] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[95] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.

Hao-Chun,Yang (Student Member, IEEE) received the B.S. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016. He is currently working toward a Ph.D. degree with the Electrical Engineering Department, NTHU, Hsinchu Taiwan. His research interests are in affective multimedia, physiological signals, computational neuroscience, and context-aware machine learning. He was the recipient of ICASSP SPS Travel Grants, ACLCLP Scholarship, Adbertech AI Scholarship, and NTHU Presidents Scholarship. He is also a Student Member of the IEEE Signal Processing Society.

Chi-Chun Lee (M'13, S'20) is an Associate Professor at the Department of Electrical Engineering with joint appointment at the Institute of Communication Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective multimedia, health analytics, and behavior computing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), and a TPC member for APSIPA IVM and MLDA committee. He serves as an area chair for INTERSPEECH 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, sponsorship and special session chair for ISCSLP 2018, 2020, and a guest editor in Journal of Computer Speech and Language on special issue of Speech and Language Processing for Behavioral and Mental Health.

He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in INTERSPEECH 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in INTERSPEECH 2019. He is a coauthor on the best paper award/finalist in INTERSPEECH 2008, INTERSPEECH 2010, IEEE EMBC 2018, INTERSPEECH 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is an IEEE senior member and a ACM and ISCA member.