ALIS: Learning Affective Causality Behind Daily Activities From a Wearable Life-Log System

Byung Hyung Kim, Sungho Jo, and Sunghee Choi

***Abstract*—Human emotions and behaviors are reciprocal components that shape each other in everyday life. While the past research on each element has made use of various physiological sensors in many ways, their interactive relationship in the context of daily life has not yet been explored. In this work, we present a wearable affective life-log system (ALIS) that is robust as well as easy to use in daily life to accurately detect emotional changes and determine the cause-and-effect relationship between emotions and emotional situations in users' lives. The proposed system records how a user feels in certain situations during long-term activities using physiological sensors. Based on the long-term monitoring, the system analyzes how the contexts of the user's life affect his/her emotional changes and builds causal structures between emotions and observable behaviors in daily situations. Furthermore, we demonstrate that the proposed system enables us to build causal structures to find individual sources of mental relief suited to negative situations in school life.**

***Index Terms*—Affective causality, daily activities, EEG, emotion recognition, lifelog, physiological signals, wearable.**

I. Introduction

PEOPLE experience various emotions from a single event in different situations. For instance, today's coffee is not always the same as yesterday's coffee. The cup of coffee we drank today may not be as enjoyable as the cup of coffee we drank yesterday. While drinking coffee generally helps to reduce a person's stress, the stress-relieving effects of coffee may vary from day to day for many reasons. For a person who likes calm and quiet surroundings, a cup of coffee drunk today in a crowded coffee shop with distracting background noise is likely to be less enjoyable than a cup of coffee drunk yesterday in the quiet kitchen of one's own home. The lesson was stored in memory along with some affective residue associating the unhappy emotion with the regretted, less enjoyable coffee experience. Later, the affective residue became activated in a similar situation and led to a change in subsequent behavior. This instance shows that a person can have different emotional responses to the same life events in different circumstances.

Why and how does a person experience various emotions from a single event in different situations? Answering this question could improve human life in a variety of ways, such as by improving physical health. People with depression are more vulnerable to heart disease than are people with no history of depression [1]. Therefore, discovering life elements such as drinking coffee related to depression and offering guidance to avoid such elements could help individuals prone to depression to experience fewer depression-related issues. In response to this question, the recent research on recognizing the human affect has used a variety of physiological sensors in many ways [2], [3]. Using these sensors, features, such as heart rate (HR) variability, pulse oximetry, and galvanic skin response, have been used to capture emotional changes and thereby, help elucidate the etiology of mental health pathologies such as stress.

However, discovering the process of human affect and action in different situations in daily life has not yet been explored. In our study, understanding this process in daily life primarily refers to analyzing how affective residue shapes behavior in everyday life. The human affect may often be sufficient to guide the current behavior. All daily activities (habitual, familiar, or new) that involve emotional outcomes leave affective residue within the individual. The individual then anticipates possible emotional outcomes and behaves accordingly. The affective residue provides the impetus to support future behavior change [4]. Hence, the affective residue of prior emotional outcomes is likely to contribute to the process of human affect and action in daily life. To understand this process, it requires understanding the affect-elicitation mechanisms and their effect on emotional responses.

There are two main challenging problems in real-world environments as follows.

1) *Limited and Unbalanced Labeling Problem in Emotion Recognition (C.1):* First, obtaining accurate affect recognition is extremely challenging when physiological signals accompany unreliable labels. Most studies [5] have been limited to laboratory environments assessed by self-report tools [6] or specialized questionnaires. However, these methods only provide evidence of

Manuscript received September 30, 2020; revised April 7, 2021 and June 30, 2021; accepted August 13, 2021. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant (2017-0-00432), and in part by the National Research Foundation of Korea (NRF) Grant (2021R1C1C2012437) funded by the Korea Government (MSIT). This article was recommended by Associate Editor J. Han. *(Corresponding authors: Sungho Jo; Sunghee Choi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) in Human Subjects Research (KH2016-84).

Byung Hyung Kim is with the Department of Artificial Intelligence, Inha University, Incheon 22212, Republic of Korea.

Sungho Jo and Sunghee Choi are with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (e-mail: shjo@kaist.ac.kr; sunghee@kaist.ac.kr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3106638.

Digital Object Identifier 10.1109/TCYB.2021.3106638

the immediate affect from a single event and permit only limited understanding of affective dynamics. Furthermore, in real-world scenarios, only a few labels are available when humans rate their emotions in indoor environments using self-report tools such as self-assessment manikin (SAM) [6]. The limited and unbalanced class labels deteriorate the performance of classifiers in a broad range of emotion classification problems.

2) *Sparsity Problem in Affective Causality Identification (C.2):* It is unclear whether affective causality exists in real-world situations, during which a user encounters life events and behaves over extended temporal sequences. In general, it is challenging to identify causal direction and discover the influence of latent factors, as the human affect is usually influenced by various complex and subtle factors in daily life.

To solve these challenges, we introduce a wearable affective life-log system (ALIS), which helps to bridge the gap between the low-level physiological sensor representation and the high-level context-sensitive interpretation of affect. Notably, ALIS aims to determine affective causality by analyzing two-channel EEG and photoplethysmogram (PPG) signals together with visual life-content encountered in various affective situations, such as hanging out with friends and reading books.

Most EEG-based emotion recognition systems have extracted and selected EEG-based features through electrode selection based on neuroscientific assumptions [7]–[12]. Emotional states can be well differentiated by assessing frontal EEG asymmetry, and a multidimensional directed-information approach to causality between right and left hemispheres has revealed emotional lateralization in the frontal and temporal lobes [13]. Despite the low-spatial resolution of EEG, its very high temporal resolution, noninvasiveness, and mobility are valuable in real-world environments [14]–[17]. EEG-based emotion recognition systems have often shown improved results when different modalities were used [13], [18]–[20]. Among the many peripheral physiological signals, PPG, which measures blood volume, is widely used to compute HR. Although its accuracy is considered lower than that of electrocardiogram (ECG), due to its simplicity, PPG has been used to develop wearable biosensors for clinical applications, such as detecting mental stress in daily life [21]. HR, as well as HR variability (HRV), has been shown to be useful for emotion assessment [22]–[24].

Without loss of generality, we treat the affective causality problem as a combination of two-stage problems. The first stage is the affect recognition problem. Given physiological signals, the proposed ALIS is built upon our previous work on recognizing the human affect, namely, a deep physiological affect network (DPAN), whose outputs are discrete valence and arousal values on the affect dimension, underlying emotional lateralization and HRV [13]. Although DPAN has merit for discriminating physiological signals associated with different valence and arousal labels, nontrivial and challenging issues exist when the model is applied to demonstrate its efficacy in daily life (i.e., various noises, low signal-to-noise ratio (SNR) of physiological signals, intersubject and intrasubject variability, and usability). To alleviate these issues, ALIS comprises a subnetwork, which learns to reduce label noise and predict more accurate labels. The proposed physiological affect network learns affective dynamics with minimal supervision. Unlike previous works [5], [25] that focused on a specific personalized assessment for every event, our model captures affective states and continuously traces their changes, exploiting unlabeled and unbalanced real-world data through semisupervised learning.

The second stage is the affective causality learning problem. The problem is formulated as a graph to derive causality by analyzing affective changes and the user's relevant situations. The computation is based on conditional independence testing to detect the relationship with latent confounders underlying the two observational sequences. We present an asymmetric measure by which the causal relationship is identified between the affective content and human emotion in daily life. The model for the first time allows users to understand when, what, and how their surroundings affect them unconsciously in their daily life. Understanding the causal direction is essential to predict the consequence of any intervention from a group of observation samples, and is critical to many applications, including within biology and social science [26]–[29]. We note that causal learning is different from mainstream statistical learning methods in that the former aims to discover the data-generation mechanism instead of characterizing the joint distribution of the observed variables; this represents the most significant difference between causality and correlation. Discovering the emotional influence has been a focus for testing whether the affective correlation exists in real-world applications [27]–[30].

We examined the robustness of the proposed ALIS when applied to two datasets: 1) a public dataset and 2) a synthetic dataset, for the quantitative evaluation of affect recognition (C.1) and causality identification (C.2). By applying our proposed model to real-world scenarios, the results show that our approach is able to find meaningful causal connections between emotions and behaviors by tracking how affective residue shapes behavior, even in the presence of confounder variables that potentially affect human emotions and behaviors.

The remainder of this article is organized as follows. Section II presents the problem formulation of modeling affective causality in daily life and covers the preliminaries of our previous work and causal inference. In Section III, we present our system, describing the system design and framework. Sections IV and V evaluate our system on synthetic, public, and real datasets. Finally, we conclude this article with a discussion of future work in Section VI.

II. Preliminaries

A. Problem Statements

Our goal is to determine affective causality in daily life by analyzing emotional states and life contents encountered in various situations. We present the problem formulation of modeling the causal relation between life contents and

emotional changes in the proposed ALIS. Without loss of generality, we treat the problem as a combination of two-stage problems. The first stage is the emotion recognition problem. Given physiological signals, the problem is the identification of the correct emotional state as

$$\hat{y} = \text{argmax}_{y\in\mathcal{Y}} P(y|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t) \tag{1}$$

where $\mathcal{X}_t$ is a segment of physiological signals at time $t \in T$ and $\mathcal{Y}$ is the set of emotional states, such as happiness, surprise, anger, fear, and sadness.

The second stage is the affective causality learning problem. The problem is formulated as a graph

$$\mathcal{G} = (\mathcal{M}, \mathcal{C}, E) \tag{2}$$

where $\mathcal{M} \in \mathbb{R}^{N_s \times N_e \times T}$ is the emotion occurrence tensor, N_s is the number of situations, and N_e is the number of emotional states. $\mathcal{C} \in \mathbb{R}^{N_s \times N_c \times T}$ is the life contents occurrence tensor, where N_c is the number of contents and $e_{ij} \in E$ indicates the affective causal effect of sequence $\mathcal{C}_i$ on sequence $\mathcal{M}_j$.

B. Deep Physiological Affect Network for Recognizing Human Emotions

DPAN describes affect elicitation mechanisms used to detect emotional changes reflected by physiological signals. It takes two-channeled EEG signals underlying brain lateralization and a PPG signal as inputs and outputs a 1-D vector representing emotional states scaled from 1 to 9. Suppose that DPAN obtains the physiological signals at time N over a spectral-temporal region represented by an $M \times N$ matrix with P different modalities. From the two modalities of EEG and PPG sensors, physiological features B_t and H_t are extracted from the respective sensors as follows:

$$B_t = \xi_{rl} \circ \frac{(\zeta_l - \zeta_r)}{(\zeta_l + \zeta_r)} \tag{3}$$

where "$\circ$" denotes the Hadamard product. $[(\zeta_l - \zeta_r)/(\zeta_l + \zeta_r)]$ represents the spectral asymmetry and the matrix ξ_{rl} is the causal asymmetry between the r and l EEG bipolar channels. The brain asymmetry feature B_t describes the directionality and magnitude of emotional lateralization between the two hemispheres.

DPAN extracts the HR features H_t over the $M \times N$ spectral-temporal domain, where frequencies with peaks in the PSD of the PPG signal are regarded as candidates of the true HR from the PPG signal P_t at each time frame t. These data form a candidate set over time. The observation at a given time can then be represented by a tensor $\mathcal{X} \in \mathbb{R}^{M \times N \times P}$, where $\mathbb{R}$ denotes the domain of the observed physiological features. Then, the learning problem is the identification of the correct class based on the sequence of tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t$

$$\hat{y} = \arg\max_{y\in\mathcal{Y}} P(y|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_t) \tag{4}$$

where $\mathcal{Y}$ is the set of valence-arousal classes. Fig. 1 shows the entire overview of DPAN for the recognition of emotions. To solve the learning problem, DPAN feeds spectral-temporal tensor-based physiological features into ConvLSTMs to compute affective scores of emotions via the proposed loss model, temporal margin-based loss (TM-loss). The proposed TM-loss aims to learn the progression patterns of the emotions in training for developing reliable affect models. The TM-loss is a new formulation based on the temporal margin between the correct and incorrect emotional states. The reasoning for using the formulation is as follows.

> *When more of a particular emotion is observed, the model should be more confident of the emotional elicitation as the recognition process progresses.*

The function constrains the affective score of the correct emotional state to discriminate its margin, which does not monotonically decrease with all the others while the emotion progresses

$$\mathcal{L}_t = -\log s_t(y) + \lambda \max\left(0, \max_{t'\in[t_0, t-1]} m_{t'}(y) - m_t(y)\right) \tag{5}$$

where $-\log s_t(y)$ is the conventional cross-entropy loss function commonly to train deep-learning models, y is the ground truth of emotion rating, $s_t(y)$ is the classified affective score of the ground-truth label y for the time t, and $m_t(y)$ is the discriminative margin of the emotion label y at time t

$$m_t(y) = s_t(y) - \max\{s_t(y')|y' \in \mathcal{Y}, y' \neq y\}. \tag{6}$$

$\lambda \in \mathbb{Z}^+$ is a relative term to control the effects of the discriminative margin. As described in (5) and (6), a model becomes more confident in discriminating between the correct state and the incorrect states over time. With this function, DPAN is encouraged to maintain monotonicity in the affective score as the emotion training progresses. As shown in Fig. 1(b), after the time t_c, the loss becomes nonzero due to the violation of the monotonicity of the margin. Note that the margin $m_t(y)$ of the emotion y spanning $[t_0, t]$ is computed as the difference between the affective score $s_t(y)$ for the ground truth y and the maximum classification scores $\max_{y'\neq y} s(y')$ for all incorrect ratings at each time point in $[t_0, t]$.

C. Conditional Independence Test

To discover the causality between affective dynamics and life contents and answer (C.2), we test the conditional independence with three sequences A_i, A_j, and A_k. Suppose the three sequences have the same length T, the test is to verify the statistical significance of the statement $A_i \perp A_j | A_k$. Considering each triplet $(A_i(t), A_j(t), A_k(t))$ for each $1 \leq t \leq T$ as a sample over three variables, a 3-D contingency table C records the number of triplet samples o, p, and q on the three variables at each entry such that

$$C_{\text{opq}} = \left|\left\{1 \leq t \leq T \middle| A_i(t) = o, A_j(t) = p, A_k(t) = q\right\}\right|. \tag{7}$$

Note that the expectation of C_{opq} under the null hypothesis can be estimated by

$$E(C_{\text{opq}}) = (C_{*pq} C_{o*q})/(C_{**q}) \tag{8}$$

where C_{*pq}, C_{o*q}, and C_{op*} are the marginals of the counts with A_i, A_j, and A_k, respectively. For the test, we use the standard G^2 conditional independence test, which returns the

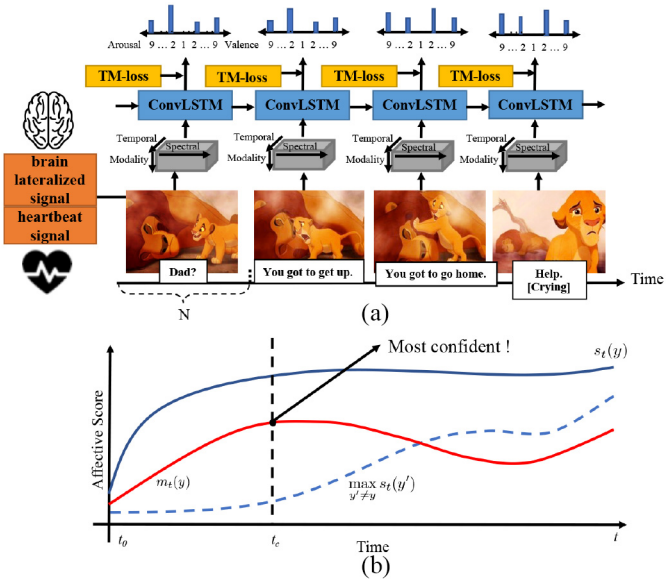

Fig. 1. (a) Overview of DPAN. After every time interval N, the proposed DPAN first extracts two physiological features (brain lateralized and heartbeat features) and constructs a spectral-temporal tensor. These features are then fed into ConvLSTM to compute affective scores of emotions via our proposed loss model, TM-loss. The output at the final sequence is selected to represent an emotion over a 2-D valence-arousal model for the entire sequence. (b) Discriminative margin $m_t(y)$ (red line) of an emotion y started at t_0. The margin $m_t(y)$ is computed as the difference between the ground-truth affective score $s_t(y)$ (blue line) and the maximum scores $\max_{y' \neq y} s_t(y')$ (dashed blue line) of all incorrect emotion states between t_0 and t. The model becomes more and more confident in classifying emotion states until the time t_c. However, after the time t_c, $\mathcal{L}_t$ are nonzero due to the violation of the monotonicity of the margin.

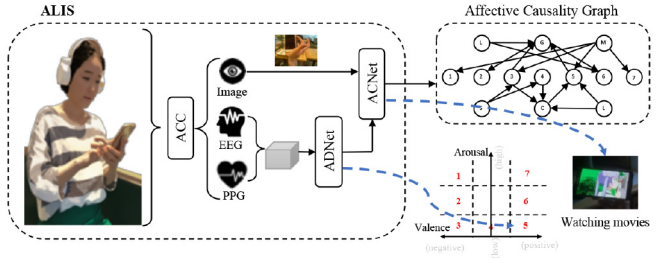

Fig. 2. Overview of ALIS, which consists of an ACC, ADNet, and ACNet. ACC collects a user's contextual information in situations with frontal images and emotion measured by EEG and PPG signals. Given this information, ADNet detects emotional changes, which are used as an input with frontal images for ACNet to discover the causal relationship between emotions and situations.

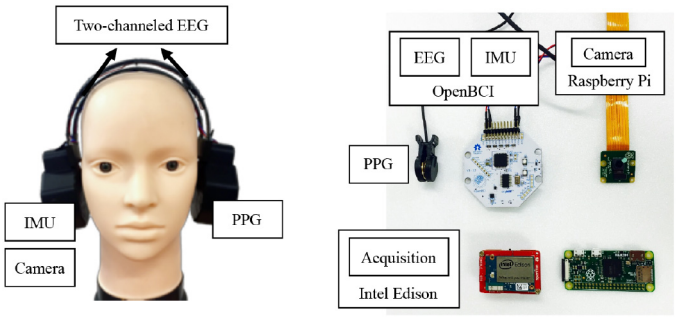

Fig. 3. ACC and its components; IMU, EEG, PPG sensors, and a tiny frontal camera. The location of two EEG electrodes (F3 and F4) on the 10–20 international system.

Kullback–Leibler divergence between the distributions of C_{opq} and $E(C_{\text{opq}})$ over all three variables

$$G^2 = 2 \sum_{o,p,q} C_{\text{opq}} \ln \frac{C_{\text{opq}}}{E(C_{\text{opq}})} \tag{9}$$

which follows a χ^2 distribution with degree of freedom $(|A_i| - 1) * (|A_j| - 1) * |A_k|$. For removing the sparsity in the table C, the degree of freedom is penalized by the number of zero cells as in [31].

III. Wearable Affective Lifelog System

ALIS consists of three main parts: 1) affective contents collector (ACC); 2) affective dynamics network (ADNet); and 3) affective causality network (ACNet). Fig. 2 shows the entire framework of ALIS. First, ACC gathers contextual information continuously surrounding the wearer in daily life. The logged data are then transferred into ADNet, which aims to find answers about under which situations and to what extent a human feels the elicited affect. ACNet discovers the dynamic causal relationship between the situation faced and the human emotion explored by ADNet. It provides an intuitive understanding of affect dynamics for users. The following sections describe the details of each component.

A. Affective Contents Collector for Logging Data

ACC is a simple device designed to be easily wearable for users to act freely in everyday situations (See Fig. 3) as well as to collect the human affect correctly. Since the human affect is sophisticated and subtle, it is vulnerable to personal, social, and contextual attributes. The noticeability and visibility of wearable devices could elicit unnecessary and irrelevant emotions. Therefore, recording human affect should be unobtrusive when measured in the natural environment. To design an unnoticeable device, we imitated the design of existing easy-to-use wireless headsets. We note that the term "unobtrusive device" in this article means that it is not easily noticed or does not draw attention to itself. The term does not imply that our device aims to be small or concealable. This easy-to-use device provides comfort and performance to users during long-term activities.

Everyday technology requires wearable systems to have the unprecedented ability to perform the comfortable, long-term, and in situ assessment of physiological activities. However, the development of practical applications is challenging because of the cumbersomeness of the equipment that requires multiple channels to obtain reliable signals and the complexity of setting up experiments. To use body sensors with a guarantee of both reliability and simplicity, we designed the proposed device with a tiny PPG sensor and a minimum configuration of a two-channel EEG, the lowest number of channels necessary to learn patterns of lateralized frontal cortex activity involved in human affect.

While satisfying the two criteria, our device consists of multimodal sensors to capture various emotions surrounding daily life as follows.

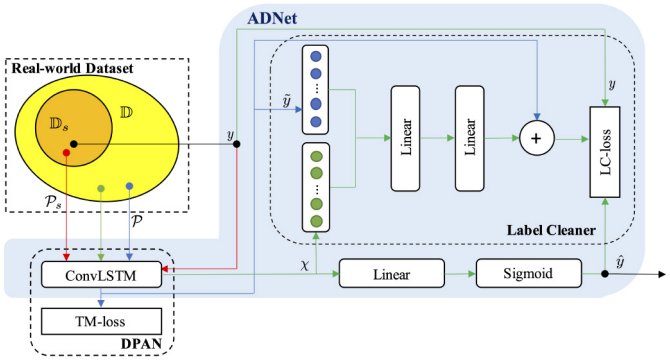

Fig. 4. Overview of ADNet. ADNet first uses DPAN to train physiological signals associated with their labels y on the dataset $\mathbb{D}_s$ (red lines). The model then uses the learned parameters on the dataset $\mathbb{D}$ for predicting noisy pseudolabels $\tilde{y}$ (blue lines), which are fed into a subnetwork label cleaner conditioned on physiological features χ from ConvLSTMs. Within the label cleaner network, a residual architecture learns the difference between the noisy $\tilde{y}$ and clean labels y (green lines). Finally, the model predicts cleaned labels $\hat{y}$ penalized by a joint loss function (LC-loss).

1) *Frontal Camera for Collecting Visual Contents:* Visual information has been widely used to detect situations faced by the user. The study of recognizing scenes and activities by analyzing images from a camera has provided an understanding of contextual information. Hence, in our system, a small frontal viewing camera with a 30 fps sampling rate (Raspberry Pi Zero Camera) was used to record images.
2) *Small Physiological Sensor to Capture Human Affect:* The analysis of patterns of physiological changes has been increasingly studied in the context of affect recognition. To capture this information, we used a two-channel EEG sensor on OpenBCI on the left and right hemispheres and a small ear-PPG sensor, with sampling rates of 250 and 500 Hz, respectively.

Physiological signals from EEG and PPG sensors and frontal images collected by ACC comprise a large real-world dataset $\mathbb{D}$. The dataset has a subset $\mathbb{D}_s$ where a small number of ground-truth labels y by self-reporting are available. The real-world dataset in our experiment is called the affective lifelog dataset, which is further described in Section V. Given the dataset $\mathbb{D}$ and its subset $\mathbb{D}_s$, the proposed ADNet jointly learns to reduce the label noise and predict more accurate labels $\hat{y}$ on the dataset $\mathbb{D}$.

B. Affective Dynamics Network for Recognizing Emotions

ADNet aims to solve the problem in (1), addressing the challenging questions (C.1) on a real-world large dataset. Our network first learns the representations of the physiological signals $\mathcal{P}$ according to the affective labels y on the subset $\mathbb{D}_s$ by DPAN [13] and produces noisy pseudolabels $\tilde{y} = (\tilde{\mathcal{V}}, \tilde{\mathcal{A}})$ on the dataset $\mathbb{D}$. Given the dataset $\mathbb{D} = \{(\mathcal{P}_i, \tilde{y}_i), \ldots, \}$ and its subset $\mathbb{D}_s = \{(\mathcal{P}_i, \tilde{y}_i, y_i), \ldots, \}$, the proposed framework jointly learns to reduce the label noise and predict more accurate labels $\hat{y}$ on the dataset $\mathbb{D}$ (see Fig. 4).

While DPAN has shown its superiority in classifying emotions, its predicted labels have contained noise. To overcome this issue, ADNet comprises a subnetwork, which learns to map noisy labels $\tilde{y}$ to clean labels y, conditioned on physiological features from ConvLSTMs in DPAN. We denote this subnetwork as the label cleaner. The subnetwork is supervised by the human-reported labels y and follows a residual architecture so that it only needs to learn the difference between the noisy and clean labels. In particular, to handle the sparsity in noisy labels, the label cleaner encodes the emotional occurrence $\tilde{y}$ of each of the d classes in valence $\tilde{\mathcal{V}}$ and arousal $\tilde{\mathcal{A}}$ ratings into a pair of d-dimensional vector $[0, 1]^d$. Similarly, the model projects the physiological features $\mathcal{X}$ into a low-dimensional embedding, and then all embedding vectors from the two modalities are concatenated and transformed with a hidden linear layer followed by a projection back into the high-dimensional label space.

Simultaneously, the primary network ADNet shares the physiological features extracted by ConvLSTMs (the last unit before TM-Loss) in DPAN and learns to directly predict labels $\hat{y}$ following a sigmoid function. The predicted labels $\hat{y}$ are supervised by either the output of DPAN or a human from label cleaner. To train the ADNet, we formulate a joint loss function as follows:

$$\mathcal{L}_c = \sum_{i \in \mathbb{D}_s} |\bar{y}_i - y_i| - \sum_{j \in \mathbb{D}} \left[u_j \log(\hat{y}_j) + (1 - u_j) \log(1 - \hat{y}_j) \right] \tag{10}$$

where u_j is y_j when the SAM-ratings are available, otherwise, $\bar{y}_j$. The LC-loss $\mathcal{L}_c$ is the combination of: 1) the difference between the cleaned labels and the corresponding ground-truth verified labels and 2) the cross-entropy to capture the difference between the predicted labels and the target labels. The cross-entropy term is only propagated to $\hat{y}_j$. The cleaned labels $\bar{y}$ are treated as constants with respect to the classification and only incur gradients from the LC-loss. We note that DPAN is the pretrained model from the dataset $\mathbb{D}_s$ prior to ADNet. The shared parameters of ConvLSTM are updated with the LC-loss function, not with the TM-loss function in DPAN.

C. Affective Causality Network

ACNet aims to solve the problem in (2). The goal is to derive the affective causality by analyzing emotional changes and the user's relevant situations. Suppose the emotion sequence $\mathcal{M}(1:T)$ is a binary stochastic process during a discrete time interval $[1, T]$, where T is the maximal length of the interval. Then, an element $\mathcal{M}_j$ in the sequence $\mathcal{M}(1:T)$ is a binary indicator of the occurrence of a certain emotion j detected by ADNet at time t. With this notation, $\mathcal{M}_j^\eta$ is generated with each concatenated element $\mathcal{M}_j^\eta(t) = (\mathcal{M}_j(t - \eta), \ldots, \mathcal{M}_j(t))$ at each timestamp $t \leq T$. In the same way, an element $\mathcal{C}_i$ in a situation sequence $\mathcal{C}(1:T)$ is a binary indicator of the occurrence of certain situation i observed by ACC at time t. Based on these notations, the time-varying affective network and the affective causality are defined with the formulation of the related learning task.

Definition 1: The time-varying affective network is denoted as $\mathcal{G} = (\mathcal{M}, \mathcal{C}, E)$, where $\mathcal{C}_i$ and $\mathcal{M}_j$ are the situation i and the

emotional state j sequences, and $e_{ij} \in E$ indicates the affective causal effect of the sequence $\mathcal{C}_i$ on the sequence $\mathcal{M}_j$.

Definition 2: Given n situation and emotion sequences $\{\mathcal{C}_1, \ldots, \mathcal{C}_i, \mathcal{M}_1, \ldots, \mathcal{M}_j\}$ from a user, the directed graph $\mathcal{G}$ is constructed in the underlying affective causality model while addressing confounding factors.

By representing each of the emotion and situation variables as a node, ACNet could be formulated as a directed graph $\mathcal{G}$ over the variable $\mathcal{C}_i \cup \mathcal{M}_j$ such that each edge $\mathcal{C}_i \to \mathcal{M}_j$ indicates the affective causal effect of sequence $\mathcal{C}_i$ on sequence $\mathcal{M}_j$. In the network, there is a parameter associated with a node or an edge between $\mathcal{C}_i$ and $\mathcal{M}_j$. Based on the descriptions above, a network generally consists of two components: 1) a directed graph $\mathcal{G}$ used to model the causal structure among the sequences and 2) association parameters on the edges used to model the causal phenomena. These two components characterize the global and local properties of the interaction between emotions and life contents, respectively, the combination of which provides a general guideline ruling the events happening on the two sequences.

Affective causality is a computational asymmetric measure to determine the causal relationship between affective dynamics and situations. The computation is based on conditional independence testing to detect the relationship with latent confounders underneath the observational two sequences. The causal structure learning method is inspired by Cai *et al.*'s work [26], but differs in that the ACNet has two different variables $\mathcal{M}$ and $\mathcal{C}$ as input nodes, considering sparse relations. The proposed model analyzes this causal problem, considering: 1) with and 2) without confounding factors.

1) Learning Without Confounding Factors: The interaction without any latent factors can observe that the state $\mathcal{C}_i$ at timestamp $t-1$ affects the state of $\mathcal{M}_j$ at timestamp t, while the reverse does not hold. This observation can be formalized in the language of statistical testing with Lemma 1.

Lemma 1: Given two dependent sequences $\mathcal{C}_i \to \mathcal{M}_j$ without a connected latent variable, the following asymmetric dependence relations hold: 1) there exists a delay η_c satisfying $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$ and 2) there does not exist a delay η_m satisfying $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$.

2) Learning With Confounding Factors: Sequences under the existence of confounding factors: 1) behave similarly because of common characteristics or 2) interact independently over the timeline. Both the situation $\mathcal{C}_i$ and emotion $\mathcal{M}_j$ sequences are affected by a constant latent variable. In such a case, $\mathcal{C}_i(t)$ is thus dependent on $\mathcal{M}_j(t-1)$ given any previous states of $\mathcal{C}_i$. Similarly, $\mathcal{M}_j(t)$ also depends on $\mathcal{C}_i(t-1)$ given any previous states of $\mathcal{M}_j$. The confounding factor, which is itself an independent variable over time, is underneath the two sequences under test. When this case occurs, a positive statistical dependence between $\mathcal{C}$ and $\mathcal{M}$ is observed. But the states of $\mathcal{C}_i(t)$ could be completely independent of $\mathcal{M}_j(t-1)$ given its previous states $\mathcal{C}_i(t-\eta_c-1 : t-1)$, and $\mathcal{M}_j(t)$ could also be independent of $\mathcal{C}_i(t-1)$ given its previous states $\mathcal{M}_j(t-\eta_m-1 : t-1)$. The following lemma could be recognized using the same group of conditional independence tests.

Lemma 2: If there is a latent factor between $\mathcal{C}_i$ and $\mathcal{M}_j$, the following symmetric relations hold: 1) there does not exist any delay η satisfying $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$ nor $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$ and 2) there exists delay η_m and η_c satisfying $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$ and $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$, respectively.

Algorithm 1 Affective Causality Direction Learning

Input: $\mathcal{C}_i$: situation i sequence
$\mathcal{M}_j$: emotion state j sequence
α : confidence threshold
T : maximal timestamp
$\mathcal{F}$: affective pair set $\{\mathcal{C}_i, \mathcal{M}_j\}$
Output: The directed graph $\mathcal{G}$

```
Initialization : set G as empty set;
for each i and j in a pair F in a situation do
    if ∄F' ∈ F − {C_i, M_j}, C_i ⊥ M_j | F' then
        Test S_1 on C_i and M_j;
        Test S_2 on M_j and C_i;
        if S_1 ∧ ¬S_2 then
            Add i → j into G;
        else if ¬S_1 ∧ S_2 then
            Add j → i into G;
        else
            Add i ← H → j into G;
        end if
    end if
end for
return G
```

3) Affective Causality Direction Learning Algorithm: Algorithm 1 describes the asymmetric relations on all dependent sequence pairs of situation $\mathcal{C}$ and emotional state $\mathcal{M}$ and detects the directions of causal edges on the underlying affective model $\mathcal{G}$ by applying the following theorem based on the two lemmas.

Theorem 1: Given two sequences $\mathcal{C}_i$ and $\mathcal{M}_j$, the following propositions on the causal structure between the two sequences hold: 1) $\mathcal{C}_i \to \mathcal{M}_j$ in $\mathcal{G}$, if $S_1 \wedge \neg S_2$; 2) $\mathcal{C}_i \leftarrow \mathcal{M}_j$ in $\mathcal{G}$, if $\neg S_1 \wedge S_2$; and 3) there is a latent factor between $\mathcal{C}_i$ and $\mathcal{M}_j$, if $S_1 \wedge S_2$ or $\neg S_1 \wedge \neg S_2$.

S_1 and S_2 in the theorem are used for hypothesis tests on any pair of the two sequences.

S_1: $\exists \eta_c$ satisfying $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$.

S_2: $\exists \eta_m$ satisfying $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$.

In this algorithm, the affective sequence set $\mathcal{F} = \{\mathcal{C}, \mathcal{M}\}$ along with the maximal timestamp T, the number of sequence N, and the confidence threshold α is used as inputs for the test. Given the inputs, each affective pair is tested by applying Theorem 1 to see whether each is independent of the other conditioned on other variables. The output $\mathcal{G}$ is a set of pairwise relations. The complexity of Algorithm 1 is determined by the dependent pair detection and the test of S_1 and S_2. All proofs are provided in the Appendices.

IV. Evaluation on Synthetic Dataset

In the following two sections, we examine the robustness of the proposed ALIS on the two datasets: 1) a public dataset and 2) a synthetic dataset, for the quantitative evaluation of ADNet in emotion recognition and ACNet in causality identification.

A. Affective Dynamics Network Performance

For quantitative evaluation of ADNet in emotion classification (C.1), we performed realistic experiments by deliberately manipulating labels on a public dataset called DEAP [19].

1) DEAP Dataset: DEAP is a public dataset of physiological signals to analyze emotions quantitatively on a 2-D plane along with valence and arousal as the horizontal and vertical axes. Its physiological signals are recorded from 32-channel EEGs at a sampling rate of 512Hz using active AgCl electrodes placed according to the international 10–20 system and 13 other peripheral physiological signals from 32 participants while they watched 40 1-min-long excerpts of music videos. The dataset rates emotions with respect to continuous valence, arousal, liking, and dominance scaled from 1 to 9 and discrete familiarity on scales from 1 to 5 using SAMs [19].

2) Dataset Configuration: Since ACC gathers EEG signals from a pair of electrodes and a PPG signal, we retrieved data from the eight selected pairs of electrodes and a plethysmograph on the DEAP dataset. The 64 combinations of physiological signals per video generate 81 920 physiological data points. The data were high-pass filtered with a 2-Hz cutoff frequency using EEGlab and the same blind source separation technique as in [13] for removing eye artifacts in EEG signals. A constrained independent component analysis (cICA) algorithm was applied to remove motion artifacts in PPG signals. We built a subject-independent dataset in which physiological signals were retrieved from all participants to evaluate the proposed model as a subject-independent classifier directly applied to any users without personalized optimization. To make the evaluation independent from the effect of personalization, we split the dataset into fifths: one-fifth for testing and the remaining. Then, the remaining (= four-fifths) was split into fifths again: four-fifths for training and one-fifth for validation. The validation data were randomly chosen from the remaining while keeping the distribution of the label ratings balanced ($= 1/d^2$) for fair evaluation per label.

We used the balanced datasets to evaluate our models' performance and other methods for solving the limited amount of clean-label problems. To evaluate the unbalanced labeling problem, we stochastically changed the number of physiological signals associated with a label while varying the distribution of labels in the training dataset. The unbalanced dataset comprised p percentage of physiological signals according to a pair of random labels in valence and arousal, with the others distributed equally. The test data remained unperturbed to allow us to validate and compare our model to other methods. The highlighted 1-min EEG and plethysmography signals were split into six frames of 10 s each. They were downsampled to 256 Hz, and their power spectral features were extracted.

3) Evaluated Methods and Metrics: We evaluated the performance of our ADNet, comparing the results with state-of-the-art methods that have shown their performance on the DEAP dataset in terms of loss functions and the number of layers. We would first note that our previous work, TM-Loss in DPAN [13], had improved emotion recognition performance, constraining the affective score of the correct emotional state to discriminate its margin, which does not monotonically decrease with all others while the emotion progresses. Second, ADNe, in this work, consists of three components: 1) ConvLSTM, which learns physiological signals in a supervised way; 2) DPAN, which produces pseudolabels; and 3) label cleaner, which cleans label noises.

Given the configuration, to evaluate ADNet, it is necessary to investigate how each part contributes to differentiate multiple emotions, handling the limited and unbalanced label issues. Three models, called Model A, B, and C, are designed for the comparative study as follows.

1) *Model A:* It is a 1-layered ConvLSTMs with a softmax layer as a baseline classifier. The results from the model represents a performance of simplified version of deep neural networks (DNNs).
2) *Model B:* It is a 1-layered ConvLSTMs with the TM-loss function as in [13]. The model has shown its superiority in recognizing human emotions, increasing the distinctiveness of physiological characteristics between correct and incorrect labels. Pseudolabels $\tilde{y}$ are essential to produce the cleaned labels $\bar{y}$, which cover all samples in the dataset $\mathbb{D}$. We chose the above simplified version to focus on studying the effectiveness of TM-loss, excepting other potential factors.
3) *Model C:* It is a 4-layered ConvLSTMs with a softmax layer. The model was implemented based on [32].

Our proposed system and the three comparative models consist of 256 hidden states and 5×5 kernel sizes for the input-to-state and state-to-state transition. They were trained by learning batches of 32 sequences and backpropagation through time for ten time steps. The momentum and weight decay were set to 0.7 and 0.0005, respectively. The learning rate starts at 0.01 and is divided by 10 after every 20 000 iterations. We also performed early stopping on the validation set. All models learn physiological signals to classify $d^2 (d = 2, 3, 4)$ affective states. For instance, when $d = 4$, 16 affective states are combined of two to four valence and arousal states rated from 1 to 3, 3 to 5, 5 to 7, and 7 to 9. We choose the mean average precision (MAP) as a metric to evaluate the performance of our system with additional *Precision*, *Sensitivity*, and *Specificity* to report how commonly occurring classes in a training set affect the model performance.

4) Evaluation Results: Fig. 5 shows that our proposed system performed better than the alternatives on the balanced dataset over all dimensions ($d = 2, 3, 4$) unless the amount of clean labels was very large (> 0.7). Model C increased its performance rapidly as the number of SAM-rated labels increased. When the fraction was greater than 0.7, the model achieved the next best classification results. Model A consistently reported the lowest precision consistently over all configurations. This consistency could imply that Model A was overfit and overconfident to some labels. Concerning overfitting, although having a specialized loss function (Model B) and increasing the number of layers in DNNs (Model C) improved the discriminative power to classify different emotions, the limited number of clean labels seems to lead to overfitting for many classes.

Fig. 6 provides a closer look at how different label frequencies affect the performance in emotion recognition

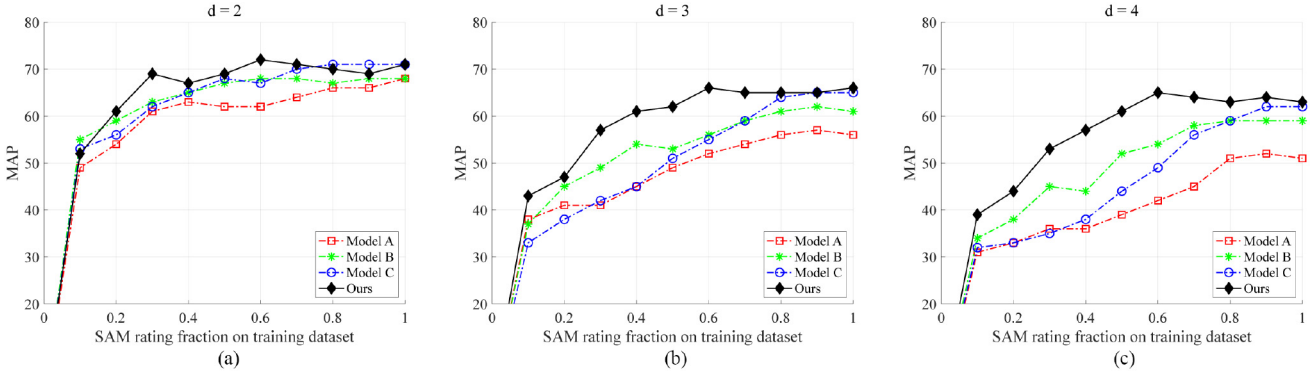

Fig. 5. Test classification performance on the DEAP dataset with varying the fractions of SAM rated labels over different number of classes (valence and arousal). (a) d = 2. (b) d = 3. (c) d = 4.

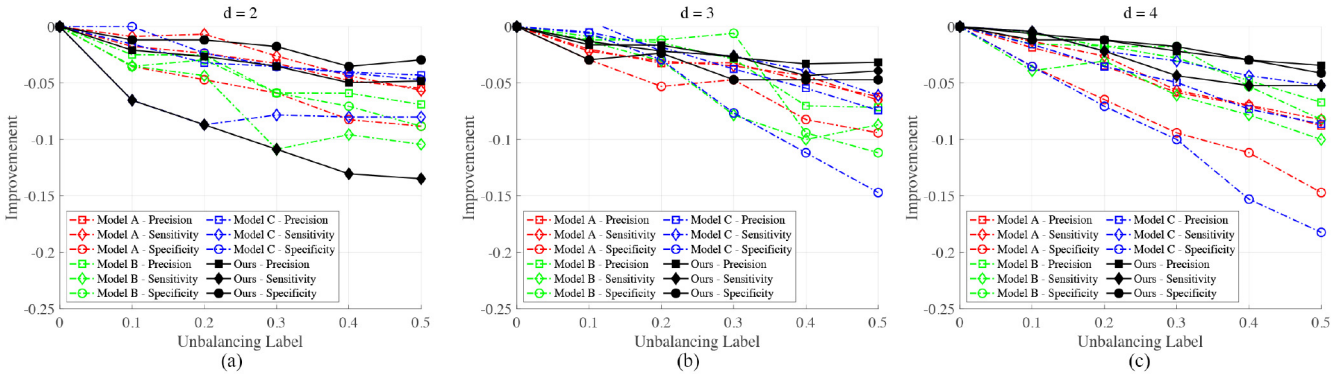

Fig. 6. Test classification performance on the DEAP dataset with varying label frequencies over different classes. (a) d = 2. (b) d = 3. (c) d = 4.

when the label balances were collapsed by p. Our method's three scores converged around -0.03 except sensitivity, whereas other methods decreased specificity, predicting incorrectly negative samples. Furthermore, the proposed model recovered the worsen performance in sensitivity up to -0.03 ($d = 4$) from -0.13 ($d = 2$). Increasing the number of labels deteriorated the classification performance, including our model [$d > 2$, Fig. 6(b) and (c)]. All models became overfit to common labels and reduced overall accuracies. Despite alleviated overfitting, our model yielded the best results. In contrast, the other methods lost their ability down to approximately 15% (Model A) and 18% (Model C, Specificity) for the balanced and unbalanced datasets, respectively. The vertically stacked layers in Model C led it to misunderstand the physiological characteristics associated with a major number of labels. Along with the same line from the above results, these results imply that our method effectively learns invariant features for classifying multiple classes.

B. Affective Causality Network Performance

1) Experimental Setup: To evaluate the affective causality identification (C.2) on the proposed ACNet, a synthetic dataset was generated. In the dataset, pairs of affective sequences $\mathcal{F}$ are generated by simulating a Poisson point process with 10 min per timestamp and ϵ as occurrence frequency for situations $\mathcal{C}$ and emotions $\mathcal{M}$ per day. Each sequence is influenced by its causal nodes as the time-dependence influence function with exponential probability $p(\Delta_t, \eta) = \eta e^{(-\Delta_t \eta)}$, where Δ_t is the time interval between t and the causal sequence's most recent state, and η is the average influence lag.

2) Evaluated Methods and Metrics: We evaluated the performance of ACNet, comparing it against the transfer entropy (TR) method [33] and Granger's causality (GC) method [34]. Two groups of causal structures were designed for experiments. The first group has the causal structures of directed graphs without latent variables, and the second group consists of directed graphs with latent factors. Given n sequences and the average in-degree d_g of the graph, a pair of two emotion and situation nodes and a directed edge between the pairs into the graph are selected until there are $d_g \cdot n$ edges in the graph. For the second group, confounding factors are selected by n_c independent pair of nodes and an additional latent factor $\mathcal{H}$ underlying two edges is added into the causal graph. We choose *Precision*, *Recall*, and *F1*-score as metrics to evaluate each type of causality.

3) Results: Fig. 7 shows the performance of ACNet and the two other methods on different causal structures and data generation parameters with varying occurrence frequencies ϵ and average influence lag η. Overall, ACNet consistently outperformed the other methods. As shown in Fig. 7(a), the occurrence frequency reflects the sparsity of the sequences such as the number of timestamps with recorded situations. The performance of TR and GC declines with increasing sparsity. On the other hand, ACNet maintained its performance at around 0.7 even when there was only one occurrence in 3 h. This result implies that our model is effective in solving the causal discovery problems on the sparse affective situation

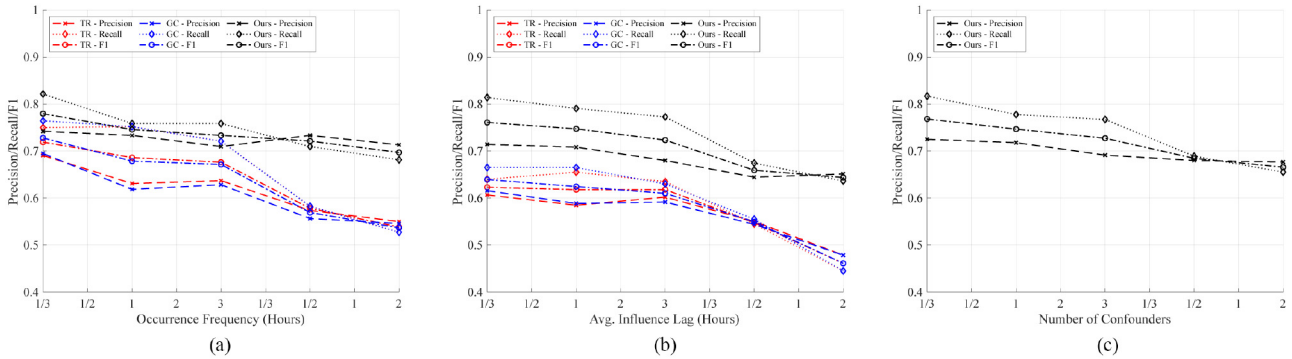

Fig. 7. Affective identification performance on the synthetic dataset varying occurrence frequency, average influence lag, and numbers of confounder parameters. (a) Occurrence frequency. (b) Averaged influence lag. (c) Number of confounders.

sequences. Fig. 7(b) shows the sensitivity with different influence lags. The recall of our model decreased with the increasing average influence lags because larger influence lags mean longer dependence on the previous states. The precision of our model was not sensitive to the influence lag, keeping a performance between 0.65 and 0.8, while the precision of other methods decreased as the influence lag increased. These results imply our model is capable of catching long-term dependence for learning causal direction.

V. Evaluation on Real-World Dataset

Underlying the quantitative analysis of our model on the public and synthetic datasets, we further demonstrated the capability of this model via a long-term series of life logging over several days in real-world scenarios. We first built a real-world dataset called the affective lifelog dataset, where participants used our ACC in their daily life. We then evaluated the performance of the proposed ALIS with respect to physiological discrimination in emotion recognition and user agreement in causal identification.

A. Affective Lifelog Dataset

1) Data Acquisition: The dataset consists of two modalities of physiological signals, accelerometer signals, and frontal images obtained by ACC from 16 male and 5 female university students aged from 21 to 35 (26.4 $\pm$ 4.87) years. Our requirement for participation was to perform at least one common task of a university student, such as conducting research, taking classes, or having a discussion with colleagues. We required them to wear the device over 6 h per day for up to 45 days with $10 compensation per day. This experiment was approved by the institutional review board (IRB) in Human Subjects Research.

To identify individual specific affective contents and assign the SAM ratings to them as ground-truth labels, we asked the participants if they had any affective contents that had elicited a specific feeling, and how the contents changed their emotion before and after the elicitation. The life contents were perceived as breakthroughs to change their mentality. The changes were rated by the SAM scaled from 0 to 6 and -3 to 3 for arousal and valence ratings. Furthermore, we retrieved the following affective contents manually, which are considered to potentially affect mental status stress: 1) watching movies; 2) drinking coffee; 3) drinking green tea; 4) hanging out with friends; 5) playing games; 6) drawing a picture; 7) taking a walk; 8) reading a book; 9) eating food; 10) studying at a desk; 11) reading research papers on a computer monitor; 12) conducting researches in a laboratory; and 13) playing with media devices. In such situations, the participants performed the SAM ratings every five days.

TABLE I
Affective Lifelog Dataset Contents

Number of Participants		21
Averaged Duration per Situation (Minute)		18.3
Rating	Valence Arousal	-3 to 3 0 to 6
Signals		2-channel EEG 1-channel PPG Accelerometer Frontal images
Days	Average (Max, Min)	45 (13, 23.31)
Situation	Average (Max, Min)	378 (154, 195.69)
Data Size (Segment)	Average (Max, Min)	163k (78k, 114k)

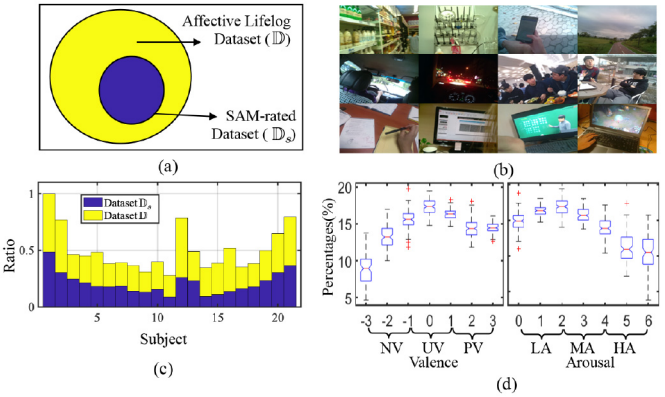

Fig. 8. Overview of the affective lifelog dataset. (a) Affective lifelog dataset $\mathbb{D}$ and its subset $\mathbb{D}_s$ which contains SAM-rated situations. (b) Example images in the dataset $\mathbb{D}$. (c) Proportion of the subset $\mathbb{D}_s$ and the dataset $\mathbb{D}$. $N_s = 378$ for the subject 1.(d) Distribution of the SAM-rated situations in valence and arousal labels on the subset $\mathbb{D}_s$.

2) Affective Lifelog Dataset: Fig. 8 and Table I summarize the affective lifelog dataset $\mathbb{D}$. The situations rated by the SAM

consist of a subset $\mathbb{D}_s$ with pairs of emotion labeling $y = (\mathcal{V}, \mathcal{A})$ scaled from -3 to 3 and from 0 to 6 for valence $\mathcal{V}$ and arousal $\mathcal{A}$ ratings, respectively. The labeling $y = (\mathcal{V}, \mathcal{A})$ is used as ground truth to evaluate the performance of our proposed system. As shown in Fig. 8(c), the dataset has a limited amount of labeling data. The ratios of the dataset $\mathbb{D}$ and the subset $\mathbb{D}_s$ are 0.418 from all participants. Furthermore, the distribution of labels is unbalanced. These challenge issues are consistent with our understanding describe in (C.1) and (C.2). Fig. 8(b) shows some situations in the dataset $\mathbb{D}$ from our real-world experiments. Participants have experienced various situations in daily life, such as driving a car, reading a research paper, playing games, and taking a walk.

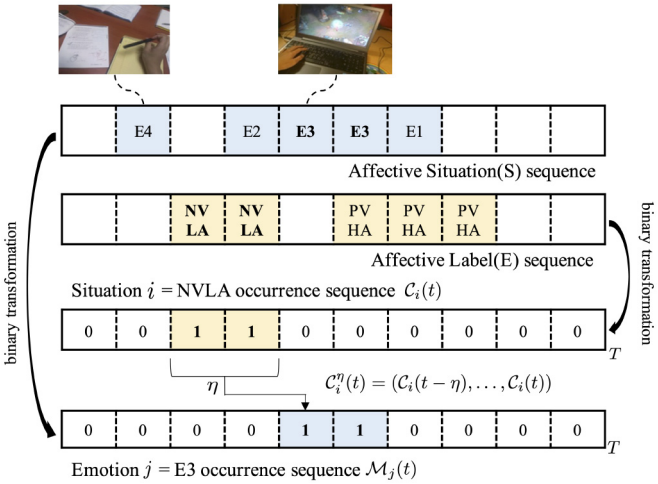

Fig. 9. Illustration of the affective causal effect of sequence $\mathcal{C}_i$ on sequence $\mathcal{M}_j$. Testing the affective causal effect of the affective sequence NVLA (may be occurred by the behavior "Studying on a desk") on the behavior sequence "Playing Games" indexed as E3 with a delay η.

TABLE II
PERFORMANCE OF ADNET FOR CLASSIFYING THE SEVEN EMOTIONAL STATES AND ACNET FOR IDENTIFYING THE AFFECTIVE CAUSALITIES ON THE DATASET $\mathbb{D}$

ACNet	ADNet						
	NVHA	NVMA	NVLA	UVLA	PVLA	PVMA	PVHA
0.74	0.61	0.59	0.55	0.64	0.52	0.56	0.55

B. Experimental Setup

1) Train/Test/Evaluation on the Dataset $\mathbb{D}$: From the real-world dataset $\mathbb{D}$, we grouped affective labels $(\mathcal{V}, \mathcal{A})$ in valence and arousal into seven discrete affective states: 1) NVHA; 2) NVMA; 3) NVLA; 4) UVLA; 5) PVLA; 6) PVMA; and 7) PVHA, comprising low (LA), mid (MA), and high (HA) arousal and negative (NV), neutral (UV), and positive (PV) valence ratings. The three classes were determined by dividing the 6-point rating scale of the participants' valence and arousal ratings into three classes (low, mid, and high), with each class containing two points [Fig. 8(d)]. We should note that two states (UVMA and UVHA) were omitted since their occurrence was extremely low. We retrieved 10 000 pieces of physiological data per affective state, which is 70 000 pieces of physiological data on total of seven states for every participant on the dataset $\mathbb{D}$. The test data remained unperturbed to validate and compare our model to other methods.

Since physiological signals, in particular, EEG signals, are vulnerable to motion artifacts [35], we developed a strategy to improve the quality of EEG signals by abandoning EEG signals highly correlated with motion artifacts rather than separating and removing motion artifacts in EEG signals occurring due to body movement [36], [37]. To pursue this strategy, we subdivided the EEG signals into two groups separated by varying the accelerometer data. From each of the two groups, we extracted the following EEG features: 1) mean power; 2) maximum amplitude; 3) standard deviation of the amplitude; 4) kurtosis of the amplitude; and 5) skewness of the amplitude. The features have been widely used to measure the quality of clean EEG signals [38]. After representing the features into 2-D space using the principal component analysis (PCA), we consider the average of the data points of the features in the two groups as a differentiator point. That is, PCA maximizes the Bhattacharyya distance of the projected points on the best fit line from the differentiator point. Then, we only used EEG signals when their features are belonging to the fitted line between cleaned group and the differentiator.

2) Network Settings: ADNet is composed of two-layered networks with 512 and 256 hidden states and has 5×5 size of kernels for the input-to-state and state-to-state transitions. To train the network, we used learning batches of 32 sequences, set the learning rate as 0.01 initially, and divided the rate by ten after every 20 000 iterations. The weight decay and momentum were set to 0.0005 and 0.6, respectively. Backpropagation through time was performed for ten time steps. We also performed early stopping on the validation set. For ACNet, we only selected pairs of the affective sequence $\mathcal{F}$, which were associated with clean physiological signals from all participants with every 10 min as a timestamp (see Fig. 9).

C. Evaluation Results

1) Performance of ALIS on Recognizing Emotions and Identifying Affective Causalities: Table II reports the average precision in recognizing emotions and identifying affective causalities over all participants on the dataset $\mathbb{D}$. The performance on the UVLA state showed the highest results over all participants. This implies that when participants are in a UVLA state, such as calm and relaxed feelings, their physiological signals fluctuate in common patterns, which helps ALIS to learn their characteristics. It can be also attributed that the percentages of affective situations rated with the labels UV and LA were higher than others on the dataset $\mathbb{D}$. On the other hand, the performance in classifying valence ratings associated with low arousal ratings besides LAUV, such as NVLA and PVLA states, was relatively lower than other arousal ratings. In particular, PVLA had the lowest accuracy. This observation may imply that classifying emotions by valence can be improved by considering their associations with arousal. Affective causalities in the 13 situations were identified with a precision of 0.74. Although ACNet identified affective causalities by regarding the prior results of

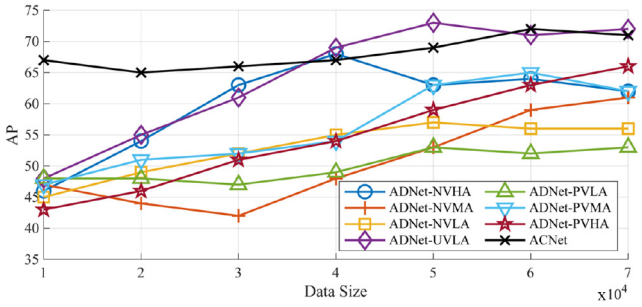

Fig. 10. Performance in emotion recognition and affective causality identification on the real-world dataset $\mathbb{D}$ increasing the dataset size over different emotional states and affective situations.

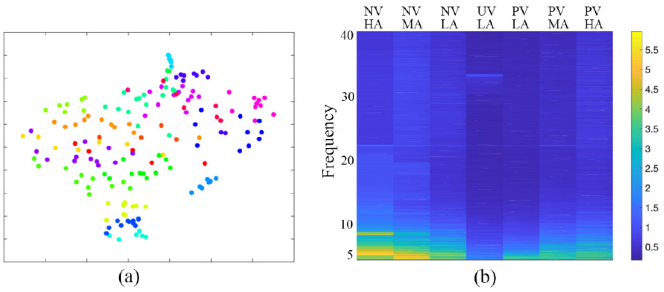

Fig. 11. (a) Clustering of the UVLA states from the participants using t-SNE applied to physiological features $\mathcal{X}$. (b) Shared physiological representation visualized by the grand average of the feature $\mathcal{X}$ over frequency bands in the seven affective states.

ADNet recognizing affective states, the performance was consistently higher than on detecting emotional states. This can be attributed to ACNet working well on eliminating false causal relationships built on wrong emotions.

Unlike the DEAP dataset, our real-world dataset $\mathbb{D}$ records everyday activities with physiological signals. Hence, the interday and intraday variability in physiological signals can determine the performance of ALIS in understanding affective dynamics. Fig. 10 shows the performance in emotion recognition and causality identification on different amounts of physiological data in days. The accuracy generally improved when sufficient data were available. ADNet classified HA states better when more of their associated signals were provided. On the other hand, classifying negative states, such as NVLA and PVLA, showed the smallest improvement with increased data sizes. These results could imply that the affective dynamics in valence requires the development of elaborate deep learning architectures more than the provision of sufficient physiological data. Since causal identification depends on the prior emotion recognition, the performance in emotion recognition is another key in causality. However, this experiment shows that our system does not face this problem. It consistently achieved high scores with small increments.

D. Discussion

1) Analysis of Emotion Recognition: Fig. 11(a) shows the multimodal physiological features $\mathcal{X}$ in a 2-D space obtained using t-distributed stochastic neighbor embedding (t-SNE) [39], a popular technique for unsupervised dimensional reduction and visualization. The features of different participants in an affective state cluster in the projected space, revealing their high variety. Despite their heterogeneity, ADNet is capable of capturing some shared physiological characteristics. To investigate the physiological phenomena observed when emotions are classified using ADNet, we visualize the grand average of brain lateralization features in (3) across participants in valence and arousal ratings. We should note that heart-related features have served as essential elements reflecting the function of ANS. However, in this section, we focus more on brain lateralization, as it has relatively large intersubject and intrasubject variability.

Fig. 11(b) shows the commonly shared frequencies over all participants in the seven emotional states. We found that alpha (7–15 Hz) and beta (15–30 Hz) bands were activated when most emotional states except UVLA and PVLA were elicited. Strong negative-related feelings, such as NVHA and NVMA, led to an increase in physiological changes in alpha and beta. The other two states were characterized by either the theta or gamma band. This indicates that emotional reactions in real-world situations lead to the activation of physiological signals in alpha and beta while the stability in emotion maintains physiological signals in the theta or gamma bands. Although several alternatives have been suggested in reports on the neurophysiological correlates of affective states, this result is in line with our previous work [13].

Our findings may also be justified by the fact that when some negative but approach-related emotions, such as "anger" that would be lateralized to the left hemisphere, are induced, they lead to increase in the alpha band activity in the left anterior and the left temporal regions in the beta bands. We also found that emotional lateralization has also affects on reacting emotions in cases of arousal correlated with valence. When increasing arousal from low to high within the same level of valence states, there are slight increments in the theta band. This observation may reflect the intercorrelations between valence and arousal, as reported in [19].

2) Causality Discovery: To better understand the overall causal pairs in daily life, we report three types of causal networks of the four frequent stress relievers on the dataset $\mathbb{D}$. We choose the four most frequent situations (approximately 67% of all situations): 1) "studying at a desk," 2) "playing games," 3) "drinking coffee/green tea in a cafeteria," and 4) "watching movies." From the 13 situations, the four frequent events were manually browsed in the results of our system by three annotators (interclass correlation $= 0.79$). Given the four situations, we reported the affective causality graph $\mathcal{G}$ resulted from ALIS. Fig. 12 shows the causality structure over all participants and individual causality structures. The top three participants who had the largest data in the real-world dataset have been selected. These results show every one has different causality between emotions and behaviors. For instance, 15 participants ($= 15/21$) had an asymmetric causality from the situation "studying on a desk" to NVLA. Note that the graph does not include cascade flows; namely, in the case of $1 \rightarrow 2 \rightarrow 3$, it has only a causal relation between node $1 \rightarrow 2$ and $2 \rightarrow 3$, but the causality does not propagate $1 \rightarrow 3$.

Participants change their behaviors when they feel specific emotions. First, we found that while studying at a desk, most

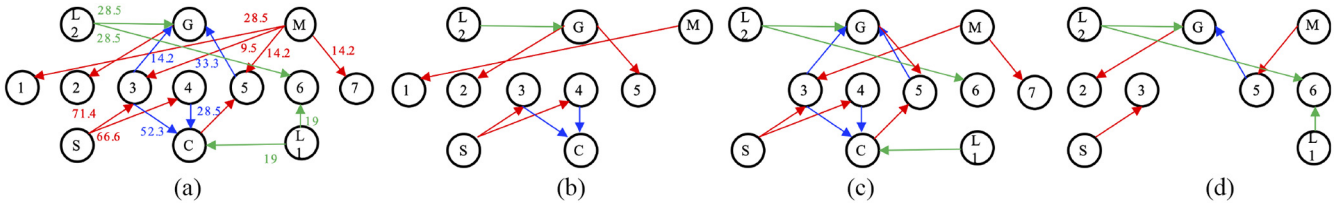

Fig. 12. Causal Structures for (a) all and (b), (c), and (d) the top 3 participants between the four situations and the seven emotional states on the real-world dataset $\mathbb{D}$. Nodes from 1 to 7 represent affective states NVHA, NVMA, NVLA, UVLA, PVLA, PVMA, and PVHA. Situation nodes are denoted as S, G, C, M, and L for *studying on a desk*, *playing games*, *drinking coffee/green tea in a cafeteria*, *watching movies*, and a latent factor. Red lines indicate the asymmetric causality from the situation to the emotion states. Blue lines indicate the opposite causalities. Green lines show the latent factors between the situation and emotional states.

users felt the emotional states NVLA and UVLA, which are close to calm and stress, but any feeling did not affect students' desire to study. This phenomenon indicates that the activity can be a major emotional cause regulating their feelings of being stressed in repeated school routines. In other words, the participants have studied either habitually or for no particular emotional reason, but they were stressed by studying. Interestingly, these types of stressed feelings led them to change their behaviors. Some people had coffee when their feelings rated as having negative valence and low arousal. Furthermore, in line with the affective causality, drinking coffee had a causal effect on overcoming emotional negativity, increasing valence. While most activities cause a given emotion to a particular feeling, "watching a movie" affected multiple emotional states. This can be attributed to individual emotional acceptance of a movie or characteristics of different movie genres. Similarly, when users play games, they feel either NVMA or PVMA emotional states. The polarity in valence from the two emotional states implies playing games is accompanied by emotional elements of fatigue, while it helps to lead positive feeling. From a few users, we found there exist two hidden factors between emotions and situations. When users had the latent factor $L2$, they played the game while feeling excited. Similarly, the latent factor $L1$ bridged users to drink coffee/green tea with a happy feeling. Neither connection had been established without the two factors.

VI. Conclusion

We presented a new wearable system to detect emotional changes and find casual relationships in daily life, based on the new affective model of interaction behavior. By applying our proposed model to a real-world dataset, our approach can find meaningful causal connections between emotions and behaviors, even when confounder variables potentially affect human emotions and behaviors. In the future, we will explore the possibilities of social interaction behavior caused by personal emotional changes. It is also interesting and even more challenging to effectively implement causality identification in the complex human behaviors, such as facial microexpressions [40] and situational analysis of daily life.

Appendix

Proof of Lemma 1

Proof: Suppose η_m and η_c are the influence lags of $\mathcal{M}$ and $\mathcal{C}$, respectively. In the case of 1), the state of $\mathcal{C}_i$ at t time is determined only by its previous states $\mathcal{C}_i^{\eta_c}$. Hence, $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$ naturally holds. In the case of 2), there is no variable η_m to $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$ because the state of $\mathcal{M}_j$ at t time is directly influenced by the previous state of $\mathcal{C}_i$ at $t-1$. ■

Proof of Lemma 2

Proof: For 1), they are dependent on each other in the given condition set without the latent factor because $\mathcal{C}_i$ and $\mathcal{M}_j$ are dependent on the latent factor. For 2), suppose η_c and η_m be the self-influence lag of $\mathcal{C}_i$ and $\mathcal{M}_j$, respectively, the latent factor at time t is independent of the latent factor at time $t-1$. Therefore, $\mathcal{C}_i \perp \mathcal{M}_j^1 | \mathcal{C}_i^{\eta_c}$ and $\mathcal{M}_j \perp \mathcal{C}_i^1 | \mathcal{M}_j^{\eta_m}$ hold. ■

Proof of Theorem 1

Proof: Assuming there are only three directions between two sequences $\mathcal{C}_i$ and $\mathcal{M}_j$: 1) $\mathcal{C}_i \rightarrow \mathcal{M}_j$; 2) $\mathcal{M}_j \rightarrow \mathcal{C}_i$; and 3) $\mathcal{C}_i \leftarrow \mathcal{H} \rightarrow \mathcal{M}_j$. $\mathcal{H}$ is a latent factor. For the first case, $S_1 \wedge \neg S_2 \Rightarrow \mathcal{C}_i \rightarrow \mathcal{M}_j$, and the causal structure of $S_1 \wedge \neg S_2$ based on Lemma 1 cannot be $\mathcal{C}_i \leftarrow \mathcal{M}_j$. There is no confounding factor between the two sequences. Therefore, the causal structure must be $\mathcal{C}_i \rightarrow \mathcal{M}_j$. ■

References

[1] R. M. Carney, K. E. Freedland, and R. C. Veith, "Depression, the autonomic nervous system, and coronary heart disease," *Psychosomatic Med.*, vol. 67, pp. S29–S33, May/Jun. 2005.

[2] R. W. Picard, S. Fedor, and Y. Ayzenberg, "Multiple arousal theory and daily-life electrodermal activity asymmetry," *Emotion Rev.*, vol. 8, no. 1, pp. 62–75, 2016.

[3] B. H. Kim, S. Jo, and S. Choi, "A-Situ: A computational framework for affective labeling from psychological behaviors in real-life situations," *Sci. Rep.*, vol. 10, Sep. 2020, Art. no. 15916.

[4] R. F. Baumeister, K. D. Vohs, C. N. DeWall, and L. Zhang, "How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation," *Pers. Soc. Psychol. Rev.*, vol. 11, no. 2, pp. 167–203, 2007.

[5] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[6] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Exp. Psychiat.*, vol. 25, no. 1, pp. 49–59, 1994.

[7] N.-S. Kwak and S.-W. Lee, "Error correction regression framework for enhancing the decoding accuracies of ear-EEG brain–computer interfaces," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3654–3667, Aug. 2020.

[8] L. Piho and T. Tjahjadi, "A mutual information based adaptive windowing of informative EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 722–735, Oct.–Dec. 2020.

[9] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, "A case-based reasoning model for depression based on three-electrode EEG data," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 383–392, Jul.–Sep. 2020.

[10] Y. Ding, X. Hu, Z. Xia, Y. J. Liu, and D. Zhang, "Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 92–102, Jan.–Mar. 2021.

[11] E. T. Pereira, H. M. Gomes, L. R. Veloso, and M. R. A. Mota, "Empirical evidence relating EEG signal duration to emotion classification performance," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 154–164, Jan.–Mar. 2021.

[12] Y.-J. Suh and B. H. Kim, "Riemannian embedding banks for common spatial patterns with EEG-based SPD neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 854–862.

[13] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 230–243, Apr.–Jun. 2020.

[14] S. Smith, "EEG in the diagnosis, classification, and management of patients with epilepsy," *J. Neurol. Neurosurg. Psychiat.*, vol. 76, no. 2, pp. II2–II7, 2005.

[15] J. W. Choi, S. Huh, and S. Jo, "Improving performance in motor imagery BCI-based control applications via virtually embodied feedback," *Comput. Biol. Med.*, vol. 127, Dec. 2020, Art. no. 104079.

[16] N. Kaongoen, J. Choi, and S. Jo, "Speech-imagery-based brain–computer interface system using ear-EEG," *J. Neural Eng.*, vol. 18, no. 1, 2021, Art. no. 016023.

[17] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Aug. 27, 2019, doi: 10.1109/TAFFC.2019.2937768.

[18] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan.–Mar. 2015.

[19] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[20] X. Zhang *et al.*, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Trans. Cybern.*, early access, May 14, 2020, doi: 10.1109/TCYB.2020.2987575.

[21] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 522–531, Feb. 2015.

[22] H. Chigira, A. Maeda, and M. Kobayashi, "Area-based photoplethysmographic sensing method for the surfaces of handheld devices," in *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, 2011, pp. 499–508.

[23] D. Sun, P. Paredes, and J. Canny, "MouStress: Detecting stress from mouse motion," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI)*, 2014, pp. 61–70.

[24] Y. Lyu *et al.*, "Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI)*, 2015, pp. 857–866.

[25] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics," *Sci. Rep.*, vol. 4, p. 4998, May 2014.

[26] R. Cai, Z. Zhang, Z. Hao, and M. Winslett, "Understanding social causalities behind human action sequences," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1801–1813, Aug. 2017.

[27] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, "Inferring emotional tags from social images with user demographics," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1670–1684, Jul. 2017.

[28] Y. Yang, J. Jia, B. Wu, and J. Tang, "Social role-aware emotion contagion in image social networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 65–71.

[29] T. Zhang, Y. Li, and C. P. Chen, "Edge computing and its role in industrial Internet: Methodologies, applications, and future directions," *Inf. Sci.*, vol. 557, pp. 34–65, May 2021.

[30] T. Zhang, X. Gong, and C. L. P. Chen, "BMT-Net: Broad multitask transformer network for sentiment analysis," *IEEE Trans. Cybern.*, early access, Mar. 4, 2021, doi: 10.1109/TCYB.2021.3050508.

[31] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[32] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," in *Proc. Innov. Appl. Artif. Intell. (IAAI)*, 2017, pp. 4746–4752.

[33] G. Ver Steeg and A. Galstyan, "Information transfer in social media," in *Proc. ACM Int. Conf. World Wide Web (WWW)*, 2012, pp. 509–518.

[34] S. Seth and J. C. Principe, "Assessing granger non-causality using nonparametric measure of conditional independence," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 47–59, Jan. 2012.

[35] I. Daly, M. Billinger, R. Scherer, and G. Müller-Putz, "On the automated removal of artifacts related to head movement from the EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 3, pp. 427–434, May 2013.

[36] I. Daly *et al.*, "What does clean EEG look like?" in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, 2012, pp. 3963–3966.

[37] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul.–Sep. 2019.

[38] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.–Sep. 2014.

[39] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[40] T. Zhang *et al.*, "Cross-database micro-expression recognition: A benchmark," *IEEE Trans. Knowl. Data Eng.*, early access, Apr. 6, 2020, doi: 10.1109/TKDE.2020.2985365.

Byung Hyung Kim received the master's degree in computer science from Boston University, Boston, MA, USA, in 2010, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2018.

He is an Assistant Professor with the Department of Artificial Intelligence, Inha University, Incheon, Republic of Korea. Before he joined Inha University, he was a Research Assistant Professor with the School of Computing, KAIST. His research interests include algorithmic transparency, interpretability in affective intelligence, computational emotional dynamics, cerebral asymmetry and the effects of emotion on brain structure for affective computing, brain–computer interface, and assistive and rehabilitative technology.

Sungho Jo received the B.S. degree in mechanical and aerospace engineering from Seoul National University, Seoul, South Korea, in 1999, and the S.M. degree in mechanical engineering and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2001 and 2006, respectively.

He is a Professor with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. His research interests include intelligent robots, neural interfacing computing, and wearable computing.

Sunghee Choi received the B.S. degree in computer engineering from Seoul National University, Seoul, South Korea, in 1995, and the M.S. and Ph.D. degrees in computer science from the University of Texas at Austin, Austin, TX, USA, in 1997 and 2003, respectively.

She is an Associate Professor with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. Her research interests include geometric computing, computational geometry, geometric modeling, and computer graphics.