Emotion Dictionary Learning with Modality Attentions for Mixed Emotion Exploration

Fang Liu, Pei Yang, Yezhi Shu, Fei Yan, Guanhua Zhang, and Yong-Jin Liu, *Senior Member, IEEE*

Abstract—Multi-modal emotion analysis, as an important direction in affective computing, has attracted increasing attention in recent years. Most existing multi-modal emotion recognition studies are targeted at a classification task that aims to assign a specific emotion category to a combination of several heterogeneous input data, including multimedia signals and physiological signals. Compared to single-class emotion recognition, a growing number of recent psychological evidence suggests that different discrete emotions may co-exist at the same time, which promotes the development of mixed-emotion recognition to identify a mixture of basic emotions. Although most current studies treat it as a multi-label classification task, in this work, we focus on a challenging situation where both positive and negative emotions are presented simultaneously, and propose a multi-modal mixed emotion recognition framework, namely *EmotionDict*. The key characteristics of our EmotionDict include the following. (1) Inspired by the psychological evidence that such a mixed state can be represented by combinations of basic emotions, we address mixed emotion recognition as a label distribution learning task. An emotion dictionary has been designed to disentangle the mixed emotion representations into a weighted sum of a set of basic emotion elements in a shared latent space and their corresponding weights. (2) While many existing emotion distribution studies are built on a single type of multimedia signal (such as text, image, audio, and video), we incorporate physiological and overt behavioral multi-modal signals, including electroencephalogram (EEG), peripheral physiological signals, and facial videos, which directly display the subjective emotions. These modalities have diverse characteristics given that they are related to the central or peripheral nervous system, and the motor cortex. (3) We further design auxiliary tasks to learn modality attentions for modality integration. Experiments on two datasets show that our method outperforms existing state-of-the-art approaches on mixed-emotion recognition.

Index Terms—Emotion distribution learning, multi-modal, mixed emotion, modality attention.

1 Introduction

EMOTION recognition has emerged as an important topic in the affective computing field not only because it is a basis of a wide range of downstream tasks and applications (e.g., media analytical tasks [1], human-computer interaction [2], and mental health treatment [3]), but also because emotion plays a critical role in people's mental states [4]. The emotion space is primarily described by two models: (1) the discrete model that maps an emotional state to a set of basic emotion categories, such as happiness, sadness, surprise, fear, anger, and disgust [5], and (2) the dimensional model that divides the space into valence-arousal (VA) [6], [7] or valence-arousal-dominance (VAD) dimensions [8], where valence indicates whether the emotion is positive or negative, arousal reflects the intensity of the emotion, and dominance refers to whether the subject can control the emotion. Although recent studies have achieved promising emotion recognition results, there still exists an important issue: most works of emotion recognition only identify the dominant emotion from the input signals [9], [10], while existing studies have shown that humans can experience a co-occurrence of two or more emotional feelings with different intensities at the same time [11], [12], [13]. Current single-class emotion recognition studies do not account for the diversity, complexity, and ambiguity of human emotions.

Mixed-emotion recognition is a specialized sub-field of emotion recognition that focuses on identifying and understanding complex emotional states involving multiple emotions simultaneously. A growing number of studies have found that people can experience a co-occurrence of two or more emotional feelings with different intensities [14]. Such mixed emotions are common in everyday scenarios, where individuals often experience conflicting or blended emotions in response to complex life events, relationships, or decisions. Recognizing and understanding these mixed emotions provide more fine-grained insights into the psychological complexities of human emotions. Moreover, it has a range of practical applications in various fields, such as human-computer interaction [15], and healthcare [16].

In this paper, we improve emotion recognition based on the discrete model by incorporating the intensities of multiple basic emotions. As such, mixed-emotion recognition is regarded as an emotion distribution learning (EDL) task. Note that EDL differs from multi-label emotion learning [17], [18], [19], although both tasks address mixed-emotion recognition. The goal of multi-label learning is only to identify whether specific emotions exist or not, while EDL further detects the intensity of each emotion.

Although several attempts at mixed human emotion recognition and EDL have been carried out in the last

- *F. Liu, P. Yang, Y. Shu, and Y.-J. Liu are with BNRist, the Department of Computer Science and Technology, Tsinghua University, and MOE-Key Laboratory of Pervasive Computing, Beijing 100084, China. E-mail: {lfang@, yangpei20@mails., shuyz19@mails., liuyongjin@}tsinghua.edu.cn.*
- *F.Yan is with the School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China. E-mail: yanfei@cust.edu.cn.*
- *G. Zhang is with the Institute for Visualisation and Interactive Systems, University of Stuttgart, Germany. E-mail: guanhua.zhang@vis.uni-stuttgart.de.*
- *Y.-J. Liu is the corresponding author.*

few years, most of them were based on single-modal signals, including facial expressions [20], text [21], eye movements [22], peripheral physiological signals [23], or electroencephalogram (EEG) [9]. Specifically, visual emotion distribution learning aims to recognize multiple emotions with different description degrees evoked by images [24]. Affective text emotional analysis maps text sentences to emotion distributions of multiple emotions and their respective intensities [25]. Another related field is multi-label emotion recognition. Mixed-emotion recognition differs from conventional multi-label emotion recognition mainly in two aspects: (1) Mixed-emotion recognition deals with emotions that are not limited to single, discrete categories (such as happy, sad and angry). Instead, it aims to capture the complexities of emotions that can exist at the same time, which are harder to categorize into traditional emotion labels. (2) Mixed-emotion recognition detects the intensity of each emotion, while multi-label emotion recognition only identifies if each emotion exists.

Existing works based on single-modal signals usually only utilizes one specific type of information while disregarding the fact that emotions are multifaceted subjective feelings including subjective experiences, external manifestations, and physiological responses [26]. Starting from this point, we study multi-modal emotion recognition in this paper. The most related works on multi-modal emotion distribution learning are MEDL [27] and EDL networks [28]. MEDL conducted EDL with audio and video modalities, while the EDL network is built on peripheral physiological signals. In contrast, our method unitizes overt behavioral facial videos and non-behavioral physiological signals, including EEG, photoplethysmogram (PPG), and galvanic skin response (GSR). Though multi-modal signals can offer supplementary emotional information, integrating these multi-modal signals related to different parts of nervous systems (motor cortex, central and peripheral nervous systems) makes our EDL task more challenging.

Our motivation in this paper has two aspects. (1) Since multi-modal information can reflect more aspects of mixed emotions from different perspectives, compared to most existing studies which usually utilize single type of multimedia information to analyze mix emotion, in this paper, we propose to address the mixed-emotion analysis problem as a label distribution learning task with multi-modal information. Subjects' objective feelings are considered to be reflected in their behaviors and physiological performance (also called *physiological arousal*) [29], and we use the fusion of EEG, peripheral physiological signals (i.e., PPG and GSR), and facial videos to conduct a mixed-emotion analysis. (2) We are motivated by the view of the *General Psychoevolutionary Theory of Emotion* that states (i) there is a small number of basic, primary, or prototype emotions; (ii) all other emotions are mixed or derivative states, that is they occur as combinations, mixtures, or compounds of the primary emotions. Moreover, inspired by the latent representation composition and learning widely studied in the deep learning field [30], [31], we propose to model each emotion distribution with a combination of a set of basic latent vectors and their weights. Specifically, we present an emotion distribution learning framework named *EmotionDict*, which aims to learn efficient emotion features for EDL by learning an *emotion dictionary* consisting of a set of basic emotion elements in the latent feature space. In this way, the emotion features of an mixed emotional state could be disentangled into a weighted combination of the set of basic emotion elements and their associated weights in the latent representation space.

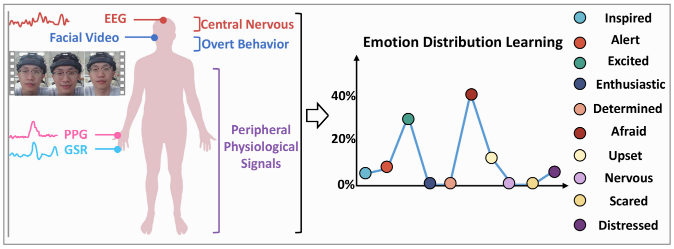

Fig. 1. Multi-modal emotion distribution learning. Existing mixed emotion recognition studies rather focus on multi-label classification or analyzing with overt behaviors, while we aim to integrate both subject's physiological signals and overt behavior (facial videos) to conduct emotion distribution learning in this paper.

One of the most important issues in the multi-modal analysis is to model the contributions of each modality and set up an appropriate fusion mechanism, which is even more difficult in a mixed-emotion situation. For example, as shown in Figure 1, though visual modality is easier to get, it is likely to cause emotional misjudgment, and physiological signals have the advantage of being difficult to disguise. To effectively utilize the diverse information covering subjects' overt behavior and physiological arousal, we have designed a multi-modal integration module, where two auxiliary modality integration tasks are set as attention mechanisms for the emotion dictionary.

The main contributions of this paper are three-fold:

1) We address the challenge of mixed emotion analysis as a distribution learning task utilizing multi-modal signals, including subjects' overt behavior and physiological responses, in which way external manifestations and physiological responses are integrated to offer complementary information for emotion recognition. An end-to-end mixed-emotion recognition model, namely *EmotionDict*, is proposed for the EDL task, which can combine heterogeneous signals to improve the performance of EDL.
2) Inspired by the idea that mixed emotions can be represented by a set of basic emotions, we design an *emotion dictionary* module in our EDL method. This module aims to disentangle emotion features of an emotional state into a weighted combination of a set of basic emotion elements and their associated weights in a latent representation space.
3) We design two auxiliary tasks as the explicit supervision to assign attention to the emotion dictionary, which exploits the feature correlations of multi-modal signals to help learn emotional features and further enhance the emotion distribution learning performance. Moreover, the two auxiliary tasks improve the multi-modal feature fusion process by integrating the consistency and diverse information of the heterogeneous modalities (behavioral and physiological signals).

The task setting of our EmotionDict, which utilizes the combination of subjects' overt behavior (facial videos) and physiological response (EEG, PPG, and GSR) signals as the input to conduct EDL, comes from the inherent characteristics of human emotion. These multi-modal signals can represent different and supplement aspects of human emotions, i.e. behavioral and physiological arousal. Experiments on two datasets show that our proposed EDL model achieves superior performance compared to existing state-of-the-art methods on both subject-dependent and subject-independent protocols. Additional experiments on our method, modified with various multi-modal fusion strategies, have also conducted to further validate the effectiveness of our attention-based emotion dictionary.

2 Related Work

Our work is related to prior works on (1) label distribution learning, (2) human emotion recognition with multi-modal signals, (3) emotion distribution learning, and (4) multi-modal multi-label emotion recognition, which are briefly summarized below.

2.1 Label Distribution Learning

Label distribution learning (LDL) was systematically presented in [32] to address the label ambiguity problem where both the labels and their weights are necessarily needed. Pioneering LDL methods are primarily divided into three types according to the LDL strategies, i.e., Problem Transformation (PT), Algorithm Adaptation (AA), and Specialized Algorithms (SA). Several specific LDL approaches based on deep learning have been developed. A typical example is the deep label distribution learning (DLDL) method [33], which is an end-to-end LDL framework built by utilizing the label ambiguity in both feature learning and classifier learning. Moreover, a lightweight ranking method is presented in [34] to jointly learn age distribution and predict the user's age. Chen et al. [35] proposed a Label Distribution Learning on Auxiliary Label Space Graphs (LDL-ALSG) to leverage the topological information of the labels from related and distinct tasks, such as action unit recognition and facial landmark detection. Different from the above-mentioned general LDL methods, in this paper, we design an end-to-end LDL network by exploiting the comprehensiveness, diversity and ambiguity of multi-modal overt behavioral and physiological signals. Our LDL framework is specifically built for mixed emotions, where an emotion dictionary module is designed to exploit the basis of mixed emotion decomposition.

2.2 Human Emotion Recognition with Multi-Modal Signals

Human emotion recognition has been long studied with subjects' behavior and physiology signals, ranging from facial expressions [36], body language [37], audio [38] to EEG [7] and peripheral physiological signals [39]. Since emotions are complex subjective experiences involving behavioral responses and related to physiological reactions [40], in this paper, people's facial videos and physiological signals are combined to conduct the EDL task. Similar to us, Yang et al. [29] integrate behavioral (i.e. facial expressions, speech, keystroke) and physiological (i.e. blood volume, electrodermal activity, skin temperature) signals by an attention-based LSTM system to conduct emotion recognition, but their goal is to use signals collected from portable devices (i.e. smartphones and wristbands) to detect single emotion category in an unobtrusive manner, while this paper aims to study EDL via the fusion of multi-modal signals. Compared with the physiological signals that are only from the peripheral nervous system in [29], we include EEG responses of the central nervous system in this work to make use of more emotion-related information.

2.3 Emotion Distribution Learning

While the dominant emotion recognition research has been devoted to single-class emotion recognition, EDL has gained increasing attention nowadays. Emotion distribution which is based on the discrete emotion model, is utilized to provide a characterization of emotional states. Existing discrete models contains a diverse numbers of emotional categories. For instance, the classic Plutchik's model [41] contains eight distinct emotional categories, while the newly proposed EDL dataset [28] includes seven basic emotions. In our DMER dataset, ten affect items selected from a simplified version of the original 20-item PANAS [42] are used as fundamental emotions for emotion distribution. These basic emotion categories depict various facets of the emotional experience and are essential and reliable for a comprehensive representation of emotional states.

Major EDL approaches have been developed to recognize emotions and their intensities across various media data [1], [21], [24]. For text data, a multi-task CNN framework is applied to learn text emotion distribution [21]. In the context of visual data, a circular-structured representation has been presented for visual emotion distribution learning in [24]. Additionaly, the Emotion Wheel Attention-based Emotion Distribution Learning (EWA-EDL) model is proposed in [43], which generates a prior emotion distribution describing the relevance of emotional psychology for each basic emotion. Similarly, another EDL framework has been proposedin [44] to investigate the subjectivity in visual emotion distribution and the divergence between individuals. Moreover, Xu et al. [45] investigated the relationships between different image regions and the arousal of each emotion. They propose a region-wise attention-based multi-feature fusion framework for emotion discrete probability distribution prediction.

Moreover, EDL has also been studied with a combination of multiple modalities. For example, MEDL [27] combines audio and video for fine-grained emotion recognition. In addition, Shu et al. [28] use four peripheral physiological signals, i.e., galvanic skin response (GSR), skin temperature (SKT), electrocardiogram (ECG), and heart rate (HR), and establish a CNN-based network for EDL. In this paper, we focus on the mixed emotion analysis of humans, and take advantage of the comprehensive and abundant information offered by multi-modal signals, including EEG from central nervous, PPG, and GSR of peripheral physiological signals, and facial videos of overt behavior, to construct an EDL framework for mixed emotion analysis.

2.4 Multi-Modal Multi-label Emotion Recognition

Another relevant domain is multi-modal multi-label emotion recognition, whose goal is to assign an arbitrary number of emotion category labels to an input sample. Multi-label emotion recognition has also been studied with a wide range of media data, including text [46], [47], facial expressions [17], videos [48], EEG [49], etc. Kostiuk et al. [50] employ multiple classifiers to address the multi-label emotion recognition task from music videos. Li et al. construct a multi-label expression database RAF-ML [17] and propose a deep bi-manifold convolutional network for multi-label classification from expressions. In addition, multi-modal signals have been also ensembled detecting multiple emotions. The MESGN, proposed in [51], is a variation of the transformer network designed for multi-label emotion detection using textual, visual, and acoustic modalities. Similarly, Zhang et al. [19] fuse text, visual, and audio modalities by integrating transformer networks and multi-head modality attention to predict emotions. More recently, Yu et al. [52] design a self-supervised multi-task learning strategy to enhance the consistency and difference of multi-modal emotion feature representations. Based on textual, acoustic, and visual modalities, Zhang et al. [18] present a heterogeneous hierarchical graph message passing network to simultaneously model the feature-to-label, label-to-label, and modality-to-label dependencies.

Stepping forward from the above studies, our multi-modal EDL approach not only considers a set of emotion categories, but also detects the intensity of each category.

3 Method

In this section, we first define the EDL task of this paper and then introduce our proposed *EmotionDict* framework, which mainly consists of a feature preprocessing module, a multi-modal feature encoder, an emotion dictionary module, two auxiliary tasks, and an attention-based classifier. The overall architecture of EmotionDict is shown in Figure 2. The feature encoder module contains one public Transformer [53]-based multi-modal encoder, which embeds all four modalities (EEG, PPG, GSR, and video), and four modality-specific encoders, each dedicated to one of the four signal types. The emotion dictionary module is the key component of our EmotionDict, which learns an emotion feature representation with a set of basic emotion vectors and their corresponding weights in a latent representation space. Furthermore, two auxiliary tasks are designed to utilize the relationships between overt behavior and physiological signals to offer modality-attention supervision for the emotion dictionary module. Finally, an attention-based classifier is utilized to predict the emotion distribution.

3.1 Preliminary and Task Definition

We follow the task definition of previous emotion distribution learning studies in [20], [28], [32]. Let $Y = \{y_1, y_2, ..., y_L\}$ be the set of predefined emotion labels, where L is the total number of emotion categories.

Given a training set $S = \{(x_i, D_{x_i})|i = 1, ..., N\}$, where $D_{x_i} = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, ..., d_{x_i}^{y_L}\}$ is the emotion distribution corresponding to the sample x_i indicating a signal in a time window. N means the total number of samples. $d_{x_i}^{y_j}$ represents that, to which extent the class y_j describes the emotion state of sample x_i. Therefore, it satisfies: (1) $0 \leq d_{x_i}^{y_j} \leq 1$; (2) $\sum_{j=1}^{L} d_{x_i}^{y_j} = 1$.

Furthermore, the sample data x_i is composed of EEG, PPG, GSR, and facial video signals, i.e., $x_i = \{x_i^{\text{EEG}}, x_i^{\text{PPG}}, x_i^{\text{GSR}}, x_i^{V}\}$. Our multi-modal EDL aims at finding a label distribution $\hat{D}_{x_i}$=$\{\hat{d}_{x_i}^{y_1}, \cdots, \hat{d}_{x_i}^{y_L}\}$ that is optimally close to D_{x_i} from the multi-modal input x_i. Specifically, our goal is to learn a conditional probability mass function $p(\hat{D}_{x_i}|x_i; \theta)$, where θ is the learnable network parameter set. We compute the optimal parameter θ^* by solving the problem as follows:

$$\theta^* = argmin_{\theta} \sum_{i=1}^{N} Dist(\hat{D}_{x_i}, D_{x_i}) \tag{1}$$

where $Dist(\cdot)$ is the distance function measuring the similarity of the predicted emotion distributions and the ground truth distributions.

3.2 Data Preprocessing and Feature Extraction

We first employ signal preprocessing and feature extraction operations to the original multi-modal raw data (refer to Section 4.2 for more details).

EEG preprocessing and feature extraction. We extract two of the most widely used EEG features from the EEG signals, i.e. the Power Spectral Density (PSD) [10] and the Differential Entropy (DE) [54] features. Specifically, a bandpass filter (1-50 Hz) and a 50 Hz notch filter are first used to remove noise, then an independent principal component analysis are conducted to eliminate artifacts. The signals are down-sampled to 100 Hz to speed up the computation. A time sliding window of 1 second is set for the PSD and DE feature extraction. Within each window, the Short Time Fourier Transform (STFT) is used to extract the PSD and DE features in each EEG channel over five frequency bands, which are δ band: 1-4 Hz, θ band: 4-8 Hz, α band: 8-14 Hz, β band: 14-31 Hz and γ band: 31-45 Hz.

Peripheral physiological signal preprocessing and feature extraction. The GSR signals are first filtered with a bandpass filter with a lower cutoff frequency of 0.01 Hz and a higher cutoff frequency of 49 Hz, and the PPG signals are filtered with a bandpass filter of 0.01 Hz-1.9 Hz. Following the PPG and GSR feature extraction procedures in [28], [55], we extract a total number of 27 and 28 features from PPG and GSR, respectively. These features are statistical features from both the time domain and frequency domain. The sliding window for peripheral physiological signal feature extraction is set as 5 seconds with an overlap of 4 seconds between two consecutive windows. The full details of PPG and GSR features are listed in Table A1 in Appendix A.

Facial video preprocessing and feature extraction. For facial videos, we first conduct face detection on 1 out of 30 video frames collected in 1s. Then the facial images are resized to a resolution of 128×128. We further extract a 32-dimensional feature for each image using VGG-16 [56] pre-trained on ImageNet [57].

Time alignment. Finally, we conduct time alignment of the four modalities according to the end of each trial. As a 5-second sliding window with an overlap of 4 seconds

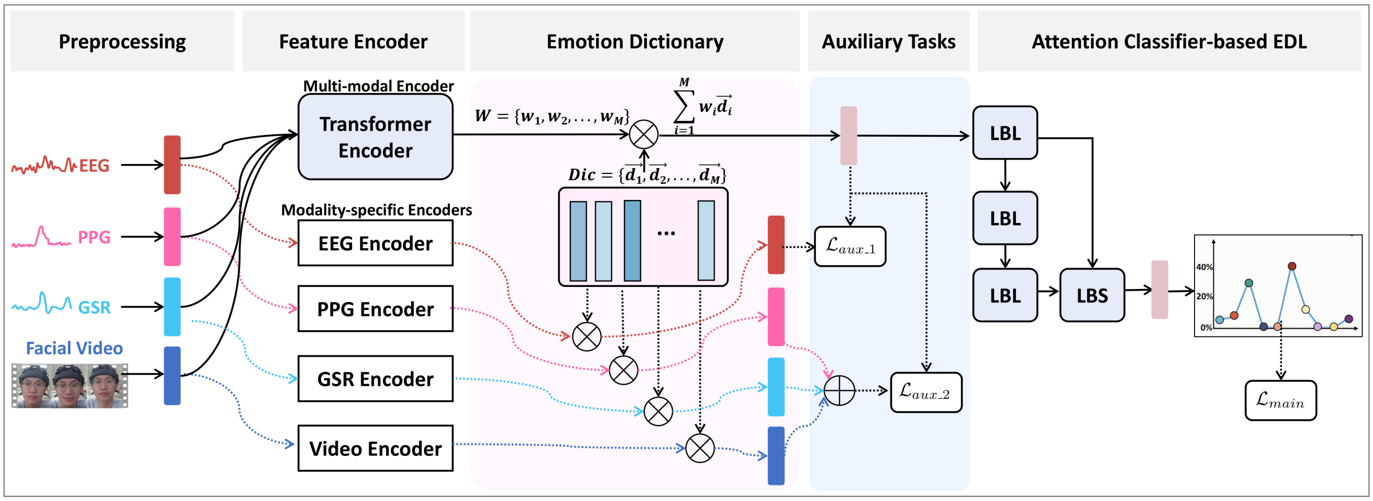

Fig. 2. The framework of our proposed EDL method. We first apply an *emotion dictionary* module which decomposes mixed emotion into the combination of a set of basic emotion vectors and their weights for emotion representation learning (see Section 3.3). Then we set multi-modal integration as auxiliary tasks to offer attention to the emotion dictionary, to enhance the overall EDL performance. At last, an attention-based classifier, which is composed of three **LBL** sub-modules and one **LBS** sub-module, is applied to obtain the final predicted emotion distribution. Both the **LBL** and **LBS** sub-modules are constructed with **L**inear (fully-connected) layers, **B**atch normalization layers, and activation functions.

between two consecutive windows has been set for the PPG and GSR feature extraction and the sliding window for the EEG and facial video feature extraction is 1 second with no overlap, we time-align all features using the backend alignment method, where the EEG features of the first 4 seconds of each trial are discarded.

In this way, we get a sample input x composed of four components, i.e., $x^{\text{EEG}} \in R^{t \times n_{\text{EEG}} \times d_{\text{EEG}}}$, $x^{\text{PPG}} \in R^{t \times n_{\text{PPG}} \times d_{\text{PPG}}}$, $x^{\text{GSR}} \in R^{t \times n_{\text{GSR}} \times d_{\text{GSR}}}$, $x^{V} \in R^{t \times d_V}$, where t is the length of the time window, n_{EEG}, n_{PPG}, n_{GSR} and n_V are the number of signal channels, and d_{EEG}, d_{PPG}, d_{GSR} and d_V are the corresponding dimensions of the extracted features. The four types of features are concatenated and reshaped to form a fused representation $\mathcal{X}$.

3.3 EDL with Latent Emotion Dictionary

On one hand, prior works [58] have shown that complex emotions are constructed by a set of basic emotions. On the other hand, several recently proposed studies in the deep learning field leverage the concept of "latent representation decomposition" for better representation learning [31], [59], [60]. Our introduction of the emotion dictionary is deeply inspired by the effective linear deep feature decomposition mechanism in the computer vision field [30]. Similar to the basic emotions that make up the mixed emotions, we hypothesize that there also exists a set of basic emotion elements in the latent space of emotional features that can be used to represent any emotion state. Thus, instead of directly learning entangled and unexplainable multi-modal feature representations from $\mathcal{X}$, we assume that the latent representation h of any emotion state can be represented by a linear combination of a set of *basic emotion elements* with different weights:

$$h = \sum_{i=1}^{M} w_i \vec{d_i} \tag{2}$$

where $\vec{d_i} \in Dic = \{\vec{d_1}, \vec{d_2}, ..., \vec{d_M}\}$ represents a set of basic emotion elements in the latent space, $w_i \in W = \{w_1, w_2, ..., w_M\}$ is the weight of element $\vec{d_i}$ in the emotion dictionary. The weights were obtained via a Transformer encoder [53] applied on the preprocessed multi-modal sample $\mathcal{X}$:

$$W = \text{Transformer}(\mathcal{X}) \tag{3}$$

In the latent space, we apply the QR decomposition to obtain the emotion dictionary matrix **Dic**, where the vectors in the matrix **Dic** are all orthogonal. That is, the inner product of d_i and d_j is 1 if $i = j$, and otherwise is 0. Specifically, **Dic** $\in \mathcal{R}^{K*M}$, where K is the size of $\vec{d_i}$, which keeps consistency with the dimension of the Transformer encoder output. Both K and M are hyper-parameters, and we set them as 128 and 32 respectively in our experiments, reaching a stable EDL performance. The influence of different numbers of the basic emotion elements M on the performance of our method is shown in Section 4.5.5.

The emotion dictionary is trained with the other modules of the entire network to learn the set of basic emotion elements. In the training process, weights are computed based on input multi-modal signals, and the corresponding emotion distributions serve as supervision. Once the network training process is completed, the set of basic emotion elements remains fixed. During the testing stage, the basic emotion elements are used to embed the input multi-modal features, enabling the final emotion distribution learning.

Moreover, we integrate the emotion dictionary with a self-attention classifier module (as shown in the right part of Figure 2), which is composed of three **LBL** sub-modules and one **LBS** sub-module. Both the **LBL** and **LBS** sub-modules are constructed with **L**inear (fully-connected) layers, **B**atch normalization layers, and activation functions (i.e., the **L**eaky ReLU function for **LBL** and the **S**igmoid function for **LBS**). These two kinds of network sub-modules are denoted as:

$$\mathbf{LBL}(x) = \sigma_1(BN(FC(x))) \tag{4}$$

$$\mathbf{LBS}(x) = \sigma_2(BN(FC(x))) \tag{5}$$

where $FC(\cdot)$ is the fully-connected layer, $BN(\cdot)$ is the batch normalization, σ_1 and σ_2 are two activation functions, i.e., Leaky ReLU and Sigmoid.

The main function of the **LBS** is to conduct feature extraction, and the **LBS** sub-module is set as the attention to **LBL**, so as to normalize the feature to $[0, 1]$ for final classification.

Then, the latent representation h is fed to the self-attention module:

$$h^1 = \mathbf{LBL}(h) \tag{6}$$

$$h^2 = \mathbf{LBS}(\mathbf{LBL}(\mathbf{LBL}(h^1))) \tag{7}$$

where h^2 is the attention weights learned from the **LBS** sub-module and is normalized to $[0, 1]$.

Then the embedding extracted by the first **LBL** sub-module (i.e., h^1) is multipled by the attention weights h^2:

$$o = (o_1, \cdots, o_L) = h^1 * h^2 \tag{8}$$

where $o = (o_1, \cdots, o_L)$ is output of the last layer of the attention-based classifier module before the softmax layer. Finally, the predicted output $\hat{D}_x$=$(\hat{d}_x^{y_1}, \hat{d}_x^{y_2}, \cdots, \hat{d}_x^{y_L})$ are obtained by normalizing the features of the last layer in Equation 8 using a softmax function, i.e.,

$$\hat{d}_x^{y_i} = Softmax(o_i) = \frac{e^{o_i}}{\sum_{c=1}^{L} e^{o_c}} \tag{9}$$

The Kullback-Leibler divergence (KLD) [61] is adopted as the loss function to measure the distance between the predicted distribution and the ground truth:

$$\mathcal{L}_{main} = KLD(\hat{D}_x, D_x) = \sum_{j=1}^{L} d_x^{y_j} log \frac{d_x^{y_j}}{d_x^{\hat{y_j}}} \tag{10}$$

where D_x=$(d_x^{y_1}, d_x^{y_2}, \cdots, d_x^{y_L})$ are the ground truth of the emotion distribution.

3.4 Modality Integration as Attentions

To further refine and improve the emotion dictionary learning, we propose a multi-modal alignment module as auxiliary tasks (see the "Auxiliary Tasks" module in Figure 2). We have categorized the modality-specific encoders into two groups for the two auxiliary tasks: one for EEG and the other for the remaining three modalities. This classification stems from the understanding that emotions encompass subjective feelings, overt behavior, and physiological responses [62]. EEG signals, being activations of the central nervous system (CNS), significantly influence subjects' overt behavior (e.g., facial videos) and physiological signals (e.g., PPG and GSR signals). Moreover, our emotion dictionary is based on the assumption that any emotion can be represented by the combination of the learned emotion dictionary and the corresponding weights in the latent space. Therefore, in addition to the fusion of multi-modal emotion features, the latent features of individual modalities, as well as combinations of several modalities, can also be effectively represented using the learned emotion dictionary.

Based on the above considerations, we set two modality alignment and integration-based auxiliary tasks as the attention mechanism for emotion dictionary learning. The tasks are (1) emotion representation learning of EEG, and (2) emotion representation learning of joint behavioral and peripheral physiological signals.

Specifically, for the pre-extracted features of each modality x^{EEG}, x^{PPG}, x^{GSR}, x^{V} (see Section 3.2), we use separated modality-specific encoders to learn the modality-specific emotion dictionary weights for each modality as:

$$W^k = \mathbf{E}_k(x^k), k \in \{\text{EEG}, \text{PPG}, \text{GSR}, \text{V}\} \tag{11}$$

where $\mathbf{E}_k(\cdot)$ are four modality-specific feature encoders, and $W^k = \{w_1^k, w_2^k, ..., w_M^k\}$ are the weights of basic emotion elements in the emotion dictionary for the four modalities. Four different Transformer encoders are adopted (with separate parameters for each modality) as the backbones. Similar to Equation 2, we obtain the latent features of the four modalities as:

$$h^k = \sum_{i=1}^{M} w_i^k \vec{d}_i \tag{12}$$

where $h^k (k \in \{\text{EEG}, \text{PPG}, \text{GSR}, \text{V}\})$ are emotion feature representations learned by the emotion dictionary for the four modalities.

The main idea of the two auxiliary tasks is that all the feature representations of the signals of (i) the central nervous system (EEG), (ii) the peripheral nervous system and the motor cortex (PPG, GSR, and facial video), (iii) the central and peripheral nervous systems, and the motor cortex (EEG, PPG, GSR, and facial video), have similar distributions after encoded by the emotion dictionary in the latent space. Thus, the two auxiliary loss functions are:

$$\mathcal{L}_{aux\_1} = \mathbf{Dis}(h, h^{EEG}) \tag{13}$$

$$\mathcal{L}_{aux\_2} = \mathbf{Dis}(h, h^{PPG} + h^{GSR} + h^{V}) \tag{14}$$

where the $\mathbf{Dis}(\cdot)$ is the distance function between two feature representations. In our model, the Kullback-Leibler divergence [61] is adopted for measuring the similarity between two features.

Finally, the overall learning of our EmotionDict model is performed by minimizing:

$$\mathcal{L} = \mathcal{L}_{main} + \mathcal{L}_{aux\_1} + \mathcal{L}_{aux\_2} \tag{15}$$

The main loss $\mathcal{L}_{main}$ is responsible for achieving the desired emotion dictionary, and the auxiliary losses are effective for modality integration. We will discuss them in Section 4.5.3.

The training procedure of our proposed EmotionDict emotion distribution learning framework is shown in Algorithm 1. More details about the experiment setting are shown in Section 4.2.

Algorithm 1 Training of EmotionDict

```
Require:
    Input:
    The training set of current batch $S = \{(x_i, D_{x_i})|i = 1, ..., N\}$, where $D_{x_i} = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \cdots, d_{x_i}^{y_L}\}$ is the emotion distribution corresponding to the sample $x_i$;
    The predefined emotion labels set $Y = \{y_1, y_2, ..., y_L\}$;
    The total number of samples$N$.
    Parameter:
    The initialization of the parameters of the whole model $\theta = \{\theta_{\textbf{En}}, \theta_{\textbf{Dic}}, \theta_{\textbf{Classifier}}\}$, where $\theta_{\textbf{En}}$ contains parameters for a multi-modal encoder and four modality-specific encoders, i.e. $\theta_{\textbf{En}} = \{\theta_{\textbf{E}}, \theta_{\textbf{E}_{\textbf{EEG}}}, \theta_{\textbf{E}_{\textbf{PPG}}}, \theta_{\textbf{E}_{\textbf{GSR}}}, \theta_{\textbf{E}_{\textbf{V}}}\}$;
    The hyper-parameters, such as the number of basic emotion elements in the emotion dictionary and the size of each basic emotion element, the learning rate.
    Output:
    The parameters of the whole model $\theta$.
1: Randomly initialize $\theta$
2: for all samples do
3:     Evaluate $\mathcal{L}_{main} = KLD(\hat{D}_x, D_x)$ with multi-modal features from the multi-modal encoder with parameters $\theta_{\textbf{E}}$
4:     Evaluate $\mathcal{L}_{aux\_1}$ with multi-modal features from the multi-modal encoder with parameters $\theta_{\textbf{E}}$ and features from the EEG modality-specific encoder with parameters $\theta_{\textbf{E}_{\textbf{EEG}}}$
5:     Evaluate $\mathcal{L}_{aux\_2}$ with multi-modal features from the multi-modal encoder with parameters $\theta_{\textbf{E}}$ and features from the PPG, GSR and video modality-specific encoders with parameters $\theta_{\textbf{E}_{\textbf{PPG}}}, \theta_{\textbf{E}_{\textbf{GSR}}}, \theta_{\textbf{E}_{\textbf{V}}}$
6: end for
7: Update the parameters $\theta$ of the EmotionNet with $\mathcal{L}_{main}$, $\mathcal{L}_{aux\_1}$, and $\mathcal{L}_{aux\_2}$
8: Go to step 2 until convergence, or the algorithm reaches the maximum number of epochs
9: Return $\theta = \{\theta_{\textbf{En}}, \theta_{\textbf{Dic}}, \theta_{\textbf{Classifier}}\}$ as the trained parameters
```

4 Experiments

We evaluate our model on two recent and publicly available datasets: DMER [55] and EDL [28] datasets. We first briefly introduce the two datasets, then describe the implementation details and the evaluation metrics. Finally, we compare our proposed model with eleven state-of-the-art methods on the two datasets with both subject-dependent and subject-independent protocols.

4.1 Datasets

Most publicly available emotion datasets either contain single-modal signals or multi-category labels instead of distribution labels. Exceptions are two lately proposed multi-modal emotion distribution learning datasets, namely DMER [55] and EDL [28], which are summarized in Table 1. They evoke emotion through audio-visual (AV) materials and recorded multi-modal signals.

TABLE 1
Summary of the two datasets used in our experiments.

Dataset	Subjects	Emotion Classes	Modalities
DMER [55]	28	10	EEG, PPG, GSR, Video
EDL [28]	38	7	ECG, HR, GSR, SKT

DMER dataset [55] contains four modalities (EEG, GSR, PPG and frontal facial videos) collected from 28 subjects. Each subject watched 32 video clips, which elicited a mixed-emotion state composed of 10 basic emotions. Five of them (inspired, alert, excited, enthusiastic and determined) can be further categorized as positive emotions and the other five (afraid, upset, nervous, scared, and distressed) as negative emotions. The length of each video clip is 20 s-30 s. Each subject conducted four groups of experiments, each containing 8 trials. In each trial, the subject first watched one video clip, then completed the self-report for the emotional adjectives (10-item short positive affect and negative affect schedules (PANAS [63])). The score for each basic emotion ranged from 1 (none at all) to 5 (very strong) and was then transformed into emotion distributions. The positions of the 10 emotional words were randomly assigned. Then there was a 5-second break before starting the next trial.

The three physiological modalities (EEG, PPG, and GSR) were collected via portable devices. 21-channel EEG signals were recorded at a sampling rate of 300 Hz, while the sampling rates of PPG and GSR were 4 Hz and 100 Hz, respectively. Facial videos were recorded by a built-in camera with a resolution of 640×480 and a frame rate of 30 fps.

All procedures of the study were reviewed and approved by the Ethics Committee of Tsinghua University. Before the experiment, all participants were thoroughly briefed about the experiment, including its purpose, procedure and the utilization of the collected data.

EDL dataset [28] is an emotion distribution dataset, containing four types of peripheral physiological signals (ECG, HR, GSR, and SKT) from 38 subjects. The distribution was generated from seven basic emotions (anger, disgust, sadness, fear, tenderness, joy, and amusement), which were evoked by 14 emotional video clips. Each trial in the experiment procedure mainly consists of four steps: (1) experiment instruction display; (2) a 1 min go/no go task served as a distraction to eliminate the effects of previous emotion; (3) an 80-second rest; (4) video clip presentation. Videos were arranged in random order and labeled according to the subjects' self-reports. The labels were further transformed into emotion distributions. All the signals were recorded using an MP150 data recording system (BIOPAC Systems Inc.).

4.2 Implementations

Data processing in DMER. The multi-modal data preprocessing and feature extraction details are shown in Section 3.2. After feature extraction and time alignment of the four modalities, the shapes of the extracted features are $t_i \times 18 \times 5$, $t_i \times 1 \times 27$, $t_i \times 1 \times 28$, and $t_i \times 32$ for EEG, PPG, GSR, and facial video, respectively, where t_i indicates the time length of the ith trial and $i \in \{1, 2, ..., 32\}$. Then zero-padding is applied to the PPG, GSR, and facial video features by adding zeros at the end of each feature to ensure uniform dimensions. The shapes of the EEG, PPG, GSR, and facial video features are $t_i \times 18 \times 5$, $t_i \times 1 \times 36$, $t_i \times 1 \times 36$, and $t_i \times 36$ respectively.

Data Preprocessing and Feature Extraction in EDL. We follow the data preprocessing, feature extraction, and feature selection in [28]. The GSR, SKT, and ECG signals are filtered by a second-order Butterworth filter. Then a total number of 89 features (39 from GSR, 4 from SKT, 39 from ECG and 7 from HR) have been extracted. Then 50 features have been selected and used in the EDL algorithms. The shapes of the extracted features are $t_i \times 39$, $t_i \times 4$, $t_i \times 39$, and $t_i \times 7$ for GSR, SKT, ECG, and HR respectively, where t_i indicates the time length of the ith trial and $i \in \{1, 2, ..., 14\}$. Detailed configurations of the network layers of our EmotionDic framework and the corresponding output dimensions of samples from the DMER and EDL datasets are listed in Table 2.

TABLE 2
Network details. We present the details of the network layers in each module of our EmotionDic framework, as well as the dimensions of the latent embeddings of the samples from the DMER and EDL datasets. "BN" represents the 1d batch-normalization layer. "B" represents batch-size.

Module	Layer	Output Shape	
		DMER	EDL
Feature Encoder	Input concatation & transpose	[11, B, 18]	[89, B, 1]
	Linear	[11, B, 24]	[89, B, 24]
	Positional Encoding	[11, B, 24]	[89, B, 24]
	Transformer (hidden-dim 32, multi-head num=4, layer num=3)	[11, B, 24]	[89, B, 24]
	Linear	[B, 32]	[B, 32]
Emotion Dictionary	Emotion_Dict (weight shape [32,128])	[B, 128]	[B, 128]
Attention Classifier	Linear+BN+LeakyReLU	[B, 128]	[B, 128]
	Linear+BN+LeakyReLU	[B, 64]	[B, 64]
	Linear+BN+LeakyReLU	[B, 128]	[B, 128]
	Sigmoid	[B, 128]	[B, 128]
	Linear	[B, 10]	[B, 7]
	Softmax	[B, 10]	[B, 7]

Model Training Protocol. We apply both subject-dependent and subject-independent evaluations on the two datasets. For subject-dependent evaluation, data of each subject are spilt into the training set (80%) and testing set (20%) randomly. For subject-independent evaluation, we conduct leave-one-subject-out cross-validation, where the data from one subject is used as the testing set and data from the others is used as the training set in each fold. The final performance is reported by the average across all the folds. The whole network is trained using Adam optimizer [64] in an end-to-end manner for 100 epochs. The learning rate was set to 0.001 and the batch size is 32. Our algorithm is implemented in PyTorch and our experiments are carried out on a Tesla A100 GPU. The training time is around 1 h for each subject on the DMER dataset. For the auxiliary task construction in the EDL dataset (Section 3.4), the four peripheral physiological signals (ECG, HR, GSR, and SKT) are joined to construct the L_{aux_2}, and there is no L_{aux_1} since EEG modality is not contained in the EDL dataset in our experiments.

4.3 Evaluation Metrics

We adopted six distribution-based measurements that are commonly used in emotion distribution learning [20], [24], [28], [32], including four distance metrics (Chebyshev (↓), Clark (↓), Canberra (↓), and Kullback-Leibler (KL) (↓)) and two similarity measurements (Cosine (↑) and Intersection (↑)). ↓ indicates the metric is "the smaller the better", and ↑ means "the larger the better". Moreover, we reported *Average Rank* denoting the mean rank of the six metrics following [24], [28].

4.4 State-of-the-art Methods

To evaluate the effectiveness of our method, we conducted extensive experiments to compare with eleven state-of-the-art methods on the DMER and EDL datasets. These baseline methods can be categorized into four types:

- **Label Distribution Learning Methods.** Classical LDL methods could be further sorted into three categories: Problem Transformation (PT), Algorithm Adaptation (AA), and Specialized Algorithms (SA). (1) **PT-Bayes** [32] and **PT-SVM** [32] are based on Bayes and SVM classifiers with problem transformation strategy, which converts LDL to single-label distribution problems via transforming the training data into weighted single-label data. (2) **AA-BP** [32] and **AA-KNN** [32] are two adaptation strategy-based methods, which extend classic machine learning algorithms, i.e., k-Nearest Neighbor (k-NN) and Back Propagation (BP) neural network, to handle LDL problems. (3) **SA-IIS** [32] and **SA-CPNN** [65] are specialized algorithms for LDL. SA-IIS assumes the maximum entropy model as the parametric model, and SA-CPNN is built on a three-layer conditional probability neural network.
- **Multi-modal Multi-label Emotion Recognition Methods (denoted as MMER).** Since our multi-modal EDL problem can be seen as an extension of the multi-modal multi-label emotion recognition task, we also compare it with two representative multi-label emotion recognition baselines. **TAILOR** [66] addresses the commonality and diversity among multiple modalities and enhances the discriminative capability of label representations via adversarially extracting private and common modality representations. **MISA** [67] projects each modality to modality-invariant and modality-specific subspaces for multi-modal representation learning. Both the two baselines have obtained good performance on visual, audio, and text modalities. Before comparing our method with these multi-modal multi-label recognition baselines, we apply a softmax function to the output of the last layer of the networks to obtain the predicted emotion distributions.
- **Single-modal Emotion Distribution Learning Methods (denoted as Single-modal EDL).** We also compare the performance of our approach with two state-of-the-art single-modal emotion distribution learning methods. **EDL-LRL** [20] captures local-level label correlations to tackle the emotion label distribution learning problem on facial expressions. **DLDL** [33] addresses label ambiguity in both feature learning and classifier learning, which learns the label distribution by minimizing a Kullback-Leibler divergence between the

TABLE 3
Subject-dependent comparison of experimental results of our method and 11 baseline algorithms on six measures on the DMER and EDL datasets. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The best results are in **bold**,and the parentheses show the corresponding ranks on each evaluation metric and the *Average Ranks*.

Dataset	Measure	PT		AA		SA		MMER		Single-modal EDL		Multi-modal EDL	
		PT-Bayes	PT-SVM	AA-KNN	AA-BP	SA-IIS	SA-CPNN	TAILOR	MISA	EDL-LRL	DLDL	EDLConV	**Ours**
DMER	Chebyshev (↓)	0.7733(12)	0.1025(11)	0.0789(5)	0.0890(10)	0.0786(4)	0.0871(9)	0.0819(7)	0.0807(6)	0.0769(2)	0.0785(3)	0.0823(8)	**0.0717(1)**
	Clark (↓)	2.8673(12)	0.7020(11)	0.5409(2)	0.6490(10)	0.5556(5)	0.6100(9)	0.5968(8)	0.5672(6)	0.5411(3)	0.5423(4)	0.5706(7)	**0.4729(1)**
	Canberra (↓)	8.9723(12)	1.8544(11)	1.4263(2)	1.7358(10)	1.5049(5)	1.7313(9)	1.6380(8)	1.5344(7)	1.4632(3)	1.4785(4)	1.5089(6)	**1.2297(1)**
	KL (↓)	12.6201(12)	0.1378(11)	0.0831(7)	0.1133(10)	0.0777(4)	0.0979(9)	0.0788(6)	0.0707(2)	0.0742(3)	0.0778(5)	0.0854(8)	**0.0592(1)**
	Cosine (↑)	0.4003(12)	0.8776(11)	0.9237(7)	0.9102(10)	0.9288(6)	0.9119(9)	0.9335(3)	0.9418(2)	0.9320(5)	0.9289(4)	0.9202(8)	**0.9419(1)**
	Intersection (↑)	0.1961(12)	0.8025(11)	0.8500(3)	0.8242(9)	0.8445(6)	0.8228(10)	0.8410(7)	0.8535(2)	0.8488(4)	0.8471(5)	0.8406(8)	**0.8686(1)**
	Average Rank (↓)	12(12)	11(11)	4.33(5)	9.83(10)	5(6)	9.16(9)	6.5(7)	4.16(3)	3.33(2)	4.16(3)	7.5(8)	**1(1)**
EDL	Chebyshev (↓)	0.7568(12)	0.2193(10)	0.1919(5)	0.2245(11)	0.1970(7)	0.1940(6)	0.2111(9)	0.2027(8)	0.1833(2)	0.1849(3)	0.1857(4)	**0.1813(1)**
	Clark (↓)	2.4055(12)	1.0603(10)	0.8819(6)	1.0657(11)	0.8849(8)	0.8808(5)	0.8957(9)	0.8823(7)	0.8274(2)	0.8446(3)	0.8495(4)	**0.8201(1)**
	Canberra (↓)	6.2638(12)	2.4477(10)	2.0920(5)	2.4484(11)	2.1441(6)	2.1452(7)	2.1512(9)	2.1494(8)	**1.9785(1)**	2.0097(2)	2.0154(3)	2.0195(4)
	KL (↓)	12.0230(12)	0.4034(10)	0.2931(9)	0.4255(11)	0.2626(7)	0.2413(4)	0.2765(8)	0.2538(6)	0.2238(2)	0.2385(3)	0.2426(5)	**0.2093(1)**
	Cosine (↑)	0.3862(12)	0.7276(11)	0.7739(8)	0.7359(10)	0.7979(6)	0.8056(5)	0.7722(9)	0.7940(7)	0.8199(2)	0.8102(4)	0.8065(3)	**0.8333(1)**
	Intersection (↑)	0.2182(12)	0.6463(11)	0.6870(7)	0.6537(10)	0.6898(6)	0.6912(5)	0.6798(9)	0.6861(8)	**0.7131(1)**	0.7068(3)	0.7049(4)	0.7115(2)
	Average Rank (↓)	12(12)	10.33(10)	6.67(6)	10.67(11)	6.67(6)	5.33(5)	8.83(9)	7.33(8)	**1.66(1)**	3(3)	3.83(4)	**1.66(1)**

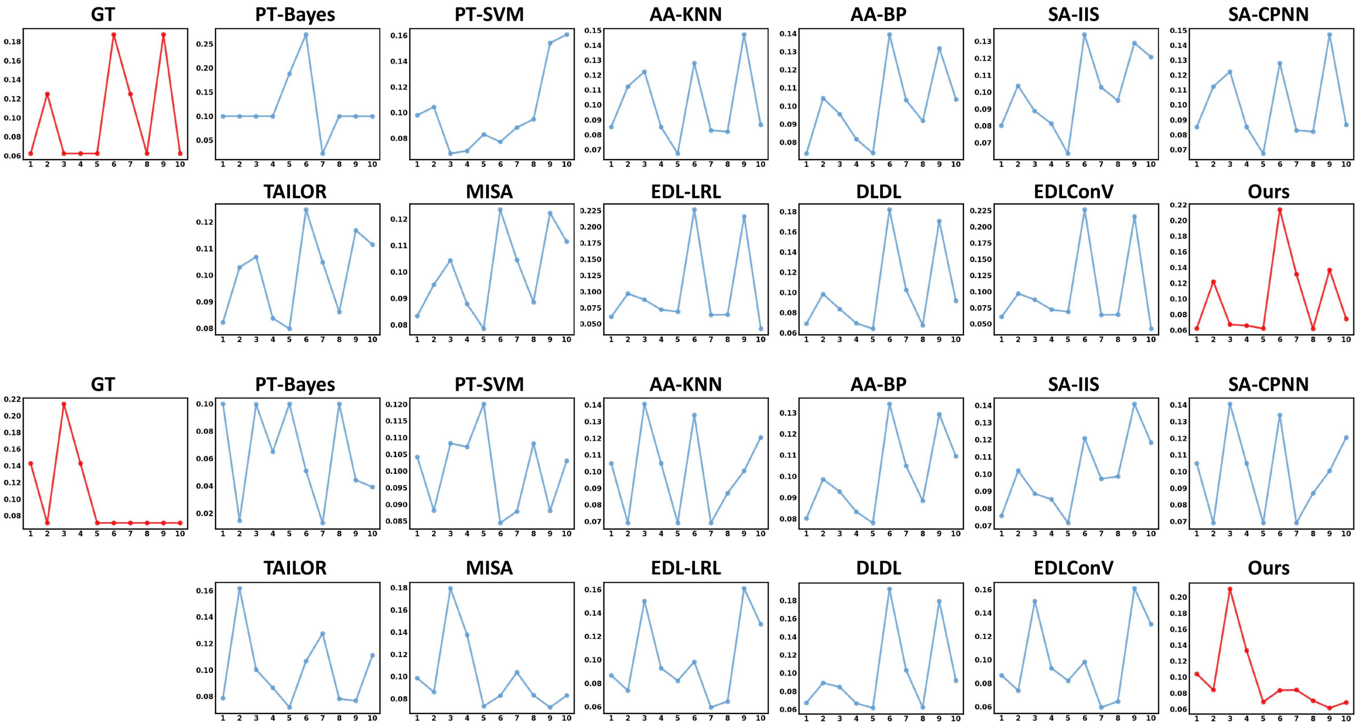

Fig. 3. Predicted emotion distributions with baseline approaches and our method. We show two panels of two test samples of Subject #28. GT indicates the ground truth distribution. The numbers 1 to 10 correspond to emotions inspired, alert, excited, enthusiastic, determined, afraid, upset, nervous, scared, and distressed.

predicted and ground-truth label distributions.

- **Multi-modal Emotion Distribution Learning Method (denoted as Multi-modal EDL).** Emotion distribution learning network [28] (denoted as **EDLConV**) is the most related work to ours, which unitizes a CNN-based network with four types of peripheral physiological signals (i.e., ECG, HR, GSR, and SKT) for the EDL task. When comparing with EDLConV, we use the same facial video data preprocessing and feature extraction procedures as our proposed EmotionDict method, where face detection is firstly conducted on each 1 out of 30 video frames, and 32-dimensional features are extracted with a VGG-16 [56] backbone pre-trained on ImageNet [57] on the resized facial images.

4.5 Evaluation Results

4.5.1 Subject-Dependent Evaluation

The results are summarized in Table 3. In addition to the six evaluation metrics mentioned in Section 4.3, we also follow the previous EDL studies [28] to compute the rankings of the twelve methods on each metric and adopt the mean value of rankings of all six metrics (*Average Rank*).

We can conclude that (1) our method has the best overall performance. Except for the Canberra and Intersection metrics on the EDL dataset, our method performs best on all six metrics on both datasets. (2) Generally, multi-modal multi-label emotion recognition (MMER), single-modal emotion distribution learning methods (Single-modal EDL) and multi-modal emotion distribution learning (Multi-modal EDL) methods achieve better results than the classic LDL methods (PT, AA, and SA). An exception is the AA-

TABLE 4
Subject-independent comparison of experimental results of our method and 11 baseline algorithms on six measures on the DMER and EDL datasets. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The best results are in **bold**,and the parentheses show the corresponding ranks on each evaluation metric and the *Average Ranks*.

Dataset	Measure	PT		AA		SA		MMER		Single-modal EDL		Multi-modal EDL	
		PT-Bayes	PT-SVM	AA-KNN	AA-BP	SA-IIS	SA-CPNN	TAILOR	MISA	EDL-LRL	DLDL	EDLConV	**Ours**
DMER	Chebyshev (↓)	0.6993(12)	0.1035(9)	0.1027(8)	0.0881(4)	0.0878(2)	0.0882(5)	0.1109(10)	0.1116(11)	0.0880(3)	0.0970(6)	0.0994(7)	**0.0858(1)**
	Clark (↓)	2.7320(12)	0.7445(9)	0.7074(8)	0.6157(3)	0.6236(5)	0.6152(2)	0.8287(11)	0.6507(6)	0.6158(4)	0.6800(7)	0.7777(10)	**0.5975(1)**
	Canberra (↓)	8.4200(12)	2.0409(9)	1.9207(8)	1.7192(3)	1.7256(4)	1.7664(6)	2.3248(11)	**1.5958(1)**	1.7262(5)	1.8900(7)	2.1334(10)	1.6284(2)
	KL (↓)	8.8751(12)	0.1438(10)	0.1301(7)	0.0949(2)	0.0953(3)	0.1011(5)	0.1632(11)	0.1390(9)	0.0953(3)	0.1193(6)	0.1344(8)	**0.0898(1)**
	Cosine (↑)	0.3978(12)	0.8761(9)	0.8802(8)	0.9137(2)	0.9132(4)	0.9087(5)	0.8543(10)	0.8536(11)	0.9134(3)	0.8934(6)	0.8820(7)	**0.9172(1)**
	Intersection (↑)	0.2352(12)	0.7881(9)	0.7963(8)	0.8231(2)	0.8220(4)	0.8190(5)	0.7596(11)	0.8145(6)	0.8224(3)	0.8039(7)	0.7839(10)	**0.8299(1)**
	Average Rank (↓)	12(12)	9.16(10)	7.83(8)	2.66(2)	3.66(4)	4.66(5)	10.66(11)	7.5(7)	3.5(3)	6.5(6)	8.66(9)	**1(1)**
EDL	Chebyshev (↓)	0.7170(12)	0.2267(11)	0.1969(7)	0.2156(10)	0.2036(9)	0.2005(8)	0.1116(2)	0.1262(3)	0.1928(6)	0.1822(4)	0.1910(5)	**0.1093(1)**
	Clark (↓)	2.3580(12)	1.0265(11)	0.8649(7)	1.0134(10)	0.8969(9)	0.8866(8)	0.7247(3)	0.6154(2)	0.8479(6)	0.7995(4)	0.8392(5)	**0.5736(1)**
	Canberra (↓)	6.0691(12)	2.3925(11)	2.0515(6)	2.3642(10)	2.2107(9)	2.2025(8)	2.0942(7)	1.4494(2)	2.0414(5)	0.1942(3)	2.0319(4)	**1.3631(1)**
	KL (↓)	8.9546(12)	0.3815(10)	0.2868(9)	0.3842(11)	0.2701(8)	0.2483(6)	0.1293(3)	0.1077(2)	0.2500(7)	0.2040(4)	0.2322(5)	**0.0863(1)**
	Cosine (↑)	0.3906(12)	0.7272(11)	0.7766(9)	0.7531(10)	0.7912(8)	0.8009(6)	0.8855(3)	0.9312(2)	0.8004(7)	0.8355(4)	0.8139(5)	**0.9445(1)**
	Intersection (↑)	0.2389(12)	0.6474(11)	0.6895(7)	0.6622(10)	0.6793(9)	0.6826(8)	0.7837(3)	0.8192(2)	0.6982(6)	0.7212(4)	0.7035(5)	**0.8296(1)**
	Average Rank (↓)	12(12)	10.83(11)	7.5(8)	10.17(10)	8.67(9)	7.33(7)	3.5(3)	2.17(2)	5.83(6)	3.83(4)	4.83(5)	**1(1)**

KNN which performs relatively better, owing to the video data samples in the DMER dataset having relatively fewer changes, and there also exist some similar samples in the EDL dataset. Among the classic LDL methods, the AA-KNN and SA-IIS methods outperform the PT methods by a large margin. (3) Although the deep learning methods (MISA, EDL-LRL, and DLDL) achieve good results on all six measures, they are more likely to output similar distributions for different signal samples. In comparison, our proposed method and EDLConv algorithm tend to produce more diverse distributions. Visualization examples of the predicted emotion distributions using all the methods and the ground truths on the DMER dataset are shown in Figure 3. These qualitative results show that our method can predict emotion distributions most similar to the ground truths.

4.5.2 Subject-Independent Evaluation

We compare our method with the eleven baselines on the two datasets. Results are shown in Table 4, presenting that (1) our proposed method has achieved the best performance compared to all the eleven baseline methods on the subject-independent protocol on all the metrics, indicating that our method is effective to capture the invariant deep emotion features between individuals. (2) Compared with the subject-dependent setting (as shown in Table 3), the superiority of our EmotionDict method over the baseline methods is more obvious in subject-independent experiments. In other words, compared to the baseline methods, our method has more significant superiority to handle the subject-independent EDL task, which is more challenging than the subject-dependent situation. (3) The performance of the TAILOR varies dramatically on the two datasets (with an Average Rank of 10.66 on the DMER dataset and 3.5 on the EDL dataset), which is also the case for AA-BP, MISA, EDL-LRL, DLDL, and EDLConV. Compared with these baselines, our method has reached a more stable EDL performance.

4.5.3 Ablation Study

Investigation of network components. Table 5 shows the results of our ablation study on the proposed emotion dictionary, and multi-modal auxiliary losses on the DMER and EDL datasets. We first evaluate the impact of our emotion feature representation learning based on the emotion dictionary composition in the latent space. By removing the emotion dictionary module, and only preserving the Transformer encoder and attention classifier (denoted as "w/o Emotion Dictionary"), (i.e., removing the **Emotion Dictionary** part in Figure 2), all six metrics are computed on both datasets. Then we removed the $\mathcal{L}_{aux_1}$ and $\mathcal{L}_{aux_2}$ losses in turn to investigate the effectiveness of the multi-modal auxiliary tasks. We observed significant drops in the performance on both datasets with both subject-dependent and subject-independent experimental settings. Therefore, we conclude that both the emotion dictionary and multi-modal attentions necessarily contribute to the final performance of our model.

Investigation of modalities. We also assessed the impact of each modality by separately removing each modality from our full model. These experiments were conducted in both subject-dependent and subject-independent settings. The results, presented in Table 6, indicate that each component contributes significantly to the overall performance of multi-modal emotion distribution learning.

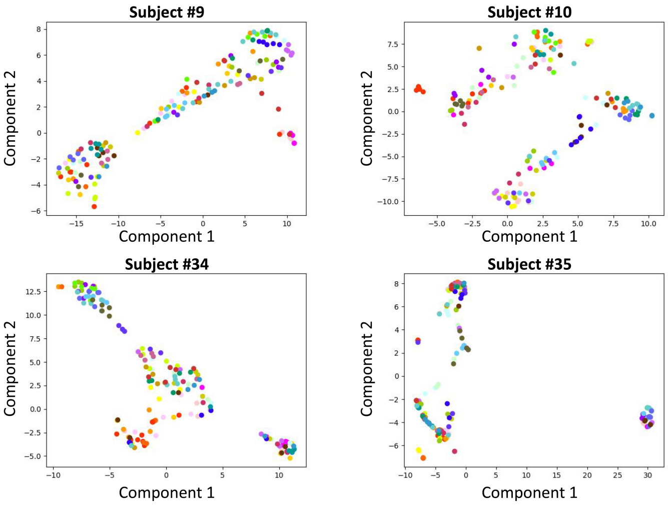

Fig. 4. Visualization of feature embedding with the *emotion dictionary* using t-SNE on the testing set of DMER dataset. The feature spaces of four subjects are shown in the four panels, each of which shows the subject conducting 32 trials. Points with the same color represent samples from the same trial, and 5 samples with lengths 1-s from each trial are displayed.

TABLE 5
Ablation study. We investigate the importance of each module to the performance of our final method by removing the loss function of each module separately. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The best results are in **bold**.

Settings	Measure	DMER				EDL		
		w/o Emotion Dictionary	w/o $\mathcal{L}_{aux_1}$	w/o $\mathcal{L}_{aux_2}$	**Ours**	w/o Emotion Dictionary	w/o $\mathcal{L}_{aux}$	**Ours**
Subject-dependent	Chebyshev (↓)	0.0849	0.0870	0.0772	**0.0717**	0.1952	0.1905	**0.1813**
	Clark (↓)	0.0849	0.6216	0.5489	**0.4729**	1.0496	1.0057	**0.8201**
	Canberra (↓)	1.5917	1.6316	1.4734	**1.2297**	2.8569	2.7746	**2.0195**
	KL (↓)	0.0916	0.0964	0.0762	**0.0592**	0.2327	0.2493	**0.2273**
	Cosine (↑)	0.9168	0.9140	0.9304	**0.9419**	0.7818	0.7915	**0.8333**
	Intersection (↑)	0.8342	0.8311	0.8477	**0.8686**	0.7038	0.7010	**0.7115**
Subject-independent	Chebyshev (↓)	0.0881	0.0871	0.0856	**0.0858**	0.1355	0.1267	**0.1093**
	Clark (↓)	0.6393	0.6283	0.5975	**0.5975**	0.6717	0.6401	**0.5736**
	Canberra (↓)	1.7464	1.7123	1.6827	**1.6284**	1.6309	1.5434	**1.3631**
	KL (↓)	0.0986	0.0966	0.0913	**0.0898**	0.1325	0.1170	**0.0863**
	Cosine (↑)	0.9108	0.9122	0.9171	**0.9172**	0.9022	0.9163	**0.9445**
	Intersection (↑)	0.8196	0.8224	0.8273	**0.8299**	0.7791	0.7952	**0.8296**

TABLE 6
Investigation of modalites. The EDL results of our EmotionDict framework using different combinations of signals are shown. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The values in parentheses are the standard deviations. The best results are in **bold**.

Modality	Subject-dependent					Subject-independent				
	w/o EEG	w/o PPG	w/o GSR	w/o Video	**Ours**	w/o EEG	w/o PPG	w/o GSR	w/o Video	**Ours**
Chebyshev (↓)	0.0856 (0.0128)	0.0873 (0.0170)	0.0792 (0.0126)	0.0849 (0.0156)	**0.0717** (0.0149)	0.0859 (0.0168)	0.0993 (0.0098)	0.1006 (0.0085)	0.0998 (0.0085)	**0.0858** (0.0134)
Clark (↓)	0.6254 (0.1092)	0.6328 (0.1106)	0.5581 (0.0748)	0.5883 (0.1187)	**0.4729** (0.1417)	0.6076 (0.1296)	0.6173 (0.1283)	0.6115 (0.1246)	0.6191 (0.1378)	**0.5975** (0.1438)
Canberra (↓)	1.6350 (0.3055)	1.6872 (0.3296)	1.4939 (0.2337)	1.5550 (0.3527)	**1.2297** (0.4207)	1.6537 (0.4002)	1.6987 (0.4481)	1.6713 (0.4091)	1.6976 (0.4664)	**1.6284** (0.4248)
KL (↓)	0.0933 (0.0295)	0.0979 (0.0292)	0.0776 (0.0167)	0.0866 (0.0284)	**0.0592** (0.0227)	0.1015 (0.0311)	0.1004 (0.0280)	0.0997 (0.0264)	0.1021 (0.0291)	**0.0898** (0.0223)
Cosine (↑)	0.9159 (0.0221)	0.9109 (0.0237)	0.9246 (0.0147)	0.9189 (0.0225)	**0.9419** (0.0175)	0.9004 (0.0250)	0.9037 (0.0170)	0.9040 (0.0171)	0.9023 (0.0184)	**0.9172** (0.0163)
Intersection (↑)	0.8299 (0.0297)	0.8229 (0.0320)	0.8415 (0.0192)	0.8368 (0.0322)	**0.8686** (0.0376)	0.8189 (0.0369)	0.8183 (0.0353)	0.8208 (0.0338)	0.8181 (0.0383)	**0.8299** (0.0375)

TABLE 7
Results of different fusion strategies. We investigate the effectiveness of different multi-modal feature fusion mechanisms by modifying the Emotion Dictionary and Auxiliary Tasks in our full EmoDict EDL framework. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The best results are in **bold**.

Settings	Fusion Strategy	DMER				EDL			
		w/o Emotion Dictionary	Concatenation	Transformer	**Ours**	w/o Emotion Dictionary	Concatenation	Transformer	**Ours**
Subject-dependent	Chebyshev (↓)	0.0849	0.1382	0.0723	**0.0717**	0.1952	0.1958	0.1934	**0.1813**
	Clark (↓)	0.0849	1.1138	0.4927	**0.4729**	1.0496	0.8684	0.8579	**0.8201**
	Canberra (↓)	1.5917	2.9528	1.3266	**1.2297**	2.8569	2.155	2.1296	**2.0195**
	KL (↓)	0.0916	0.2990	0.0664	**0.0592**	0.2327	0.2341	0.2286	**0.2273**
	Cosine (↑)	0.9168	0.8053	0.9406	**0.9419**	0.7818	0.8120	0.8168	**0.8333**
	Intersection (↑)	0.8342	0.7085	0.8659	**0.8686**	0.7038	0.6901	0.6946	**0.7115**
Subject-independent	Chebyshev (↓)	0.0881	0.1573	0.0976	**0.0858**	0.1355	0.1976	0.1959	**0.1093**
	Clark (↓)	0.6393	1.1232	0.6952	**0.5975**	0.6717	0.8604	0.8666	**0.5736**
	Canberra (↓)	1.7464	3.0116	1.9799	**1.6284**	1.6309	2.1372	2.1547	**1.3631**
	KL (↓)	0.0986	0.3334	0.1128	**0.0898**	0.1325	0.2320	0.2332	**0.0863**
	Cosine (↑)	0.9108	0.7739	0.9073	**0.9172**	0.9022	0.8127	0.8130	**0.9445**
	Intersection (↑)	0.8196	0.6912	0.8059	**0.8299**	0.7791	0.6918	0.6907	**0.8296**

4.5.4 Analysis of the Multi-modal Fusion Strategy

As the multi-modal fusion strategy based on the emotion dictionary module is the most important part of our EmotionDict framework, we conducted a comprehensive analysis to assess its effectiveness. We further compare it with three other conventional fusion mechanisms: (1) removing the emotion dictionary module directly from our full method, which is the same condition as "w/o Emotion Dictionary" setting in the investigation of network components in our ablation study. (2) Substituting the **Emotion Dictionary** component in Figure 2 by concatenating all the multi-modal features (i.e., the features learned from the Transformer encoders) and the four modality-specific features (i.e., the features from the EEG, PPG, GSR and video encoders) to form a fusion feature for subsequent classification (denoted as "Concatenation"). (3) Instead of directly removing the **Emotion Dictionary** part in Figure 2 as in (1) or simply concatenating the multi-modal features as in (2), we employ a Transformer-based attention network [53] to fuse both the multi-modal and modality-specific features (denoted as "Transformer"). These experiments were conducted under both subject-dependent and subject-independent settings.

The results on the DMER and EDL datasets, presented in Table 7, shows that our multi-modal fusion strategy, leveraging the emotion dictionary and auxiliary tasks as attentions, achieves the best EDL performance in both subject-dependent and subject-independent settings.

4.5.5 More Analysis of the Emotion Dictionary

We further explore the effectiveness of the emotion dictionary module by visualizing the latent representation learned by the emotion dictionary, i.e., the h feature vector in Equation 2. Feature distributions of four subjects (Subject #9, Subject #10, Subject #34, and Subject #35) from the DMER dataset are shown in 2-dimensional space with t-SNE. For each subject, we randomly visualize five latent feature vectors of length 1-s from each trial in the testing set. The results are shown in Figure 4, where every five points with the same color represent latent feature vectors from the same trial. There are a total of 32 color points in Figure 4 (5 points for each color), corresponding to the 32 trials or video clips in DMER. It can be observed that (1) the distance between the points with the same color is relatively close, indicating that

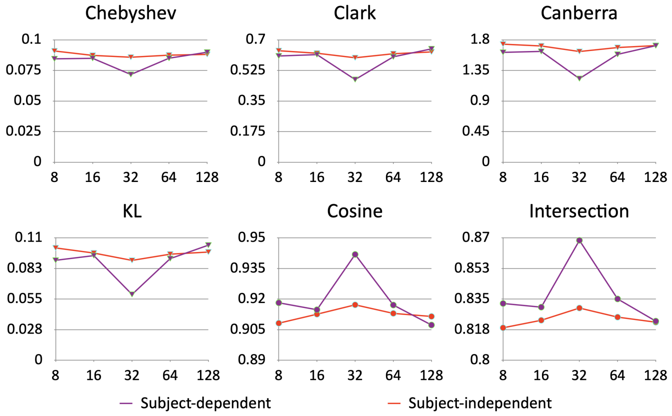

Fig. 5. Emotion distribution learning performance with respect to the number of *basic emotion elements* $\vec{d_i}$ in the latent emotion dictionary in Equation 2. The experimental results with numbers of basic emotion elements as 8, 16 32, 64, and 128 with both subject-dependent and subject-dependent protocols are shown. ▽ indicates "the smaller the better", and ● indicates "the larger the better".

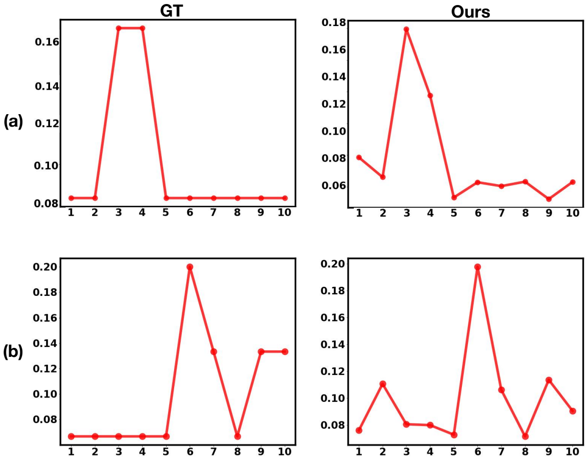

Fig. 6. Failure cases. We show two panels of two test samples, each of which the ground truth (GT) and the predicted emotion distributions with our method are shown. GT indicates the ground truth distribution. The numbers 1 to 10 correspond to emotions inspired, alert, excited, enthusiastic, determined, afraid, upset, nervous, scared, and distressed.

the learned latent feature vectors can distinguish the unique emotion distribution of different trials. (2) Points of arbitrary numbers of colors are gathered into clusters because, for each subject, the similarities of the emotion distributions in different trials are different. Overall, the latent features have smaller distances within the same trial and larger distances between different trials, which proves the effectiveness of our method.

Moreover, we conduct an experiment with our EmotionDict framework with varying numbers of *basic emotion elements* $\vec{d_i}$ in Equation 2 to explore the influence of the size of emotion dictionary on the emotion distribution learning performance. Figure 5 shows the EDL results with the number of *basic emotion elements* as 8, 16, 32, 64, 128 with both subject-dependent and subject-independent experimental protocols on the DMER dataset, indicating that our EmotionDict can obtain stable EDL results even though the size of emotion dictionary changes. The number of *basic emotion elements* in our final EmotionDict is set as a hyper-parameter of 32, leading to relatively better performance.

5 Discussions

In this paper, we address the challenge of recognizing mixed emotions in situations where both positive and negative emotions co-exist. Our task is formulated as a label distribution learning task. In our approach, we extract and fuse features from multi-modal signals related to both subjects' overt behavior (facial video) and physiological signals (EEG, PPG, GSR, and video). We have designed an emotion distribution learning framework primarily centered around learning an emotion dictionary with two auxiliary tasks. Extensive experiments demonstrate the superiority of our proposed EmotionDict method over all 11 baseline methods on two datasets. Every key component in our framework contributes significantly to the final performance of our method.

Moreover, we conducted an in-depth investigation of our emotion dictionary, including the multi-modal fusion strategy and the impact of the number of basic emotion elements in the dictionary. Experimental results indicate that our multi-modal fusion strategy, utilizing the emotion dictionary and auxiliary tasks as attentions, outperforms the baselines. Additionally, our EmotionDict framework maintains stable performance with varying numbers of basic emotion elements. Finally, the visual results depicting the latent representation learned by the emotion dictionary further verify the effective emotion representation learning capabilities of our method.

Limitations. While our EmotionDict framework demonstrates superior performance compared to the 11 baseline methods, as indicated by both the quantitative results (Table 3 and Table 4) and the qualitative results (Figure 3), there are certain limitations to our approach. Despite the fact that our method generally aligns well with the ground truth in most cases, occasional fluctuations in the prediction results are observed. Figure 6 provides two examples depicting the ground truth alongside the corresponding predicted distributions generated by our approach. It is evident that our method performs well in predicting dominant emotions, i.e., the basic emotional categories characterized with significantly high intensities. However, for basic emotional categories with relatively lower intensities (such as the emotion categories 5 to 10 in (a) and 1 to 5 in (b) of Figure 6), there exists a moderate discrepancy between the predicted emotional intensities and the ground truth.

Future works. In our future research, we plan to investigate the distinctions in multi-modal behavior and physiological signals among individuals when they undergo similar emotional states. This exploration aims to enhance our understanding and modeling of mixed emotions. Given that our method is tailored to improve representations from emotional features extracted from multi-modal signals, we aim to enhance our capability to capture the inherently unique traits of individuals. Therefore, our future work will involve leveraging the similarities and differences among individuals to construct a more generalized EDL framework.

6 Conclusions

In this study, we studied the multi-modal emotion distribution learning task, and propose a multi-modal EDL framework, *EmotionDict*, that learns an "emotion dictionary" of a set of basic emotion representations in a latent space. Further, we enhance the emotion dictionary by learning attention via a multi-modal integration module that is trained with two auxiliary tasks, i.e., (1) learning emotion representation from EEG signals, and (2) learning emotion representation from joint behavioral and peripheral physiological signals (PPG, GSR and facial videos). Experiments on two multi-modal emotion distribution datasets show that the proposed method effectively handles mixed emotions recognition, outperforming eleven state-of-the-art approaches in both subject-dependent and subject-independent settings. Experiments have also been conducted to verify the effectiveness of the key components of our proposed model, including the Emotion Dictionary, the two auxiliary losses.

References

[1] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 1247–1250.

[2] M. Jeon, "Emotions and affect in human factors and human–computer interaction: Taxonomy, theories, approaches, and methods," *Emotions and affect in human factors and human-computer interaction*, pp. 3–26, 2017.

[3] S. Mukhtar, "Mental health and emotional impact of covid-19: Applying health belief model for medical staff to general public of pakistan," *Brain, Behavior, and Immunity*, vol. 87, pp. 28–29, 2020.

[4] C. D. Salzman and S. Fusi, "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex," *Annual Review of Neuroscience*, vol. 33, pp. 173–202, 2010.

[5] E. L. Van den Broek, "Ubiquitous emotion-aware computing," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 53–67, 2013.

[6] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[7] G. Zhang, M. Yu, Y. Liu, G. Zhao, D. Zhang, and W. Zheng, "Sparsedgcnn: Recognizing emotion from multichannel EEG signals," *IEEE Transactions on Affective Computing*, 2021.

[8] P. J. Lang, "The emotion probe: Studies of motivation and attention." *American Psychologist*, vol. 50, no. 5, p. 372, 1995.

[9] W. Zheng and B. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016, pp. 2732–2738.

[10] X. Du, C. Ma, G. Zhang, J. Li, Y. Lai, G. Zhao, X. Deng, Y. Liu, and H. Wang, "An efficient lstm network for emotion recognition from multichannel EEG signals," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1528–1540, 2020.

[11] P. Williams and J. L. Aaker, "Can mixed emotions peacefully coexist?" *Journal of Consumer Research*, vol. 28, no. 4, pp. 636–649, 2002.

[12] J. T. Larsen and A. P. McGraw, "Further evidence for mixed emotions." *Journal of Personality and Social Psychology*, vol. 100, no. 6, p. 1095, 2011.

[13] G. Zhao, Y. Zhang, G. Zhang, D. Zhang, and Y. Liu, "Multi-target positive emotion recognition from EEG signals," *IEEE Transactions on Affective Computing*, vol. DOI: 10.1109/TAFFC.2020.3043135, 2020.

[14] M. Svašek, "Introduction: Emotions in anthropology," in *Mixed Emotions*. Routledge, 2020, pp. 1–23.

[15] W. Celniak and P. Augustyniak, "Eye-tracking as a component of multimodal emotion recognition systems," in *International Conference on Information Technologies in Biomedicine*. Springer, 2022, pp. 66–75.

[16] H. E. Hershfield, S. Scheibe, T. L. Sims, and L. L. Carstensen, "When feeling bad can be good: Mixed emotions benefit physical health across adulthood," *Social psychological and personality science*, vol. 4, no. 1, pp. 54–61, 2013.

[17] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 884–906, 2019.

[18] D. Zhang, X. Ju, W. Zhang, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 338–14 346.

[19] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion detection with modality and label dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 3584–3593.

[20] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9841–9850.

[21] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, "Text emotion distribution learning via multi-task convolutional neural network," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 4595–4601.

[22] Y. Lu, W. Zheng, B. Li, and B. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 1170–1176.

[23] R. Somarathna, A. Quigley, and G. Mohammadi, "Multi-componential emotion recognition in VR using physiological signals," in *Proceedigns of the Australasian Joint Conference on Artificial Intelligence*. Springer, 2022, pp. 599–613.

[24] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, "A circular-structured representation for visual emotion distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4237–4246.

[25] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion distribution learning from texts," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 638–647.

[26] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling, and physiology." 2000.

[27] X. Jia and X. Shen, "Multimodal emotion distribution learning," *Cognitive Computation*, vol. 14, no. 6, pp. 2141–2152, 2022.

[28] Y. Shu, P. Yang, N. Liu, S. Zhang, G. Zhao, and Y. Liu, "Emotion distribution learning based on peripheral physiological signals," *IEEE Transactions on Affective Computing*, vol. DOI: 10.1109/TAFFC.2022.3163609, 2022.

[29] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[30] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," *arXiv preprint arXiv:2203.09043*, 2022.

[31] Q. Zhang, C. Yang, Y. Shen, Y. Xu, and B. Zhou, "Towards smooth video composition," *arXiv preprint arXiv:2212.07413*, 2022.

[32] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[33] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[34] B. Gao, H. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning." in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2018, pp. 712–718.

[35] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 984–13 993.

[36] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[37] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2012.

[38] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, 2020.

[39] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[40] D. B. Lindsley, "Emotion." 1951.

[41] R. E. Plutchik and H. R. Conte, *Circumplex models of personality and emotions.* American Psychological Association, 1997.

[42] K. Kercher, "Assessing subjective well-being in the old-old: The panas as a measure of orthogonal dimensions of positive and negative affect," *Research on Aging*, vol. 14, no. 2, pp. 131–168, 1992.

[43] X. Zeng, Q. Chen, X. Fu, and J. Zuo, "Emotion wheel attention-based emotion distribution learning," *IEEE Access*, vol. 9, pp. 153 360–153 370, 2021.

[44] J. Yang, J. Li, L. Li, X. Wang, Y. Ding, and X. Gao, "Seeking subjectivity in visual emotion distribution learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 5189–5202, 2022.

[45] Z. Xu and S. Wang, "Emotional attention detection and correlation exploration for image emotion distribution learning," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 357–369, 2023.

[46] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 523–536, 2023.

[47] Y. Zhou, X. Kang, and F. Ren, "Prompt consistency for multi-label textual emotion detection," *IEEE Transactions on Affective Computing*, pp. 1–10, 2023.

[48] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, "Emotion understanding in videos through body, context, and visual-semantic embedding loss," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 747–755.

[49] X. Du, X. Deng, H. Qin, Y. Shu, F. Liu, G. Zhao, Y.-K. Lai, C. Ma, Y.-J. Liu, and H. Wang, "Mmpose: Movie-induced multi-label positive emotion classification through eeg signals," *IEEE Transactions on Affective Computing*, pp. 1–14, 2022.

[50] B. Kostiuk, Y. M. Costa, A. S. Britto, X. Hu, and C. N. Silla, "Multi-label emotion classification in music videos using ensembles of audio and video features," in *Proceedings of the 31st IEEE International Conference on Tools with Artificial Intelligence*. IEEE, 2019, pp. 517–523.

[51] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based label set generation for multi-modal multi-label emotion detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 512–520.

[52] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems*. IEEE, 2017, pp. 5998–6008.

[54] R. Duan, J. Zhu, and B. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proceedings of the 6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[55] P. Yang, "Dmer," https://zenodo.org/record/7385297, 2022, accessed: 2022-12-31.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[58] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[59] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/496e05e1aea0a9c4655800e8a7b9ea28-Paper.pdf

[60] J. Crabbé, Z. Qian, F. Imrie, and M. van der Schaar, "Explaining latent representations with a corpus of examples," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 154–12 166, 2021.

[61] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[62] F. H. Wilhelm and P. Grossman, "Emotions beyond the laboratory: Theoretical fundamentals, study design, and analytic strategies for advanced ambulatory assessment," *Biological Psychology*, vol. 84, no. 3, pp. 552–569, 2010.

[63] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the positive and negative affect schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample," *Personality and Individual differences*, vol. 27, no. 3, pp. 405–416, 1999.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[65] X. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[66] Y. Zhang, M. Chen, J. Shen, and C. Wang, "Tailor versatile multi-modal learning for multi-label emotion recognition," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022, pp. 9100–9108.

[67] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

PLACE PHOTO HERE

Fang Liu received her Ph.D. degree from the University of the Chinese Academy of Sciences (UCAS), Beijing, China, in 2021. She is currently a Postdoctoral Fellow at Tsinghua University. Her research interests include computer vision, sketch interaction, and affective computing.

PLACE PHOTO HERE

Pei Yang is is currently a Ph.D student in Department of Computer Science and Technology, Tsinghua University. He received his bachelor's degree from Inner Mongolia University, China, in 2009 and his master's degree from Minzu University of China in 2016. His research interests include emotion recognition and machine learning.

PLACE PHOTO HERE

Yezhi Shu is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. She received her B.Eng. degree from Shandong University, China, in 2019. Her research interests include computer vision, deep learning algorithms and applications.

PLACE
PHOTO
HERE

Fei Yan is an associate professor at the School of Computer Science and Technology at Changchun University of Science and Technology, China. He is now a visiting scholar at the Department of Computer Science and Technology, Tsinghua University, China. He received his doctorate in engineering and worked as a postdoctoral fellow afterward in the Department of Computational Intelligence and Systems Science at Tokyo Institute of Technology, Japan. He has published more than 80 papers in the fields of quantum information processing, affective computing, and medical image analysis.

PLACE
PHOTO
HERE

Guanhua Zhang is a PhD student at the Institute for Visualisation and Interactive Systems, University of Stuttgart, Germany. She received her B.Sc. degree in computer science and technology from Beijing University of Posts and Telecommunications in 2017, and her M.Sc. degree in computer science from Tsinghua university in 2020. Her current research interests include human computer interaction, affective computing and user behaviour modelling.

PLACE
PHOTO
HERE

Yong-Jin Liu is a Professor with Department of Computer Science and Technology, Tsinghua University, Beijing, China. He received the B.Eng. degree from Tianjin University, Tianjin, China, in 1998, and the M.Phil. and Ph.D. degrees from the Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. His research interests include computational geometry, computer vision, cognitive computation and pattern analysis. For more information, visit http://cg.cs.tsinghua.edu.cn/people/∼Yongjin/Yongjin.htm