PIRNet: Personality-Enhanced Iterative Refinement Network for Emotion Recognition in Conversation

Zheng Lian, Bin Liu, *Member, IEEE*, and Jianhua Tao, *Senior Member, IEEE*

***Abstract*— Emotion recognition in conversation (ERC) is important for enhancing user experience in human–computer interaction. Unlike vanilla emotion recognition in individual utterances, ERC aims to classify constituent utterances in a dialog into corresponding emotion labels, which makes contextual information crucial. In addition to contextual information, personality traits also affect emotional perception based on psychological findings. Although researchers have proposed several approaches and achieved promising results on ERC, current works in this domain rarely incorporate contextual information and personality influence. To this end, we propose a novel framework to integrate these factors seamlessly, called "Personality-enhanced Iterative Refinement Network (PIRNet)." Specifically, PIRNet is a multistage iterative method. To capture personality influence, PIRNet leverages personality traits to mimic emotional transitions and generates personality-enhanced results. Then we exploit sequence models to capture contextual information in conversations. To verify the effectiveness of our proposed method, we conduct experiments on three benchmark datasets for ERC, that is, IEMOCAP, CMU-MOSI, and CMU-MOSEI. Experimental results demonstrate that our PIRNet succeeds over currently advanced approaches to emotion recognition.**

***Index Terms*— Contextual information, emotion recognition in conversation (ERC), iterative method, Personality-enhanced Iterative Refinement Network (PIRNet), personality influence.**

I. Introduction

EMOTION recognition is a cutting-edge interdisciplinary subject of information technology and psychology [1]. It aims to integrate multisource information and identify the emotional state of each utterance. With the increasing amounts of conversations on social media platforms, emotion recognition in conversation (ERC) has attracted interest from researchers [2]. It can be widely utilized in diverse areas such as dialog generation [3], public opinion mining [4], and social media analysis [5]. As an extension of traditional emotion recognition, ERC aims to identify the emotional states of all constituent utterances in a dialog, which makes contextual information vitally important [6].

Besides contextual information, personality traits also affect human perception and expression of emotional information [7]. Previously, Winter and Kuiper [8] built on neuropsychological and cognitive perspectives to highlight the importance of personality factors in emotional experience. Based on their psychological findings, Li and Lee [9] integrated individual personality embeddings into an attention-based model for multimodal emotion recognition. Meanwhile, Li *et al.* [10] proposed a multitask learning framework to enhance emotion recognition performance by learning the commonalities and differences between personality trait detection and emotion detection. However, these works [9], [10] mainly focus on emotion recognition in individual utterances. They ignore context-sensitive dependencies in conversations, thus limiting their performance in ERC.

To this end, we propose a novel framework to integrate contextual information and personality influence seamlessly. Instead of modeling these two factors via a single architecture, we abstract ERC into two separate modules: an emotion recognition module and a refinement module. The emotion recognition module aims to predict the emotional state of each utterance. The refinement module aims to correct some errors in these preliminary results, further improving the emotion recognition performance. Specifically, we propose a novel framework for emotion refinement, called "Personality-enhanced Iterative Refinement Network (PIRNet)." Fig. 1 shows the overall structure of our model. To capture personality influence, we exploit personality traits to mimic emotional transitions and generate personality-enhanced results. Then, we utilize a bidirectional gated recurrent unit (GRU) to capture contextual information in both directions. The final outputs are fed into a fully-connected layer, followed by a softmax activation for emotion prediction. To make full use of emotion labels during training, we further equip PIRNet with a multistage iterative method. At each stage, PIRNet takes the optimized results of the previous stage as the inputs and ground-truth labels as the outputs. During inference, PIRNet leverages the multistage refinement process to generate final emotion results.

Recently, researchers have proposed an effective emotion refinement model for ERC [11]. It utilizes graph networks to model human interactions, followed by attention-based

Manuscript received 10 September 2021; revised 4 May 2022; accepted 16 July 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61831022, Grant U21B2010, Grant 61901473, and Grant 62101553; and in part by the Open Research Projects of Zhejiang Laboratory under Grant 2021KH0AB06. *(Corresponding authors: Bin Liu; Jianhua Tao.)*

Zheng Lian and Bin Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lianzheng2016@ia.ac.cn; liubin@nlpr.ia.ac.cn).

Jianhua Tao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China (e-mail: jhtao@nlpr.ia.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3192469.

Digital Object Identifier 10.1109/TNNLS.2022.3192469

GRUs for context-sensitive modeling. However, this model does not consider personality influence in emotion recognition. In this article, we further take full advantage of personality traits through a multistage iterative strategy. To verify the effectiveness of our method, we evaluate PIRNet on three benchmark datasets for ERC, that is, IEMOCAP, CMU-MOSI, and CMU-MOSEI. Experimental results demonstrate that our PIRNet outperforms currently advanced approaches to emotion recognition and emotion refinement. The main contributions of this article can be summarized as follows.

1) We propose a novel emotion refinement framework, PIRNet. It incorporates personality influence and contextual information to improve emotion recognition performance through a multistage iterative strategy.
2) We systemically investigate the importance of each component in PIRNet, including personality trails and the multistage iterative approach.
3) Extensive experiments on three benchmark datasets demonstrate the effectiveness of our method. PIRNet can achieve performance improvements in both unimodal and multimodal conditions, succeeding over existing models on emotion recognition and emotion refinement.

The remainder of this article is organized as follows. In Section II, we provide a brief review of related works. In Section III, we formalize the problem statement and describe our proposed method. In Section IV, we present experimental datasets and setup in detail. Experimental results and analysis are illustrated in Section V. Finally, we conclude this article and discuss future work in Section VI.

II. Related Works

A. Emotion Recognition in Conversation

To generate more human-like conversations, we need to incorporate empathy into the design of dialog systems [12], [13]. Emotion as an important component in building empathetic dialog systems has attracted increasing interest from researchers [14], [15]. For example, Zhou *et al.* [16] leveraged emotion category embedding, internal emotion memory, and external memory to model emotion influence in conversations. Cui *et al.* [17] exploited multitask learning to model semantic and emotional relationships in multiturn emotional conversations. Liang *et al.* [18] further proposed a heterogeneous graph-based model to predict appropriate emotions in response. Accurate perception of emotional states is the first step in the success of empathetic dialog systems. In this article, we focus on ERC to enhance the perception of emotional states in conversations.

Unlike vanilla emotion recognition in individual utterances, ERC processes the constituent utterances of a dialog consecutively. Therefore, contextual information plays an important role in emotion recognition [6]. Previously, researchers have proposed some methods for context-sensitive modeling, but the most popular strategy is the recurrent neural network (RNN). Due to the well-designed gating mechanisms, RNNs such as long short-term memory (LSTM) [19] and GRU [20] are capable of modeling long-term emotional dynamics in conversations. For example, Zhang *et al.* [21] utilized an LSTM-based model to capture contextual information obtained from historical utterances while classifying the target utterance. But, in addition to historical utterances, contextual information can also come from future utterances [22]. To this end, Poria *et al.* [2] proposed a bidirectional model to capture contextual information in both directions. To further consider interactions between different speakers, Hazarika *et al.* [23], [24] exploited distinct GRUs to learn the context of each speaker separately. More recently, Li *et al.* [25] employed low-rank matrix approximation on bidirectional LSTMs to build a parameter-efficient model. Ma *et al.* [26] further leveraged an RNN-based hierarchical structure to integrate word-level dependencies between utterances and utterance-level dependencies in the context.

Hence, we decide to leverage RNNs for context-sensitive modeling. It should be noted that our PIRNet is different from previous approaches for ERC. In addition to contextual information, we tightly integrate ERC with personality traits because of their impact on emotion recognition [27].

TABLE I
Big-Five Personality Traits and Associated Adjectives [31]

Personality Trait	Adjectives
Openness	Openminded, Imaginative, Curious
Conscientiousness	Efficient, Organized, Planful
Extraversion	Active, Assertive, Energetic
Agreeableness	Appreciative, Trusting, Cooperative
Neuroticism	Anxious, Self-pitying, Tense, Unstable

B. Personality Effect on Emotion

Personality is an important psychological concept that can be characterized by some stable attributes [28]. There are different approaches for personality description [29], [30], but the most popular one is called the Big-Five Model [31]. This model captures individual differences through five personality traits, including *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* [31]. Table I further explains these traits by their associated adjectives. In this article, we also use the Big-Five Model to measure personality traits.

Different people have different ways of perceiving and reacting to emotional information, and such difference is related to their personality traits [27]. Therefore, it is necessary to reveal the personality effects on emotions [32]. Previously, Barford and Smillie [33] conducted a theoretically grounded investigation of personality and emotion. They found that *neuroticism* predicted higher negative emotions and lower positive emotions, while *agreeableness* exhibited the opposite phenomenon to *neuroticism*. Mohammadi and Vuilleumier [34] summarized existing works and presented similar conclusions to Barford and Smillie [33]. They found that *extraversion*, *agreeableness*, *conscientiousness*, and *openness* were positively correlated with positive emotions. On the contrary, *neuroticism* was positively correlated with negative emotions and negatively correlated with positive emotions. Based on their psychological findings, Wen *et al.* [35] automatically selected emotions in response by considering personality differences. Liang *et al.* [36] further proposed a graph model to encode

multisource knowledge (including the personalities of the speaker, the dialog history, and the facial expression) and then predicted appropriate emotions for feedback. Inspired by their success, we propose to equip our emotion recognition system with personality traits and enable it to capture personality influence.

C. Iterative Method

The iterative method is a process that utilizes initial values to generate a series of improved approximate solutions for optimization problems [37]. For example, the conjugate gradient method [38] is an iterative method that uses a linear combination of the current residual and the search direction to update the current iteration. BFGS [39] is another popular iterative method that utilizes curvature information to determine the search direction. Most iterative methods terminate when the improvement is sufficiently small [40].

Recently, iterative methods have been widely utilized in various areas. Typically, Goodfellow *et al.* [41] proposed the generative adversarial network to capture data distributions through the iterative method. It contained a generator and a discriminator. The generator learned to capture data distributions, while the discriminator distinguished the data distribution produced by the generator from the real data distribution. Kolotouros *et al.* [42] leveraged the iterative method for human pose estimation. They first estimated the initial values of the optimization routine. Then they fitted the body model and utilized the fitted estimates to supervise the network. In this article, we draw inspiration from the iterative method. We find that this concept can be integrated with emotion refinement seamlessly, resulting in our multistage PIRNet. To the best of our knowledge, it is the first emotion refinement model using the multistage iterative method.

III. Methodology

In this section, we formalize the problem statement and describe our proposed framework in detail. The overall structure of our PIRNet is shown in Fig. 1.

A. Problem Definition and Notation

Our task is to classify all constituent utterances of a conversation into corresponding emotion labels. We conduct experiments in the offline condition, that is, we have access to the entire conversation during both training and inference phases, in line with previous works [2], [43]. Suppose we have a conversation $\mathcal{U} = \{(u_t, y_t)\}_{t=1}^{N}$ consisting of N utterances. Here, u_t is the tth utterance in the conversation. $y_t \in \{1, 2, \ldots, c\}$ denotes the emotion label of u_t, where c represents the number of discrete labels in the corpus. Let us define a function $\mathcal{S}(\cdot)$ that maps the index of utterance into its associated speaker. Therefore, each utterance u_t is uttered by the speaker $\mathcal{S}(t) \in \mathbb{S}$, where $\mathbb{S}$ denotes the set of speakers.

PIRNet is an iterative method with multiple stages. For convenience, we use the superscript $^{[k]}$ to denote the index of each stage. In the initial stage (i.e., $k = 0$), we extract feature vectors $f_t^{[0]} \in \mathbb{R}^{d_f}$ for each utterance u_t, where d_f represents the input feature dimension.

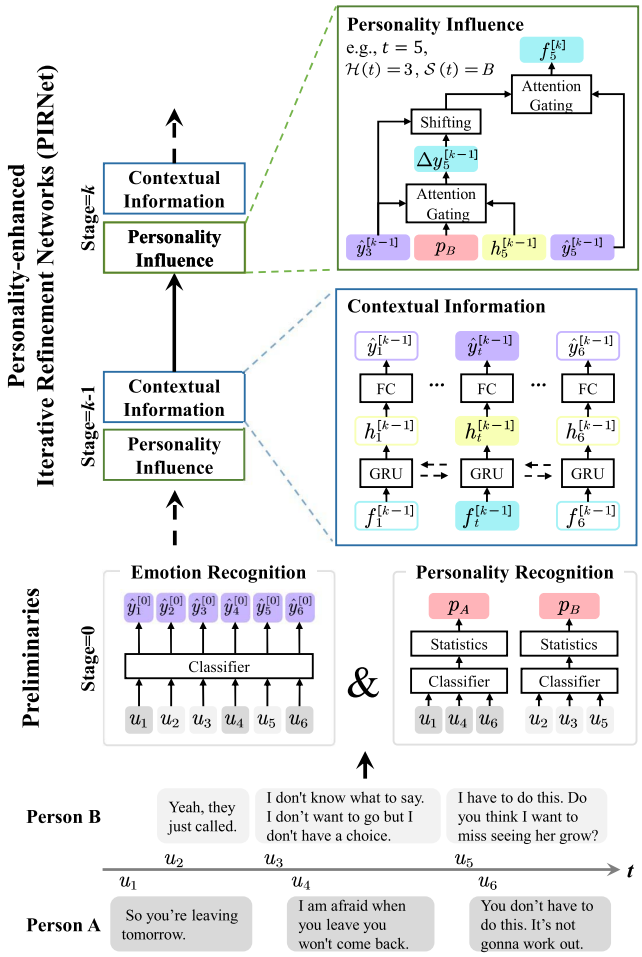

Fig. 1. Overall structure of PIRNet. It takes preliminary emotion recognition results and personality detection results as the inputs, aiming to improve emotion recognition performance by leveraging personality influence and contextual information through a multistage iterative strategy.

B. Preliminaries

1) Emotion Recognition: Since contextual information is crucial in ERC, we exploit a bidirectional GRU to capture contextual information in both directions [44], [45]. For each direction, GRU takes the input feature of the current step, along with the hidden state carried from the previous step, and outputs a new hidden state for the next step. The calculation formula of the forward GRU is shown as follows:

$$r_t^{[0]} = \sigma\left(W_e^r f_t^{[0]} + U_e^r \overrightarrow{h}_{t-1}^{[0]} + b_e^r\right) \tag{1}$$

$$z_t^{[0]} = \sigma\left(W_e^z f_t^{[0]} + U_e^z \overrightarrow{h}_{t-1}^{[0]} + b_e^z\right) \tag{2}$$

$$g_t^{[0]} = \tanh\left(W_e^g f_t^{[0]} + U_e^g \left(r_t^{[0]} \circ \overrightarrow{h}_{t-1}^{[0]}\right) + b_e^g\right) \tag{3}$$

$$\overrightarrow{h}_t^{[0]} = \left(1 - z_t^{[0]}\right) \circ \overrightarrow{h}_{t-1}^{[0]} + z_t^{[0]} \circ g_t^{[0]} \tag{4}$$

for each time step t, $\overrightarrow{h}_{t-1}^{[0]} \in \mathbb{R}^{d_h}$ is the hidden state of the previous step, and $f_t^{[0]} \in \mathbb{R}^{d_f}$ is the input feature of the current step. Here, d_h denotes the feature dimension of the hidden state. $r_t^{[0]} \in \mathbb{R}^{d_h}$ and $z_t^{[0]} \in \mathbb{R}^{d_h}$ are the reset gate and the update gate, respectively. $W_e^m \in \mathbb{R}^{d_h \times d_f}$, $U_e^m \in \mathbb{R}^{d_h \times d_h}$ and $b_e^m \in \mathbb{R}^{d_h}$ are the trainable parameters, where $m \in \{r, z, g\}$. Here, σ represents the sigmoid activation function and $\circ$

denotes elementwise multiplication. $\overrightarrow{h}_t^{[0]}$ is the final output of the forward GRU. Repeating such a process, we can also extract $\overleftarrow{h}_t^{[0]}$ for the backward GRU. Finally, we concatenate them together and generate feature representations $h_t^{[0]} = [\overrightarrow{h}_t^{[0]}, \overleftarrow{h}_t^{[0]}] \in \mathbb{R}^{2d_h}$. These features are fed into a fully-connected layer, followed by a softmax function to predict emotion-class probabilities $\hat{y}_t^{[0]} \in \mathbb{R}^c$ for each utterance u_t:

$$\hat{y}_t^{[0]} = \text{softmax}\left(W_e^y h_t^{[0]} + b_e^y\right) \tag{5}$$

where $W_e^y \in \mathbb{R}^{c\times 2d_h}$ and $b_e^y \in \mathbb{R}^c$ are the trainable parameters. We predict $\hat{y}_t^{[0]}$ for all utterances in the conversation, generating $\{\hat{y}_t^{[0]}\}_{t=1}^N$. To optimize all trainable parameters, we choose the cross-entropy loss function. Minimizing this loss function ensures that we can learn more discriminative features for emotion recognition. Suppose we have an emotion corpus $\mathcal{D}_e = \{\mathcal{U}_i\}_{i=1}^{N_c}$, where $\mathcal{U}_i = \{(u_{i,j}, y_{i,j})\}_{j=1}^{N_i}$. Here, N_c denotes the number of conversations in the corpus and N_i denotes the number of utterances in the conversation $\mathcal{U}_i$. The cross-entropy loss $L_e^{[0]}$ is calculated as follows:

$$L_e^{[0]} = -\frac{1}{\sum_{i=1}^{N_c} N_i} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} y_{i,j} \log\left(\hat{y}_{i,j}^{[0]}\right) \tag{6}$$

where $y_{i,j} \in \mathbb{R}^c$ and $\hat{y}_{i,j}^{[0]} \in \mathbb{R}^c$ are the ground-truth emotion label and the predicted emotion label for the jth utterance in the conversation $\mathcal{U}_i$, respectively.

2) Personality Recognition: Personality traits affect human perception and expression of emotional information [27]. But most emotion datasets do not have personality annotations. To compute personality scores on emotion datasets, we first train a personality detection model on the ChaLearn First Impression Database [46], a benchmark database for personality recognition. This corpus leverages the Big-Five Model to annotate each sample, resulting in a 5-D personality score. The value of each dimension is in the range of [0, 1].

Suppose we have a personality corpus $\mathcal{D}_p = \{(u_i^p, y_i^p)\}_{i=1}^{N_p}$ containing N_p samples. Here, u_i^p is a sample in the corpus and $y_i^p \in \mathbb{R}^5$ is the personality score of u_i^p. In our personality detection model, we first extract feature vectors for each sample. These features are then fed into stacked fully-connected layers, followed by ReLU for personality prediction. During training, we choose the absolute loss to optimize all trainable parameters. The loss L_p is calculated as follows:

$$L_p = \frac{1}{N_p} \sum_{i=1}^{N_p} \left| y_i^p - \hat{y}_i^p \right| \tag{7}$$

where $\hat{y}_i^p \in \mathbb{R}^5$ is the predicted result for the sample u_i^p. We train and evaluate personality recognition performance on the ChaLearn First Impression Database. We choose the best model as our pretrained personality detection model.

On emotion datasets, each speaker $s \in \mathbb{S}$ usually has multiple utterances during conversations. To extract speaker-level personality scores on emotion datasets, we first leverage the pretrained personality detection model to extract 5-D scores from each utterance. Then we draw inspiration from recent work [9] and aggregate these scores by five statistical functions (i.e., mean, standard deviation, median, maximum, and minimum). Finally, we obtain a 25-D personality feature $p_s \in \mathbb{R}^{d_p}$ for each speaker $s \in \mathbb{S}$. Here, d_p is the feature dimension.

C. Personality-Enhanced Iterative Refinement Network

Fig. 1 shows the overall structure of our proposed framework. It is a multistage iterative method. Each stage leverages personality influence and contextual information to improve the emotion prediction results carried from the previous stage. PIRNet terminates the iteration when the improvement is less than a sufficiently small value ϵ. In this section, we assume that K is the total number of stages, and we take the calculation formula of stage $k \in \{1, \ldots, K\}$ as an example.

1) Personality Influence: Personality traits affect emotional transitions in conversation [35]. Emotional transitions measure the variation between the current emotion and the preceding emotion of the same speaker. We decide to model this transition process to capture the personality effect on emotions. First, we need to define a function $\mathcal{H}(\cdot)$ to map the index of utterance u_t to its immediate preceding utterance of the same speaker. This function should satisfy the following conditions:

$$\begin{aligned} &\max\ \mathcal{H}(t) \\ &\text{s.t. } \mathcal{H}(t) < t \\ &\quad \mathcal{S}(\mathcal{H}(t)) = \mathcal{S}(t) \end{aligned} \tag{8}$$

since the function $\mathcal{S}(\cdot)$ maps the index of utterance to its corresponding speaker, $\mathcal{S}(\mathcal{H}(t)) = \mathcal{S}(t)$ ensures that $u_{\mathcal{H}(t)}$ and u_t belong to the same speaker. Therefore, the maximum $\mathcal{H}(t)$ that satisfies $\mathcal{H}(t) < t$ ensures that $u_{\mathcal{H}(t)}$ is the immediate preceding utterance of u_t with the same speaker.

Emotional transitions are triggered by contextual information and are affected by personality traits [35]. At each stage k, we take the outputs of the previous stage (i.e., $k-1$) as the inputs. To simulate the transition process of u_t, we leverage the personality trait $p_{\mathcal{S}(t)} \in \mathbb{R}^{d_p}$ along with the contextual feature $h_t^{[k-1]} \in \mathbb{R}^{2d_h}$ and the preceding emotion $\hat{y}_{\mathcal{H}(t)}^{[k-1]} \in \mathbb{R}^c$ to compose the variation in emotional transitions. Since these features have different feature dimensions, we first equalize their dimensions through separate fully-connected layers

$$q_p^{[k-1]} = \tanh\left(W_p^q p_{\mathcal{S}(t)} + b_p^q\right) \tag{9}$$

$$q_h^{[k-1]} = \tanh\left(W_h^q h_t^{[k-1]} + b_h^q\right) \tag{10}$$

$$q_y^{[k-1]} = \tanh\left(W_y^q \hat{y}_{\mathcal{H}(t)}^{[k-1]} + b_y^q\right) \tag{11}$$

since $\hat{y}_{\mathcal{H}(t)}^{[k-1]} \in \mathbb{R}^c$ has the minimum feature dimension, we equalize the dimension of all features to c, thus generating fixed-size outputs $q_m^{[k-1]} \in \mathbb{R}^c, m \in \{p, h, y\}$. Here, $\{W_m^q, b_m^q\}, m \in \{p, h, y\}$ are the trainable parameters. Different features contribute differently to emotional transitions. To prioritize important features and prevent being overwhelmed by other unimportant features, we exploit the attention mechanism to calculate the variation $\Delta y_t^{[k-1]} \in \mathbb{R}^c$ in emotional

transitions. The calculation formula is shown as follows:

$$\alpha_m^{[k-1]} = \frac{\exp\left(W_m^\alpha q_m^{[k-1]} + b_m^\alpha\right)}{\sum_i \exp\left(W_i^\alpha q_i^{[k-1]} + b_i^\alpha\right)}, \quad m \in \{p, h, y\} \tag{12}$$

$$\Delta y_t^{[k-1]} = \sum_m \alpha_m^{[k-1]} q_m^{[k-1]} \tag{13}$$

where $\alpha_m^{[k-1]} \in \mathbb{R}^1, m \in \{p, h, y\}$ is the attention score for each input. Here, $\{W_m^\alpha, b_m^\alpha\}, m \in \{p, h, y\}$ are the trainable parameters in the attention mechanism. For each utterance u_t, we generate the personality-enhanced result $\widetilde{y}_t^{[k-1]} \in \mathbb{R}^c$ by shifting the preceding emotion with the variation

$$\widetilde{y}_t^{[k-1]} = \hat{y}_{\mathcal{H}(t)}^{[k-1]} + \Delta y_t^{[k-1]}. \tag{14}$$

Finally, we use this personality-enhanced result $\widetilde{y}_t^{[k-1]} \in \mathbb{R}^c$ to improve the prediction result carried from the previous stage $\hat{y}_t^{[k-1]} \in \mathbb{R}^c$. To focus on important information, we integrate these features through the attention mechanism, generating $f_t^{[k]} \in \mathbb{R}^c$. This output is generated for all utterances in the conversation, marked as $F^{[k]} = \{f_t^{[k]}\}_{t=1}^N$.

2) Contextual Information: To successfully predict the emotional state of each utterance in a conversation, we should take full advantage of its contextual information in the surrounding utterances [2]. Therefore, we further feed $F^{[k]} = \{f_t^{[k]}\}_{t=1}^N$ into a bidirectional GRU for context-sensitive modeling and generate $H^{[k]} = \{h_t^{[k]}\}_{t=1}^N$. Then we leverage the fully-connected layer and the softmax activation to predict emotion probabilities $\hat{Y}^{[k]} = \{\hat{y}_t^{[k]}\}_{t=1}^N$. Here, $\hat{y}_t^{[k]} \in \mathbb{R}^c$ is the enhanced prediction result. Same as the initial stage, we choose the cross-entropy loss to optimize all trainable parameters. The loss function of stage k is calculated as follows:

$$L_e^{[k]} = -\frac{1}{\sum_{i=1}^{N_c} N_i} \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} y_{i,j} \log\left(\hat{y}_{i,j}^{[k]}\right) \tag{15}$$

where $y_{i,j} \in \mathbb{R}^c$ and $\hat{y}_{i,j}^{[k]} \in \mathbb{R}^c$ are the ground-truth emotion label and the predicted emotion label for the jth utterance in the conversation $\mathcal{U}_i$, respectively. The pseudo-code of the training process is summarized in Algorithm 1.

Algorithm 1 Pseudo-Code of Training Process

Require: Emotion dataset $\mathcal{D}_e$ and Personality dataset $\mathcal{D}_p$
function EMOTION RECOGNITION($\mathcal{D}_e$)
  Model parameters $\theta_e^{[0]}$ and the learning rate r_e
  Initialize model parameters with random values
  while loss decreases **do**
    Randomly select a batch B_e from $\mathcal{D}_e$
    Forward B_e to predict emotion labels
    Obtain loss $L_e^{[0]}$ and gradient descent $\nabla L_e^{[0]}(\theta_e^{[0]})$
    Update parameters $\theta_e^{[0]} \leftarrow \theta_e^{[0]} - r_e \nabla L_e^{[0]}(\theta_e^{[0]})$
  end while
  return Trained model $\mathcal{F}_e^{[0]}$ with $\theta_e^{[0]}$
end function

function PERSONALITY RECOGNITION($\mathcal{D}_p$)
  Model parameters θ_p and the learning rate r_p
  Initialize model parameters with random values
  while loss decreases **do**
    Randomly sample a batch B_p from $\mathcal{D}_p$
    Forward B_p to predict personality traits
    Obtain loss L_p and its gradient descent $\nabla L_p(\theta_p)$
    Update parameters $\theta_p \leftarrow \theta_p - r_p \nabla L_p(\theta_p)$
  end while
  return Trained model $\mathcal{F}_p$ with θ_p
end function

function PIRNET($\mathcal{D}_e$, $\mathcal{D}_p$)
  $\mathcal{F}_e^{[0]}$ = EMOTION RECOGNITION ($\mathcal{D}_e$)
  $\mathcal{F}_p$ = PERSONALITY RECOGNITION ($\mathcal{D}_p$)
  Use $\mathcal{F}_e^{[0]}$ to calculate emotion results in $\mathcal{D}_e$
  Use $\mathcal{F}_p$ to calculate personality scores in $\mathcal{D}_e$
  $\mathcal{F}_e^{[0]}$ calculates WAF on $\mathcal{D}_e$'s test set, denoted as $res^{[0]}$
  Initialize a sufficiently small value ϵ
  Initialize the refinement stage $k = 1$
  while loss decreases **do**
    Model parameters $\theta_e^{[k]}$ and the learning rate r_e
    Randomly initialize trainable parameters
    Randomly sample a batch B_e from $\mathcal{D}_e$
    Create $\mathscr{L}$ for parameter update
    for each conversation $\mathcal{U} = \{(u_t, y_t)\}_{t=1}^N$ in B_e **do**
      Capture personality influence and generate personality-enhanced results $\{f_t^{[k]}\}_{t=1}^N$
      Capture contextual information and generate improved emotion results $\{\hat{y}_t^{[k]}\}_{t=1}^N$
      Obtain loss $L_e^{[k]}$ and stack it to $\mathscr{L}$
    end for
    Calculate the gradient to obtain $\nabla \mathscr{L}(\theta_e^{[k]})$
    Update parameters $\theta_e^{[k]} \leftarrow \theta_e^{[k]} - r_e \nabla \mathscr{L}(\theta_e^{[k]})$
  end while
  Generate the trained model $\mathcal{F}_e^{[k]}$ with $\theta_e^{[k]}$
  $\mathcal{F}_e^{[k]}$ calculates WAF on $\mathcal{D}_e$'s test set, denoted as $res^{[k]}$
  Check the convergence condition $res^{[k]} - res^{[k-1]} < \epsilon$
  if not converged **then**
    Update stage $k = k + 1$
    Go to step 33 until convergence
  end if
end function

IV. Experimental Databases and Setup

A. Emotion Datasets

To evaluate the performance of PIRNet, we conduct experiments on three popular benchmark datasets in ERC, that is, IEMOCAP [47], CMU-MOSI [48], and CMU-MOSEI [49]. Unlike laboratory-controlled datasets, these datasets either mimic real-world conditions or are collected from social media platforms, making them challenging for ERC. In Table II, we present statistics and data partitions of these datasets.

IEMOCAP contains multiple conversations between two speakers. Each conversation is divided into small utterances, and each utterance is annotated with discrete emotion labels. For a fair comparison, we follow two popular label process manners, resulting in four-class and six-class datasets. In the four-class dataset, we consider four labels containing *anger*, *happiness*, *sadness*, and *neutral*, where the *excitement* category is merged into the *happiness* category [2], [23]. In the six-class dataset, we consider six labels, including *anger*, *happiness*, *sadness*, *neutral*, *excitement*, and *frustration* [43], [50].

CMU-MOSI is a collection of movie review videos from online websites. Each utterance is annotated with a sentiment intensity score ranging from $[-3, +3]$. Here, -3 represents the extreme score of negative sentiment and $+3$ represents the extreme score of positive sentiment.

TABLE II
STATISTICS AND DATA PARTITIONS OF EMOTION DATASETS

Dataset	# conversations		# utterances	
	train & val	test	train & val	test
IEMOCAP (four-class)	120	31	4290	1241
IEMOCAP (six-class)	120	31	5810	1623
CMU-MOSI	62	31	1513	686
CMU-MOSEI	2549	676	18197	4659

CMU-MOSEI is an extended version of CMU-MOSI with more utterances and a greater variety of topics. Following the annotation method in CMU-MOSI, each utterance is labeled ranging from $[-3, +3]$ to reflect the sentiment intensity.

B. Evaluation Metrics

In this article, we focus on the classification performance of different methods. Since IEMOCAP is annotated with discrete emotion labels, we can naturally evaluate its classification performance. CMU-MOSI and CMU-MOSEI are labeled with regression values ranging from $[-3, 3]$. For these datasets, we threshold the regression values to categorical values. Specifically, we assign <0 and >0 scores to positive and negative emotion classes, respectively. Due to the inherent imbalance across various emotions [47]–[49], we choose weighted average F1-score (WAF) as the primary evaluation metric and weighted average accuracy (WAA) as the secondary evaluation metric, in line with previous works [51], [52]. For these evaluation metrics, higher values indicate better performance.

C. Multimodal Features

Compared with handcrafted features, deep features have recently achieved remarkable performance in emotion recognition [11]. In this section, we focus on deep features and describe the feature extraction process in detail.

1) Acoustic Features: We extract acoustic features using the pretrained wav2vec [53], a multilayer convolutional neural network trained on large amounts of unlabeled data. Recently, wav2vec has proved its ability in various tasks such as speech recognition, speaker verification, and language identification [53], [54]. Inspired by its success, we attempt to leverage wav2vec for emotion recognition. Specifically, we utilize the pretrained *wav2vec-large* model to extract 512-D acoustics features for each utterance.

2) Lexical Features: We extract lexical features via the pretrained RoBERTa [55]. It is an improved approach for training BERT [56], including using bigger batches over more data and removing the next sentence prediction objective. Recently, RoBERTa has achieved performance improvements in multiple tasks such as reading comprehension, question answering, and sentiment analysis [57], [58]. Inspired by its success, we attempt to leverage RoBERTa for ERC. Specifically, we leverage the pretrained *RoBERTa-large* model to extract 768-D lexical features for each utterance.

3) Multimodal Features: To focus on important modalities, we draw inspiration from recent work [22] and utilize the attention mechanism for multimodal fusion. Concretely, we first equalize the feature dimension of all unimodal features. Then we learn attention weights to evaluate the importance of each modality, followed by the weighted sum for multimodal fusion. Finally, we extract 100-D multimodal features.

D. Implementation Details

Based on the emotion recognition performance, we set hyperparameters as follows. In this work, we first predict the emotional state of each utterance and the personality score of each speaker. The "emotion recognition" module in Section III-B contains a bidirectional GRU layer, and the dimension of hidden representations is set to $d_h = 100$. The "personality recognition" module in Section III-B consists of two fully-connected layers, and the output feature dimension is set to 5. Then we feed these preliminary results into PIRNet, aiming to improve the emotion recognition performance. PIRNet contains a bidirectional GRU layer, and we set the number of hidden units to c. Here, c is the number of discrete emotion labels in the corpus. We terminate the iteration when the improvement of WAF is less than $\epsilon = 0.001$.

We use the Adam scheme [59] to optimize all trainable parameters. For the "personality recognition" module, we set the mini-batch size to 64 and the learning rate to $r_p = 0.0001$. To help convergence and improve generalization, we also employ Dropout [60] with a rate of 0.5 behind fully-connected layers. For the "emotion recognition" module and PIRNet, we set the learning rate to $r_e = 0.0001$. We also employ L2 regularization with a weight of 0.00001 to alleviate overfitting. Different conversations have distinct numbers of utterances, resulting in variable-length inputs. To implement our model in TensorFlow [61], we pad the conversations of the same mini-batch to have the same number of utterances. Bit masking is also utilized to remove the effect of padding.

E. Baseline Models

Since contextual information plays an important role in ERC, researchers have proposed various contextual models. To evaluate the performance of the proposed method, we implement these contextual models as our baselines: **BC-LSTM** [2] builds an LSTM-based model that enables utterances to capture contextual information from their surroundings. **CHFusion** [62] and **CAT-LSTM** [63] are two extended versions of **BC-LSTM**, including the use of the hierarchical structure and the attention mechanism. To further capture interactions between different speakers, **CMN** [23], **ICON** [24], **DialogueRNN** [43], and **A-DMN** [64] employ RNNs to learn separate contexts for each speaker. Different from these methods, **HiTrans** [65] exploits a pairwise speaker verification task to make the model speaker-sensitive.

Besides contextual models, we also compare the proposed method with the following multimodal fusion strategies: **TFN** [66] introduces a tensor fusion network to learn both intramodality and intermodality dynamics. However, **TFN** is generally constrained by computational and memory costs. To this end, **LMF** [67], **LMFN** [51], **HFFN** [68], and **Dual-LMF** [69] utilize various approaches to improve

TABLE III
MULTIMODAL EMOTION RECOGNITION RESULTS ON THE TEST SET OF IEMOCAP. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Approaches	IEMOCAP (four-class)		Approaches	IEMOCAP (six-class)	
	WAF	WAA		WAF	WAA
BC-LSTM [2]	–	75.6	ARGF [50]	59.53	60.69
CMN [23]	–	77.62	DialogueRNN [43]	62.75	63.40
CHFusion [62]	76.0	76.1	ICON [24]	63.2	63.8
RLC [5]	79.1	–	A-DMN [64]	64.1	64.9
HFFN [68]	80.60	80.38	HiTrans [65]	64.50	–
M3ER [77]	82.4	82.7	DialogXL [79]	65.94	–
LMFN [51]	82.54	82.45	DialogueTRM [75]	69.23	68.92
PIRNet (Ours)	**86.96**	**86.95**	PIRNet (Ours)	**72.18**	**72.21**

the efficiency of **TFN**. In addition to the above methods, modality factorization and modality translation are also widely utilized in multimodal fusion. Modality factorization methods decompose multimodal features into distinct subspaces (such as **MFM** [70], **MISA** [71], and **MMB2** [72]). Modality translation methods learn joint representations between two modalities (such as **MCTN** [73], **ARGF** [50], and **CIA** [74]). For descriptions of the remaining baselines, we recommend readers refer to **GMFN** [49], **RLC** [5], **DialogueTRM** [75], **MulT** [76], **ICCN** [52], **M3ER** [77], **QMF** [78], and **DialogXL** [79].

V. RESULTS AND DISCUSSION

We first conduct comparative experiments between PIRNet and currently advanced emotion recognition models. Then we investigate the importance of different modalities and study the generalization performance of personality recognition. Subsequently, we reveal the necessity of each component in PIRNet, including personality trails and iterative methods. To further verify the effectiveness of our proposed method, we also compare PIRNet with other refinement models for ERC. In this section, we treat the "emotion recognition" module in Section III-B as a comparison system. It does not include any emotion refinement process, referred to as **BIGRU-ERC**.

A. Performance on Emotion Recognition

Tables III–V present the experimental results on different datasets. For IEMOCAP (four-class), PIRNet achieves new state-of-the-art records with 86.96% on WAF and 86.95% on WAA, which shows an absolute improvement of 4.42% on WAF and 4.25% on WAA. For IEMOCAP (six-class), our method outperforms the currently advanced approaches by 2.95% on WAF and 3.29% on WAA. Experimental results in Tables IV and V demonstrate that PIRNet also exhibits performance improvement on CMU-MOSI and CMU-MOSEI. To sum up, we can achieve competitive results on all datasets, highlighting the capabilities of PIRNet for emotion recognition. These improvements are caused by the following reasons.

1) In Tables III–V, we observe that PIRNet succeeds over existing contextual models on emotion recognition (such as **BC-LSTM**, **CAT-LSTM**, and **DialogueRNN**). Compared with these contextual models, PIRNet can take full advantage of personality traits to improve emotion recognition performance through a multistage iterative strategy. These results demonstrate the effectiveness of our proposed method.
2) In Tables III–V, experimental results demonstrate that PIRNet also exhibits performance improvement over existing multimodal fusion methods (such as **TFN**, **MFM**, and **LMF**). These multimodal fusion methods do not explicitly model contextual information in conversations, while contextual information plays an important role in ERC. In this article, we exploit bidirectional recurrent layers to capture contextual information in both historical and future utterances, resulting in better emotion recognition performance. These results demonstrate the importance of contextual information and verify the effectiveness of our contextual modeling strategy.

TABLE IV
MULTIMODAL EMOTION RECOGNITION RESULTS ON THE TEST SET OF CMU-MOSI. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Approaches	WAF	WAA
CHFusion [62]	–	79.1
BC-LSTM [2]	–	79.33
MMB2 [72]	75.1	75.2
LMF [67]	75.7	76.4
GMFN [49]	77.0	76.9
TFN [66]	77.9	77.1
MFM [70]	78.1	78.1
HFFN [68]	78.29	78.06
Dual-LMF [69]	78.3	78.4
CIA [74]	78.98	79.15
MCTN [73]	79.1	79.3
CAT-LSTM [63]	80.1	–
ICCN [52]	80.56	80.59
LMFN [51]	80.92	80.85
PIRNet (Ours)	**81.63**	**81.55**

TABLE V
MULTIMODAL EMOTION RECOGNITION RESULTS ON THE TEST SET OF CMU-MOSEI. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Approaches	WAF	WAA
GMFN [49]	77.0	76.9
QMF [78]	77.48	79.71
CIA [74]	77.80	80.06
MulT [76]	82.3	82.5
ICCN [52]	82.96	82.80
MISA [71]	85.3	85.5
PIRNet (Ours)	**85.59**	**85.64**

B. Importance of Different Modalities

To investigate the importance of different modalities, we present unimodal and multimodal results in Table VI. From Table VI, we observe that PIRNet achieves performance improvement in both unimodal and multimodal conditions. For IEMOCAP (four-class), PIRNet shows an absolute improvement of 2.27%–2.87% on WAF and 2.18%–2.82% on WAA over **BIGRU-ERC**. For IEMOCAP (six-class), PIRNet succeeds over **BIGRU-ERC** by 1.63%–3.62% on WAF and 1.60%–3.69% on WAA. We can also find the same phenomenon in other datasets. Since personality influence

TABLE VI

PERFORMANCE OF PIRNet USING DIFFERENT MODALITY COMBINATIONS. BEST RESULTS ARE HIGHLIGHTED IN BOLD. Δ SHOWS THE IMPROVEMENT OR REDUCTION OVER THE COMPARISON SYSTEM. IMPROVEMENTS ARE HIGHLIGHTED IN GREEN

Model	Modality	IEMOCAP (four-class)		IEMOCAP (six-class)		CMU-MOSI		CMU-MOSEI	
		WAF	WAA	WAF	WAA	WAF	WAA	WAF	WAA
BIGRU-ERC	A	79.05	79.05	62.77	62.54	60.26	60.21	70.21	70.64
PIRNet (Ours)	A	**81.50**	**81.55**	**66.17**	**65.80**	**62.80**	**64.02**	**71.79**	**72.62**
Δ	A	↑2.45	↑2.50	↑3.40	↑3.26	↑2.54	↑3.81	↑1.58	↑1.98
BIGRU-ERC	L	82.43	82.59	66.41	66.24	79.18	79.42	82.51	82.86
PIRNet (Ours)	L	**85.30**	**85.41**	**70.03**	**69.93**	**80.59**	**80.64**	**84.54**	**84.65**
Δ	L	↑2.87	↑2.82	↑3.62	↑3.69	↑1.41	↑1.22	↑2.03	↑1.79
BIGRU-ERC	A+L	84.69	84.77	70.55	70.61	80.11	80.18	84.23	84.20
PIRNet (Ours)	A+L	**86.96**	**86.95**	**72.18**	**72.21**	**81.63**	**81.55**	**85.59**	**85.64**
Δ	A+L	↑2.27	↑2.18	↑1.63	↑1.60	↑1.52	↑1.37	↑1.36	↑1.44

and contextual information are crucial in ERC, PIRNet exploits these two factors to improve prediction results of **BIGRU-ERC**, resulting in better emotion recognition performance.

In Table VI, we observe that lexical results outperform acoustic results in emotion recognition. This phenomenon can also be found in previous works [24]. Since text tends to have less noise than audio in our emotion datasets [23], we can learn more emotion-salient features from the text. Meanwhile, we observe that multimodal results succeed over unimodal results in all cases. Humans express emotions through various modalities. To better understand emotional states, we exploit the attention mechanism to integrate multimodal information. These results demonstrate the importance of multisource information and the effectiveness of our fusion strategy.

Meanwhile, experimental results in Table VI demonstrate that the improvement brought by PIRNet is different for distinct datasets. Compared with CMU-MOSI and CMU-MOSEI, PIRNet brings more performance improvement to IEMOCAP. These results suggest that personality influence and contextual information in PIRNet are more useful in IEMOCAP. Furthermore, we observe that PIRNet brings more performance improvement to unimodal results than multimodal results. The reason lies in that multimodal results have already gained performance improvement from our multimodal fusion strategy, which may weaken the effectiveness of PIRNet.

C. Generalization Performance of Personality Recognition

Since emotion datasets do not have personality annotations, it is hard to study the generalization performance of personality detection models on emotion datasets. To this end, we conduct experiments between two benchmark personality detection datasets: ChaLearn First Impression Database [46] and Essays [80]. The first one is collected from YouTube, while the second one contains multiple anonymous essays in a controlled environment. These datasets leverage the Big-Five Model to measure personality. To investigate the generalization performance of personality detection, we conduct experiments in both within-corpus and cross-corpus settings. Experimental results are listed in Table VII.

1) *Within-corpus results:* In this setting, we conduct experiments on ChaLearn First Impression Database. Specifically, we train the model on its training set and evaluate the performance on its test set.
2) *Cross-corpus (w/o FC) results:* In this setting, we conduct cross-corpus experiments between ChaLearn First Impression Database and Essays. Specifically, we train the model on Essays and evaluate the performance on the test set of ChaLearn First Impression Database.
3) *Cross-corpus (w FC) results:* Personality scores from different datasets are correlated, but there are also inevitable differences. In this cross-corpus setting, we further utilize a fully-connected layer to reduce annotation bias between different datasets.

TABLE VII

GENERALIZATION PERFORMANCE OF PERSONALITY RECOGNITION

	Within-corpus	Cross-corpus (w/o FC)	Cross-corpus (w FC)
Extraversion	0.8846	0.8694	0.8773
Neuroticism	0.8854	0.7969	0.8768
Agreeableness	0.8980	0.8163	0.8932
Conscientiousness	0.8864	0.8654	0.8739
Openness	0.8905	0.8830	0.8827
Average	0.8890	0.8462	0.8808

In Table VII, we observe that cross-corpus results are slightly worse than within-corpus results. It confirms that there is annotation bias between different datasets. With the help of the fully-connected layer, we can reduce the discrepancy between within-corpus and cross-corpus results. Therefore, we also leverage the fully-connected layer [see (9)] to reduce annotation bias in our implementation.

D. Importance of Personality Traits

In this section, we implement a comparison system to verify the importance of personality traits in our proposed method. Experimental results are listed in Table VIII.

1) *PIRNet:* To capture the personality effect on emotions, this model leverages personality traits, contextual representations, and preceding emotions to simulate emotional transitions and generate personality-enhanced results.
2) *PIRNet-NP:* It comes from PIRNet but omits personality traits. Specifically, this model only utilizes contextual

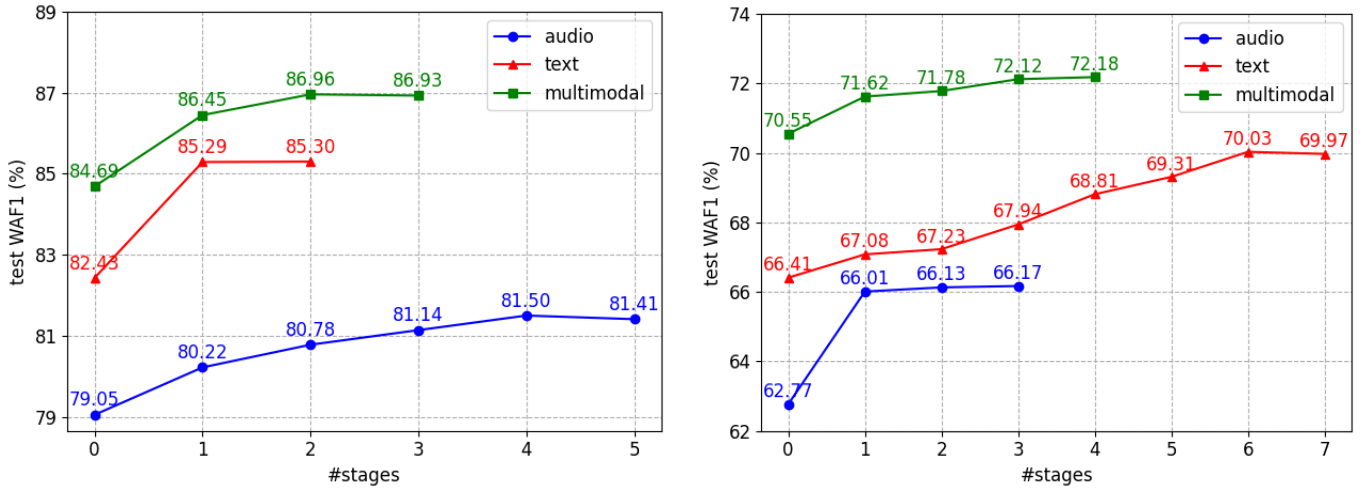

Fig. 2. Effect of number of stages ($\epsilon = 0.001$). (a) Experiments on IEMOCAP (four-class). (b) Experiments on IEMOCAP (six-class).

TABLE VIII
IMPORTANCE OF PERSONALITY TRAITS IN PIRNet. BOLD FRONT REPRESENTS THE BEST PERFORMANCE

Modality	Model	IEMOCAP (four-class)		IEMOCAP (six-class)	
		WAF	WAA	WAF	WAA
A	PIRNet-NP	80.32	80.34	63.80	63.34
A	PIRNet (Ours)	**81.50**	**81.55**	**66.17**	**65.80**
L	PIRNet-NP	85.00	85.09	66.57	66.30
L	PIRNet (Ours)	**85.30**	**85.41**	**70.03**	**69.93**
A+L	PIRNet-NP	86.49	86.46	**72.32**	**72.27**
A+L	PIRNet (Ours)	**86.96**	**86.95**	72.18	72.21

representations and preceding emotions to compute the variation in emotional transitions.

From Table VIII, we observe that the proposed method outperforms **PIRNet-NP** in most cases. Personality traits affect human perception and expression of emotional information [27]. In PIRNet, we can capture personality influence through a multistage iterative strategy. These results highlight the importance of personality information in emotion recognition.

E. Effectiveness of Iterative Methods

To verify the effectiveness of our iterative method, we present experimental results of different stages in Fig. 2. Specifically, we utilize PIRNet-RK to represent the number of stages. Therefore, PIRNet-R0 represents a system without emotion refinement (i.e., **BIGRU-ERC**). From Fig. 2, we observe that increasing the number of stages can improve the emotion recognition performance to a certain point. However, further increases result in limited performance improvement or performance degradation and then trigger the termination criterion when the improvement is less than a sufficiently small value $\epsilon = 0.001$. The reason lies in that too many stages may increase the risk of over-fitting. These results also verify the effectiveness of our iterative method in PIRNet.

F. Comparison With Other Refinement Models

In addition to PIRNet, we can also utilize other models for emotion refinement. In this section, we implement three comparison systems to verify the effectiveness of our method. Experimental results are listed in Table IX.

1) *LSTM-CRF:* It uses a bidirectional LSTM layer and a CRF layer for emotion refinement. LSTM captures contextual information in both directions. CRF exploits past and future emotion labels to decode the best label sequence for the entire conversation.
2) *LSTM-GCN:* It employs LSTM to capture contextual information, followed by GCN for interaction modeling. GCN symbolizes conversations into heterogeneous graphs. The nodes in the graph represent individual utterances. The edges between a pair of nodes represent dependencies between the speakers of those utterances.
3) *DECN:* It employs graph networks to model human interactions, followed by an attention-based GRU layer to capture context-sensitive dependencies.
4) *PIRNet:* It combines personality influence and contextual information through a multistage iterative strategy.

In Table IX, we observe that with the help of emotion refinement models, we can achieve better performance than **BIGRU-ERC** in most cases. These results confirm the necessity of the emotion refinement process. Meanwhile, compared with other methods, PIRNet can achieve the best performance in most cases. Existing methods do not consider the influence of personality on emotion recognition, thus limiting their performance. In this article, PIRNet further captures personality influence through a multistage iterative strategy and achieves better emotion recognition performance.

In Fig. 3, we visualize the confusion matrices of **BIGRU-ERC** and PIRNet on IEMOCAP(four-class). We observe that improvements in classification performance can be seen for all emotion categories. With the help of PIRNet, we can incorporate personality influence and contextual information to correct some misclassified samples in preliminary emotion results and further improve the

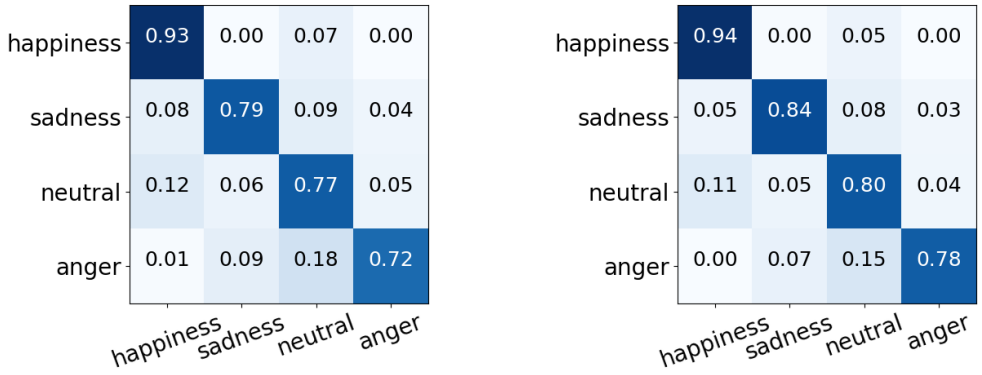

Fig. 3. Visualization confusion matrices on the test set of IEMOCAP (four-class). (a) Experiments on BIGRU-ERC. (b) Experiments on PIRNet.

TABLE IX
COMPARED WITH OTHER REFINEMENT MODELS ON THE TEST SET OF IEMOCAP. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD AND THE SECOND HIGHEST RESULT IS LABELED BY †

Modality	Model	IEMOCAP (four-class)		IEMOCAP (six-class)	
		WAF	WAA	WAF	WAA
A	BIGRU-ERC	79.05	79.05	62.77	62.54
A	LSTM-CRF [81]	79.64	79.61	63.44	63.09
A	LSTM-GCN [82]	80.26	80.26	65.00	65.13
A	DECN [11]	80.64†	80.74†	**66.37**	**66.17**
A	PIRNet (Ours)	**81.50**	**81.55**	66.17†	65.80†
L	BIGRU-ERC	82.43	82.59	66.41	66.24
L	LSTM-CRF [81]	83.65	83.72	67.51†	67.41†
L	LSTM-GCN [82]	83.08	83.16	66.31	66.05
L	DECN [11]	85.05†	85.17†	67.15	66.97
L	PIRNet (Ours)	**85.30**	**85.41**	**70.03**	**69.93**
A+L	BIGRU-ERC	84.69	84.77	70.55	70.61
A+L	LSTM-CRF [81]	84.33	84.37	70.88	70.98
A+L	LSTM-GCN [82]	85.19	85.25	71.85†	71.84†
A+L	DECN [11]	86.49†	86.46†	71.26	71.29
A+L	PIRNet (Ours)	**86.96**	**86.95**	**72.18**	**72.21**

emotion recognition performance. These results verify the effectiveness of our method.

VI. CONCLUSION

In this article, we propose "PIRNet," a novel framework for emotion refinement. It is an iterative method with multiple stages. Each stage exploits personality influence and contextual information to improve emotion recognition performance. Experimental results on three benchmark datasets demonstrate the effectiveness of our method. PIRNet can improve the emotion recognition performance in both unimodal and multimodal conditions. Meanwhile, we confirm the necessity of each component in PIRNet, including personality traits and the multistage iterative strategy. Furthermore, we observe that PIRNet also succeeds over existing emotion refinement methods. All these results demonstrate the effectiveness of our PIRNet.

In the future, we will extend the applications of our proposed method. Besides ERC, we will leverage PIRNet for other types of conversation understanding tasks. Meanwhile, besides lexical and acoustic modalities, we will further exploit visual information to improve emotion recognition performance. Furthermore, in addition to the personality effect on emotions, we will also utilize emotion results to improve personality recognition performance in our future work.

REFERENCES

[1] J. Tao and T. Tan, *Affective Information Processing*. London, U.K.: Springer-Verlag, 2009.
[2] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 873–883.
[3] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-LM: A neural language model for customizable affective text generation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 634–642.
[4] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017.
[5] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5060–5066.
[6] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
[7] W. Revelle and K. R. Scherer, "Personality and emotion," in *Oxford Companion to Emotion and the Affective Sciences*, vol. 1. Oxford Univ. Press, 2009, pp. 304–306.
[8] K. Winter, "Individual differences in the experience of emotions," *Clin. Psychol. Rev.*, vol. 17, no. 7, pp. 791–821, Nov. 1997.
[9] J.-L. Li and C.-C. Lee, "Attention learning with retrievable acoustic embedding of personality for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 171–177.
[10] Y. Li, A. Kazameini, Y. Mehta, and E. Cambria, "Multitask learning for emotion and personality detection," 2021, *arXiv:2101.02346*.
[11] Z. Lian, B. Liu, and J. Tao, "DECN: Dialogical emotion correction network for conversational emotion recognition," *Neurocomputing*, vol. 454, pp. 483–495, Sep. 2021.
[12] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, Dec. 2020.
[13] K. Wołk, A. Wołk, D. Wnuk, T. Grześ, and I. Skubis, "Survey on dialogue systems including slavic languages," *Neurocomputing*, vol. 477, pp. 62–84, Mar. 2022.
[14] T. Fu, S. Gao, X. Zhao, J.-R. Wen, and R. Yan, "Learning towards conversational AI: A survey," *AI Open*, vol. 3, pp. 14–28, Jan. 2022.
[15] G. Tu, J. Wen, C. Liu, D. Jiang, and E. Cambria, "Context- and sentiment-aware networks for emotion recognition in conversation," *IEEE Trans. Artif. Intell.*, early access, Feb. 7, 2022, doi: 10.1109/TAI.2022.3149234.
[16] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 730–739.

[17] F. Cui, H. Di, L. Shen, K. Ouchi, Z. Liu, and J. Xu, "Modeling semantic and emotional relationship in multi-turn emotional conversations using multi-task learning," *Int. J. Speech Technol.*, vol. 52, no. 4, pp. 4663–4673, Mar. 2022.

[18] Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, and J. Zhou, "Emotional conversation generation with heterogeneous graph neural network," *Artif. Intell.*, vol. 308, Jul. 2022, Art. no. 103714.

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

[20] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, 2014, pp. 1–9.

[21] R. Zhang, A. Ando, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue," in *Proc. Interspeech*, 2017, pp. 1094–1097.

[22] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 985–1000, 2021.

[23] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 2122–2132.

[24] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.

[25] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, Jan. 2022.

[26] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowl.-Based Syst.*, vol. 236, Jan. 2022, Art. no. 107751.

[27] E. G. Kehoe, J. M. Toomey, J. H. Balsters, and A. L. W. Bokde, "Personality modulates the effects of emotional arousal and valence on brain activation," *Social Cognit. Affect. Neurosci.*, vol. 7, no. 7, pp. 858–870, Oct. 2012.

[28] D. J. Ozer and V. Benet-Martínez, "Personality and the prediction of consequential outcomes," *Annu. Rev. Psychol.*, vol. 57, no. 1, pp. 401–421, Jan. 2006.

[29] M. Carlyn, "An assessment of the myers-briggs type indicator," *J. Personality Assessment*, vol. 41, no. 5, pp. 461–473, Oct. 1977.

[30] H. J. Eysenck, *A Model for Personality*. Berlin, Germany: Springer-Verlag, 1981.

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

[32] E. Komulainen *et al.*, "The effect of personality on daily life emotional processes," *PLoS ONE*, vol. 9, no. 10, pp. 1–9, 2014.

[33] K. A. Barford and L. D. Smillie, "Openness and other big five traits in relation to dispositional mixed emotions," *Personality Individual Differences*, vol. 102, pp. 118–122, Nov. 2016.

[34] G. Mohammadi and P. Vuilleumier, "A multi-componential approach to emotion recognition and the effect of personality," *IEEE Trans. Affect. Comput.*, early access, Oct. 1, 2020, doi: 10.1109/TAFFC.2020.3028109.

[35] Z. Wen, J. Cao, R. Yang, S. Liu, and J. Shen, "Automatically select emotion for response via personality-affected emotion transition," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 5010–5020.

[36] Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, and J. Zhou, "Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 15, pp. 13343–13352.

[37] C. T. Kelley, *Iterative Methods for Optimization* (Frontiers in Applied Mathematics). Philadelphia, PA, USA: SIAM, 1999.

[38] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bureau Standards*, vol. 49, no. 6, pp. 409–436, Dec. 1952.

[39] D. F. Shanno, "Conditioning of quasi-Newton methods for function minimization," *Math. Comput.*, vol. 24, no. 111, pp. 647–656, Jul. 1970.

[40] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*. Philadelphia, PA, USA: SIAM, 1995.

[41] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[42] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2252–2261.

[43] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6818–6825.

[44] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[45] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[46] V. Ponce-López *et al.*, "ChaLearn LAP 2016: First round challenge on first impressions-dataset and results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Switzerland: Springer, 2016, pp. 400–418.

[47] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.

[48] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

[49] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2236–2246.

[50] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 164–172.

[51] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2020.

[52] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8992–8999.

[53] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[54] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," 2020, *arXiv:2012.06185*.

[55] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186.

[57] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.

[58] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. (NLP)*, 2018, pp. 353–355.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[61] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[62] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[63] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.

[64] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, early access, Jun. 29, 2020, doi: 10.1109/TAFFC.2020.3005660.

[65] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, "HiTrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4190–4200.

[66] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[67] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[68] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.
[69] T. Jin, S. Huang, Y. Li, and Z. Zhang, "Dual low-rank multimodal fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process., Findings*, 2020, pp. 377–387.
[70] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. 7th Int. Conf. Learn. Represent.*, 2019, pp. 1–20.
[71] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[72] P. P. Liang, Y. C. Lim, Y.-H.-H. Tsai, R. Salakhutdinov, and L.-P. Morency, "Strong and simple baselines for multimodal utterance embeddings," in *Proc. Conf. North*, 2019, pp. 2599–2609.
[73] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6892–6899.
[74] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5651–5661.
[75] Y. Mao *et al.*, "DialogueTRM: Exploring the Intra- and inter-modal emotional behaviors in the conversation," 2020, *arXiv:2010.07637*.
[76] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[77] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1359–1367.
[78] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Inf. Fusion*, vol. 65, pp. 58–71, Jan. 2021.
[79] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 13789–13797.
[80] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *J. Personality Social Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999.
[81] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.
[82] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 154–164.

Zheng Lian received the B.S. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2016, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2021.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include affective computing, deep learning, and multimodal emotion recognition.

Bin Liu (Member, IEEE) received the B.S. and M.S. degrees from the Beijing Institute of Technology (BIT), Beijing, China, in 2007 and 2009, respectively, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, in 2015.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include affective computing and audio signal processing.

Jianhua Tao (Senior Member, IEEE) received the M.S. degree from Nanjing University, Nanjing, China, in 1996, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He has authored or coauthored more than 80 articles in major journals and proceedings, including the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His current research interests include speech recognition, speech synthesis and coding methods, human–computer interaction, multimedia information processing, and pattern recognition.

Dr. Tao was a recipient of several awards from important conferences, such as Eurospeech and NCMMSC. He is the Chair or a Program Committee Member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, and NCMMSC. He is also a Steering Committee Member of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, an Associate Editor of *Journal on Multimodal User Interface* and *International Journal on Synthetic Emotions*, and the Deputy Editor-in-Chief of *Chinese Journal of Phonetics*.