Dynamic Confidence-Aware Multi-Modal Emotion Recognition

Qi Zhu, Chuhang Zheng, Zheng Zhang, Wei Shao and Daoqiang Zhang, *Senior Member, IEEE*

Abstract—Multi-modal emotion recognition has attracted increasing attention in human-computer interaction, as it extracts complementary information from physiological and behavioral features. Compared to single modal approaches, multi-modal fusion methods are more susceptible to uncertainty in emotion recognition, such as heterogeneity and inconsistent predictions across different modalities. Previous multi-modal approaches ignore systematic modeling of uncertainty in fusion and revelation of dynamic variations in emotion process. In this paper, we propose a dynamic confidence-aware fusion network for robust recognition of heterogeneous emotion features, including electroencephalogram (EEG) and facial expression. First, we develop a self-attention based multi-channel LSTM network to preliminarily align the heterogeneous emotion features. Second, we propose a confidence regression network to estimate true class probability (TCP) on each modality, which helps explore the uncertainty at modality level. Then, different modalities are weighted fused according to above two types of uncertainty. Finally, we adopt self-paced learning (SPL) mechanism to further improve the model robustness by alleviating negative effect from the hard learning samples. The experimental results on several multi-modal emotion datasets demonstrate the proposed method outperforms the state-of-the-art methods in emotion recognition performance and explicitly reveals the dynamic variation of emotion with uncertainty estimation. Our code is available at: https://github.com/xbrainnet/CAFNet.

Index Terms—Multi-modal fusion, emotion recognition, self-attention mechanism, uncertainty, self-paced learning.

1 Introduction

EMOTION recognition is receiving increasing attention as it provides critical information for the perception and decision-making of human-computer interaction (HCI) systems [1], [2], which enables such systems to respond appropriately to the user's emotional state. Emotion is a complex psychological dynamic process generated by various feelings and thoughts [3], and people can express emotions through different behavioral patterns due to linguistic and cultural differences. In addition, the process of emotion data collection and recognition are often affected by varying degrees of uncertainty impact. Therefore, it is still a challenging task to develop robust recognition model that can effectively extract the intrinsic states of emotion.

The commonly used emotion traits include physiological signals, such as electroencephalogram (EEG) [4], electromyogram (EMG) [5] and electrocardiogram (ECG) [6], and behavioral signals, such as voice signals [7], facial expression images [8], [9], and body posture [10]. Among the physiological emotion traits, EEG has advantages of non-invasiveness, low cost and high temporal resolution [11]. Zhang et al. [12] proposed a self-attention network to jointly model the local and global temporal information of the EEG features. Li et al. [13] built an effective emotion recognition model that performs multi-category emotion recognition with multiple emotion-related spatial network topology patterns by learning discriminative graph topologies in EEG brain networks. Peng et al. [14] proposed a graph-based model by unifying the adaptive graph learning and emotion recognition into a single objective. Furthermore, facial expression in behavioral signals is one of the most intuitive forms of emotions, and many facial image-based emotion recognition methods have been proposed. Jia et al. [15] exploited low-rank structure to capture the local label correlations implicitly of emotion distribution. Zhao et al. [16] presented a facial expression recognition

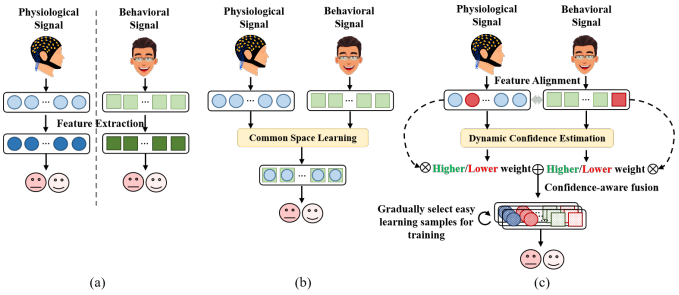

Fig. 1. The single modal and multi-modal emotion recognition method schematic diagrams. (a) Recognition based on single modal emotion data. (b) Multi-modal emotion data fusion via common space representation. (c) The proposed confidence-aware multi-modal data fusion strategy with multi-level uncertainty estimation.

- *This work is supported by National Natural Science Foundation of China (Nos. 62076129, 62371234, 62136004, and 62272226), Natural Science Foundation of Jiangsu Province (No. BK20231438) and Key Research and Development Plan of Jiangsu Province (No. BE2022842).*
- *Qi Zhu, Chuhang Zheng, Wei Shao and Daoqiang Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106. (E-mail: zhuqinuaa@163.com, {h.zheng, shaowei20022005, dqzhang}@nuaa.edu.cn).*
- *Zheng Zhang is with School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China, and also with Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: darrenzz219@gmail.com).*

(Corresponding authors: Zheng Zhang, Wei Shao and Daoqiang Zhang)

network to extract the local and global-salient facial features. However, the aforementioned methods rely on single modal analysis, which cannot fully reflect emotional states, thus making it difficult to achieve satisfactory levels of accuracy and stability.

Benefiting from the complementary information between multiple modalities, integrating physiological and behavioral signals can enhance the performance of emotion recognition [17], [18]. For example, Wang et al. [19] utilized transformer encoders with attention-based fusion to fuse EEG and eye movement data for emotion recognition. Ngai et al. [20] developed a multi-branch deep convolutional neural network using EEG signals, facial expression, and eye movement data for emotion recognition. Ma et al. [21] proposed the multi-modal residual LSTM to learn the correlation between EEG and peripheral physiological to improve the performance of recognition. The complex physiology of emotions makes uncertainty an important factor in emotion recognition. Specifically, the presence of uncertainty in various forms within multi-modal data heightens the difficulty in accurately recognizing emotions [22], [23].

In recent years, some uncertainty learning-based emotion recognition methods have been proposed. For example, Deng et al. [24] captured uncertainty from multiple emotion descriptors by adopting deep ensemble models, and applied iterative self-distillation to improve the performance of emotion recognition and uncertainty estimation. Yang et al. [25] design a new emotional representation feature, i.e., intensity quantity of information (IQI), which can be used to describe the uncertainty of signal strength in different frequency bands. Wang et al. [26] designed a self-cured network (SCN) to weight each sample in training through self-attention mechanism. It is worth noting these uncertainty analysis methods are all developed for single modal emotion prediction and cannot alleviate the more varied uncertainties in multi-modal emotion data and fusion model. In this paper, we attribute the uncertainty in multi-modal emotion data and model to the following three aspects: First, multi-modal emotion data are usually with different semantic representation, and the significance of each time window to emotion state varies due to the mechanism of emotional generation, which reflects the uncertainty at temporal level. Second, the ambiguity of the decision-making produced by different modalities leads to the uncertainty at fusion level. Third, the difficulty or confidence of recognizing emotion samples from different subjects or trails by the same model varies, reflecting the uncertainty at the sample level. The above uncertainties jointly pose challenge to multi-modal emotion recognition. Therefore, it is urgent to develop effective fusion methods to alleviate the above uncertainties and improve the performance of multi-modal emotion recognition.

To address the above problems, in this paper, we propose a dynamic confidence-aware fusion network for multi-modal emotion recognition, which systematically model uncertainties in multi-modal emotional data and reveals the dynamic variation of emotion process with uncertainty estimation. Specifically, we first introduce a multi-channel LSTM network with shared weights to achieve heterogeneous data alignment. Simultaneously, self-attention mechanism is adopted into the model to adaptively estimate the significance of different time windows to emotion, by which the negative impact of high uncertainty time windows on recognition can be reduced. Second, to avoid the problem of overconfidence in neural network prediction, we introduce true class probability (TCP) [27] to measure modality level uncertainty by training a confidence regression network. Then, the extracted features of different modalities are dynamically weighted by TCP for confidence-aware fusion. Meanwhile, KL-divergence is adopted in the model to preserve the common features between modalities. Finally, we exploit self-paced learning (SPL) in the optimization of the model, by which the samples with low uncertainty are gradually feed in the model to improve its robustness. The comparisons of our approach and existing emotion recognition methods are shown in Fig.1. Fig.1 (a) shows the single modal-based emotion recognition methods, which cannot take advantage of complementary information of different modalities. Fig. 1 (b) denotes the conventional fusion methods, and they often focus on capturing the common subspace of multi-modal data, ignoring suppressing the uncertainty in fusion. Fig. 1 (c) represents the proposed confidence-aware fusion network strategy, which can simultaneously effectively fuse the heterogeneous features and alleviate uncertainty in multi-modal emotion recognition. In brief, our proposed method has the following contributions:

1) To the best of our knowledge, it is the first work to systemically model the uncertainty of multi-modal emotion recognition, enabling confidence-aware fusion of multi-modal emotion data and explicitly revealing the dynamic variation of emotions.

2) We develop a multi-channel LSTM feature network with an attention mechanism that simultaneously aligns heterogeneous multi-modal features and adaptively predicts the emotion uncertainty at temporal level.

3) A true class probability-based confidence regression network is proposed to estimate the emotion prediction uncertainty at modality level. It allows for more interpretable and reliable multi-modal emotion fusion recognition with confidence weighting.

4) We adopt the self-paced learning to improve the robustness of the proposed model during the optimization of above fusion network. The experiment results on several multi-modal emotion datasets demonstrate the proposed method outperforms the state-of-the-art algorithms for emotion prediction.

The rest of this paper is organized as follows. In Section II, we introduce the related works. Then, we present the details of proposed method and its optimization in Section III. The experimental results are given in Section IV. In Section V, the ablation studies and several visualizations are presented. Finally, we draw a summary in Section VI.

2 Related Work

2.1 Uncertainty Modeling

Uncertainty can be classified into two categories, epistemic (sampling-based) and aleatoric (sampling-free) uncertainty. Sampling-based methods train prediction model through weighted sampling or model integration, leveraging inconsistencies in multiple results to quantify uncertainty, such as Bayesian neural network [28], Monte Carlo dropout [29],

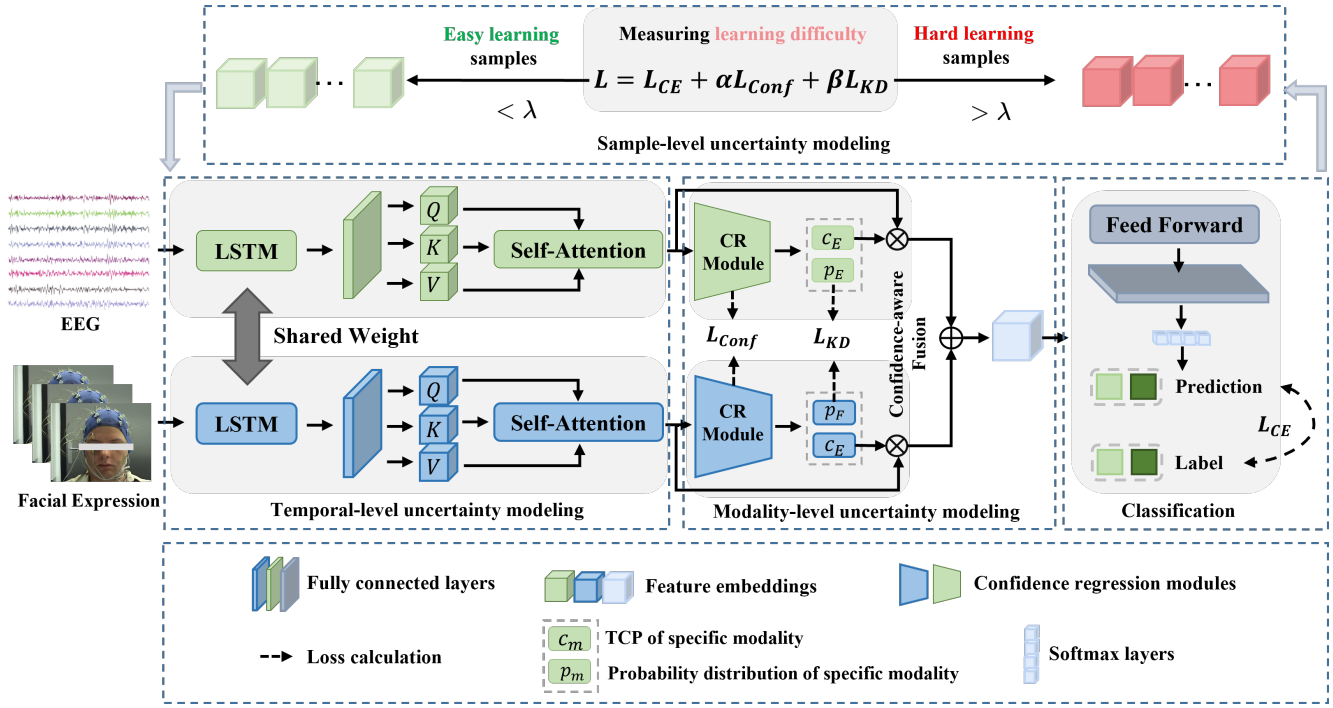

Fig. 2. The proposed multi-channel dynamic confidence-aware fusion network for emotion recognition.

and bootstrap ensemble [30]. However, the high memory and computational cost make them challenging to apply in real-time applications. Alternatively, sampling-free methods use deterministic parameters in neural network predictions to estimate uncertainty, which is more computationally efficient. For example, Dirichlet prior network [31] and evidential neural network [32] enable the quantification of uncertainty without the need for repeated evaluations. In contrast with uncertainty estimation methods, confidence calibration methods [27], [33] directly calibrate the classification results to measure confidence, which can quantify predictive uncertainty of model in a more efficient manner. In this paper, we propose a confidence-aware fusion network that models uncertainties from multiple levels and alleviates the adverse impact of uncertainty in emotion recognition.

2.2 Self-paced Learning

The concept of self-paced learning [34] is inspired by human learning processes, it is able to adjust the learning process based on samples with a simple-to-hard pattern. The goal of SPL is to learn both the model parameters by minimizing the following objective function:

$$\min_{\boldsymbol{w};v\in[0,1]^N} E(w,v;\lambda)\sum_{i=1}^{N} v_i l_i + g(v,\lambda) \tag{1}$$

where w represents the model parameters, v denotes the latent weight variables of samples, l_i is the training loss, $g(v,\lambda)$ represents the self-paced regularization term and λ controls the learning pace. By gradually increasing the parameter λ, the samples from easy to hard learning can be in turn added into the training process. Jiang et al. [35] found SPL can mitigate the issue of getting trapped in poor local optima. Zhou et al. [36] introduced self-paced learning to deep neural networks, allowing samples to be adaptively incorporated into the training process. Moreover, Fan et al. [34] proposed a generalized implicit regularization framework, which provided insights on the working mechanism of SPL. As the dynamic process, some emotion states are very subtle. For example, it is difficult for human to accurately judge the emotions from some facial expressions. In this work, we leverage the self-paced learning strategy to alleviate the negative effect from low confidence samples, which can improve the robustness of our model.

3 Proposed Method

In this section, we introduce the proposed dynamic confidence-aware multi-channel fusion network for multimodal emotion recognition, and the architecture of the proposed network is shown in Fig. 2. First, we use a self-attention based multi-channel LSTM network to align EEG and facial expression features and reveal temporal uncertainty. The weights of the multi-channel LSTM are shared with each other, while those of the self-attention modules are not. Second, we develop a confidence regression network to measure the true class probability (TCP) on each modality, which estimates the uncertainty at the modality level. Then, different modalities are weighted fused according to the uncertainties estimated above rather than being treated equally. Finally, self-paced learning is used to enhance the model's robustness by mitigating the impact of hard learning samples at sample level. We provide a detailed introduction of each module of the proposed network as follows.

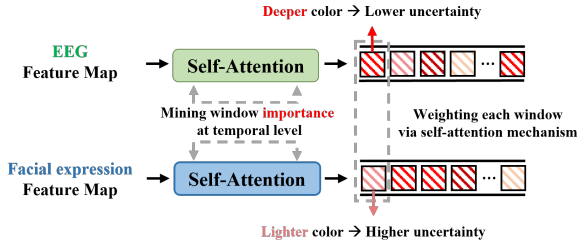

Fig. 3. Illustration of the self-attention mechanism based multi-modal emotion uncertainty learning at temporal level. Self-attention mechanism enables the window importance weighting to model the temporal level uncertainty.

3.1 Temporal Uncertainty based Multi-Modal Feature Alignment

Suppose the feature matrices of EEG and facial expression are $x^E \in R^{T\times D_E}$, $x^F \in R^{T\times D_F}$, respectively, where T denotes the number of time window, and D_E, D_F denote the feature dimension of these two modalities. To exploit the potential temporal dependencies presented by continuously stimuli between EEG and facial expression, we employ a multi-channel long short-term memory (MC-LSTM) network as proposed in [21], [37] to jointly extract features from EEG and facial expression. By sharing weights, the multi-channel LSTM can learn the correlation across the modalities. Formally:

$$\left(x'^m_{i,t}, c^m_{i,t}, h^m_{i,t}\right) = \text{MC-LSTM}\left(x^m_{i,t}, c^m_{i,t-1}, h^m_{i,t-1}\right) \quad (2)$$

where $c^m_{i,t-1}$, $h^m_{i,t-1}$ and $x^m_{i,t}$ represent the memory cell, hidden state and the input vector at $(t-1)$-th or t-th time window of the i-th sample for modality $m \in \{E, F\}$, respectively. $x'^m_{i,t}$ denotes the output of MC-LSTM model and $c^m_{i,t}$, $h^m_{i,t}$ are the memory cell and hidden state at t-th time window, respectively. It is worth nothing that the weights of hidden state are shared across both modalities and time windows, while the weights of input vector are only shared across time windows. The multi-channel LSTM is able to capture potential correlations in multi-modal sequential data and enables heterogeneous alignment.

Due to the variances in emotional patterns, it can be challenging to guarantee a consistent emotional prediction among different modalities even during prolonged and continuous stimulation, leading to potential uncertainty in certain time windows of the physiological or behavioral sequence. To overcome this limitation and enhance the specificity of temporal information, we adopt self-attention mechanism in above model to identify more reliable time windows within the sequential data, shown in Fig. 3. By applying the self-attention mechanism, the proposed model gains the ability to dynamically adjust its focus on different time windows. This adaptability is important for handling temporal-level uncertainty and understanding the temporal dynamics in the sequence. We perform separate linear mapping transformations on x'^E and x'^F produced by multi-channel LSTM:

$$Q_E = x'^E W^E_Q, K_E = x'^E W^E_K, V_E = x'^E W^E_V \quad (3)$$

$$Q_F = x'^F W^F_Q, K_F = x'^F W^F_K, V_F = x'^F W^F_V \quad (4)$$

where Query Q_m, Key K_m and Value V_m are three groups of feature representation through linear projection for modality $m = \{E, F\}$. W^m_Q, W^m_K, W^m_V are parameter matrices used to generate queries, keys and values, respectively, which are updated by network backpropagation during the training of model. We obtain the attention coefficients between time slices by computing the scalar dot product correlation between Q_m and K_m, which is input to the SoftMax function. Then, the features with self-attention are obtained by the product of self-attention weight and V_m:

$$f_m = \text{softmax}\left(\frac{Q_m K^T_m}{\sqrt{d}}\right) V_m \quad (5)$$

where $f_m \in R^{T\times D}$ is the feature map of the EEG and facial expression modalities, respectively, and d is a normalization parameter equal to the dimension of K_m.

3.2 Modality Uncertainty based Confidence-Aware Fusion

3.2.1 Uncertainty estimation

Although multi-modal fusion usually enhances the performance of emotion recognition due to the complementary information between modalities, the inconsistency of the decision-making from different modalities makes an impact on multi-modal fusion. To effectively integrate information from multiple modalities, it is crucial to estimate the prediction confidence of each modality. The higher the confidence of the prediction of the emotion modality is, the lower the uncertainty of the corresponding model's prediction is. Therefore, an appropriate confidence criterion should assign incorrect predictions with low confidence value and achieve prediction with high confidence value. Maximum class probability (MCP) can be used to represent the confidence of model's prediction by:

$$\text{MCP}(x) = \max_{k\in y} P(Y = k \mid w, x) = P(Y = \hat{y} \mid w, x) \quad (6)$$

where w denotes the network parameter set and $\hat{y}$ is the predicted class corresponding to sample x. Although MCP is effective for classification tasks, it may lead to overconfidence, especially for incorrect predictions. The use of maximum SoftMax probability leads to the high confidence score of both correct and incorrect predictions, which indicates that the confidence criterion is unreliable. To address this issue, in the proposed method, we introduce true class probability (TCP) as a confidence criterion for emotion recognition.

Different from MCP, which uses the maximum SoftMax output as the prediction confidence, TCP uses the output probability of the SoftMax corresponding to the true label as the prediction confidence. Formally, the TCP for each modality can be expressed as:

$$\text{TCP}^m\left(x^m, y\right) = P\left(Y = y \mid w, x^m\right) \quad (7)$$

where x^m denotes the high-dimensional feature vector of specific modality m of sample x, and y denotes its category.

3.2.2 Confidence-aware multi-modal fusion

We extract EEG and facial expression modality features using self-attention mechanism, which can be represented as $f_E \in R^{T\times D}$, $f_F \in R^{T\times D}$, respectively. During the

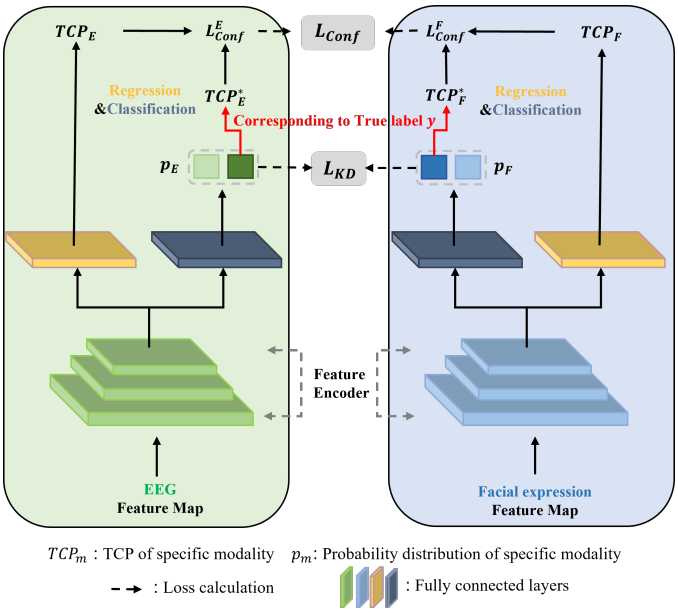

Fig. 4. The proposed multi-modal confidence regression network module, which contains two pairs of regression sub-network and classification sub-network. During the network training process, for a specific modality, the classification sub-network is used to obtain the the ground truth of TCP, and then the regression sub-network is trained to predict the TCP value by the loss function $\mathcal{L}_{conf}$.

training process, we utilize a confidence regression network to dynamically estimate the prediction confidence of each modality, which is employed as guidance for the weighted fusion of different emotion modalities. It can enhance the perdition from the modality with lower uncertainty by assigning higher weight and suppress those with high uncertainty with lower weight by:

$$f_{joint} = c_E f_E \oplus c_F f_F \tag{8}$$

where c_E, c_F represent the confidence of EEG modality and facial expression modality generated by confidence regression network, respectively, and $\oplus$ denotes the feature concatenation. By modelling uncertainty from modality level, our model can make more reliable multi-modal fusion prediction.

3.3 Training Objective

For multi-modal tasks, a binary cross entropy (BCE) is often used to supervise the learning of each branch, but it will heavily penalize the samples that cannot be classified from a specific modality. In this case, the model may overfit the biases in the data, which leads to overfitting of model. Therefore, based on the joint feature representation obtained by confidence-aware fusion, we use cross entropy (CE) loss to supervise multi-modal learning:

$$\mathcal{L}_{CE} = -Y \log P\left(f_{joint}\right) - (1-Y)\log\left(1 - P\left(f_{joint}\right)\right) \tag{9}$$

where $P(f_{joint})$ is the probability of dividing the joint feature vector into a certain class, and Y is the label of sample.

When the classification result is correct, both TCP and MCP are the maximum SoftMax output. However, when the classification is incorrect, TCP is more likely to be close to a lower value, whereas MCP may assign a high confidence score for misclassified samples. Although TCP is able to obtain more reliable confidence, it cannot be used directly in the testing phase due to the absence of label information. For m-th modality, we design a confidence regression network g_m, which consists of two MLP feature encoders and several fully connected layers, shown in Fig. 4, to approximate TCP. The two MLP encoders are used to predict the value of TCP and the prediction probability distribution, respectively. We apply MSE loss to train the confidence regression network:

$$\mathcal{L}_{Conf} = \frac{1}{M}\sum_{m=1}^{M}\left(c_m - c_m^*\right)^2 \tag{10}$$

where $c_m = g_m(x^m)$ denotes the confidence predicted by the confidence regression network, M is the number of all modalities, and c_m^* represents the true value of the predicting confidence. Then TCP is able to use the confidence regression network for approximation. Meanwhile, to preserve the common features between multi-modal data, we apply Kullback-Leibler divergence (KL-divergence) to constraint the similarity of different modalities. We can obtain the prediction distribution $p_m\left(y \mid x^m\right) = \left[p_m^1, \cdots, p_m^K\right]$ of each modality based on the SoftMax output by the proposed confidence regression network. Therefore, KL-divergence loss (Kd-loss) calculated by the KL-divergence between modalities can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{Kd} =& \frac{1}{M}\left(\mathcal{D}_{KL}\left(p_{m1} \| p_{m2}\right) + \mathcal{D}_{KL}\left(p_{m2} \| p_{m1}\right)\right), \\ & \mathcal{D}_{KL}\left(p_{m1} \| p_{m2}\right) = p_{m1}\log\sum\frac{p_{m1}}{p_{m2}} \end{aligned} \tag{11}$$

where $\mathcal{D}_{KL}(\cdot\|\cdot)$ represents the KL-divergence and M denotes the number of modalities.

Using the above proposed loss function as an auxiliary supervision, we obtain the total loss function to be minimized:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{Conf} + \beta\mathcal{L}_{Kd} \tag{12}$$

where α and β are hyperparameters, which adjust the weights of $\mathcal{L}_{Conf}$ and $\mathcal{L}_{Kd}$, respectively.

3.4 Optimization based on Sample Uncertainty

Samples with higher training learning loss have greater potential to negatively impact the robustness of the model if they are introduced earlier in the training process. In order to learn robust representations of multi-modal inputs, we adopt self-paced learning to gradually train the above model based on the samples with low to high uncertainty. Specifically, we introduce sample weight variables $v = [v_1, \ldots, v_n]^T$ based on the training loss as a measure of sample uncertainty by minimizing the following function:

$$\min_{v\in[0,1]^n} E(w, v, \lambda) = \sum_{i=1}^{n}\left(v_i\mathcal{L} + f\left(v_i, \lambda\right)\right) \tag{13}$$

where n denotes the training sample size, and y_i is the corresponding label in the training dataset, and $f(v_i,)$ is the self-paced regularization term. Here, we use P^* to denote

the training dataset selected in each sub-process. Then, we can rewrite the above equation as follows:

$$\begin{aligned} \min_{v\in[0,1]^n} \quad & E\left(w, v, \lambda, P^*\right) \\ & = \sum_{i=1}^{n}\left(v_i\mathcal{L} + f\left(v_i, \lambda\right)\right) \\ & \text{s.t. } \left\{\left(x_1, y_1\right), \ldots, \left(x_k, y_k\right)\right\} \in P^* \end{aligned} \tag{14}$$

The self-paced regularization term $f(v_i,)$ in the above equation favors selecting low uncertainty samples from the training set, and k represents the number of samples in the current iteration. We use a hard self-paced regularization [38] in our method, denoted as follows:

$$f\left(v_i, \lambda\right) = -\lambda v_i \tag{15}$$

Then, we can obtain the corresponding approximate solution of the above objective:

$$v^*(\lambda, \mathcal{L}) = \begin{cases} 1, & \text{if } \mathcal{L} < \lambda \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

In the initial stage, we randomly select a small subset of the training data to train the model. As the training proceeds, more samples are added to the training process according to their training loss until all samples are included, which allows the model to learn a more robust multi-modal feature representation. Algorithm 1 summarizes the algorithm description of our proposed method.

Compared to the existing multi-modal methods in other fields, our proposed method not only captures common features from heterogeneous modalities but also explicitly models the uncertainty in the dynamic emotion process. Meanwhile, the proposed method takes into account the decision divergence among different modalities, reducing the negative impact of modality-level uncertainty on multi-modal fusion. In the model optimization, by alleviating the negative impact of hard learning emotion samples with high uncertainty, our method ensures that the model is trained on a manageable set of samples. This strategic adaptation reduces the risk of overfitting raised by challenging instances, leading to a more reliable and generalizable emotion recognition model. In summary, the above innovations of the proposed method collectively contribute to the advancement of multi-modal emotion recognition, offering a more comprehensive and effective solution compared to existing multi-modal methods in other fields.

Algorithm 1 Training of the proposed dynamic confidence-aware fusion network.

Input:
EEG modality samples x_E, facial expression modality Samples x_F;
Ground truth labels Y, number of epochs T;
hyperparameters λ, α, β, batch size B;
sample weight $v = [v_1, \ldots, v_n]^T, v_i \in [0, 1]$.

Output:
Model parameters w;

while $epoch < T$ **do**
  for i in x_B **do**
    $x_i'^E, x_i'^F = \text{MC-LSTM}\left(x_i^E, x_i^F\right)$
    $Q_i^E = x_i'^E W_Q^E, Q_i^F = x_i'^F W_Q^F$
    $K_i^E = x_i'^E W_K^E, K_i^F = x_i'^F W_K^F$
    $V_i^E = x_i'^E W_V^E, V_i^F = x_i'^F W_V^F$
    $f_{E_i} = \text{softmax}\left(Q_i^E {K_i^E}^T / \sqrt{d_1}\right) V_i^E$
    $f_{F_i} = \text{softmax}\left(Q_i^F {K_i^F}^T / \sqrt{d_2}\right) V_i^F$,
    where d_i is the normalization parameter.
    $c_{Ei}, p_{Ei} = \text{ConfNet}\left(f_{E_i}\right)$,
    $c_{Fi}, p_{Fi} = \text{ConfNet}\left(f_{F_i}\right)$
    $f_i = c_{Ei} f_{E_i} \oplus c_{Fi} f_{F_i}$,
    where $\oplus$ denotes the concatenate opreation.
  end for
  $c^* \leftarrow$ calculated by Eq. (7)
  $\mathcal{L}_{CE}\left(f, Y\right) \leftarrow$ calculated by Eq. (9)
  $\mathcal{L}_{conf}\left(c, c^*\right) \leftarrow$ calculated by Eq. (10)
  $\mathcal{L}_{KD}\left(p_E, p_F\right) \leftarrow$ calculated by Eq. (11)
  $\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{conf} + \beta\mathcal{L}_{KD} \leftarrow$ calculated by Eq. (12)
  Update v by Eq. (16)
  Update λ to a lager value
end while
return model parameters w;

TABLE 1
Summary of the datasets used in the experiments. Features/sample indicates the extracted feature dimensions of a sample.

	DEAP		MAHNOB-HCI		ASCERTAIN	
	EEG	facial	EEG	facial	EEG	facial
Features/sample	160	29	160	29	40	22
Trials	40		20		36	
Duration of trial	60s		35-117s		51-127s	
EEG channels	32		32		8	
Sampling rate	512		256		32	
Samples	720		480		900	

4 Experiments

4.1 Multi-Modal Emotion Databases

To evaluate our proposed method, we conduct experiments on three multi-modal emotion datasets, including Database for Emotion Analysis using Physiological signals (DEAP) [3], multi-modal database for affect recognition and implicit tagging (MAHNOB-HCI) [39], and multi-modal database for implicit personality and affect recognition using commercial physiological sensors (ASCERTAIN) [40]. Table 1 summarizes the details of the datasets used in the following experiments.

4.2 Data Pre-processing

For DEAP dataset and MAHNOB-HCI dataset, all 18 and 24 subjects with both EEG signals and full facial videos are adopted in the experiment, respectively. The same data pre-processing strategy on these two datasets, among which the last 30s of each trial of MAHNOB-HCI dataset is extracted for further feature extraction [40], [41], [42]. The 32-channel EEG signal is downsampled to 128Hz. Power spectral density (PSD) features are extracted from 3 seconds non-overlapping windows through the Welch method in EEG, and 5 frequency bands are adopted, i.e., theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30+ Hz) [12]. Referring to [43], we utilize OPEN-

FACE to extract 29 expression features from facial videos, including 3 face positions relative to the camera, 3 head positions (pitch, yaw, and roll), 6 eye gaze directions and 17 facial action units. Similar to EEG, the facial sequences are divided according to the 3-second non-overlapping sliding time window and the average value of each feature is taken. For ASCERTAIN dataset, all 25 subjects with both EEG signals and full facial videos are used in experiment. EEG and facial features corresponding to each trial over the final 50 seconds from raw data is extracted. Each EEG electrode signal is filtered with 1-16 bandpass filter. PSD features are extracted from 5 seconds non-overlapping time windows using the Welch method in EEG, and 5 frequency bands are adopted, i.e., delta (1-4 Hz), theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), and beta (12-16 Hz). The 22-dimension facial features were averaged through 5 seconds non-overlapping time windows.

4.3 Experimental Settings and Comparison Methods

1) Subject-independent Classification. In the experiments, we use the leave-one-subject-out (LOSO) cross-validation strategy to validate our method. Specifically, the samples are divided into 18, 24 and 25 non-overlapping parts based on subjects for DEAP, MAHNOB-HCI and ASCERTAIN, respectively. We select one subject's sample as the testing set and the remaining subjects' samples are treated as the training set for each cross-validation. This process is repeated 18, 24 and 25 times, and finally the average performance of the cross-validations are reported as the final result.

2) Subject-dependent Classification. We also use subject-dependent classification on each subject to validate our method. Similarly, the samples of subjects are divided into 18, 24 and 25 non-overlapping parts for DEAP, MAHNOB-HCI and ASCERTAIN. Following several previous studies [12], [55], a data augmentation is adopted by splitting each trial into smaller non-overlapping segments for the training in the deep neural network. Then, for each separate part, we use 10-fold cross-validation. Typically, we divide the dataset into 10 subsets with the same size, one subset is used as the testing set, and the other 9 subsets are used as the training set. This process is repeated 10 times. The average performance of the 10-fold cross-validation was taken as the result on each subject, and finally the average performance of 18, 24 and 25 subjects are taken as the final result.

3) Implementation Details. We implement the proposed method using Pytorch on a GeForce GTX 3080 GPU device with 12GB memory. The Adam optimization algorithm is used to optimize the objective function. The model is trained for a total of 200 epochs in both subject-independent and subject-dependent settings. In addition, the grid search is adopted to select the hyperparameters, the parameter of the loss function, α and β are both searched from the set of $\{0.1, 0.2, 0.3, \ldots, 1\}$. λ and γ are both searched from the set of $\{1, 1.05, 1.1, \ldots, 1.5\}$. In the proposed model, ReLU is used as the activation function. The hidden state size of MC-LSTM module is set to 256. The size of linear transformation layer in self-attention module is 128. The size of hidden layer for the feed forward network is 512, while the output size is set to 256.

Emotion recognition performance is measured by Accuracy (ACC) and F1-Score. To further validate the effectiveness of the method, we compare our method with several popular methods from two aspects, i.e. multi-modal based methods and uncertainty learning based method, including multi-kernel Learning (MKL) [44], canonical correlation analysis (CCA) [46], deep canonical correlation analysis (DCCA) [47], kernel canonical correlation analysis (KCCA) [48], long short-term memory network (LSTM) [45], gated multi-modal unit (GMU) [49], multi-modal residual LSTM (MMResLSTM) [21], emotion transformer fusion (ETF) [19], multi-modal adaptive emotion transformer (MAET) [50], VigilanceNet [51], Bayesian Neural Network (BNN) [28], Monte-Carlo Dropout (MC-Dropout) [29], deep ensemble (DE) [30], uncertainty-aware attention (UA) [52], evidential deep learning (EDL) [53], and trusted multi-view classification (TMVC) [54]. Then, we briefly introduce the above methods.

- **MKL method.** MKL projects each modality onto a high-dimensional space by different kernel functions and performs linear fusion, in which SVM is used for the classification of the extracted features.
- **CCA method.** CCA is a multi-modal fusion method that aims to find the maximum correlation between multiple modalities by projecting them in to a common low-dimensional subspace.
- **DCCA method.** DCCA is able to solve the linearized feature vectors of different modalities through a deep network by calculating the maximum correlation between the vectors, and the extracted deep features are fed into SVM for classification.
- **KCCA method.** KCCA maps the features of each modality to a high dimensional space by kernel function, and maximizes the correlation between modalities in fusion.
- **LSTM method.** LSTM is a temporal recurrent neural network (RNN) designed to solve the problem of long-term dependence in sequences. In our experiment, we use two parallel LSTMs to extract features of EEG and facial expression modalities, respectively, and concatenate the features for classification.
- **GMU method.** GMU is intended to be used as an internal unit in a neural network architecture to find an intermediate representation based on a combination of data from different modalities.
- **MMResLSTM method.** MMResLSTM learns the correlation between EEG and other physiological signals by sharing the weights of each modal LSTM layer to accomplish multi-modal emotion recognition.
- **ETF method.** ETF inputs data from different modalities into the corresponding Transformer encoder, and then fuses them into the joint representation space through a fusion layer based on an attention mechanism.
- **MAET method.** MAET is a flexible model that can process both uni-modal and multi-modal inputs with specialized modules, and it adopts adversarial training to alleviate subject discrepancy.
- **VigilanceNet method.** VigilanceNet introduces an intra-modality representation learning method to

TABLE 2
Classification Performance (Mean/Std) of our method and the comparison methods on DEAP dataset (%).

Method	Independent (Valence)		Independent (Arousal)		Dependent (Valence)		Dependent (Arousal)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MKL [44]	59.02/12.94	67.37/16.97	58.96/13.89	64.78/18.69	72.15/07.22	72.96/07.83	73.95/05.29	74.38/11.07
LSTM [45]	59.72/12.38	63.22/14.26	58.74/12.76	64.32/18.39	85.60/05.48	85.68/06.06	87.50/04.35	89.07/05.21
CCA [46]	60.21/10.58	64.34/13.39	59.52/11.30	65.48/19.26	83.94/06.34	85.20/05.47	84.77/05.76	83.24/05.59
DCCA [47]	61.04/11.73	65.46/12.29	59.58/11.89	65.78/16.98	84.85/07.25	83.96/05.11	85.52/07.35	85.41/05.98
KCCA [48]	59.52/09.69	66.34/12.56	58.85/10.56	65.45/17.90	84.21/04.72	85.25/05.70	84.44/05.28	85.72/06.11
GMU [49]	59.38/08.98	62.27/11.75	59.88/14.26	64.92/17.38	86.13/05.78	86.10/06.49	85.92/04.55	86.09/07.31
MMResLSTM [21]	62.67/10.57	68.36/11.50	62.25/12.38	67.32/15.92	86.39/06.23	85.29/05.15	86.28/04.90	87.21/06.10
ETF [19]	62.63/09.35	66.35/13.41	61.64/10.29	66.63/16.97	86.76/05.10	88.34/05.34	85.34/06.22	86.03/06.72
MAET [50]	65.21/09.58	65.17/11.85	67.20/09.45	71.24/13.80	91.72/07.25	92.88/05.19	91.10/05.32	92.05/05.70
VigilanceNet [51]	63.98/11.20	66.72/10.36	62.13/12.15	69.52/11.09	89.36/04.55	90.42/04.53	91.37/05.13	92.84/06.03
BNN [28]	59.58/10.82	63.40/12.59	58.28/10.55	67.85/16.82	86.20/05.43	86.31/05.68	85.76/04.72	87.56/06.19
MC-Dropout [29]	60.36/11.28	65.27/11.99	59.36/10.96	68.82/18.83	87.19/05.09	88.91/04.57	87.71/05.10	88.28/07.26
DE [30]	61.39/10.87	64.78/12.86	60.87/11.39	67.52/19.25	87.21/05.11	87.83/05.23	87.88/04.33	88.17/06.22
UA [52]	60.78/09.13	65.52/13.49	59.63/11.23	66.62/17.53	88.34/04.77	89.15/05.30	86.75/05.55	87.15/05.97
EDL [53]	58.29/09.96	63.97/12.73	60.48/10.88	65.25/17.95	86.73/05.22	85.96/05.11	87.35/05.12	87.78/08.25
TMVC [54]	59.17/09.88	68.21/13.75	58.19/10.29	67.71/18.32	86.57/06.70	85.45/07.42	89.37/04.25	90.69/05.19
Ours	**72.22/06.68**	**71.81/08.08**	**70.69/08.94**	**76.83/10.31**	**95.25/03.84**	**94.50/03.97**	**94.89/03.07**	**95.55/03.30**

TABLE 3
Classification Performance (Mean/Std) of our method and the comparison methods on MAHNOB-HCI dataset (%).

Method	Independent (Valence)		Independent (Arousal)		Dependent (Valence)		Dependent (Arousal)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MKL [44]	67.49/10.63	61.54/18.75	62.25/15.70	56.27/22.82	85.29/06.14	80.19/08.48	86.58/06.85	77.54/12.00
LSTM [45]	70.58/12.74	56.22/18.94	68.33/20.62	57.65/26.73	92.10/03.35	90.48/04.32	88.10/11.90	86.90/12.37
CCA [46]	63.95/08.96	59.51/21.25	59.31/17.89	55.21/20.17	78.05/07.92	77.41/09.06	82.47/06.75	81.60/09.35
DCCA [47]	62.70/13.43	56.49/20.63	67.49/13.43	59.97/16.60	85.82/07.10	82.52/09.32	87.95/09.23	85.83/11.10
KCCA [48]	62.21/11.91	59.74/18.39	62.54/14.25	57.66/18.75	80.05/09.28	77.00/10.49	81.83/08.22	81.71/12.59
GMU [49]	69.53/10.79	66.38/18.26	71.44/11.95	64.75/17.38	87.32/03.20	87.63/07.44	90.24/08.11	88.71/11.19
MMResLSTM [21]	71.74/12.10	63.92/17.05	68.60/13.54	66.19/19.41	92.88/03.54	91.60/04.11	92.15/06.62	88.20/13.72
ETF [19]	72.49/11.78	67.60/18.83	70.85/13.07	70.11/16.67	93.10/03.37	92.81/04.48	93.29/08.02	93.76/12.38
MAET [50]	71.09/12.45	62.17/18.43	71.96/11.32	68.04/16.62	93.71/02.59	91.42/05.95	92.18/07.40	90.88/12.59
VigilanceNet [51]	73.25/10.77	66.52/17.06	70.22/12.49	66.40/18.35	94.80/01.62	94.75/03.40	94.47/08.35	92.33/09.42
BNN [28]	66.04/14.66	62.28/15.02	69.79/13.14	64.88/20.20	93.52/02.91	92.87/03.16	91.79/06.16	92.12/05.56
MC-Dropout [29]	78.33/09.04	63.96/20.71	73.12/11.59	60.19/17.95	86.92/05.78	86.15/07.28	85.74/08.01	88.37/06.30
DE [30]	65.52/15.33	66.61/22.54	66.41/13.66	61.87/19.43	83.30/04.49	80.04/09.97	82.61/11.45	86.44/12.83
UA [52]	67.12/13.28	63.25/21.69	66.09/13.15	61.26/17.07	84.70/05.61	82.39/09.62	84.02/10.59	87.57/09.67
EDL [53]	65.48/12.64	60.49/18.05	67.79/14.70	63.24/17.52	84.06/04.77	81.45/08.36	86.58/09.24	89.52/09.44
TMVC [54]	59.16/14.11	57.44/12.51	61.66/20.19	57.14/22.67	90.89/06.92	87.11/11.34	88.08/07.46	88.64/10.07
Ours	**79.37/09.00**	**74.98/16.30**	**75.62/10.14**	**71.53/16.55**	**97.24/01.79**	**95.80/02.90**	**95.22/07.72**	**93.23/10.32**

TABLE 4
Classification Performance (Mean/Std) of our method and the comparison methods on ASCERTAIN dataset (%).

Method	Independent (Valence)		Independent (Arousal)		Dependent (Valence)		Dependent (Arousal)	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MKL [44]	52.77/08.37	45.59/13.22	55.55/09.00	64.67/11.90	65.46/05.38	55.42/11.97	70.16/08.04	73.54/11.85
LSTM [45]	65.44/07.71	53.61/17.58	67.33/10.15	77.64/08.91	87.00/03.50	84.09/05.77	88.73/04.83	90.33/04.93
CCA [46]	55.55/08.01	52.17/16.25	62.66/12.60	76.34/09.49	77.60/06.10	74.81/09.11	79.11/08.51	78.42/06.55
DCCA [47]	53.11/06.72	45.52/15.34	57.55/11.66	70.34/09.97	83.56/07.35	80.18/06.70	84.60/08.23	86.87/10.25
KCCA [48]	57.20/09.44	51.59/17.48	61.41/12.70	69.31/12.43	75.04/09.15	75.22/11.60	74.63/10.32	76.85/10.12
GMU [49]	63.54/08.17	54.06/15.72	68.12/09.75	74.39/12.40	84.10/08.95	84.73/10.12	82.99/07.64	83.10/09.14
MMResLSTM [21]	64.39/06.70	57.43/15.88	69.24/09.54	74.75/13.10	87.51/04.72	85.06/09.13	88.84/04.75	89.17/05.66
ETF [19]	66.21/05.10	57.05/16.67	67.74/09.39	77.61/10.05	87.92/07.73	85.90/10.75	89.05/05.57	90.19/05.48
MAET [50]	64.68/05.59	54.75/13.81	68.50/09.78	75.27/10.36	85.38/09.45	78.10/10.01	88.76/05.88	90.66/06.69
VigilanceNet [51]	66.57/05.23	56.94/13.44	67.28/10.35	76.92/09.33	86.92/08.14	82.01/09.95	85.21/07.53	88.72/06.40
BNN [28]	65.99/05.08	55.44/12.60	69.88/07.67	73.91/14.69	80.38/03.40	74.33/07.88	82.32/05.26	84.54/07.33
MC-Dropout [29]	66.88/05.19	57.00/09.34	70.33/09.61	77.71/08.60	78.25/08.19	74.90/10.37	81.22/07.71	82.04/09.73
DE [30]	60.47/06.28	55.36/15.11	66.73/10.96	69.88/13.94	79.47/08.26	76.35/09.44	80.77/06.46	82.61/09.17
UA [52]	61.70/07.81	59.24/14.65	68.33/09.27	71.66/10.51	81.79/09.81	76.98/11.25	81.94/05.38	84.06/09.33
EDL [53]	56.83/11.25	50.51/17.52	63.95/10.84/	75.20/12.07	83.63/07.96	80.71/09.84	85.42/05.69	86.41/07.85
TMVC [54]	60.00/07.34	48.22/14.56	64.22/08.83	69.58/16.44	81.96/04.23	77.42/07.29	83.58/03.65	85.98/05.89
Ours	**67.77/04.81**	**61.04/12.91**	**71.44/09.05**	**77.82/08.53**	**89.06/07.39**	**87.11/08.24**	**90.25/04.20**	**91.84/04.37**

gain an effective representation of each modality and uses a cross-attention based fusion strategy to capture the complementary information among multi-modal data.

- **BNN method.** BNN is to incorporate uncertainty into the model by treating the weights and biases as random variables, which makes predictions are not only based on the input data but also has relationship with the uncertainty of the model.
- **MC-Dropout method.** MC-Dropout estimates Bayesian uncertainty by using dropout before each weight layer during test process and approximating the predictions for several iterations.
- **DE method.** DE is a simple non-Bayesian model that fuses different model predictions by training several independently randomly initialized network models.
- **UA method.** UA generates attention weights that follow a Gaussian distribution with learned means and variances, capable of capturing heteroscedasticity uncertainty and producing more accurate uncertainty calibration.
- **EDL method.** EDL designs predictive distributions for classification by using a Dirichlet distribution on the class probabilities.
- **TMVC method.** TMVC is an uncertainty-based multi-view network that integrates evidence from different viewpoints, modeling the input as a Dirichlet distribution in quantifying uncertainty.

TABLE 5
Classification Performance (Mean/Std) of our method with multiple modalities and single modality, respectively, on DEAP dataset (%)

Setting	Modality	Valence		Arousal	
		ACC	F1	ACC	F1
Subject-independent	EEG	63.61/09.35	61.23/13.87	62.25/11.47	69.28/16.72
	facial expression	62.69/09.79	66.82/13.23	61.19/12.98	69.14/18.29
	EEG+facial expression	**72.22/06.88**	**71.81/08.08**	**70.69/08.94**	**76.83/10.31**
Subject-dependent	EEG	91.87/04.33	91.74/05.90	92.82/06.18	90.05/05.00
	facial expression	91.43/06.20	92.08/07.16	89.31/04.77	90.61/07.85
	EEG+facial expression	**95.25/03.84**	**94.50/03.97**	**94.89/03.07**	**95.55/03.30**

4.4 Performance on Emotion Recognition

The classification performance of our proposed method and the comparison methods in both subject-independent and subject-dependent experiment settings on DEAP, MAHNOB-HCI, ASCERTAIN are illustrated in Table 2, Table 3 and Table 4, respectively. As can be seen from the tables, our proposed method achieves the best performance among all the methods in both subject-independent and subject-dependent classification settings. The average accuracies of our method in subject-independent classification for all datasets reach 72.22%, 79.37%, 67.77% and 70.69%, 75.62%, 71.44% respectively on valence and arousal, which are obviously better than the comparison methods. Besides, in subject-dependent classification, our proposed method also outperforms other comparison methods with the accuracy of 95.25%, 97.24%, 89.06% and 94.89%, 95.22%, 91.84% respectively on valence and arousal on the three datasets. In addition, not only does our method achieve the highest average accuracy, but it also achieves the highest average F1-score compared to other algorithms. This demonstrates the effectiveness of our approach in multi-modal emotion recognition.

TABLE 6
Ablation study for Subject-Independent Classification Performance (Mean/Std) of our method on DEAP dataset (%)

Method	TUM2FA	MUCF	SUTO	Valence		Arousal	
				ACC	F1	ACC	F1
Baseline				61.11/11.35	64.81/13.40	60.91/12.45	65.10/18.99
Ours1	✓			67.05/10.26	65.09/13.10	66.63/11.21	72.16/16.89
Ours2		✓		65.83/08.96	64.80/12.64	66.05/10.70	72.05/16.42
Ours3			✓	66.94/09.35	62,85/11.78	66.80/10.27	72.89/17.31
Ours4		✓	✓	66.28/09.21	66.37/12.42	66.09/10.11	69.32/16.33
Ours5	✓		✓	68.08/10.78	65.82/12.27	67.18/10.93	71.24/17.71
Ours6	✓	✓		68.19/11.47	69.54/14.01	67.36/11.64	72.77/17.42
Ours	✓	✓	✓	**72.22/06.68**	**71.81/08.08**	**70.69/08.94**	**76.83/10.31**

The main reason for the superiority of our method is that, compared with other methods, our proposed method not only captures the common features from heterogeneous emotional modalities, but also considers the uncertainty in the dynamic emotion process and the fusion recognition model. Furthermore, SPL strategy adopted in our method improves the robustness of the optimization of the prediction model by alleviating the negative impact of hard learning emotion samples with high uncertainty. In comparison, methods such as CCA and DCCA only consider the pairwise similarity between samples and cannot extract the features at the temporal level of sequential data. MMResLSTM, ETF, MAET and VigilanceNet capture the temporal feature and correlation between multi-modal data to make use of the complementary information between multiple modalities, but ignore the impact of uncertainty. Uncertainty-based methods such as TMVC and EDL are unable to extract temporal level features of sequential data and fail to effectively fuse different modalities. In contrast, our method can comprehensively model uncertainty in multiple levels to produce confidence fusion recognition.

We also conduct an experiment to validate the effectiveness of extracting complementary information from multi-modal data using our proposed method. Firstly, we perform emotion recognition using EEG and facial expressions separately, and then fuse the two modalities to verify the effectiveness of multi-modal fusion. The results are shown in Table 5, where our method effectively fuse data from different modalities to improve the performance of emotion recognition in both subject-independent and subject-dependent classification. This demonstrates that our proposed method is able to extract complementary information that helps capture robust emotion-related discriminative features.

5 Discussion

5.1 Ablation Study

The proposed method contains the following three modules, including temporal uncertainty based multi-modal feature alignment (TUM2FA), modality uncertainty-based confidence-aware fusion (MUCF) and sample uncertainty-based training optimization (SUMO). To further verify the effectiveness of each module of our proposed method, we conduct several ablation experiments by removing the main modules in the model. The experimental results are shown in Table 6, where the classification performance greatly improves with the help of the TUM2FA module, which aligns

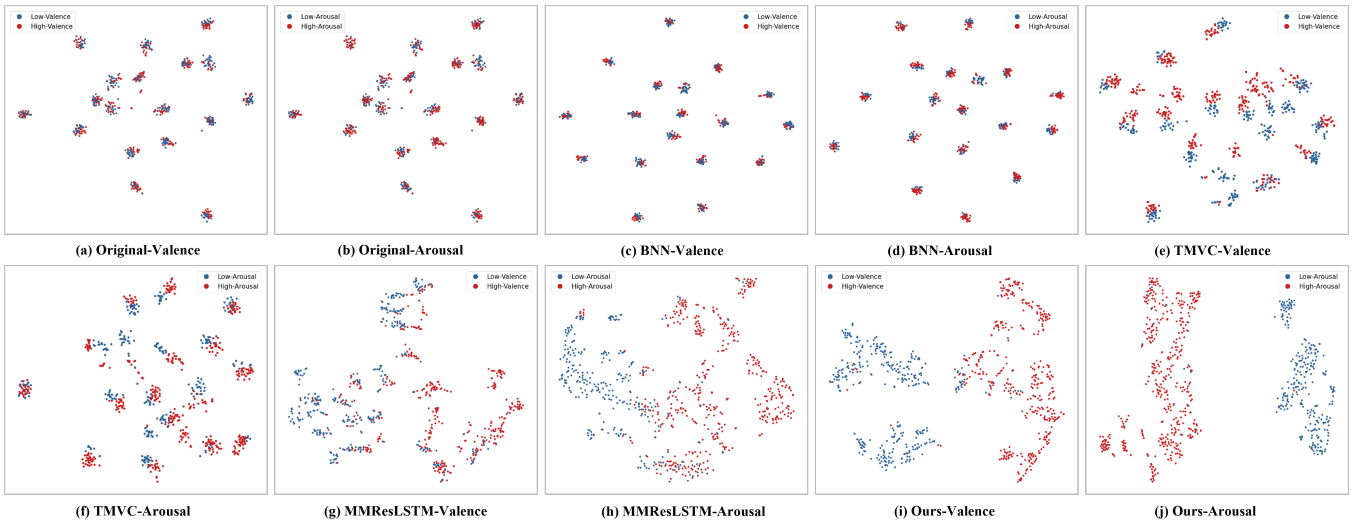

Fig. 5. T-SNE visualization for different methods: (a)-(b) are the scattered arrangements of the original samples in the representation space on DEAP dataset for the valence and arousal tasks (c)-(h) are the representation spaces of several comparison methods on valence and arousal, (i)-(j) are the representation spaces of our method on valence and arousal, respectively.

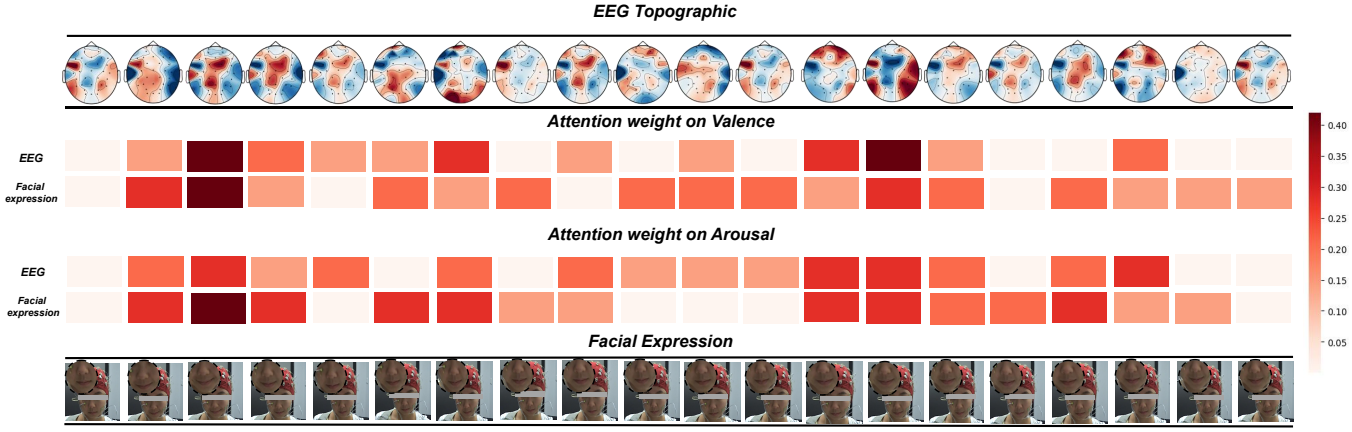

Fig. 6. Visualization the attention weights in facial expressions, EEG topographic at temporal level on DEAP dataset of all time windows.

multi-modal heterogeneous sequential data and mines uncertainty at the temporal level. In addition, confidence-aware multi-modal fusion by MUCF can alleviate the effect of multi-modal decision divergence and improve the performance of emotion recognition. Meanwhile, the introduction of SUMO also achieves better results in both accuracy and F1. Furthermore, the experimental results of using pairwise combinations of modules in Table 6 show higher classification performance than using a single module, which further proves the effectiveness of each module in our proposed method.

5.2 High-order Information Representation Ability Analysis

To verify the feature representation capability of our proposed method, we visualize the learned features of different methods with T-SNE, shown in Fig. 5. We first downscale the representation vectors learned by our method with T-SNE [56] and then project the two-dimensional vectors onto a space for visualization as shown in Fig 5. (k), (l). In addition, we also visualize the results of the original sample data and the comparison methods on the Valence and Arousal classification tasks, as shown in Fig 5 (a), (b) and Fig 5 (c)-(h). Compared with the original sample data, although the uncertainty-based deep learning methods, including TMVC and BNN, aggregate intra-class samples to a certain extent, they cannot learn the common features of different subject samples. While MMResLSTM distances the inter-class samples to a certain extent, it cannot effectively aggregate intra-class samples. In our method, the high-order relationship between EEG and facial expression is extracted utilizing cross-modal alignment of multi-channel LSTM and similarity constraints across different modalities. The final high-order emotion-related discriminative features are obtained by confidence-aware fusion, and the separability of high-order features is improved by using supervised information. Compared with other methods, our proposed method has a higher aggregation degree of samples from the same class as well as a clearer boundary between samples of different classes, which is also supported by the experimental results of classification accuracy.

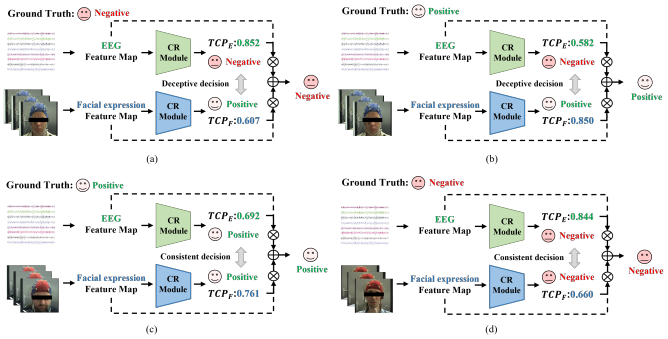

Fig. 7. Several samples in DEAP dataset with inconsistent or consistent prediction on different modalities. Through confidence-aware fusion, our proposed method can well circumvent the false predictions caused by deceptive facial expressions and emotional dispersion, so as to obtain the final correct prediction.

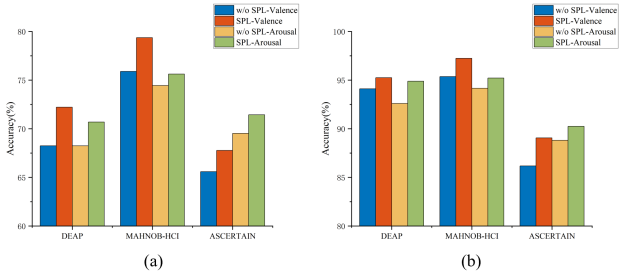

Fig. 8. The classification accuracy comparison between the proposed method and the model after removing SPL module in both subject-independent and subject-dependent classification on all three datasets. (a) Bar graph showing the classification performance in subject-independent classification. (b) Bar graph showing the classification performance in subject-dependent classification.

5.3 Uncertainty Analysis

5.3.1 Temporal level

In our method, we use the self-attention mechanism to learn the importance weights of sequential data on time segments to quantify the uncertainty at the temporal level. We visualize the facial expressions, EEG topographic, and their attention weights of the subject over several time windows, as shown in Fig. 6. The distribution of self-attention weights reflects the degree of the model's attention to different time windows in the sequence. A higher attention weight may correspond to a lower uncertainty since the model is more confident in its predictions based on the corresponding time window. The attention weights for different modalities vary significantly across specific windows, and our model can prioritize windows of modalities with higher attention weights. For example, as can be seen from the ninth window on valence in Fig. 6, the facial expression is not easy to identify, leading to a lower attention weight assignment, while the EEG modality receives higher attention. Furthermore, by considering the EEG topographic map, those windows in which the brain activation varies greatly, tend to have higher attention weights. A larger discrepancy of brain activation indicates a higher reliability and lower uncertainty. The experimental results show that the effectiveness of self-attention mechanism on capturing pivotal multi-modal temporal segments, and can alleviate the problem of emotional alienation of subjects during data collection.

5.3.2 Modality level

Due to the emotional divergence during continuous activation, there may exist highly redundant segments in EEG, which negatively impact emotion recognition. Although facial expressions are more intuitive than EEG signals to show an individual's emotions, it can be deceptive in that they do not fully accurately reflect the individual's current emotional state. To mitigate the negative effects of multi-modal fusion due to uncertainty, we introduce TCP to quantify the confidence of each modality, based on which the EEG signal and facial expressions are fused in a more reliable way, guiding the model to make more confidence prediction. For the experimental results, four cases with inconsistent predictions among different modalities are shown in Fig. 7. Specifically, for the first case shown in Fig. 7 (a), EEG branch gives the negative prediction, which is consistent to the ground truth, while facial branch produces the positive prediction.

The condition mentioned above shows the divergence of classification result from multiple modalities, which leads negative effect on multi-modal fusion. To solve this problem, our model dynamically assigns a higher confidence level to the EEG branch features and a lower confidence level to the facial expression branch, and the joint features obtained from the confidence fusion help make the correct prediction, which alleviates the effect caused by the deceptive facial expressions. In contrast, for the case in Fig. 7 (b), although the EEG branch predicts incorrectly, it is assigned a low confidence score in fusion stage, and the correct result is also obtained after the confidence-aware fusion. In addition, as can be seen from Fig. 7 (c) and Fig. 7 (d), when consistent predictions are made by multiple branches, our proposed confidence-aware fusion scheme successfully produces the correct prediction, which also shows the robustness of the proposed fusion strategy. Therefore, our method achieves more trustworthy multi-modal fusion by effectively alleviating the uncertainty brought by ambiguity predictions from different modalities.

5.3.3 Sample level

To enhance the robustness of our model, we adopt SPL strategy to constraint the training process by selecting samples from easy to hard. Different from randomly selecting samples into training, SPL firstly trains the model with the samples having lower uncertainty. As shown in Fig. 8, compared with the model without SPL, our method achieves a significant improvement in classification accuracy in both subject-independent and subject-dependent classification on all three datasets. Obviously, the experimental results shown in Fig. 8 illustrate that the SPL strategy can enhance the robustness of the proposed model in emotion recognition, which demonstrates the effectiveness of SPL strategy.

6 Conclusion

In this paper, we systematically study the uncertainty in multi-modal emotion fusion at different levels and propose an uncertainty confidence-aware multi-modal emotion

recognition model that effectively alleviates the uncertainty in fusion. First, we develop a multi-channel LSTM network with shared weights for heterogeneous feature alignment and simultaneously utilize a self-attention mechanism to learn the uncertainty of different time windows of the emotion process. Second, we propose a true class probability-based confidence regression network to quantify modality level uncertainty. Then, a confidence weighting fusion strategy is introduced to achieve more interpretable and reliable multi-modal emotion fusion recognition by alleviating the negative prediction contribution of the time windows and modalities with high uncertainty. Finally, a self-paced learning strategy is adopted to optimize the above emotion fusion prediction model, which achieves sample-level uncertainty reduction by training with easy to hard learning samples. The experimental results on three multi-modal emotion datasets show that our proposed method not only achieves promising prediction results for multi-modal emotion compared to state-of-the-art methods but also explicitly reveals the variation of emotion with uncertainty estimation.

REFERENCES

[1] J. Deng and F. Ren, "Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning," *IEEE Transactions on Affective Computing*, 2020.

[2] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453–464, 2020.

[3] 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.

[4] W. Zheng, "Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 3, pp. 281–290, 2016.

[5] B. Cheng and G. Liu, "Emotion recognition from surface emg signal using wavelet transform and neural network," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 2008, pp. 1363–1366.

[6] P. Sarkar and A. Etemad, "Self-supervised learning for ecg-based emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3217–3221.

[7] J. Hu, Y. Huang, X. Hu, and Y. Xu, "The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 17–30, 2022.

[8] R. Belmonte, B. Allaert, P. Tirilly, I. M. Bilasco, C. Djeraba, and N. Sebe, "Impact of facial landmark localization on facial expression recognition," *IEEE Transactions on Affective Computing*, 2021.

[9] R. Zhao, T. Liu, Z. Huang, D. P.-K. Lun, and K. K. Lam, "Geometry-aware facial expression recognition via attentive graph convolutional networks," *IEEE Transactions on Affective Computing*, 2021.

[10] H. Chen, H. Shi, X. Liu, X. Li, and G. Zhao, "Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1346–1366, 2023.

[11] B. Puccio, P. Vizza, and P. Veltri, "On the use of eeg functional connectivity networks in epilepsy studies," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 2945–2948.

[12] Y. Zhang, H. Liu, D. Zhang, X. Chen, T. Qin, and Q. Zheng, "Eeg-based emotion recognition with emotion localization via hierarchical self-attention," *IEEE Transactions on Affective Computing*, 2022.

[13] C. Li, P. Li, Y. Zhang, N. Li, Y. Si, F. Li, Z. Cao, H. Chen, B. Chen, D. Yao *et al.*, "Effective emotion recognition by learning discriminative graph topologies in eeg brain networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[14] Y. Peng, F. Jin, W. Kong, F. Nie, B.-L. Lu, and A. Cichocki, "Ogssl: A semi-supervised classification model coupled with optimal graph learning for eeg emotion recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1288–1297, 2022.

[15] 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.

[16] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[17] Y. Yang, Q. Gao, Y. Song, X. Song, Z. Mao, and J. Liu, "Investigating of deaf emotion cognition pattern by eeg and facial expression combination," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 589–599, 2021.

[18] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[19] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu, "Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1575–1578.

[20] W. K. Ngai, H. Xie, D. Zou, and K.-L. Chou, "Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources," *Information Fusion*, vol. 77, pp. 107–117, 2022.

[21] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[22] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2023.

[23] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Transactions on Cybernetics*, pp. 1–12, 2022.

[24] D. Deng, L. Wu, and B. E. Shi, "Iterative distillation for better uncertainty estimates in multitask emotion recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3557–3566.

[25] L. Yang, Q. Chen, Q. Zhang, and S. Chao, "Intelligent feature selection for eeg emotion classification," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 3681–3688.

[26] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.

[27] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1613–1622.

[29] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

[30] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[32] T. Joo, U. Chung, and M.-G. Seo, "Being bayesian about categorical probability," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4950–4961.

[33] S. Seo, P. H. Seo, and B. Han, "Learning for single-shot confidence calibration in deep neural networks through stochastic inferences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9030–9038.

[34] Y. Fan, R. He, J. Liang, and B. Hu, "Self-paced learning: An implicit regularization perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[35] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann, "Self-paced curriculum learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[36] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, and N. Zheng, "Deep self-paced learning for person re-identification," *Pattern Recognition*, vol. 76, pp. 739–751, 2018.

[37] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn—a multimodal lstm for speaker identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[38] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[39] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

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

[41] T. Zhang, X. Wang, X. Xu, and C. P. Chen, "Gcb-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 379–388, 2019.

[42] Z. Wang, T. Gu, Y. Zhu, D. Li, H. Yang, and W. Du, "Fldnet: Frame-level distilling neural network for eeg emotion recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2533–2544, 2021.

[43] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Multimodal gated information fusion for emotion recognition from eeg signals and facial behaviors," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 655–659.

[44] Q. Cai, G.-C. Cui, and H.-X. Wang, "Eeg-based emotion recognition using multiple kernel learning," *Machine Intelligence Research*, vol. 19, no. 5, pp. 472–484, 2022.

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

[46] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics: Methodology and Distribution*. Springer, 1992, pp. 162–190.

[47] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1247–1255.

[48] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis." *Journal of Machine Learning Research*, vol. 8, no. 2, 2007.

[49] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017

[50] W. Jiang, X. Liu, W. Zheng, and B. Lu, "Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5975–5984.

[51] X. Cheng, W. Wei, C. Du, S. Qiu, S. Tian, X. Ma, and H. He, "Vigilancenet: Decouple intra-and inter-modality learning for multimodal vigilance estimation in rsvp-based bci," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 209–217.

[52] J. Heo, H. B. Lee, S. Kim, J. Lee, K. J. Kim, E. Yang, and S. J. Hwang, "Uncertainty-aware attention for reliable interpretation and prediction," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[53] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[54] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[55] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

Qi Zhu received his B.S. degree, M.S. degree, and Ph.D. degree from Harbin Institute of Technology in 2007, 2010, and 2014, respectively. He is a professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He has published over 100 scientific articles in refereed international journal and conference proceedings. His interests include pattern recognition, feature extraction, and medical image analysis.

Chuhang Zheng received her B.S. degree in computer science and technology from Jimei University, in Jul.2022. He is currently pursuing the M.S. degree in College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His current research interests include multi-modal learning and affective computing.

Zheng Zhang (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the Harbin Institute of Technology in 2014 and 2018, respectively. He was a Research Associate at The Hong Kong Polytechnic University and a Post-Doctoral Research Fellow at The University of Queensland, Australia. Currently, he is an Associate Professor with the Harbin Institute of Technology, Shenzhen, China. His current research interests include machine learning and computer vision.

Wei Shao received his B.S. degree, M.S. degree from the Nanjng University of Technology, Jiangsu, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2018. He is an associate professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His interests include machine learning and bioinformatics.

Daoqiang Zhang (Senior Member, IEEE) received the B.S. degree, and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a Lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published over 200 scientific articles in refereed international journals such as IEEE Trans. Pattern Analysis and Machine Intelligence, IEEE Trans. Medical Imaging, IEEE Trans. Image Processing, Neuroimage, Human Brain Mapping, Medical Image Analysis; and conference proceedings such as IJCAI, AAAI, NIPS, CVPR, MICCAI, KDD, with 12,000+ citations by Google Scholar. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, won the best paper award or best student award of several international conferences such as PRICAI'06, STMI'12 and BICS'16, etc. He has served as a program committee member for several international conferences such as IJCAI, AAAI, NIPS, MICCAI, SDM, PRICAI, and ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence Pattern Recognition Society of the China Computer Federation (CCF).