A Multi-level Alignment and Cross-Modal Unified Semantic Graph Refinement Network for Conversational Emotion Recognition

Xiaoheng Zhang, Weigang Cui, Bin Hu, Fellow, IEEE, and Yang Li, Senior, IEEE

***Abstract*—Emotion recognition in conversation (ERC) based on multiple modalities has attracted enormous attention. However, most research simply concatenated multimodal representations, generally neglecting the impact of cross-modal correspondences and uncertain factors, and leading to the cross-modal misalignment problems. Furthermore, recent methods only considered simple contextual features, commonly ignoring semantic clues and resulting in an insufficient capture of the semantic consistency. To address these limitations, we propose a novel multi-level alignment and cross-modal unified semantic graph refinement network (MA-CMU-SGRNet) for ERC task. Specifically, a multi-level alignment (MA) is first designed to bridge the gap between acoustic and lexical modalities, which can effectively contrast both the instance-level and prototype-level relationships, separating the multimodal features in the latent space. Second, a cross-modal uncertainty-aware unification (CMU) is adopted to generate a unified representation in joint space considering the ambiguity of emotion. Finally, a dual-encoding semantic graph refinement network (SGRNet) is investigated, which includes a syntactic encoder to aggregate information from near neighbors and a semantic encoder to focus on useful semantically close neighbors. Extensive experiments on three multimodal public datasets show the effectiveness of our proposed method compared with the state-of-the-art methods, indicating its potential application in conversational emotion recognition. Implementation codes can be available at https://github.com/zxiaohen/MA-CMU-SGRNet.**

***Index Terms*—Emotion Recognition, Cross-modal Alignment, Multimodal Fusion, Semantic Refinement**

I. INTRODUCTION

Emotion recognition in conversations (ERC) has attracted increasing research interest due to its wide range of potential applications [1]. ERC has been particularly useful in addressing the limitations of traditional dialogue systems, which often produce responses that lack emotional depth or fail to consider the emotional state of the user [2]. For example, ERC enables dialogue systems to generate emotionally coherent and empathetic responses, which has also been utilized for opinion mining in social media analysis. Recently, despite its significant progress [2], emotion recognition has always been challenging due to interrelated reasons. First, the gap between modalities was inadequately treated due to their heterogeneity [3]. Second, as emotions were subtle, the emotion annotation was often subjective, resulting in an inevitable uncertainty [2]. Moreover, the multimodal semantic extraction was insufficient ignoring global context and speaker relationships [4].

The alignment between two modalities was crucial for the generation of the unified multimodal representation [5]. Previous works utilized unsupervised contrastive learning to alleviate this problem and obtained the promising results in several text classification tasks [6-7]. As the unsupervised contrastive learning framework commonly neglected the specificity of ERC task, the supervised contrastive learning (SCL) was investigated to the ERC task [8], where the utterances with the same emotion label were considered as positive pairs. In this way, instance-level samples with similar emotions became closer in the semantic space. However, the SCL treats two samples as a negative pair if they were with different labels [9], regardless of the quantitative semantic similarity between emotions. For instance, the happy is closer to excited than the sad. To solve this drawback, other alignments need to be explored to further enlarge the inter-class distances and to contrast the latent representations.

Previous works in the field of the conversational emotion recognition have been proposed by various fusion strategies to achieve a unified representation. These strategies included feature-level concatenation [3] which features from different modalities were concatenated together, and tensor fusion networks [10] which aimed to fuse multimodal information at a deeper level to capture more complex relationships between modalities. These approaches mentioned above addressed the challenge of the effectively integrating information from multiple modalities such as text, audio, and visual cues to improve the accuracy and robustness of the emotion recognition in conversational settings. Although these works achieved the promising results in the emotion recognition [11-12], the simple feature concatenation may easily suffer from the problem of the data sparseness. To address these limitations, recent works [13-15] mainly focused on the holistic utterance-level feature fusion and studied some powerful fusion strategies, such as a directed

This work was supported in part by the National Natural Science Foundation of China under Grant 62325301, Grant 62201023 and Grant U23A202768, and in part by the Beijing Natural Science Foundation under Grant Z220017, and in part by China Postdoctoral Science Foundation 2023M730175; and in part by the Beijing Municipal Education Commission-Natural Science Foundation under Grant KZ202110025036, and in part by the Beijing United Imaging Research Institute of Intelligent Imaging Foundation under Grant CRIBJQY202103. (Corresponding author: Yang Li.)

Xiaoheng Zhang is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. (e-mail: xiaoheng_zhang@buaa.edu.cn).

Weigang Cui is with the School of Engineering Medicine, Beihang University, Beijing 100191, China. (e-mail: cwg1994@buaa.edu.cn).

Bin Hu is with the School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China. (e-mail: bh@lzu.edu.cn).

Yang Li is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. (e-mail: liyang@buaa.edu.cn).

graph based cross-modal feature fusion [16], and the locally confined modality fusion through a local tensor fusion method [17]. However, these studies commonly ignored the impact of the uncertain factors, resulting in the inability to determine the informativeness of a modality. Higher uncertainty in a modality can reduce its informativeness, as the conveyed information becomes less reliable or harder to interpret. Conversely, a more informative and instructive modality tends to have lower levels of the uncertainty, allowing for more confident and accurate interpretation. Considering various sources of the uncertainty above [18-19], we propose a cross-modal uncertainty-aware fusion method by conducting uncertain estimation in emotion distribution. Consequently, the intrinsic ambiguity of emotions and the subjectivity of human perception, which give rise to variances in emotion labels, are inherently considered.

Since the emotion of each utterance is influenced by both previous utterances of the speaker and the responses of other interlocutors, context modeling is another key challenge for ERC [20]. Specifically, semantic understanding of the emotion shift is essential for contextual modeling. It is noted that current ERC frameworks only employed simple contextual features as representations. They seldom considered contextual semantic clues, thus resulting in an inadequate understanding of the semantics in a conversation [20]. A lack of semantics also raised difficulties for the identification of semantically similar emotions such as excited and happy. To address these issues, several approaches integrated the commonsense or external knowledge which the implicit semantic connections and dependencies were commonly ignored [21-22]. It is noted that the external knowledge can provide valuable the insight and contextual information, while the implicit semantic connections, such as relationships and subtle dependencies which significantly understand the in-depth meaning and semantic context, may not fully be captured. Thus, the node features of the semantic graph need to be further refined to capture more emotional semantic information of the dialogue context.

In this paper, we propose a multi-level alignment and cross-modal unified semantic graph refinement network (MA-CMU-SGRNet) to handle emotion recognition in conversation. First, in order to bridge the gap between acoustic and linguistic modalities, we design a multi-level alignment (MA) including the instance-level, prototype-level, and latent space alignment. A cross-modal uncertainty-aware unification (CMU) method is then investigated to deal with the issues of the emotional ambiguity and subjectivity in annotation. We further utilize the graph structure to preserve the correlations among utterances and the relations between utterances and speakers. A dual-encoding semantic graph refinement network (SGRNet) is finally introduced to consider both the syntactic structure and semantic correlation. We perform abundant experiments for our proposed model on three public datasets, achieving satisfactory recognition performance against the current methods. The main contributions of this work can be summarized as follows:

1) We incorporate the prototype-level alignment and latent space alignment to the instance-level alignment by using a novel supervised multi-level representation alignment, avoiding the problem of misalignment between modalities.

2) A cross-modal uncertainty-aware fusion strategy is proposed to obtain a unified representation which contains information from acoustic and lexical modalities, focusing on the inherent ambiguity and the subjectivity of emotions.

3) A dual-encoding semantic graph refinement network is designed to capture both the local and global underlying semantics, which includes a syntactic encoder to aggregate information from near neighbors and a semantic encoder to focus on useful semantically close neighbors, enhancing the semantic details of the context modeling.

II. RELATED WORK

A. Overview of Contrastive Representation Alignment

The contrastive learning is originally designed to map positive pairs to similar representations while pushing away those negative samples in the embedding space [23]. An innovative semantic-guided contrastive context-aware approach was presented to extract contrasting pairs of relevant and irrelevant utterances based on the conversational context of a target utterance [14]. This method established a soft semantic constraint between the target utterance and its context. Then, the success of the contrastive learning in self-supervised learning has inspired the generalization of this methodology to a much broader range, to make full use of the label information. A self-supervised batch contrastive approach was extended to the fully-supervised setting [8], which was adapted to identify similar emotions better in the ERC task [24].

However, two issues are still encountered. First, existing ERC datasets were often class-imbalanced, and samples may not be able to meet appropriate positive or negative samples in a mini-batch [25]. Compared to instance-wise contrastive methods, few works considered by using prototypical contrastive methods [26]. Besides, existing works involved a major limitation since the contrastive loss for intra-modal pairs [27] and inter-modal pairs were defined independently in separated spaces. Consequently, the contrastive loss was unaware of a substantial set of feasible combinations for the negative supervision. For instance, speech-speech pairs were not included when calculating the contrastive loss for speech-text supervision, leading to deficiencies in terms of data efficiency and feature diversity.

To address these challenges, we aim to develop a multi-level representation alignment framework. Particularly, an instance-level alignment will be defined for all conceivable intra-modal and inter-modal pairs within a unified embedding space. Moreover, the clustering strategy is proposed to generate pseudo labels, which can effectively enhance the differentiation between inter-class features while simultaneously aligning latent features.

B. Cross-modal Uncertainty-Aware Unification

Earlier research on feature fusion can be broadly categorized into two types, early fusion and late fusion. The former used the concatenation of unimodal representations, whereas the latter utilized all modalities independently to obtain a final inference. Most existing studies on ERC focused primarily on the textual modality [12]. Although they can be easily extended to multimodal paradigms [28], inter-modal interaction and fusion were not fully explored. To alleviate the problem, many efforts

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3354382

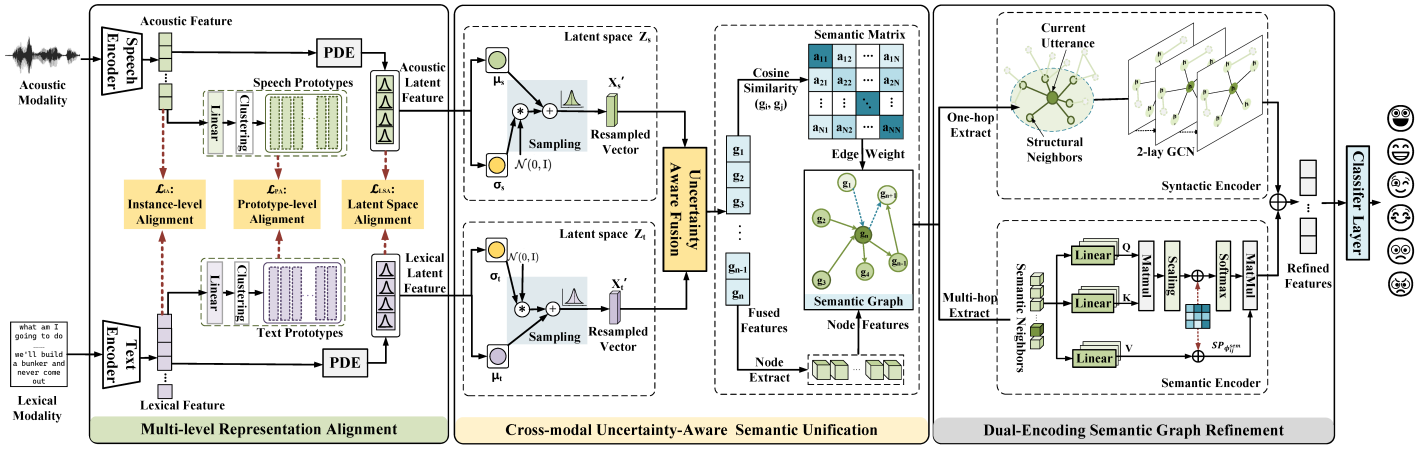


Fig. 1. The flowchart of the proposed MA-CMU-SGRNet, where the PDE is the Probability Distribution Encoder block, μ_t and μ_s denote the mean vector, σ_t and σ_s are the variance vector for text and speech distribution respectively, and $\mathcal{N}(0,1)$ represents the standard normal distribution which has the mean 0 and standard deviation 1, respectively.

have been dedicated to capture cross-modal interactions. For example, a tensor fusion network was first constructed to learn both intra-modality and inter-modality interactions via the Cartesian product [29]. Then, a low-rank multimodal fusion network was designed to improve the efficiency and reduce trainable parameters [30]. A conversational memory network also aligned features from different modalities by fusing multi-view information [11].

Nevertheless, the process of emotion recognition encompassed numerous ambiguous factors, such as the subjectivity inherent in emotion perception and evaluations [31]. Although deep neural networks detected hidden patterns, they lacked the inherent capacity to comprehend uncertainty [18]. Most previous works seldom considered the uncertainty caused by the ambiguity in emotional expression. Moreover, the informativeness of different modalities was mostly neglected, leading to an inadequate cross-modal interaction and unsatisfactory unified representation.

Therefore, we extend the existing deterministic methods to uncertainty-aware modeling, which can convey the informativeness of each modality to obtain a unified representation. We adopt the variance of unimodal latent distribution as a proxy for the informativeness of the modality in predicting the target emotion, and the inverse of variance values is used to quantify how uncertain a modality is to predict emotion labels. The potential of the variance-based uncertainty modelling for the multimodal fusion has been demonstrated [32]. Similarly, learning latent distribution variance was determined to be capable of the uncertainty modelling [33]. Inspired by these motivations, we utilize the unimodal variance values to estimate the uncertainty of the modalities which can predict the emotion labels.

C. Graph-based Semantic Context Modeling

A key challenge of ERC is to capture rich information in the dialogue context. As both the current utterance and the surrounding contexts are vital for the emotion perception, earlier works introduced recurrent neural networks (RNN) to model the intra-speaker and inter-speaker dependencies. A long short term memory (LSTM) model and an interactive conversational memory network were investigated to capture interaction and history context [34]. DialogueRNN model leveraged distinct gated recurrent units (GRU) to model multi-party relations and emotional dynamics [12]. Another branch of work leveraged the strong context modeling ability of transformer-based networks to model the global context [3]. Although the RNN-based approaches improved the accuracy compared with the earlier methods, they cannot effectively deal with the relationship between speakers and semantic contexts. To solve this problem, many efforts have been dedicated to improve graph-based neural networks. ERC is modeled as a node-classification task and solves the context propagation issues in RNN-based architectures. For instance, some research work explicitly incorporated a commonsense knowledge graph to enrich the semantic space [35], and a semantics graph attention was further employed to adjust the weight of knowledge [36]. Furthermore, besides of external knowledge, some graph networks used implicit global information to explore the semantic relationship between regional objects and concepts [37]. Another approach adopted the original transformer architecture [38] to capture the semantic relationships and dependencies between nodes.

However, the existing graph-based ERC methods also have some limitations. First, they mostly ignored the semantic similarity between context utterances leading to a lack of semantic correlation. Second, these models only focused on capturing local network structure for node embeddings, neglecting the broader context of the graph structure and global semantic features.

Inspired by previous works, we aim to develop a dual-encoding semantic graph refinement. This architecture will capture both local and global context information to solve the mentioned issues. Our approach involves a syntactic encoder for aggregating local information from nearby neighbors and a semantic encoder to focus on semantically similar neighbors globally, which can enhance mutually the local and global context information by incorporating.

III. METHODOLOGY

Given a dialogue $D = \{u_1, u_2, \ldots, u_N\}$, where N is the number of utterances. The modality-specific input features for

utterances in a dialogue are denoted as: $D^m = [u_1^m, u_2^m, \dots, u_N^m]$, $m \in \{s, t\}$, where s, t is speech and text respectively. The emotion recognition in conversation task aims to predict the emotion label for each utterance in the conversation. Each utterance involves two sources of data corresponding to acoustic and textual modalities represented as $u_i^s \in \mathbb{R}^{d_s}$ and $u_i^t \in \mathbb{R}^{d_t}$, and d_s, d_t represent original feature dimensions. We define a triplet-wise data format $\{(u_i^s, u_i^t, y_i)\}_{i=1}^N$ for a set of N utterances, where u_i^s and u_i^t are the speech and its corresponding language transcription of the i^{th} utterance, and y_i is the label indicating the emotion tag of the i^{th} utterance.

Our proposed MA-CMU-SGRNet is shown in Fig. 1 and concluded as follows: 1) Acoustic and lexical features are first embedded by speech/text encoders, followed by instance and prototype-level alignments using clustered class-specific features. Latent features are then obtained through a probability distribution encoder (PDE) block (Fig. 2) and aligned in a latent space so that the multimodal feature spaces are fully separated; 2) the input of the cross-modal uncertainty-aware semantic unification is the acoustic and lexical latent features which are resampled from Gaussian distributions to minimize the imbalance of both the within class and between class. Then a sequence of node features is generated after the uncertainty-aware fusion (UAF) block (Fig. 3) and a unified semantic graph is constructed, considering different kinds of uncertainties; 3) the semantic graph is aggregated through syntactic and semantic encoders in parallel to generate semantic-rich features; 4) the classification results are generated from the generated features of the semantic refinement. The details of the three modules are given in the following sections.

A. Multi-level Representation Alignment

For each speech $u_i^s \in \mathbb{R}^{d_s}$, a speech encoder model f_θ parameterized by θ first represents u_s as a acoustic feature vector $\tilde{u}_i^s \in \mathbb{R}^{d_e}$, $\tilde{u}_i^s = f_\theta(u_i^s)$. For each transcription $u_i^t \in \mathbb{R}^{d_t}$, we apply a text encoder f_ϕ paramerirized by ϕ to get feature vector $\tilde{u}_i^t \in \mathbb{R}^{d_e}$, $\tilde{u}_i^t = f_\phi(u_i^t)$. For the i^{th} speech $\tilde{u}_i^s$ and its transcription $\tilde{u}_j^t$, we normalize their feature vector to a hyper-sphere using $\tilde{s}_i = \frac{f_\theta(\tilde{u}_i^s)}{\|f_\theta(\tilde{u}_i^s)\|}$ and $\tilde{t}_i = \frac{f_\phi(\tilde{u}_i^t)}{\|f_\phi(\tilde{u}_i^t)\|}$.

Instance-level alignment: First, we introduce a new formulation by combining two data sources into a common speech-text-label space. In this space, we propose a new unified alignment paradigm to seamlessly prompt the synergy of two different modalities. The similarity score ${S_{i,j}}^{(0)}$ between the i^{th} embedding z_i and the j^{th} embedding z_j is defined by:

$$ {S_{i,j}}^{(0)} = exp\left(\frac{1}{\tau_1} \cdot \frac{{z_i}^T z_j}{\|z_i\|\|z_j\|}\right), z \in \{\tilde{s}, \tilde{t}\} \quad (1) $$

where the temperature τ_1 is a positive real number that can be pre-defined. The cosine similarity is divided by τ to extend its range, allowing the model to choose an appropriate scale for the convergence of a contrastive loss.

To classify an input pair (z_i, z_j) as positive or negative, we define a threshold ξ as an offset and classify it as positive if the cosine similarity between z_i and z_j is greater than ξ, and negative otherwise, which is given below:

$$ {S_{i,j}}^{(1)} = exp\left(\frac{1}{\tau_1} \cdot \frac{{z_i}^T z_j}{\|z_i\|\|z_j\|} - \xi\right), z \in \{\tilde{s}, \tilde{t}\} \quad (2) $$

For speech-text unified alignment, there are three possible combinations including speech-speech pairs, speech-text pairs, and text–text pairs. Considering that the threshold ξ can be different depending on whether the data pair is an intra-modal pair or an inter-modal pair, as in general, it would be easier to classify intra-modal positive pairs than inter-modal positive pairs. This motivates us to introduce modality-specific temperature $\tau_{m(i,j)}$, and offset $\xi_{m(i,j)}$, and propose a modality-dependent similarity score, which is represented by:

$$ S_{i,j} = exp\left(\frac{1}{\tau_{m(i,j)}} \cdot \frac{{z_i}^T z_j}{\|z_i\|\|z_j\|} - \xi_{m(i,j)}\right), z \in \{\tilde{s}, \tilde{t}\} \quad (3) $$

where $m(i,j) \in \{s, t, st\}$ denotes one of the three modality combinations. For the i^{th} encoded speech-text pair $(\tilde{s}_i, \tilde{t}_i)$ in a mini-batch $\mathcal{B}$, we regard two modalities as queries and keys alternatively to learn the correct pairings. The speech-to-text contrastive loss to align matched speeches in a batch with a given text is defined as l_i^{s2t}, while the text-to-speech contrastive loss to align matched texts to a given speech is denoted as l_i^{t2s}, which are defined by:

$$ l_i^{s2t} = -\frac{1}{|P(i)|}\sum_{j \in P(i)} \log \frac{S_{i,j}}{S_{i,j} + \sum_{n \in N(i)} S_{i,n}} \quad (4) $$

$$ l_i^{t2s} = -\frac{1}{|P(i)|}\sum_{j \in P(i)} \log \frac{S_{j,i}}{S_{j,i} + \sum_{n \in N(i)} S_{n,i}} \quad (5) $$

where $j \in P(i) = \{j | j \in \mathcal{B}, y_j = y_i, S_{i,j} > 1\}$, $\mathcal{B}$ is the mini-batch and $P(i)$, $N(i)$ represent the positive and negative set of the i^{th} embedding, respectively. The overall objective of our instance-wise alignment $\mathcal{L}_{IA}$ is the average of two losses, which can be defined by

$$ \mathcal{L}_{IA} = \frac{1}{2N_p}\sum_{i=1}^{N_p} (l_i^{s2t} + l_i^{t2s}) \quad (6) $$

where N_p is the total number of speech-text pairs.

Due to the limitation of the batch size, samples from the majority class (e.g., neutral) of the dataset may see insufficient negative samples within a batch. At the same time, it is hard for samples from the minority class to meet positive samples. To solve the issue above, we integrate a prototype-level alignment, which introduces prototype vectors of each category into the loss function mentioned above.

Prototype-level alignment: The instance-level alignment mentioned above treat two samples as a negative pair from different instances. Many pairs sharing the similar high-level semantics (e.g., emotion clues) are undesirably pushed apart in the embedding space. Therefore, we design a novel prototype-level alignment (PA) to harness the cross-modal inter-subject correspondences between speech and text.

For the i^{th} speech-text embedding pair $(\tilde{s}_i, \tilde{t}_i)$, we employ the iterative Sinkhorn-Knopp clustering algorithm to acquire two soft cluster assignment codes $q_{s,i} \in \mathbb{R}^K$ and $q_{t,i} \in \mathbb{R}^K$, by individually assigning $\tilde{s}_i$, $\tilde{t}_i$ into K clusters. These assignment codes are utilized as pseudo-labels. Meanwhile, we also pre-define K trainable cross-modal prototypes as $C = \{c_1, \dots, c_K\}$. Then, we calculate the softmax probability $p_{s,i} \in \mathbb{R}^K$, $p_{t,i} \in \mathbb{R}^K$ of the cosine similarities between $\tilde{s}_i, \tilde{t}_i$ and all cross-modal prototypes in C, which are denoted by:

$$ p_{s,i}^k = \frac{exp\,({\tilde{s}_i}^T c_k/\tau_2)}{\sum_k exp\,({\tilde{s}_i}^T c_k/\tau_2)} \quad (7) $$

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3354382

$$p_{t,i}^{k} = \frac{exp\,(\tilde{t}_i{}^T c_k/\tau_2)}{\sum_k exp\,(\tilde{t}_i{}^T c_k/\tau_2)} \quad (8)$$

where τ_2 is the prototype-level temperature parameter and k indicates the k^{th} element of the prototype vector. The cross-modal prototype-level alignment is achieved by conducting cross-modal prediction and optimizing the following two cross-entropy losses:

$$l(\tilde{s}_i, q_{t,j}) = \sum_{k=1}^{K} q_{t,i}^{(k)}\, log\, p_{s,i}^{(k)} \quad (9)$$

$$l(\tilde{t}_i, q_{s,j}) = \sum_{k=1}^{K} q_{s,i}^{(k)}\, log\, p_{t,i}^{(k)} \quad (10)$$

Here, the cross-modal prediction is implemented by using the soft text cluster assignment code $q_{t,i}$ to train the speech representation $\tilde{s}_i$, while taking that of speech $q_{s,i}$ to train the text representation $\tilde{t}_i$. Finally, the overall prototype-level alignment (PA) loss $\mathcal{L}_{PA}$ is the average of two prediction losses over all the speech-text pairs, which is defined by:

$$\mathcal{L}_{PA} = \frac{1}{2N_p}\sum_{i=1}^{N_p}(l(\tilde{s}_i, q_{t,i}) + l(\tilde{t}_i, q_{s,i})) \quad (11)$$

where N_p is the total number of speech-text pairs.

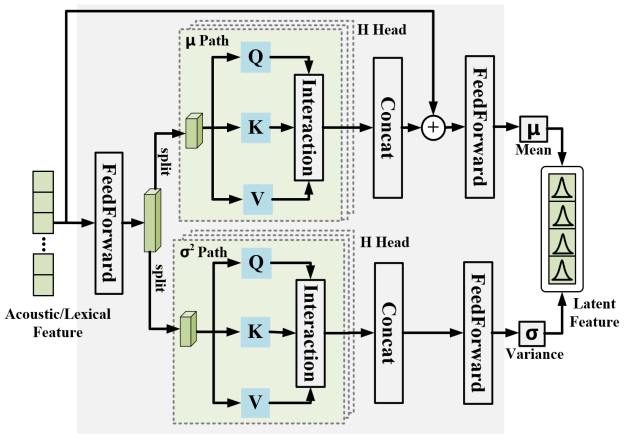


Fig. 2. Probability Distribution Encoder (PDE) block.

Latent Space Alignment: We then propose a probability distribution encoder (PDE) (Fig. 2), and considering that modeling the mean vectors and variance vectors takes feature-level and sequence-level interactions. To model the multimodal uncertainty, we further frame the input features as the multivariate Gaussian distribution. Specifically, the PDE takes an input data point and predicts a mean vector (μ) and a variance vector (σ^2) to obtain the latent features, which defines the parameters of the Gaussian distribution in the latent space. The latent space is typically a lower-dimensional continuous space and assumed to be a multivariate Gaussian distribution, where the mean and variance vector are obtained from the encoder output.

The input features of the PDE $\tilde{s}_i$ and $\tilde{t}_i$ are from the point representation space of the different modalities. Specifically, the feed forward layer is used for feature-level interactions and the multi-head operation is responsible for sequence-level interactions. The input hidden states are split into H heads, and we split the features and send them to two paths (μ, σ^2) in each head. Additionally, for the μ path, the input hidden state is projected to $Q^{(h)}, K^{(h)}, V^{(h)}$. The interaction $Int(\cdot)$ includes an activation function and a normalization function for considering sequence-level interaction. The output features of each head denoted as $Int(\frac{Q^{(h)}K^{(h)}}{\sqrt{d_e}})V^{(h)}$ are concatenated to obtain the final mean vector. Since the input point representation correlates with the mean vector, a residual connection by an add operation is employed to learn the mean vector.

Inspired by the concept of alignment for attributes and features in zero shot learning [39], the variational alignment optimizes the distance between the distribution $\mathcal{N}(\mu_s, \sigma_s)$ and $\mathcal{N}(\mu_t, \sigma_t)$ by simultaneously aligning the mean vector and variance as below:

$$\mathcal{L}_{LSA} = \frac{1}{2N_p}\sum_{i=1}^{N_p}(\|\mu_{s,i} - \mu_{t,i}\|_2^2 + \|\sigma_{s,i} - \sigma_{t,i}\|_2^2)^{\frac{1}{2}} \quad (12)$$

where $\mu_{s,i}$ and $\mu_{t,i}$ represent the mean vectors projected from the i^{th} pair of speech and text in a mini-batch. Similarly, $\sigma_{s,i}$ and $\sigma_{t,i}$ stand for the corresponding variance matrices. By minimizing $\mathcal{L}_{LSA}$, the features of various modalities are mapped to common latent embedding space and the attention mechanism module learns to assign reasonable weights to different modalities based on the importance and informativeness. The final learning objective of the multi-level alignment module is defined as follows:

$$\mathcal{L}_A = \gamma_1\mathcal{L}_{IA} + \gamma_2\mathcal{L}_{PA} + \gamma_3\mathcal{L}_{LSA} \quad (13)$$

where $\gamma_1, \gamma_2, \gamma_3$ are hyperparameters between $[0, 1]$ to balance three different level cross-modal alignments.

B. Cross-modal Uncertainty-Aware Semantic Unification

The speech-text embedding pairs are from the point representation space of different modalities. To model the multimodal uncertainty, we further frame the input features as multivariate gaussian distributions. Specifically, we predict a mean vector and a variance vector for each input feature. The mean vector represents the center position of distributions in probabilistic space, and the variance vector expresses the scope of distributions in each dimension.

Uncertainty-Aware Fusion: A multivariate normal distribution is adopted to represent the hidden state output instead of a typical deterministic embedding vector. For an input utterance with index i, the output parameters (mean and variance) of multivariate normal distributions are denoted as $\mathcal{N}(\mu_s, \sigma_s)$ and $\mathcal{N}(\mu_t, \sigma_t)$ over the audio and text temporal context vectors, respectively. As variance values are assumed to represent modality-specific certainty and to determinate how informative that modality is for predicting the target emotion. We first calculate the L_2 norms of their variance $\|\sigma_s\|$ and $\|\sigma_t\|$, then obtain the fusion weights ω_s, ω_t which are defined by:

$$\omega_s = \frac{\|\sigma_s\|}{\|\sigma_s\|+\|\sigma_t\|} \quad (14)$$

$$\omega_t = \frac{\|\sigma_t\|}{\|\sigma_s\|+\|\sigma_t\|} \quad (15)$$

After quantifying modality-related uncertainty in the temporal context space, we generate the estimated uncertainty scores as the fusion weights. $X^s \in \mathbb{R}^{N\times d_f}$ and $X^t \in \mathbb{R}^{N\times d_f}$ are resampled vectors from two Gaussian distributions based on Markov Chain Monte Carlo (MCMC) algorithm, formulated by $X^s = \mu_s + \sigma_s \odot \mathcal{N}(0,1)$ and $X^t = \mu_t + \sigma_t \odot \mathcal{N}(0,1)$, where N denotes the number of utterances and d the feature dimension. Resampling vectors based on distributions are able to minimize within-class imbalance by sampling diverse distributions and addressing bias due to the imbalanced class distribution. The resampled vectors X^s and X^t are fed into the

UAF block (Fig. 3) to generate the uncertainty-aware features $\tilde{X}^s$ and $\tilde{X}^t$, which are defined by:

$$\tilde{X}^s = \omega_s \odot X^s \quad (16)$$

$$\tilde{X}^t = \omega_t \odot X^t \quad (17)$$

where $\odot$ denotes the element-wise multiplication. We further apply an uncertainty-aware cross-modal fusion strategy to deal with the ambiguity focusing on correspondence between two modalities (Fig. 3). We extend the traditional attention to a two-stream cross-modal attention to model interactions between two modalities, which could combine the information from two weighted features $\tilde{X}^s$ and $\tilde{X}^t$ to transform the text features V_t using the feature map generated by Q_s and K_t in Eq. (21). The query Q_s, key K_t , and value V_t have been defined by

$$Q_s = \tilde{X}^s W_q^s \quad (18)$$

$$K_t = \tilde{X}^t W_k^t \quad (19)$$

$$V_t = \tilde{X}^t W_v^t \quad (20)$$

$$O_{st} = softmax(Q_s{}^T K_t) \cdot V_t \quad (21)$$

where W_q^s, W_k^t and W_v^t are trainable weight matrix. The other branch follows the same principle to obtain O_{ts}. These two features then go through a linear layer and the final output features of two branches are concatenated to obtain the final feature embeddings $\mathcal{V}^s = [g_1, g_2, \dots, g_N] \in \mathbb{R}^{N \times d_f}$, which will be used for the semantic graph construction.

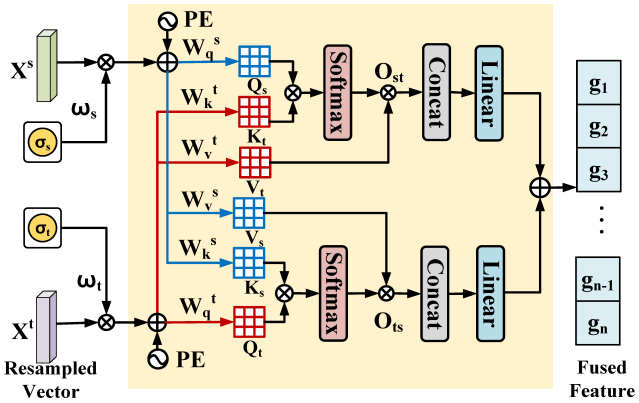


Fig. 3. Uncertainty-Aware Fusion (UAF) block.

Semantic Graph Construction: To establish semantic relations between the nearby utterances and capture both inter-speaker and intra-speaker effects, we define a semantic graph $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{M}^s)$ based on the semantic-aware dependency. Each utterance is represented by a node embedding and different connection edges represent directed relations (past and future). $\mathcal{V}^s = [g_1, g_2, \dots, g_N]$ denotes a set of utterance nodes, and $\mathcal{M}^s$ is a matrix of relations that represent the semantic similarity between the i^{th} and j^{th} utterances, which is defined below:

$$\mathcal{M}_{i,j}^s = 1 - arccos\left(\frac{g_i{}^T g_j}{\|g_i\|\|g_j\|}\right), \; i,j \in [1,N] \quad (22)$$

We define intra-relations between the utterances spoken by the same speaker $R_{intra} \in \{g^{\mathcal{S}_i} \rightarrow g^{\mathcal{S}_j}\}_{\mathcal{S}_i = \mathcal{S}_j}$ and inter-relations by different speakers $R_{inter} \in \{g^{\mathcal{S}_i} \rightarrow g^{\mathcal{S}_j}\}_{\mathcal{S}_i \neq \mathcal{S}_j}$. A context window is further considered by using Θ_p and Θ_f as hyperparameters to denote relations between the past Θ_p utterances and future Θ_f utterances for every utterance. The relational semantic graph can be regarded as a local-view modeling of relationships between utterances in a dialogue covering semantics features.

C. Dual-Encoding Semantic Graph Refinement

In this paper, we propose a dual-encoding semantic graph refinement, which consists of a syntactic encoder to aggregate information from near neighbors and a semantic encoder, focusing on useful semantically close neighbors in a global view. In particular, the well-defined semantic graph $\mathcal{G}^s$ is fed into a two-layer relational graph convolutional network (RGCN) to compute syntactic features of utterances and their interaction relations. Meanwhile, the semantic relationship between utterances are extracted by the semantic encoder.

Syntactic Encoder: A modified relational graph convolution layer is adopted to capture local dependencies defined by the relations. The node representations and edge weights are then fed into a two-layer correlation-based RGCN. Each layer can be summarized as below to aggregate structural features:

$$H_i^{(l+1)} = F\left(\sum_{r \in R} \sum_{j \in G_i^r} \frac{1}{|G_i^r|} W_r^{(l)} g_j{}^{(l)} + W_0^{(l)} g_i{}^{(l)}\right) \quad (23)$$

where G_i^r denotes the neighboring indices of g_i under the same relation $r \in R = \{R_{intra}, R_{inter}\}$, $|G_i^r|$ represents the number of G_i^r, $W_r^{(l)}$ and $W_0^{(l)}$ are learnable parameters, $F(\cdot)$ is the activation function and l represent the number of layers. In this way, each graph convolution layer models the interaction between utterances, and refines the syntactic features. The output of syntactic encoder is denoted as $h_i{}^{sy}$.

Semantic Encoder: We adopt a semantic encoder to extract global semantic information from node features. It adopts the vanilla multi-head attention into graph learning by taking into account nodes connected via edges. We define two encodings to represent semantic relationship between two nodes. The first is relative position encoding, where each vector represents the topological relation represented by their shortest path distance between the i^{th} and the j^{th} nodes denoted as $\mathcal{M}_{i,j}^p$, the second is semantic encoding defined by Eq. (22). We take an element-wise addition operation and obtain $SP_{\phi_{ij}^{sem}}$ in Eq. (24).

Previous methods only focus on encoding graph information into either the attention map or input features, while our method encodes positional and semantic information into attention map $a_{i,j}$ by Eq. (25) to consider the global semantic context. Moreover, the hidden features $h_i{}^{se}$ can be encoded via Eq. (26).

$$SP_{\phi_{ij}^{sem}} = \mathcal{M}_{i,j}^s + \mathcal{M}_{i,j}^p \quad (24)$$

$$a_{i,j} = \frac{(W_q g_i)^T (W_k g_j)}{\sqrt{d_f}} + SP_{\phi_{ij}^{sem}} \quad (25)$$

$$h_i{}^{se} = \sum_{j=1}^{N_0} softmax(a_{i,j})(W_v g_i + SP_{\phi_{ij}^{sem}}) \quad (26)$$

where W_q , W_k , and W_v are trainable parameters, N_0 is the number of nodes, d_f is the node feature dimension. Finally, the output node features $h_i \in \mathbb{R}^{d_f}$ which is the concatenation of $h_i{}^{sy}$ and $h_i{}^{se}$, is used for the emotion classification.

D. Emotion Classifier

The final output of semantic graph refinement is fed into a fully connected layer followed by a softmax layer to calculate emotion-class probabilities β_i, as shown below:

$$\tilde{h}_i = ReLU(W_f h_i + b_f) \quad (27)$$

$$\beta_i = softmax(W_p \tilde{h}_i + b_p) \quad (28)$$

where $W_f \in \mathbb{R}^{d_f \times d_f}$, $b_f \in \mathbb{R}^{d_f}$, $W_p \in \mathbb{R}^{d_f \times C_0}$ and $b_p \in \mathbb{R}^{C_0}$ are trainable parameters with C_0 the number of emotion categories. Then we select the most probable emotion class as the predicted label by

$$\hat{y}_i = argmax(\beta_i) \tag{29}$$

where $\hat{y}_i \in \mathbb{R}^1$ is the emotion label predicted for each utterance. We choose the categorical cross-entropy loss function as the classification loss $\mathcal{L}_{cl}$ during the training stage, which is shown below:

$$\mathcal{L}_{cl} = -\frac{1}{\sum_{i=1}^{L} N_i} \sum_{i=1}^{L} \sum_{c=1}^{C_0} y_{i,c}^{(j)} \cdot log\, \hat{y}_{i,c}^{(j)} \tag{30}$$

where L is the number of conversations and N_i is the number of utterances in the i^{th} conversation, $\hat{y}_{i,c} \in \mathbb{R}^1$ and $y_{i,c} \in \mathbb{R}^1$ are the predicted output of class c for the j^{th} utterance in the i^{th} conversation, respectively. The total loss function $\mathcal{L}_{total}$ for our proposed framework is defined by the combination of the classifier learning loss from Eq. (30) and the alignment loss via Eq. (13):

$$\mathcal{L}_{total} = \mathcal{L}_{cl} + \lambda\, \mathcal{L}_A \tag{31}$$

where $\lambda \in [0,1]$ is a hyperparameter to balance two terms.

IV. EXPERIMENTS AND RESULTS

In this section, we first discuss the details of three emotion recognition and sentiment analysis datasets used for evaluating the proposed MA-CMU-SGRNet. Then, we provide the corresponding performance evaluation metrics. Finally, we present the implementation details and the performance.

A. Datasets

IEMOCAP [40] dataset contains approximately 12 hours of dyadic emotional improvised and scripted conversations (10039 utterances). The labelling of each utterance was determined by 3 annotators as the following categorical labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise. Following previous work [3], utterances from the first 8 speakers are used as the training and validation sets while the others are used as the testing set.

MELD [41] is a large-scale multi-party conversational dataset which contains 13708 utterances and 1433 conversations from TV series Friends, spanning 13.7 hours of various dialogue scenarios. Each utterance is annotated with one of the following labels: anger, joy, sadness, neutral, disgust, fear and surprise by three annotators. Then a majority voting scheme is applied to select the final emotion label for each utterance. Different from IEMOCAP that contains dyadic conversations, MELD is a multi-party dataset where two or more speakers are involved in a conversation.

CMU-MOSEI [42] is a sentiment analysis dataset made up of 23,454 movie review video clips taken from YouTube. It contains 3225 dialogues and approximately 23,000 utterances. Each sample is labeled by human annotators with a sentiment score from -3 (strongly negative) to 3 (strongly positive) thus is associated with one sentiment value. In this paper, we focus on the utterance-level aligned version.

B. Evaluation Metrics

To evaluate our proposed method against previous methods, we adopt the following evaluation metrics [17]. Denote C_0 as emotion classes in the dataset, and Γ_j as the number of samples of the class $j \in [1, C_0]$. Acc_j and $F1_j$ represent the classification accuracy and the $F1$ score of the class j, respectively.

Weighted average accuracy (WAA) is a weighted mean accuracy over different emotion classes with weights proportional to the number of utterances in a particular emotion class, which is denoted by:

$$WAA = \frac{\sum_{j=1}^{C_0} \Gamma_j \cdot Acc_j}{\sum_{j=1}^{C_0} \Gamma_j} \tag{32}$$

Weighted average F1 (WAF1) is a weighted mean F1 over different emotion categories with weights proportional to the number of utterances in a particular emotion class, which is given below:

$$WAF1 = \frac{\sum_{j=1}^{C_0} \Gamma_j \cdot F1_j}{\sum_{j=1}^{C_0} \Gamma_j} \tag{33}$$

IEMOCAP and MELD are labeled in discrete emotion categories. To compare with other methods, we evaluate the performance of emotion recognition by using weighted average accuracy (WAA) and weighted average F1-score (WAF1). Due to the natural imbalance across various emotions, following [3] [11], we choose WAF1 as the primary evaluation metric.

CMU-MOSEI: In this paper, we focus on the negative and positive classification task. The positive class and the negative class are assigned for positive and negative scores, respectively. We evaluate the model performances by using binary accuracy (positive/negative sentiments) and F1 score, in agreement with those previous works [43]. We thus choose WAF1 and WAA as evaluation metrics considering the inherent imbalance.

C. Implement Details

We performed all experiments on the Pytorch framework with the Intel Core i7-12700H and the NVIDIA RTX3060 GPU. The software environment includes Python 3.9, Pytorch 1.12.1, and CUDA 11.3. Specifically, the Adam optimizer with an initial learning rate r_0 of 1×10^{-4} is used to optimize the parameters of the proposed MA-CMU-SGRNet, the weight decay λ_d is set to 0.001 and a dropout rate d_r of 0.5 is adopted to alleviate overfitting problems where the detail information of the parameter initialization can be founded in [3]. To dynamically adjust the training process, a cosine annealing strategy is utilized to update the learning rate, which is summarized in **Algorithm 1**. To ensure a fair comparison with previous works [3] [42], the entire network performs 10 cycles, where one cycle contains 150 epochs [3]. We pad the conversations of the same mini-batch to have the same number of utterances. We also add bit masks to these padded utterances to eliminate their effect during training stage.

Besides, audio features ($d_s = 100$) are extracted by using OpenSmile [45] and text features ($d_t = 768$) are extracted by using sBERT [46]. Hyperparameters are decided by applying a random search procedure [3]. Based on the validation performance, we set hyperparameters as follows: the GRU layers in the single-modal encoder map acoustic and lexical features into the fixed dimension of size $d_e = 150$. The head number is set to 4 for cross-modal uncertainty-aware fusion and 2 for semantic encoder. The outputs of the uncertainty weighted fusion structure are set to be $d_f = 100$. The number of prototypes is $K = 500$. Following the contrastive learning[47],

Algorithm 1 Pseudocode of the proposed MA-CMU-SGRNet

Input: Acoustic and textual representations as u_i^s and u_i^t, ground truth y_i, maximum epoch B, learning rate r, batch size N, the MA-CMU-SGRNet $(\cdot)$ denoted as MCSNet $(\cdot)$;
Output: The emotion classification result $\hat{y}_i$.

Initialize all parameters in MCSNet as $\theta(0)$;
Initialize data $D = \{u_1, u_2, ...u_i..., u_N\}$ with $i \in [1, N]$;
Initialize $B = 150$, $r_{max} = 1 \times 10^{-4}$, $r_{min} = 1 \times 10^{-8}$, $r_0 = r_{max}$, $q = 0$, $j_e = 0$, $\eta = 50$, $\lambda_d = 0.001$;
while $q \neq B$ **do**
  j_e++
  for $i = 1$ to N **do**
    Generate the unified semantic feature of the i^{th} utterance denoted as $g_i \in \mathbb{R}^{2d}$ with $i = 1, 2, 3...N$;
    Generate the prediction result $\hat{y}_i$ = MCSNet (θ, D);
    Calculate the loss $\mathcal{L}_{total} = \mathcal{L}_{cl} + \lambda\mathcal{L}_{\mathcal{A}}$;
    Calculate the gradient $\delta = \Delta\mathcal{L}_{total}$
    Update the parameters $\theta^{(q+1)} \leftarrow (1-\lambda_d)\theta^{(q)} - r^{(q)}\delta$
  end for
  if $j_e = \eta$ **then**
    Update the learning rate:
    $r^{(q+1)} \leftarrow r_{min} + \frac{1}{2}(r_{max} - r_{min})(1 + cos\frac{j\pi}{\eta})$;
  else
    $r^{(q+1)} \leftarrow r_{max}$;
    $j_e = 0$;
    $q++$
  end if
end while
Get the classification result $\hat{y}_i = argmax(softmax(\beta_i))$

the temperature hyperparameters are $\tau_s = 0.1$, $\tau_t = 0.3$, $\tau_{st} = 0.05$, $\tau_2 = 0.07$, which are determined by a grid search strategy via cross-validation on the training samples. For each combination of hyperparameters in the loss function, *i.e.*, $\lambda_1, \lambda_2, \lambda_3$ is [0.1, 0.2,…, 1] for the instance-level alignment, the prototype-level alignment, and the latent space alignment, we determine the parameter values via a grid search strategy and achieve $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, $\lambda_3 = 0.3$, respectively. We adopt the same hyperparameter values to the models trained on the other datasets as well.

D. Statistical Performance

In this section, to verify the effectiveness of our proposed method, we first compare the overall performance of our proposed method with the state-of-the-art baseline approaches on three datasets by using WAA, WAF1. Then, we perform comparisons on classification accuracies and F1 for each category. We implement the following state-of-the-art methods to evaluate the performance of our proposed approach:

Bc-LSTM [44] used a bi-directional LSTM to encode context information, ignoring the speaker-level dependency

DialogueRNN [12] was based on recurrent neural networks that kept the track of the individual party states throughout the con-versation and used the information for the emotion classification.

CTNet [3] utilized the transformer to obtain the multimodal repesentation by modeling cross-modal interactions.

A-DMN [11] modeled self and inter-speaker influences and then synthesizes two factors to update the memory.

I-GCN [14] utilized the graph structure to represent conversation at different time and applied the incremental graph structure to imitate the process of dynamic conversation.

GraphCFC [16] extracted various types of edges from the constructed graph for encoding, thus enabling GNNs to extract crucial contextual and interactive information more accurately when performing message passing for multimodal learning.

1) Comparison on Overall Performance: To evaluate the efficacy of the proposed method, we perform experiments on three public datasets. Table I, II and III list the performance metrics on three datasets. Since there are fewer methods that employ the CMU-MOSEI dataset, we list them separately in Table III. We list the performance of current approaches on emotion recognition and compare our proposed method with them. Experimental results demonstrate the effectiveness of our proposed MA-CMU-SGRNet. Specifically, for the IEMOCAP dataset in Table I, we obtain 72.4% on WAA and 71.6% on WAF1, which outperforms all the baselines mentioned above. Although some other methods achieve the highest F1 values for a particular emotion classification, for example, GraphCFC achieving the highest F1 value on sad (85.0%) and excited (78.9%) emotion, the performances on other emotion categories are inferior to the proposed MA-CMU-SGRNet. The obvious improvement on WAA and WAF1 shows that our proposed method can identify fine-grained emotion compared with other methods. Our proposed method succeeds over Bc-LSTM and DialogueRNN by 12.6% on WAA, 12.6% on WAF1 and 13.1% on WAA, 11.8% on WAF1, which apply context modeling with-out cross-modal fusion architecture. This observation proves the importance of cross-modal fusion strategy. In addition, it outperforms CTNet and A-DMN which utilize multimodal fusion approaches by 4.2%~4.4% on WAA and 3.5%~4.1% on WAF1. The main reason is that these existing methods only focused on the multimodal representation, ignoring the semantic relation-ship between utterances. Moreover, our proposed method also outperforms I-GCN which highlights the semantic correlation information of utterances, neglecting the representation alignment between different modalities.

For the MELD dataset, which provides a more fine-grained categorization, following previous works [33], we report the WAF1 for a fair comparison. Our proposed method obtains the highest indicator against the other methods (Table II). The WAF1 of our proposed MA-CMU-SGRNet is increased by 2% and 0.5% respectively compared with the newly proposed CMCF-CRNet. In addition, compared with Bc-LSTM and DialogueRNN depending on the text modality, our WAF1 increases by 2.0%~3.0%. This enhancement is due to our design of cross-modal unification using information from different sources. Compared with CTNet and A-DMN which consider multimodal fusion, we achieve an WAF1 increase of 1.9% by leveraging multi-level alignment to further find out the common and complementary features. Besides, compared with the graph-based method I-GCN, our proposed method gains a WAF1 improvement of 1.5%, which highlights the significance of semantic understanding in the emotion classification.

For CMU-MOSEI, our MA-CMU-SGRNet obtains 86.9% and 86.8% on WAA and WAF1, which are better than all methods in Table III. Specifically, our proposed method outper-

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3354382

TABLE I: COMPARISON WITH THE STATE-OF-THE-ART METHODS ON IEMOCAP DATASET.

Models	Year	IEMOCAP: Emotion Categories													
		Happy		Sad		Neutral		Angry		Excited		Frustrated		Average	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	WAA	WAF1
Bc-LSTM [44]	2017	22.5	35.6	58.6	69.2	56.5	53.5	70.0	66.3	58.8	61.1	**67.4**	62.4	59.8	59.0
DialogueRNN [12]	2019	31.25	33.8	66.1	69.8	63.0	57.7	61.7	62.5	61.5	64.4	59.6	59.5	59.3	59.8
CTNet [3]	2021	47.9	51.3	78.0	79.9	69.0	65.8	72.9	67.2	85.3	78.7	52.2	58.8	68.0	67.5
A-DMN [11]	2022	43.1	50.6	69.4	76.8	63.0	62.9	63.5	56.5	**88.3**	77.9	53.3	55.7	64.6	64.3
I-GCN [14]	2022	51.4	50.0	85.3	83.8	60.4	59.3	61.2	64.6	75.6	74.3	57.2	59.0	65.5	65.4
GraphCFC [16]	2023	-	43.1	-	**85.0**	-	64.7	-	71.4	-	**78.9**	-	63.7	-	68.9
Ours	2023	**52.6**	**57.1**	**78.8**	79.9	**74.3**	**71.0**	**75.2**	**71.5**	80.3	78.4	65.1	**67.5**	**72.4**	**71.6**

The improvement is statistically significant with $p \leq 0.05$ under t-test. Bold font represents the best performance. Acc. = Accuracy.

TABLE II: COMPARISON WITH THE STATE-OF-THE-ART METHODS ON MELD DATASET.

Models	Year	MELD: Emotion Categories							
		Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise	Avg
		F1	F1	F1	F1	F1	F1	F1	WAF1
Bc-LSTM [44]	2017	43.4	**23.7**	9.4	54.5	76.7	24.3	51.0	59.3
DialogueRNN [12]	2019	43.7	7.9	11.7	54.4	77.4	34.6	52.5	60.3
CTNet [3]	2021	44.6	11.2	10.0	56.0	77.4	32.5	52.7	60.5
A-DMN [11]	2022	43.9	7.2	12.0	56.7	77.1	29.1	55.1	60.4
I-GCN [14]	2022	43.5	11.8	8.0	54.7	78.0	**38.5**	51.6	60.8
CMCF-SRNet [15]	2023	43.9	10.9	11.5	55.8	77.2	36.0	52.9	61.8
Ours	2023	**44.3**	11.9	**12.1**	**56.9**	**78.4**	35.9	**53.5**	**62.3**

TABLE III PERFORMANCE ON CMU-MOSEI DATASET.

Methods	CMU-MOSEI		
	Year	WAA	WAF1
GMFN [42]	2017	76.9	77.0
MulT [48]	2019	82.5	82.3
MMIM [49]	2021	85.9	85.9
CONKI [50]	2023	86.2	86.1
Ours	2023	**86.9**	**86.8**

forms MulT by 4.2% on WAF1 and 4.4% on WAA, which constructs an architecture based on unimodal and cross-modal transformer and completes fusion process by attention. Compared with the newly proposed CONKI, the metrics of the WAA and WAF1 of our proposed method increase by 0.6% and 1.0%, respectively. This is probably due to our multi-level alignment which generate a unified representation before the semantic refinement. These improvements show the generalization capability of our proposed MA-CMU-SGRNet from the emotion recognition to sentiment analysis. Besides, all the p-values are less than 0.05, which demonstrates the improvement of the classification performance compared with recent methods is statistically significant.

2) Comparison on Performance for Each Class: In this section, we report the performance indicator corresponding to each emotion label in detail on three datasets. We also visualize the confusion matrices of the testing set in Fig. 4.

For IEMOCAP dataset, experimental results in Table I demonstrate that our proposed MA-CMU-SGRNet achieves improvements on classification accuracies in individual emotion recognition tasks in most cases. Specifically, our proposed MA-CMU-SGRNet outperforms other methods on the happiness (57.1%), neutral (71.0%), angry (71.5%) and frustrated (67.5%) emotions on F1 score. We can observe that the results of the proposed MA-CMU-SGRNet show remarkably significant improvement relative to those of baseline models for happy and frustrated, which are easily confused with other categories. It is notable that the recent I-GCN adopted the graph-based context modeling method to achieve comparable performance with our MA-CMU-SGRNet

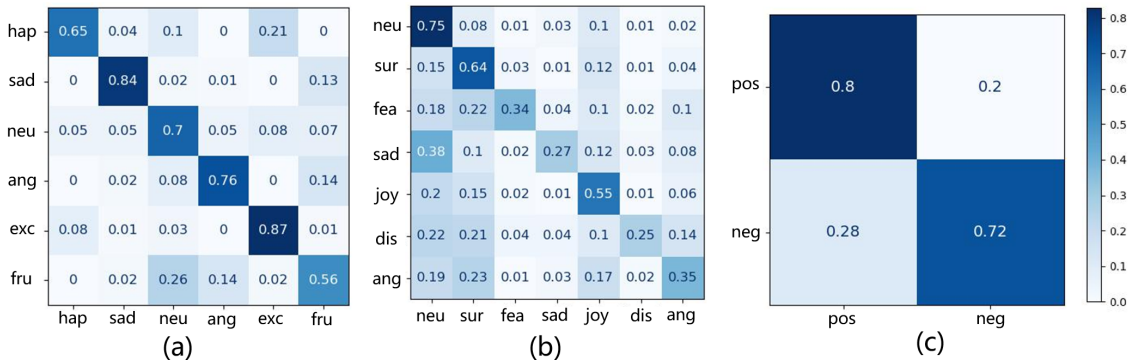


Fig. 4. The confusion matrices of the proposed MA-CMU-SGRNet on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3354382

on IEMOCAP dataset. However, the I-GCN only utilized the instance-level feature extraction when acquiring context information, resulting in the insufficient ability to differentiate between similar emotions. We can observe that the happy and excited accuracy is only 50.0% and 74.3%, which is 7.1% and 4.1% lower than the proposed MA-CMU-SGRNet.

For the MELD dataset, experimental results in Table II demonstrate that our proposed MA-CMU-SGRNet outperforms other methods on F1 scores for most emotion categories, including angry, joy, neutral, fear and surprise. A slight decrease in the F1-score of the sadness emotion can be attributed to better generalization for other emotions, we assume that is due to the negative valence and negative arousal emotion of sadness so that similar to text features, the acoustic characteristics of sadness are also implicit. Besides, our proposed MA-CMU-SGRNet has lower accuracies for disgust and fear emotion categories for the F1 score which is probably since disgust and fear emotions occupy a quite small portion in the MELD dataset. As a result, the model tends to learn less about them. The I-GCN adopted an incremental graph convolution network, which utilized the graph structure to represent conversation at different times, thereby achieving the highest F1 score on sadness emotion category (38.5%). Nevertheless, the I-GCN was 2.2% and 1.9%lower than our proposed MA-CMU-SGRNet on joy and surprise, indicating that our proposed method has more ideal classification performance than the I-GCN in terms of finer-grained emotion.

Furthermore, Fig. 4 depicts the confusion matrix of our proposed MA-CMU-SGRNet on the IEMOCAP dataset, we find excited, anger emotions can be confused with the happy, frustration emotions in some cases shown in Fig.4(a). Such phenomenon is caused by the ambiguity of the emotion classification as excitement and happiness emotions are similar for human's perception and interpretation, which is consistent with previous findings [3]. Therefore, these emotions may be misclassified. Besides, we find that the sadness, disgust and anger emotions can be confused with the neutral emotion in some cases shown in Fig.4(b) due to an imbalanced class distribution, where most of the utterances are labeled as neutral.

V. DISCUSSION

A. Ablation Study

1) *Influence of Multi-level Alignment:* To investigate our multi-level representation alignment, we separately conduct ablation studies to verify the effectiveness of three types of alignments, namely instance-level alignment (IA), prototype-level alignment (PA), and latent space alignment (LSA). The results on three datasets are illustrated in Fig. 5.

It is observed that IA, PA, and LSA can both improve the classification performance, indicating that multi-level alignments facilitate the transfer learning between two modalities. By dropping the instance-level alignment, the model encounters a performance (WAF1) drop of 0.9%~2.1% on three datasets. A similar trend is also observed after discarding another two alignments. We can also see that PA and LSA are especially helpful, which requires the model to understand specific high-level semantics for samples from similar emotion categories. A decrease of 1.9% (WAF1) on the CMU-MOSEI is observed by removing PA and a decrease of 2.2% (WAF1) on the IEMOCAP by dropping LSA, demonstrating the effectiveness of the multi-level alignment. Moreover, when combining IA and LSA, we can obtain further improvement on all datasets, indicating that the benefits of IA and LSA are complementary.

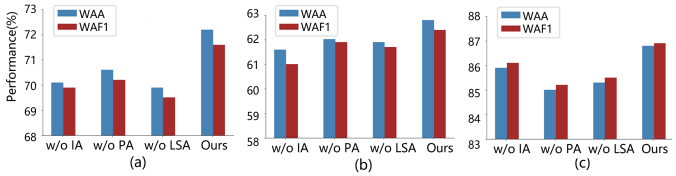


Fig. 5. Comparison of WAA and WAF1 on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI.

TABLE IV
RESULTS OF ABLATION STUDIES ON THREE DATASETS.

Methods	IEMOCAP		MELD		CMU-MOSEI	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
w/o UW	68.9±0.48†	67.6±0.61†	60.3±0.34†	59.7±0.75†	83.7±0.73†	84.1±0.52*
w/o CMI	68.1±0.61†	67.4±0.35†	60.5±0.73†	59.9±0.42*	82.8±0.54†	83.2±0.42†
w/o SyE	68.3±0.73†	66.4±0.54†	61.7±0.64*	61.4±0.52*	83.2±0.45†	83.6±0.56*
w/o SeE	68.8±0.53†	67.9±0.67†	59.5±0.32†	59.2±0.47†	82.9±0.54†	83.1±0.65†
Ours	**72.4**±0.61	**71.6**±0.73	**62.8**±0.54	**62.3**±0.62	**86.9**±0.66	**86.8**±0.92

where the symbols † and * indicate that the difference with respect to the ablation setting is statistically significant at $p < 0.001$† and $p < 0.01$* under *t*-test.

2) *Influence of Uncertainty-aware Unification:* We test our MA-CMU-SGRNet with three fusion techniques: (1) without variance-based uncertainty weights (w/o UW): it assigns equal weights to two modalities; (2) without cross-modal interaction (w/o CMI): we replace the cross-modal fusion strategy by a simple weighted summation approach. In this way, there is a lack of sufficient interaction between two modalities; (3) the proposed MA-CMU-SGRNet (Ours): our proposed method which utilizes the cross-modal uncertainty-aware fusion.

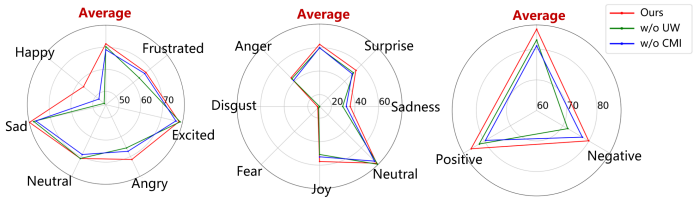


Fig. 6. Comparison of WAF1 with/without uncertainty weights and cross-modal interaction on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI, where the performance of each category and the average performance bolded in red are provided.

Experimental results in Table IV demonstrate that our fusion strategy shows an absolute improvement of 2.5%~3.5% on WAA and 2.6%~4.0% on WAF1 over the methods without uncertainty weights and an improvement of 2.3%~4.9% on WAA and 2.4%~5.4% on WAF1 compared with that without cross-modal interaction. The main reason lies in that with the uncertainty weighted cross-modal interaction fusion, we are able to determine how informative one modality is thus assign different importance. However, without UW, the feature extraction suffers from the problem of uncertainty and annotation ambiguity. People with different backgrounds might interpret differently, which has the limitation in generating a consistent

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3354382

and certain emotion label. These results verify the effectiveness of our uncertainty weighted fusion strategy. Fig. 6 clearly shows that uncertainty-aware unification plays an important role in classifying different emotions. Especially for emotions that are otherwise difficult to deduce, such as happy and angry for the IEMOCAP dataset and the sadness for the MELD dataset, the improvements caused by the cross-modal interaction and uncertainty weight are relatively significant.

3) *Influence of Semantic Graph Refinement:* To observe the effect of the graph-based semantic refinement components, we visualize the features before and after the dual-encoding semantic graph refinement on three datasets shown in Fig. 7. We easily notice a better formation of emotion clusters in the fourth column proving the necessity of capturing local and global semantic dependency in utterances.

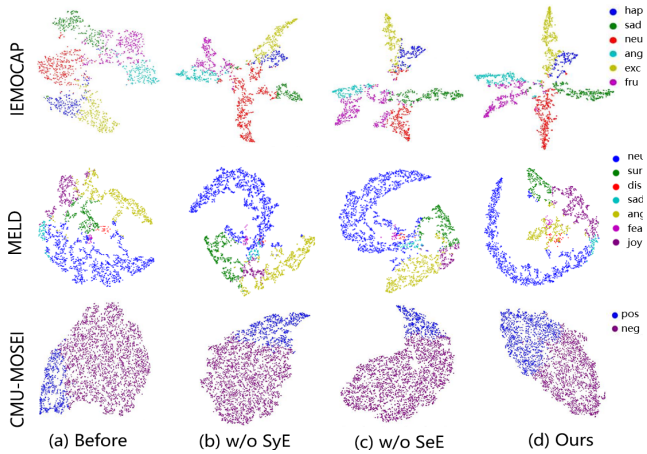


Fig. 7. The visualization of the t-SNE representations (a) Before semantic refinement (b) After semantic refinement w/o SyE (c) After semantic refinement w/o SeE (d) After semantic refinement of our proposed MA-CMU-SGRNet.

Additionally, we conduct ablation experiments on the correlation-based syntactic encoder (SyE) and semantic encoder (SeE) respectively. Specifically, we respectively remove the semantic edge weight (SEW) in the RGCN and semantic-positional encoding (SPE). After removing the SEW, both WAA and WAF1 on IEMOCAP decreased by 4.1% and 5.2%, while after removing the SPE, both WAA and WAF1 on IEMOCAP decreased by 3.6%, 3.8% respectively. From the results in Table IV, the following observations can be found: (1) our proposed method achieves better results than those without syntactic encoder (w/o SyE). The main reason is that local utterances provide more semantic information about the emotion recognition of the current utterance. (2) The impact of semantic encoder (SeE) is more significant focusing on the semantic correlation between speakers and new utterances, especially on the MELD (3.3% on WAA) and CMU-MOSEI (4.0% on WAA) dataset. (3) The combined model achieves the best results, which demonstrates that the syntactic encoder can fully focus on the local context information. Meanwhile, the semantic encoder provides a global angle to generate node embeddings, which can also provide semantic clues that is ignored by syntactic encoder.

B. Influence of Modalities

To explore the importance of each modality, we conduct experiments to compare the performance among unimodal and bimodal results. To perform single-modal experiments, we start by eliminating the bimodal alignments, including instance-level, prototype-level and latent space alignments. Subsequently, the unimodal representations obtained from the speech or text encoder are directly utilized as the node features, which are then employed to construct the semantic graph. This approach allows us to investigate the performance and characteristics of each modality independently, without the influence of cross-modal interactions. For unimodal results, results in Table V show that the lexical modality achieves better performance than that of the acoustic modality in all cases, which indicates the importance of the spoken language in conversational emotion recognition. The text tends to have lesser noisy signals compared with the audio, thus learning more effective features. This result is in line with that of previous works [3].

Furthermore, experimental results in Table V demonstrate that bimodal results outperform unimodal results in all cases. Our proposed method displays the best performance with near 1.3%~3.2% improvement in terms of both WAA and WAF1 compared with the lexical modality. The bimodal results succeed over the acoustic modality by 9.1%~21.3% on WAF1 and 7.3%~21.6% on WAA. These demonstrate the importance of integrating complementary acoustic and linguistic features.

TABLE V
COMPARISON WITH UNIMODAL SET ON THREE DATASETS.

Methods	IEMOCAP		MELD		CMU-MOSEI	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
Text only	$67.2_{\pm 0.46}^{\dagger}$	$66.1_{\pm 0.42}^{\dagger}$	$60.4_{\pm 0.57}^{\dagger}$	$59.7_{\pm 0.52}^{\dagger}$	$85.1_{\pm 0.73}^{\dagger}$	$85.5_{\pm 0.39}^{\dagger}$
Speech only	$60.6_{\pm 0.52}^{\dagger}$	$59.2_{\pm 0.76}^{\dagger}$	$55.5_{\pm 0.74}^{\dagger}$	$53.2_{\pm 0.68}^{\dagger}$	$61.3_{\pm 0.42}^{\dagger}$	$61.5_{\pm 0.48}^{\dagger}$
Ours	**72.4**$_{\pm 0.61}$	**71.6**$_{\pm 0.73}$	**62.8**$_{\pm 0.54}$	**62.3**$_{\pm 0.62}$	**86.9**$_{\pm 0.66}$	**86.8**$_{\pm 0.52}$

where the symbols † indicate that the difference with respect to the ablation setting is statistically significant at $p < 0.001$ under *t*-test.

C. Impact of Hyperparameter

The hyperparameter λ controls the trade-off between the cross-entropy loss and the multimodal alignment loss. For example, a larger λ gives greater weight to the alignment loss. To explore the impact of the λ on the emotion classification performance, we set the λ from 0.1 to 0.9 in steps of 0.1 and obtain the performance of our framework in Fig. 8. As can be seen, the proposed method yields relatively high and robust ACC and F1 concerning the $\lambda = 0.6$. An excessively high value of the λ indicates that the cross-entropy loss has little effect on

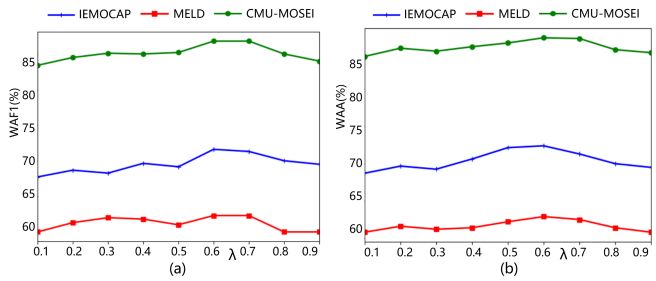


Fig. 8. Comparison of WAA and WAF1 on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI.

the model training process, which results in a large difference between the predicted labels and actual labels of training samples, i.e., underfitting model. Meanwhile, when the λ is too small, the alignment between two modalities is insufficient to counter the overfitting problem caused by the redundancy of information.

D. Limitations and Future work

In this paper, our proposed MA-CMU-SGRNet achieves optimal emotion recognition results compared with recent methods. However, there are two main deficiencies in our proposed method. First, our proposed model fails to distinguish similar emotions effectively going through the prediction results, as frustrated and anger, happy and excited (Fig. 4). Second, the proposed method tends to misclassify samples of other emotions to neutral on MELD due to the majority proportion of neutral samples in these datasets. We will address these limitations in future work by integrating a component for capturing the fine-grained emotions. Additionally, we aim to further improve the classification accuracy by using the visual information, in addition to acoustic and lexical modalities.

VI. CONCLUSION

In this paper, we propose a multi-level alignment and cross-modal unified semantic refinement network (MA-CMU-SGRNet) for the multimodal emotion recognition in conversation task. Our proposed MA-CMU-SGRNet involves three innovative modules. First, we focus on the multi-level alignment to bridge the gap between two modalities. Then, we adopt a fusion strategy that takes into account the ambiguity of emotions. Finally, a semantic graph is established and the semantic clues and context information are captured via global and local interactions. To further exploit the power of our MA-CMU-SGRNet on ERC tasks, we conduct experiments on three widely used benchmark datasets: IEMOCAP, MELD, and CMU-MOSEI. Results show that our proposed approach reaches the new state-of-the-art record for conversational emotion recognition. Additionally, experimental results on different components show the necessity of our three modules.

REFERENCES

[1] G. V. Singh, M. Firdaus, A. Ekbal, and P. Bhattacharyya, "A multimodal transformer for identifying emotions and intents in social conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 290–300, 2023.

[2] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1426–1439, 2022.

[3] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[4] C.M. Chang, G.-Y. Chao, and C.-C. Lee, "Enforcing semantic consistency for cross corpus emotion prediction using adversarial discrepancy learning in emotion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1098–1109, 2023.

[5] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," IEEE Transactions on Multimedia, vol. 25, pp. 3375–3385, 2023

[6] S. Li, H. Yan, and X. Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 002–11 010.

[7] Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou, and G. Karypis, "Anomaly detection on attributed networks via contrastive self-supervised learning," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 6, pp. 2378–2392, 2022

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 18 661–18 673.

[9] K. Liu, W. Zhu, Y. Shen, S. Liu, N. Razavian, K. J. Geras, and C. Fernandez-Granda, "Multiple instance learning via iterative self-paced supervised contrastive learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3355–3365.

[10] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, Sep. 2017, pp. 1103–1114.

[11] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1426–1439, 2022.

[12] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E.Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[13] K. Zhao, L. Ma, Y. Meng, L. Liu, J. Wang, J. M. Junior, W. N. Gonc¸alves, and J. Li, "3d vehicle detection using multi-level fusion from point clouds and images," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 146–15 154, 2022.

[14] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-gcn: Incremental graph convolution network for conversation emotion detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 4471–4481, 2022.

[15] X. Zhang and Y. Li, "A cross-modality context fusion and semantic refinement network for emotion recognition in conversation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 13 099–13 110.

[16] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Transactions on Multimedia*, pp. 1–13, 2023, doi: 10.1109/TMM.2023.3260635.

[17] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2019.

[18] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

[19] W. Wu, C. Zhang, X. Wu, and P. C. Woodland, "Estimating the uncertainty in emotion class labels with utterance-specific dirichlet priors," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2810–2822, 2023.

[20] H. Zhang and D. Song, "Towards contrastive context-aware conversational emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1879–1891, 2022.

[21] D. Jiang, R. Wei, J. Wen, G. Tu, and E. Cambria, "Automl-emo: Automatic knowledge selection using congruent effect for emotion identification in conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1845–1856, 2023.

[22] F. Wang, Z. Ding, R. Xia, Z. Li, and J. Yu, "Multimodal emotion-cause pair extraction in conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1832–1844, 2023.

[23] J. Y. Zou, D. J. Hsu, D. C. Parkes, and R. P. Adams, "Contrastive learning using spectral methods," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[24] S. Li, H. Yan, and X. Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 002–11 010.

[25] S. F. Yilmaz, E. B. Kaynak, A. Koc¸, H. Dibeklioglu, and S. S. Kozat, "Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 331–343, 2023.

[26] M. Li, C.-G. Li, and J. Guo, "Cluster-guided asymmetric contrastive learning for unsupervised person re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 3606–3617, 2022.

[27] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24 206–24 221.

[28] X. Liang, Y. Qian, Q. Guo, H. Cheng, and J. Liang, "Af: An association-based fusion method for multi-modal classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9236–9254, 2022.

[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[30] B. Wang, H. Niu, J. Zeng, G. Bai, S. Lin, and Y. Wang, "Latent representation learning model for multi-band images fusion via low-rank and sparse embedding," *IEEE Transactions on Multimedia*, vol. 23, pp. 3137–3152, 2021.

[31] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Dimensional affect uncertainty modelling for apparent personality recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2144–2155, 2022.

[32] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[33] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9074–9084.

[34] Z. Xu, C. Sun, Y. Long, B. Liu, B. Wang, M. Wang, M. Zhang, and X. Wang, "Dynamic working memory for context-aware response generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1419–1431, 2019.

[35] Z. Zhong, C. Li, and J. Pang, "Multi-grained semantics-aware graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 7251–7262, 2023.

[36] D. Guo, H. Wang, and M. Wang, "Context-aware graph inference with knowledge distillation for visual dialog," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6056–6073, 2022.

[37] W. Li, X. Liu, and Y. Yuan, "Sigma++: Improved semantic-complete graph matching for domain adaptive object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9022–9040, 2023.

[38] L. Zhang, M. Chen, A. Arnab, X. Xue, and P. H. S. Torr, "Dynamic graph message passing networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5712–5730, 2023.

[39] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8247–8255.

[40] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[41] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp. 527–536.

[42] A. B. 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2236–2246.

[43] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2019.

[44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Jul. 2017, pp. 873–883.

[45] K. El Hajal, Z. Wu, N. Scheidwasser-Clow, G. Elbanna, and M. Cernak, "Efficient speech quality assessment using self-supervised framewise embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

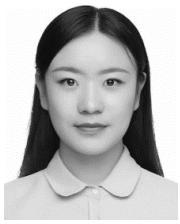
[46] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro, "Factuality enhanced language models for open-ended text generation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 34 586–34 599.

[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[48] Y. H. Tsai, S. Bai, J. Z. Kolter, L. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[49] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.

[50] Y. Yu, M. Zhao, S. Qi, F. Sun, B. Wang, W. Guo, X. Wang, L. Yang, and D. Niu, "ConKI: Contrastive knowledge injection for multimodal sentiment analysis," in *Findings of the Association for Computational Linguistics, 2023*, pp. 13 610–13 624



Xiaoheng Zhang received the bachelor's degree in information and computing science from Beihang University, Beijing, China, in 2021, where she is currently working toward the master's degree with the Department of Automation Science and Electrical Engineering. Her current research interests include affective computing and machine learning.

Weigang Cui received the bachelor's degree in mathematics and PhD degree from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2016 and 2021, respectively. He is currently doing postdoctoral research with the School of Engineering Medicine, Beihang University. His research interests include medical image analysis, machine learning, and brain functional connectivity

Bin Hu (Fellow, IEEE) received PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, China, in 1998. Since 2008, he has been a professor with the School of Information Science and Engineering, Lanzhou University, China. He had been also guest professorship in ETH Zurich, Switzerland till 2011. His research interests include pervasive computing, computational psychophysiology, and data modeling.

Yang Li received the PhD degree in automatic control and systems engineering from the University of Sheffield, Sheffield, U.K., in 2011. He did post-doctoral research with the Department of Computer and Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC, for one year. In 2013, he joined the School of Automation Sciences and Electrical Engineering, Beihang University, Beijing, China, as a professor. His current research interests include affective computing, medical image analysis, and brain–computer interface.