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

Code:https://github.com/zxiaohen/MA-CMU-SGRNet

—— IEEE Transactions on Affective Computing 2024

2024.3.17 • ChongQing

Reported by JiaWei Cheng

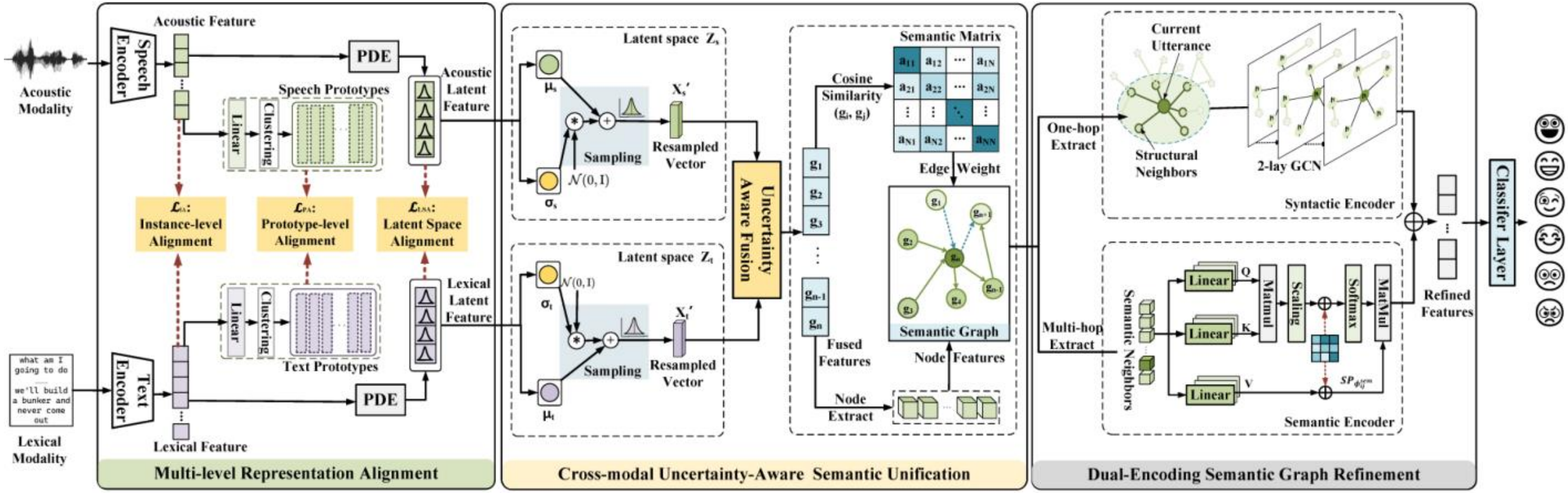
Motivation

(1) The method of supervised contrastive learning is too tough.

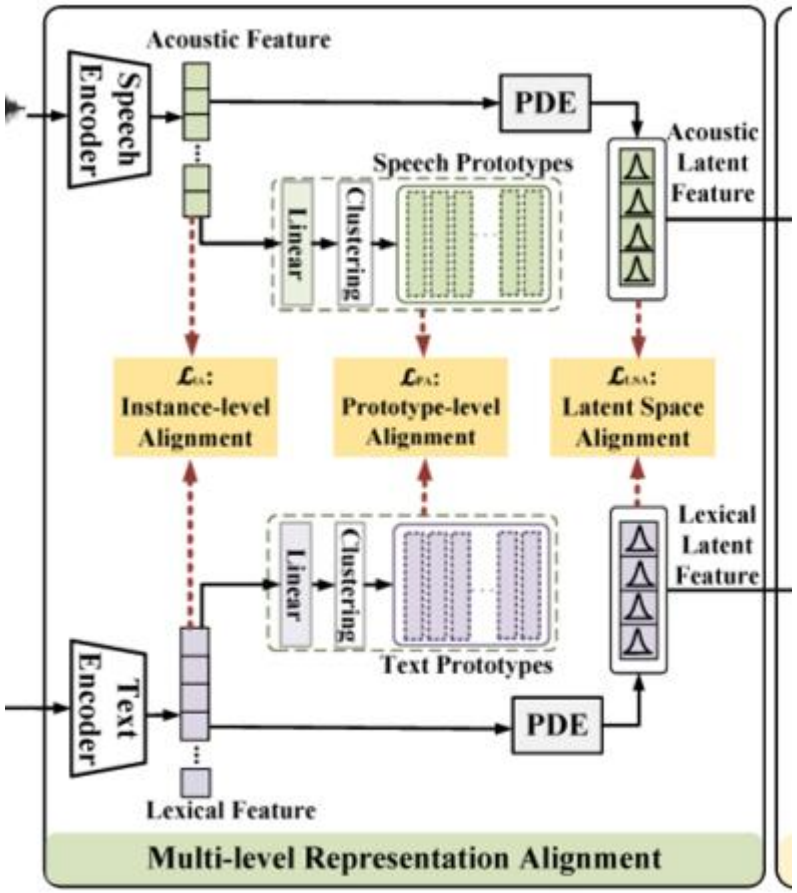
(2) Previous fusion methods did not consider the uncertainty in each modal

(3) Previous methods did not fully mine the semantic context information in the conversation

Overview


Acoustic Modality
Speech Encoder
Acoustic Feature
PDE
Acoustic Latent Feature
Speech Prototypes
Linear
Clustering
$\mathcal{L}_{IA}$: Instance-level Alignment
$\mathcal{L}_{PA}$: Prototype-level Alignment
$\mathcal{L}_{LSA}$: Latent Space Alignment
Lexical Latent Feature
Text Prototypes
what am I going to do
we'll build a bunker and never come out
Text Encoder
Lexical Feature
PDE
Lexical Modality
Multi-level Representation Alignment
Latent space Z_s
μ_s
σ_s
Sampling
$\mathcal{N}(0,1)$
Resampled Vector
X_s'
Latent space Z_t
σ_t
μ_t
$\mathcal{N}(0,1)$
Sampling
Resampled Vector
X_t'
Uncertainty Aware Fusion
Cosine Similarity (g_i, g_j)
Semantic Matrix
Edge Weight
Semantic Graph
Fused Features
Node Features
Node Extract
Cross-modal Uncertainty-Aware Semantic Unification
One-hop Extract
Current Utterance
Structural Neighbors
2-lay GCN
Syntactic Encoder
Multi-hop Extract
Semantic Neighbors
Linear
Q
K
V
Matmul
Scaling
Softmax
MatMul
Semantic Encoder
Refined Features
Classifer Layer
Dual-Encoding Semantic Graph Refinement

Method



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

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

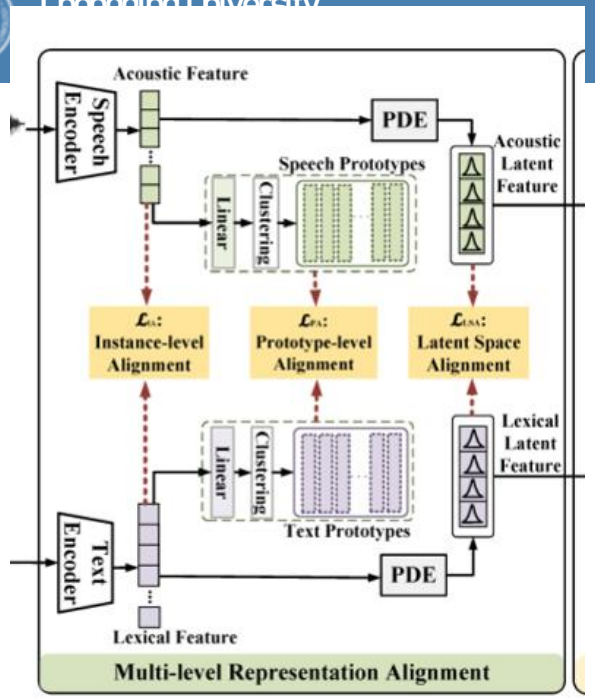
$$S_{i,j} = exp\left(\frac{1}{\tau_{m(i,j)}} \cdot \frac{z_i^T z_j}{\|z_i\|\|z_j\|} - \xi_{m(i,j)}\right), z \in \{\tilde{s}, \tilde{t}\} \quad (3)$$

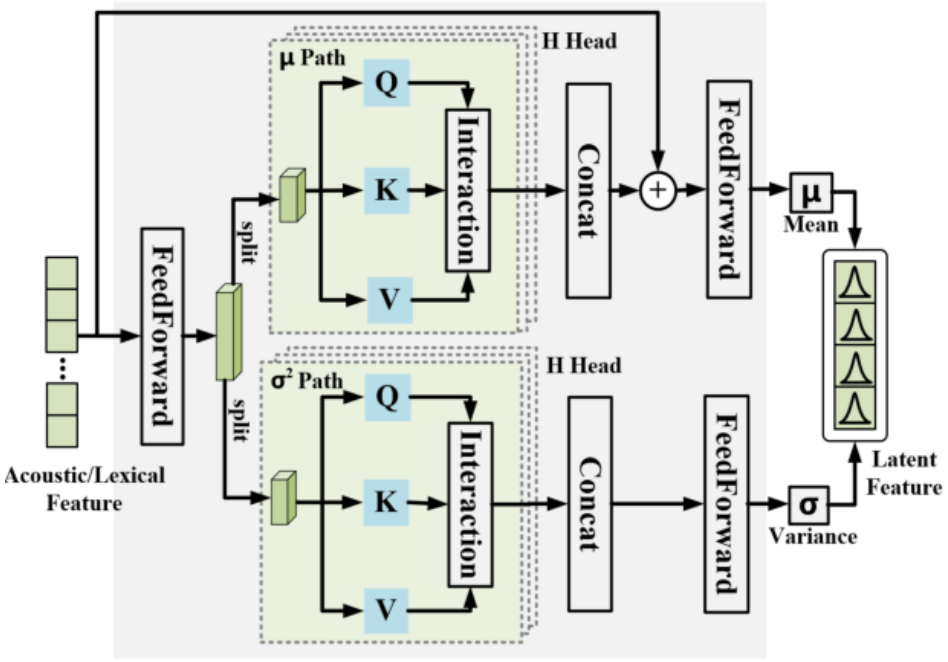
$$l_i^{s2t} = -\frac{1}{|P(i)|}\sum_{j\in P(i)} \log \frac{S_{i,j}}{S_{i,j}+\sum_{n\in N(i)} S_{i,n}} \quad (4)$$

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

$$j \in P(i) = \{j | j \in \mathcal{B}, y_j = y_i, S_{i,j} > 1\}$$

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





Method

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

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

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

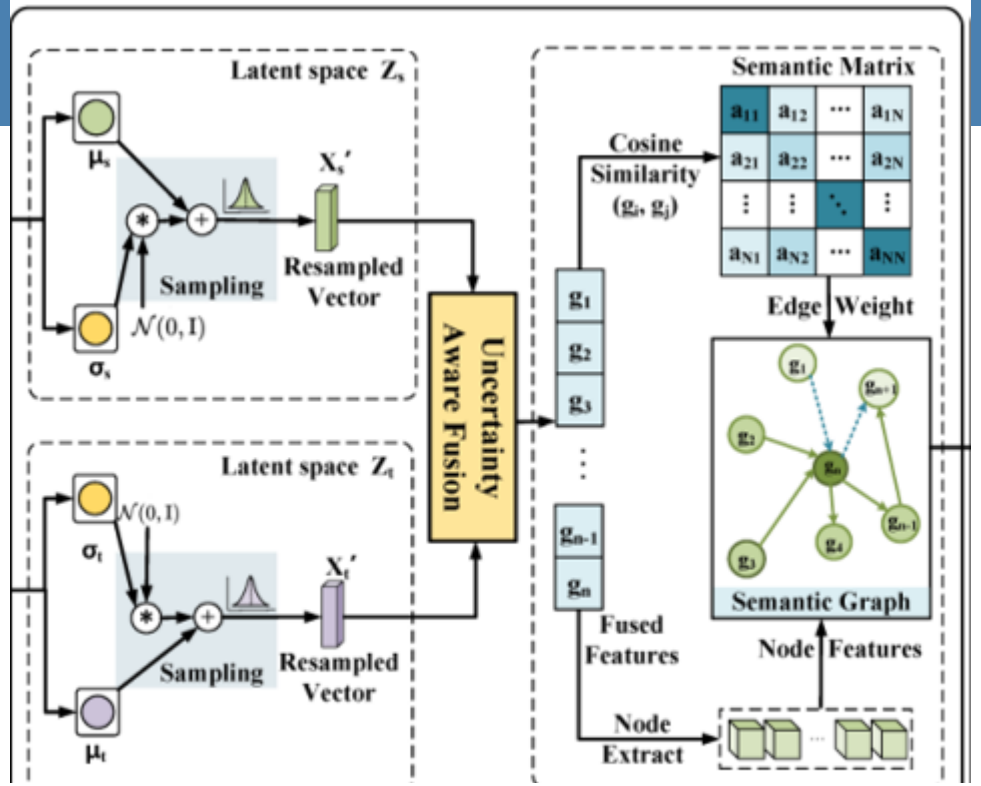
$$l(\tilde{t}_i, q_{s,j}) = \sum_{k=1}^{K} q_{s,i}^{(k)}\, log\, p_{t,i}^{(k)} \quad (10)$$

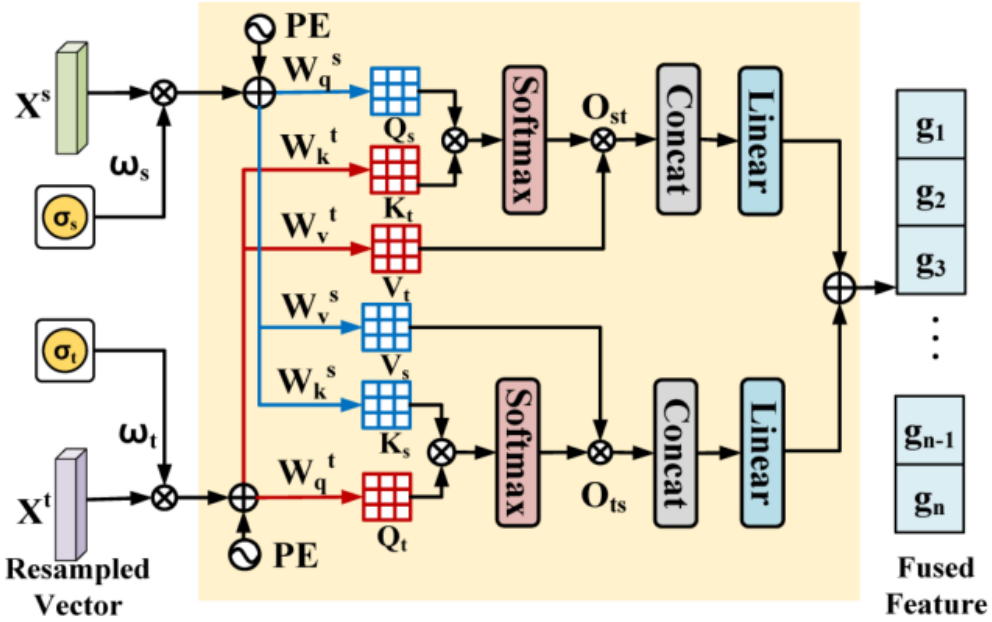
$$\mathcal{L}_{PA} = \frac{1}{2N_p} \sum_{i=1}^{N_p} (l(\tilde{s}_i, q_{t,i}) + l(\tilde{t}_i, q_{s,i})) \quad (11)$$

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

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

Method





$$\omega_s = \frac{\|\sigma_s\|}{\|\sigma_s\|+\|\sigma_t\|} \tag{14}$$

$$\omega_t = \frac{\|\sigma_t\|}{\|\sigma_s\|+\|\sigma_t\|} \tag{15}$$

$$\tilde{X}^s = \omega_s \odot X^s \tag{16}$$

$$\tilde{X}^t = \omega_t \odot X^t \tag{17}$$

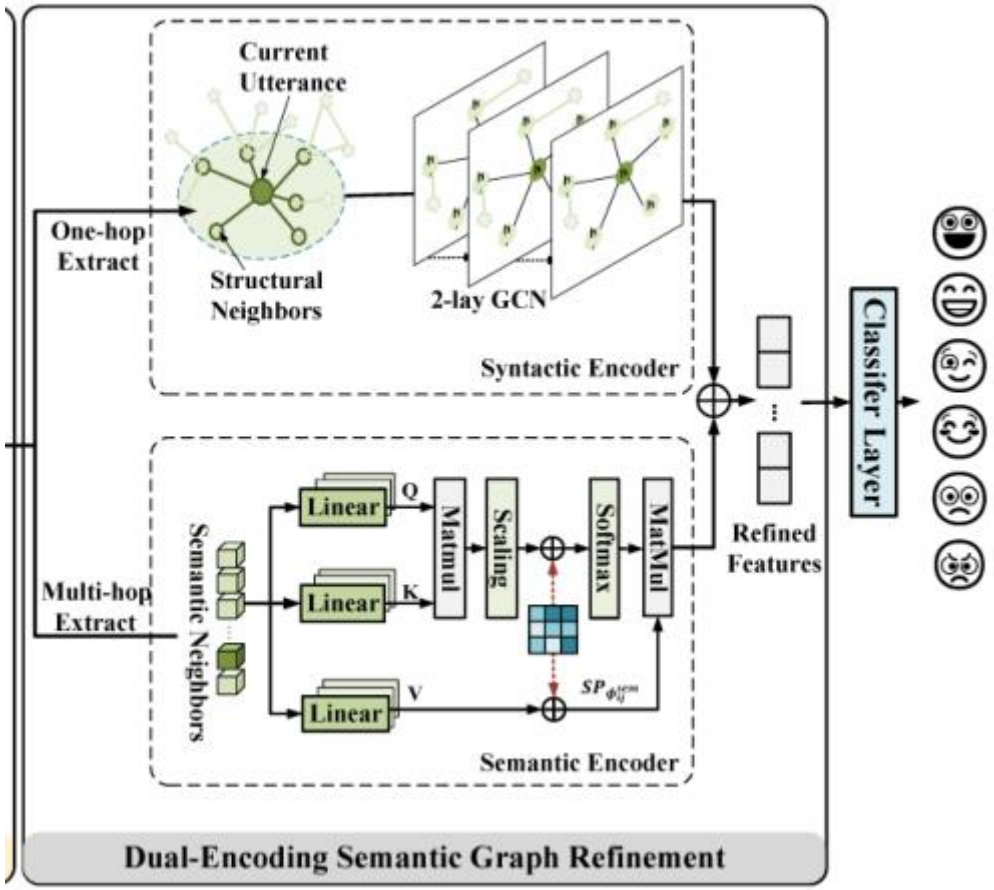
$$Q_s = \tilde{X}^s W_q^s \tag{18}$$

$$K_t = \tilde{X}^t W_k^t \tag{19}$$

$$V_t = \tilde{X}^t W_v^t \tag{20}$$

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

Method



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

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

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

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

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

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

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

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

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

$$\mathcal{L}_{total} = \mathcal{L}_{cl} + \lambda\, \mathcal{L}_A \quad (31)$$

Experiments

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.

Experiments

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**

Experiments

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**

Experiments

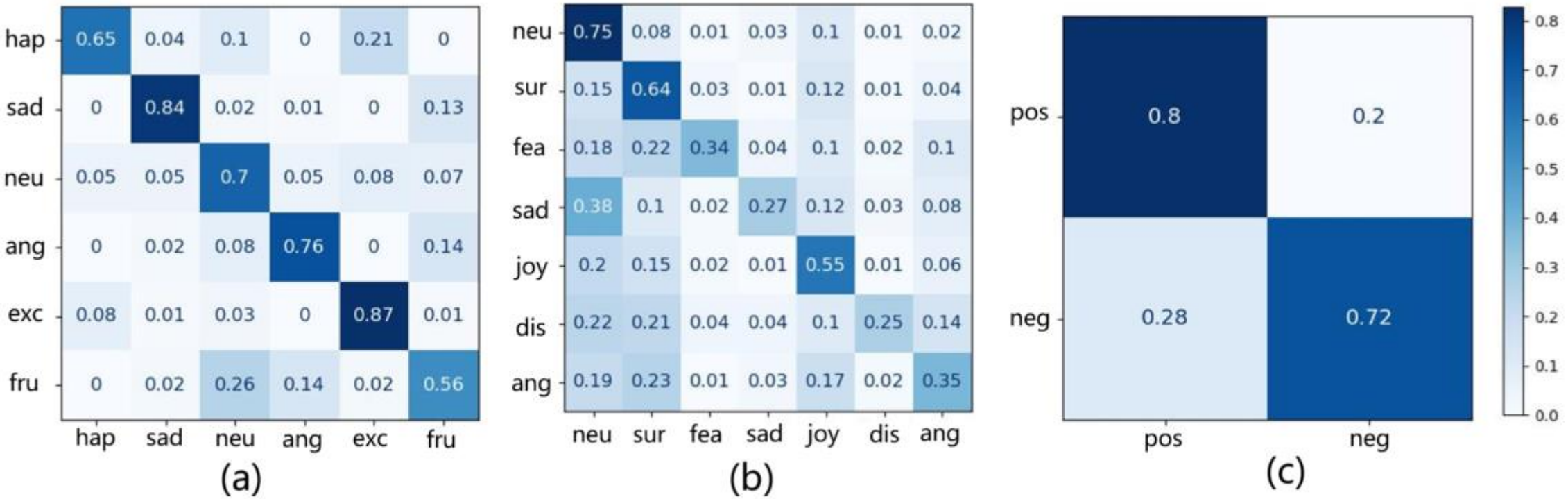


Fig. 4. The confusion matrices of the proposed MA-CMU-SGRNet on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI

Experiments

TABLE IV

RESULTS OF ABLATION STUDIES ON THREE DATASETS.

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

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

Experiments

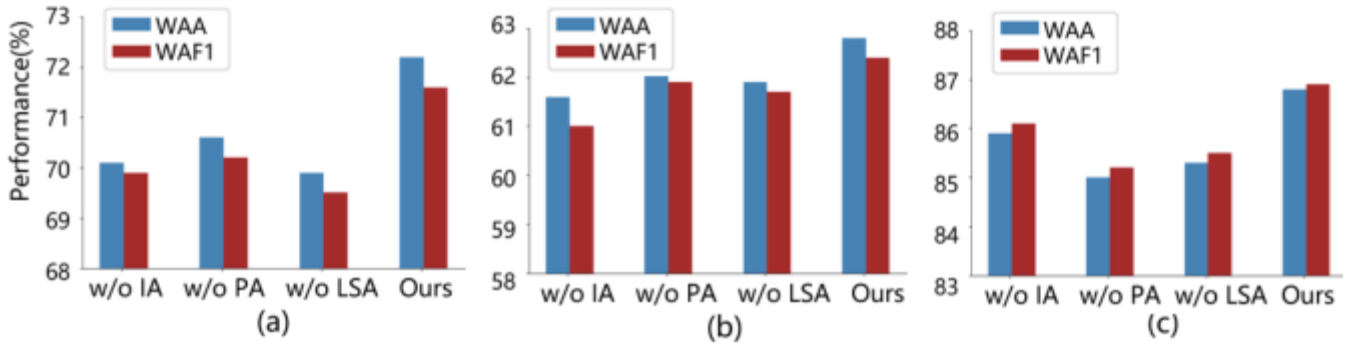


Fig. 5. Comparison of WAA and WAF1 on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI.

Experiments

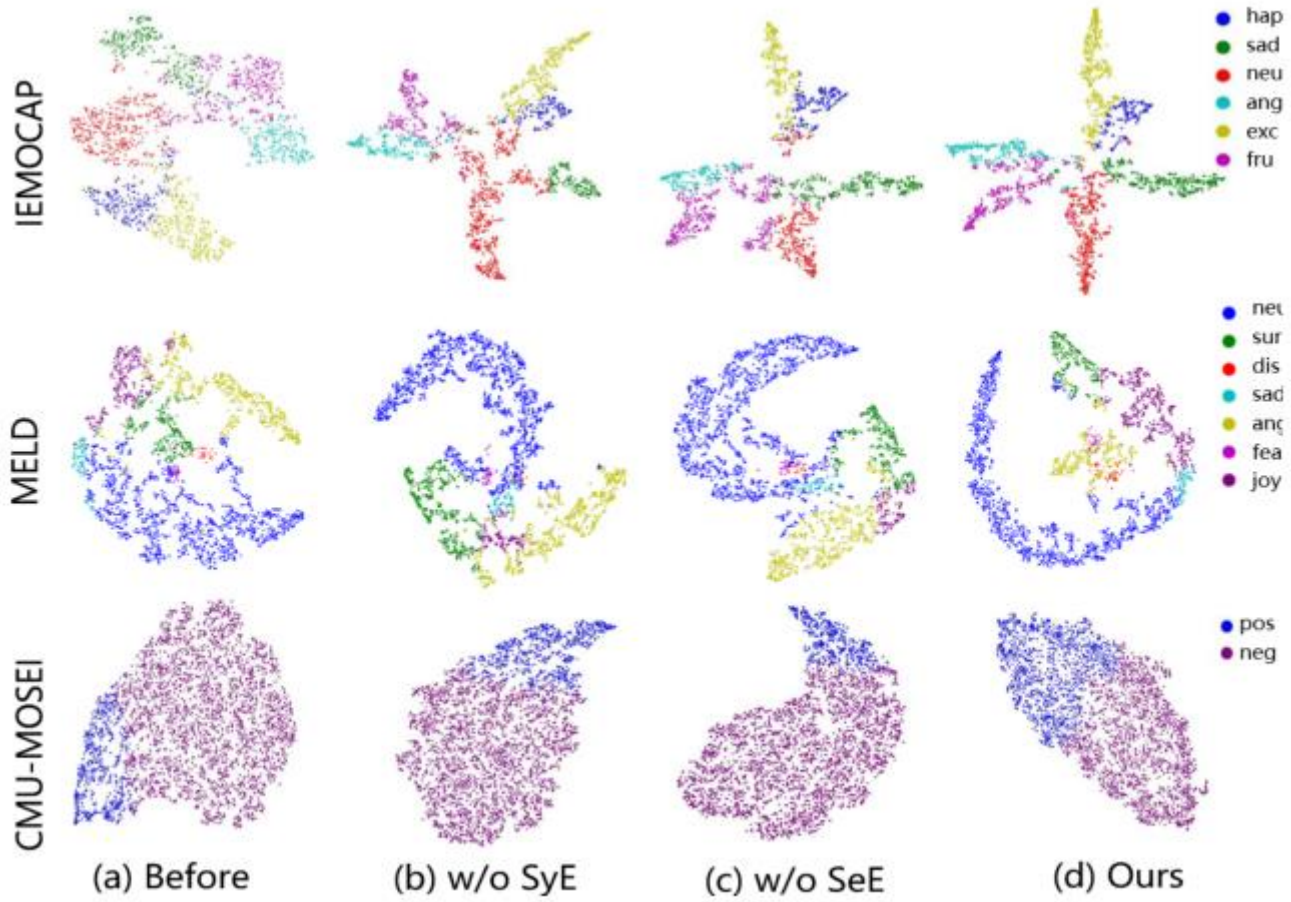


Fig. 7. The visualization of the t-SNE representations

Experiments

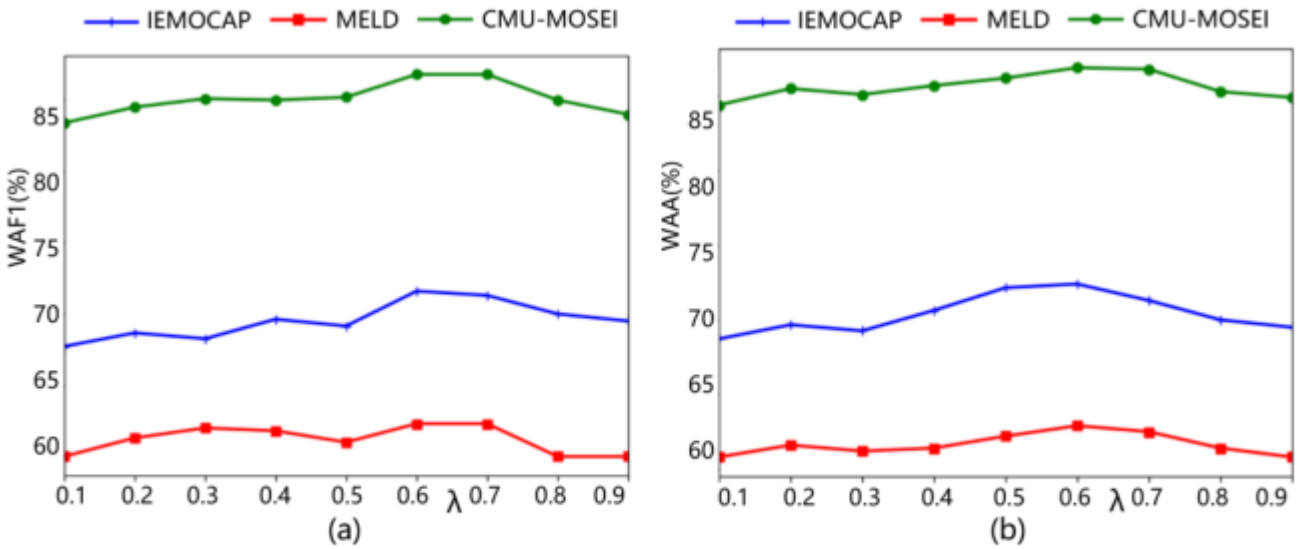


Fig. 8. Comparison of WAA and WAF1 on three datasets. (a) IEMOCAP (b) MELD (c) CMU-MOSEI.

Thanks!