A Persona-Infused Cross-Task Graph Network for Multimodal Emotion Recognition with Emotion Shift Detection in Conversations

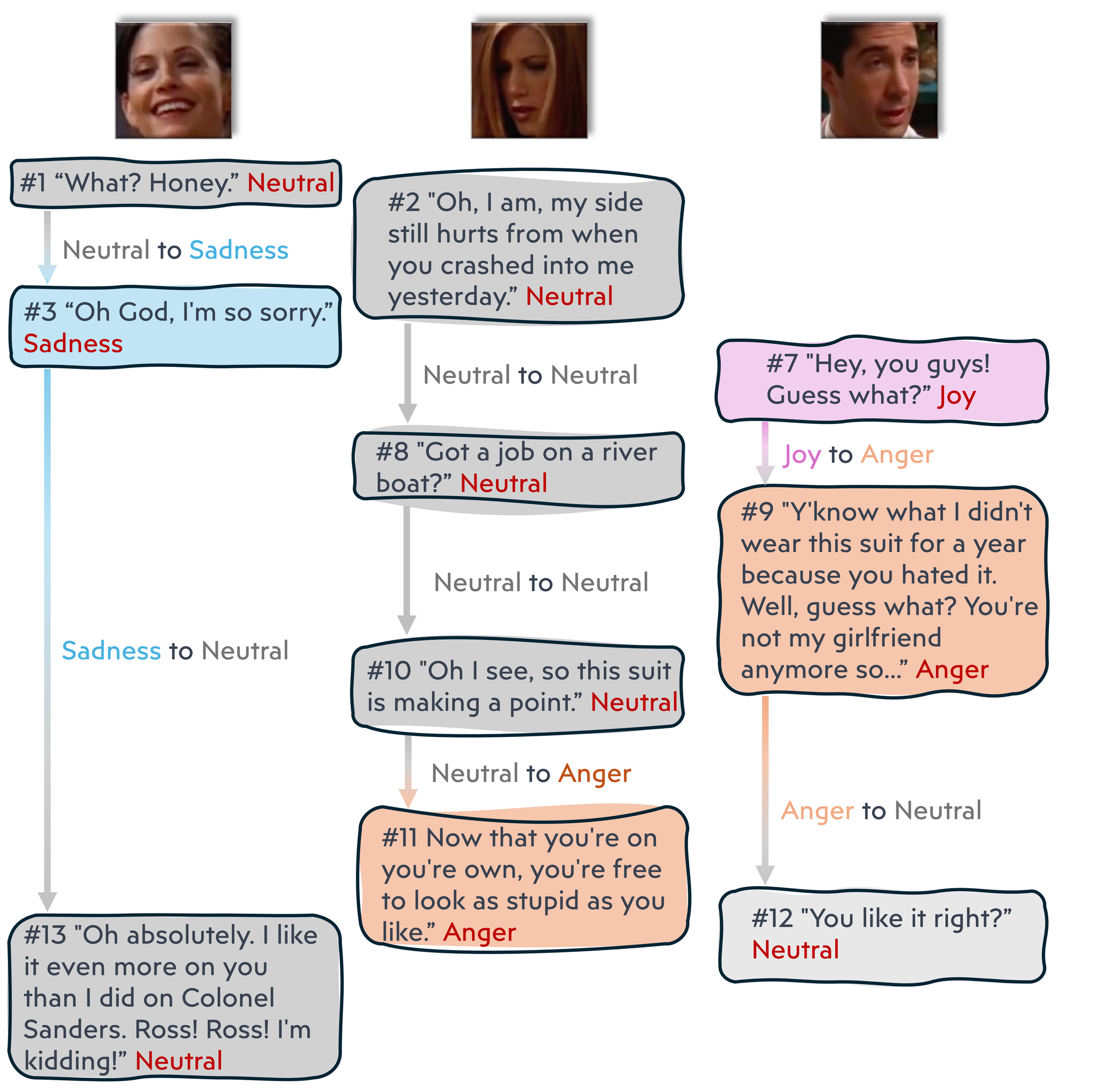
Geng Tu*, Feng Xiong*, Bin Liang, Ruifeng Xu†

23s151006@stu.hit.edu.cn Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen

1. Motivation

- Traditional MERC methods emphasize how speaker personalities affect emotional perception but typically neglect the **speaker-addressee interaction patterns**, hindering the capture of nuanced emotional exchanges.
- Additionally, the concept of **Emotion Shift**, where a speaker's emotions vary across consecutive statements, has been insufficiently explored. Previous research often includes these shifts in models without fully considering their relationship with the conversational context.

2. Motivating Example

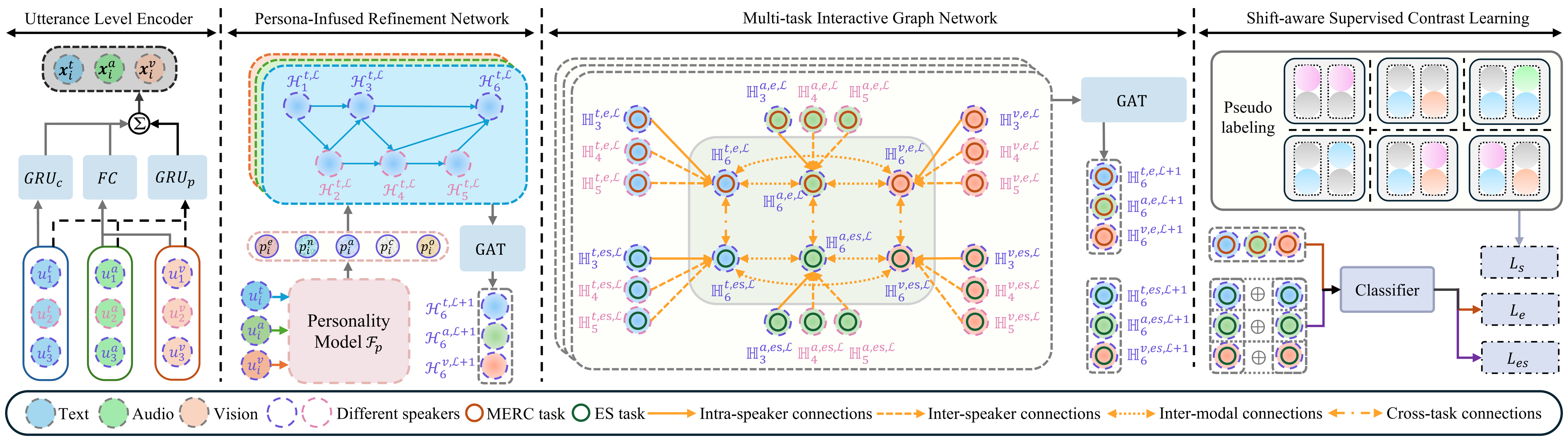


As Rachel's friend, Monica is deeply concerned about Rachel's suffering, experiencing an emotional shift from neutral to sadness upon receiving Rachel's response. Rachel, in turn, offers timely consolation to Monica. Ross, Rachel's ex-boyfriend, despite speaking with a joyful demeanor, is met with Rachel's sarcasm, leading to an emotional shift from joy to anger in Ross.

3. Framework

PCGNet consists of **three** key component: Persona-Infused Refinement Network, Multi-task Interactive Graph Network, and Shift-aware Contrastive Learning.

1. Persona-Infused Refinement Network: We introduce a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ to incorporate speaker-addressee interaction patterns, where $\mathcal{V}$ represents utterance nodes, $\mathcal{E}$ represents edges between nodes, and $\mathcal{P}$ represents personality traits including Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness. We utilize customized GAT to aggregate information, wherein the computation of attention coefficients is articulated as: $\alpha_{i,j}^{\xi,\mathcal{L}-1} = \mathrm{Softmax}\left(\mathrm{ReLU}\left(\mathbf{a}_p^T\left(\mathbf{W}_p^{\xi,\mathcal{L}-1}\mathbf{p}_i \parallel \mathbf{W}_p^{\xi,\mathcal{L}-1}\mathbf{p}_j\right)\right)\right)$, where $\mathbf{p}_i$ is the personality traits.
2. Multi-task Interactive Graph Network: We propose a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ incorporating sub-graphs for MERC and ES detection. Each utterance is represented by nodes for acoustic, visual, and textual modalities, with connections within and across modalities, speakers, and tasks. Intra-modal connections link nodes within the same modality and speaker context, inter-modal connections link nodes across different modalities, and cross-task connections link nodes between different tasks. Node representations are updated through GAT.
3. Shift-aware Contrastive Learning: To enhance model discrimination of emotional shifts (ES), we extract utterance representations from a batch and concatenate representations of consecutive utterances by the same speaker. We implement pair-aware supervised contrastive learning, forming pairs for analysis using a weighted loss function that maximizes similarity for pairs with matching pseudo-labels, which are constructed based on the emotional labels.



4. Main Results

Methods	IEMOCAP							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1
DialogueRNN$^\sharp$	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89
DialogueGCN$^\sharp$	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89
CTNet$^\natural$	51.30	79.90	65.80	67.20	**78.70**	58.80	68.00	67.50
MMGCN$^\sharp$	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26
MMDFN$^\sharp$	42.22	78.98	66.42	69.77	75.56	66.33	68.21	
SCMM$^\natural$	45.37	78.76	63.54	66.05	76.70	66.18	-	67.53
CMCF-SRNet$^\natural$	**52.20**	80.90	68.80	**70.30**	76.70	61.60	70.50	69.60
PCGNet(Ours)	49.83	**82.70**	**71.62**	69.14	76.08	**70.98**	**71.72**	**71.77**

Methods	MELD								
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	W-F1
DialogueRNN$^\sharp$	76.97	47.69	-	20.41	50.92	-	45.52	60.31	57.66
DialogueGCN$^\sharp$	75.97	46.05	-	19.6	51.2	-	40.83	58.62	56.36
CTNet$^\natural$	77.40	52.70	10.0	32.50	56.00	11.2	44.60	62.00	60.50
MMGCN$^\sharp$	76.33	48.15	-	26.74	53.02	-	46.09	60.42	58.31
MMDFN$^\sharp$	77.76	50.69	-	22.93	54.78	-	47.82	62.49	59.46
SCMM$^\natural$	-	-	-	-	-	-	-	-	59.44
CMCF-SRNet$^\natural$	-	-	-	-	-	-	-		62.30
PCGNet(Ours)	**80.25**	**61.02**	**25.88**	**41.48**	**64.65**	**25.24**	**56.09**	**67.85**	**67.02**

5. Ablation Studies

Methods	IEMOCAP				MELD			
	E-Acc	E-F1	ES-Acc	ES-F1	E-Acc	E-F1	ES-Acc	ES-F1
Ours	71.72	71.77	58.04	52.60	67.85	67.02	46.41	44.23
w/o Persona-Infused	70.86	70.73	57.78	52.04	66.78	65.99	45.33	43.11
w/o Shift-aware CL	70.92	70.96	56.76	51.27	67.05	66.13	45.82	42.91
w/o Cross-task edges	70.24	70.11	57.27	51.36	66.36	65.70	45.39	43.19

6. Results Across Modalities

Patterns	IEMOCAP	MELD
A	45.49	40.01
V	39.84	31.67
T	67.85	65.15
A + V	59.01	43.61
A + T	70.25	66.03
V + T	68.86	65.59
A + V + T	71.77	67.02

7. Conclusions

PCGNet initially models interactive relationships via a persona-infused network, then tackles ES detection and MERC through a Multi-task Interactive Graph Network, utilizing cross-task connections for correlation. It also incorporates Shift-aware Contrastive Learning to identify shift patterns effectively. Experimental results highlight PCGNet's superior performance in enhancing conversational dynamics understanding and modeling.