Decoupled Multimodal Distilling for Emotion Recognition

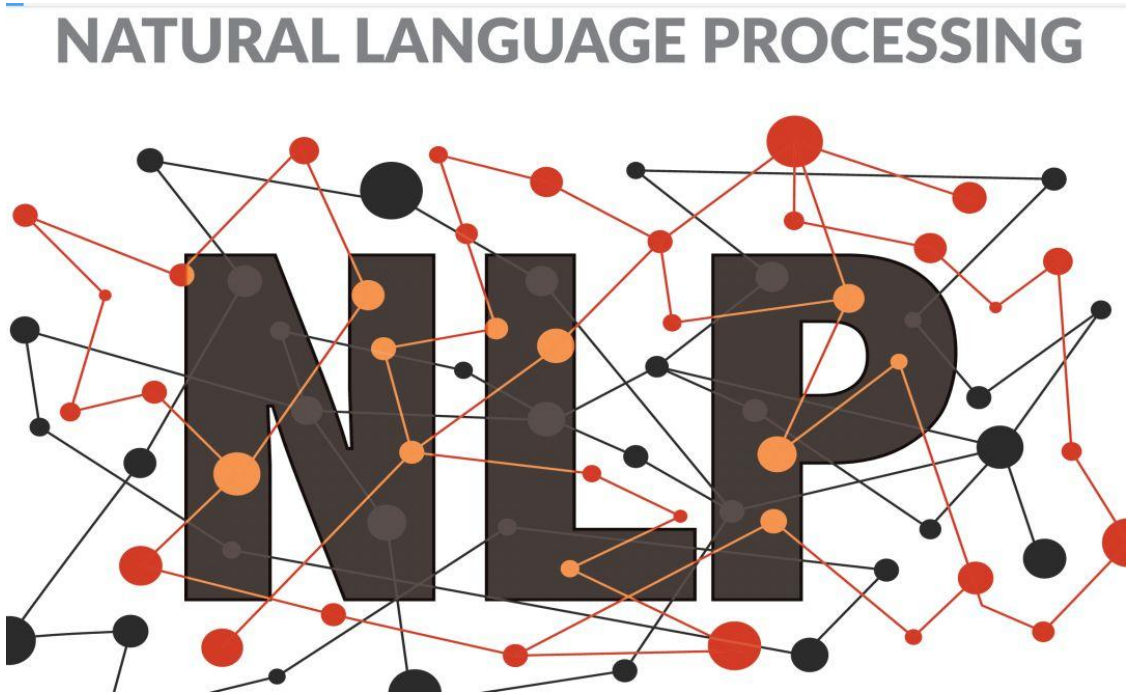
Yong Li, Yuanzhi Wang, Zhen Cui*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China.
{yong.li, yuanzhiwang, zhen.cui}@njust.edu.cn

https://github.com/mdswyz/DMD..

2023.11.11 • ChongQing

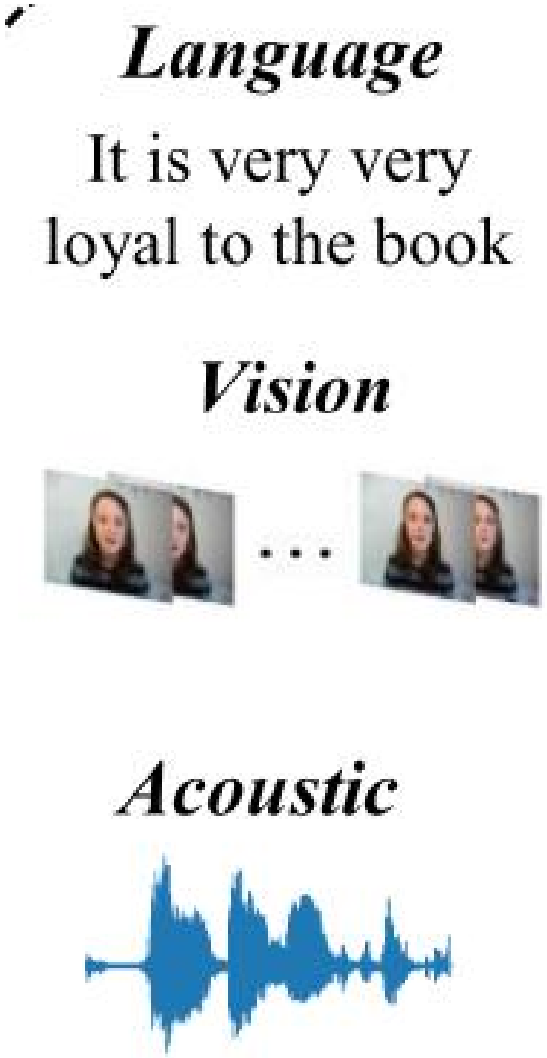
2023_CVPR

Reported by Jinyuan Zhang



1.Introduction
2.Motivation
3.Method
4.Experiments

Introduction



- Human multimodal emotion recognition (MER)

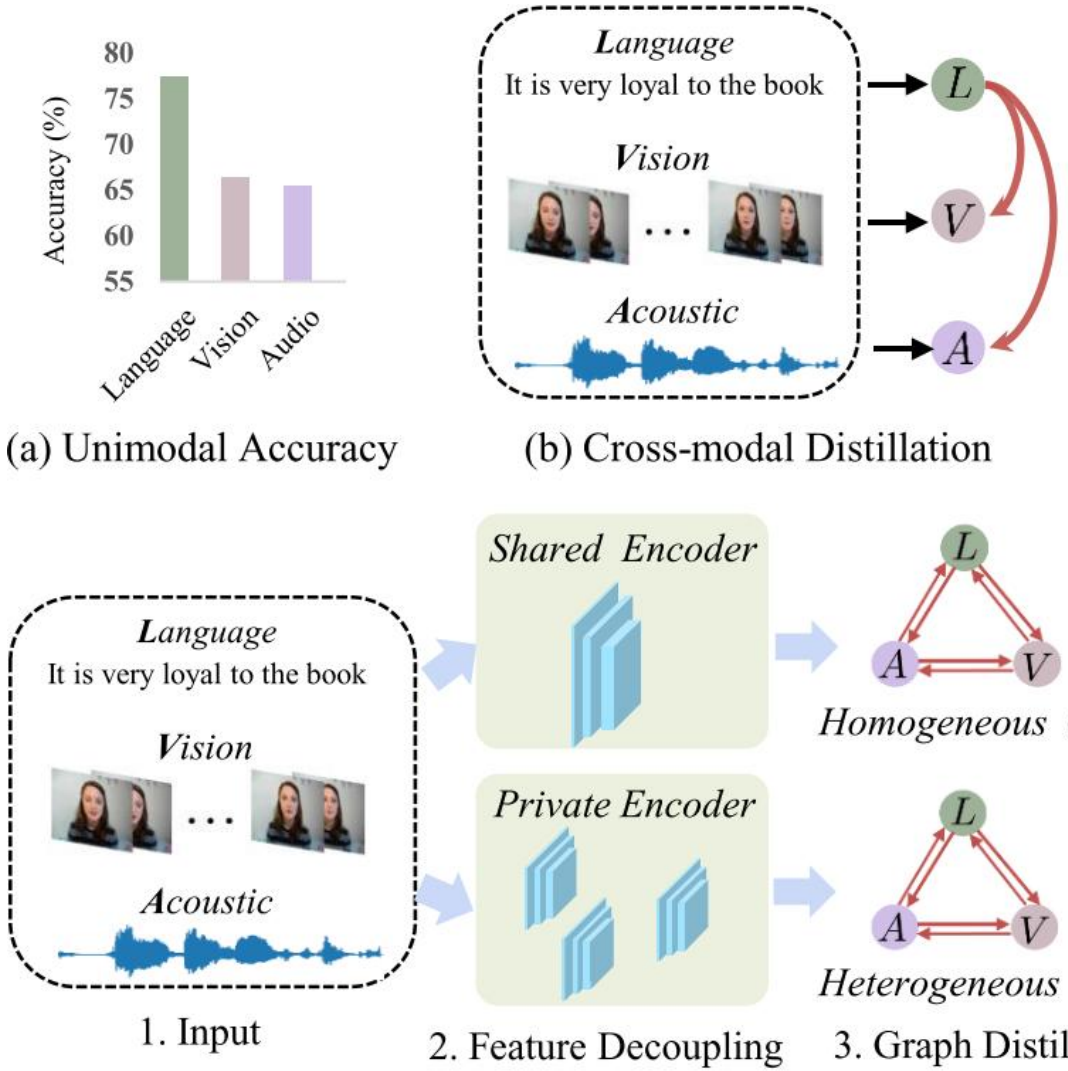
The video flows involve time-series data from various modalities, e.g., language, acoustic, and vision.

Try to perceive the sentiment attitude of humans from video clips.

For MER, different modalities in the same video segment are often complementary to each other, providing extra cues for semantic and emotional disambiguation.

The core part of MER ——the learning and fusion of multimodal representation to understand the emotion behind the raw data.

Motivation



· Limitations

the intrinsic heterogeneities among different modalities increase the difficulty of robust multimodal representation learning. Different modalities contain different ways of conveying semantic information. As(a).

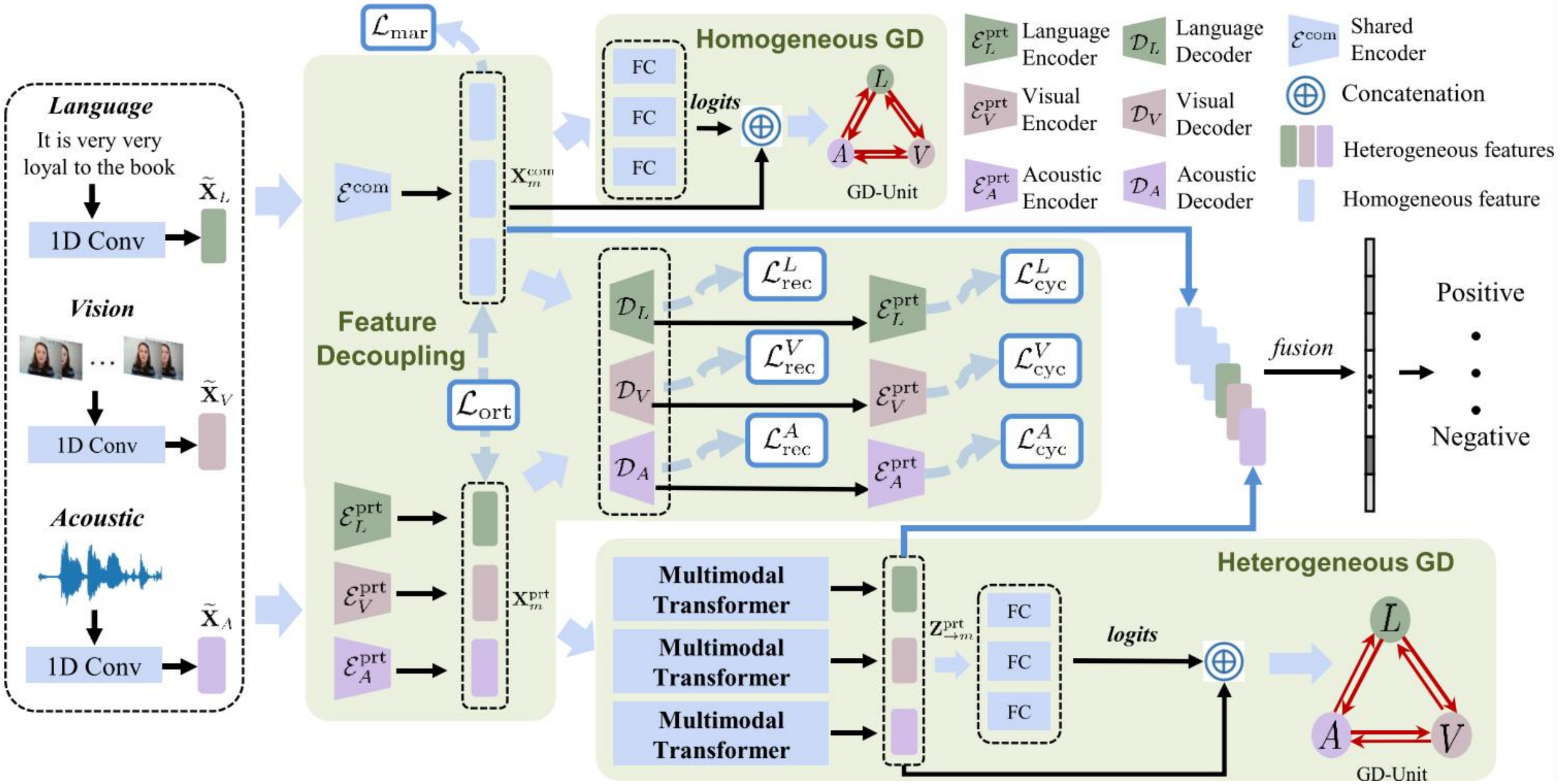
· The state-of-the-art

Before , we consider distill the reliable and generalizable knowledge from the strong modality to the weak modality. AS(b)

However, the model should learn to automatically adapt the distillation according to different examples, e.g. many emotions are easier to recognize via language while some are easier by vision

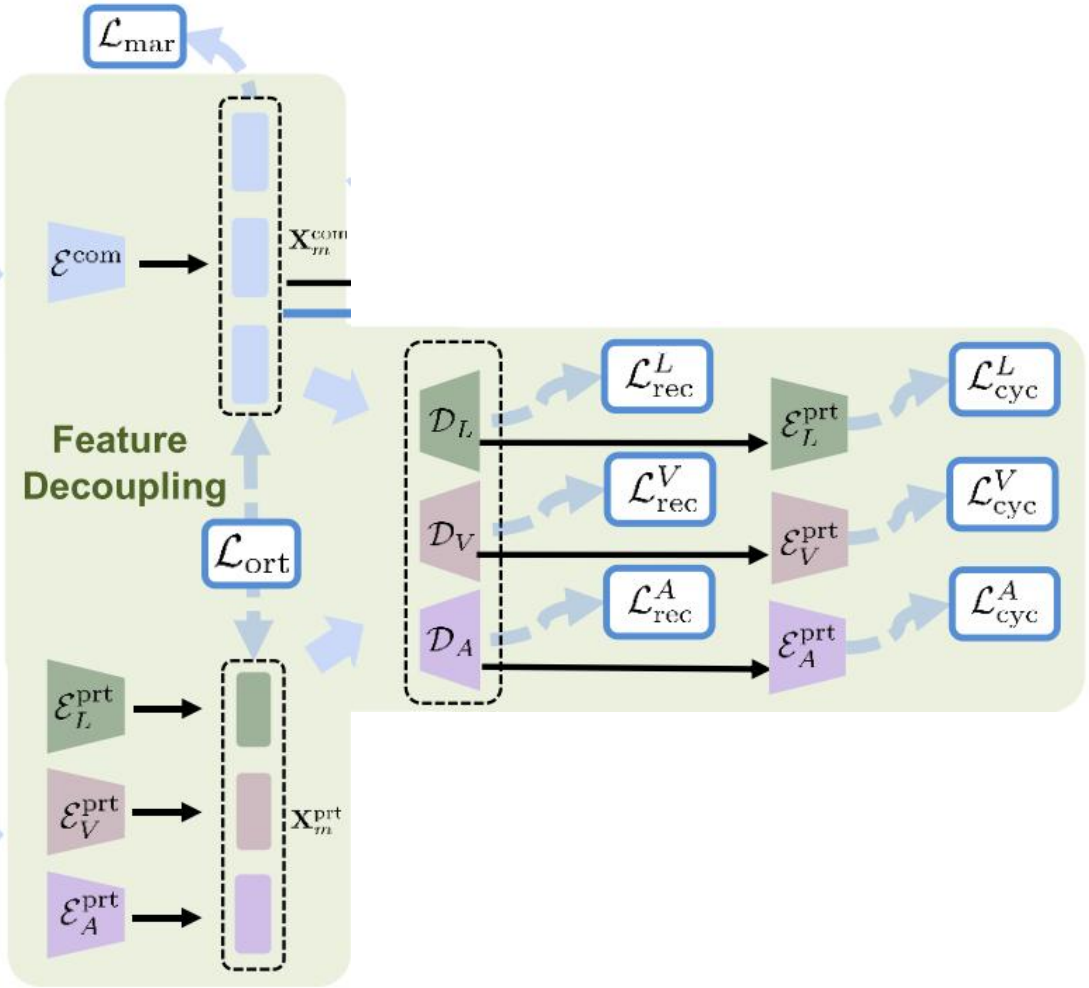
Decoupled multimodal distillation (DMD)

Method



decoupled multimodal distillation (DMD)

Method



$$\mathbf{X}_m^{\text{com}} = \mathcal{E}^{\text{com}}(\widetilde{\mathbf{X}}_m), \mathbf{X}_m^{\text{prt}} = \mathcal{E}_m^{\text{prt}}(\widetilde{\mathbf{X}}_m). \quad (1)$$

$$\mathcal{L}_{\text{rec}} = \|\widetilde{\mathbf{X}}_m - \mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}])\|_F^2. \quad (2)$$

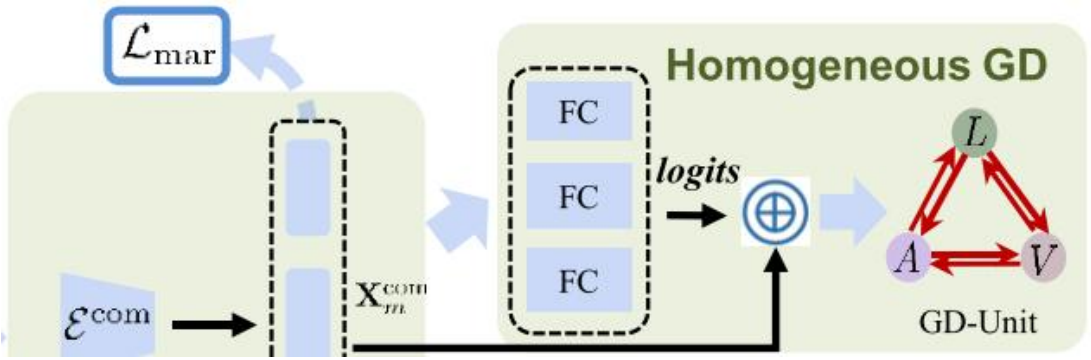
$$\mathcal{L}_{\text{cyc}} = \|\mathbf{X}_m^{\text{prt}} - \mathcal{E}_m^{\text{prt}}(\mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}]))\|_F^2. \quad (3)$$

$$\mathcal{L}_{\text{mar}} = \frac{1}{|S|}\sum_{(i,j,k)\in S} \max(0, \alpha - \cos(\mathbf{X}_{m[i]}^{\text{com}}, \mathbf{X}_{m[j]}^{\text{com}}) + \cos(\mathbf{X}_{m[i]}^{\text{com}}, \mathbf{X}_{m[k]}^{\text{com}})), \quad (4)$$

$$\mathcal{L}_{\text{ort}} = \sum_{m\in\{L,V,A\}} \cos(\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}). \quad (5)$$

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}}), \quad (6)$$

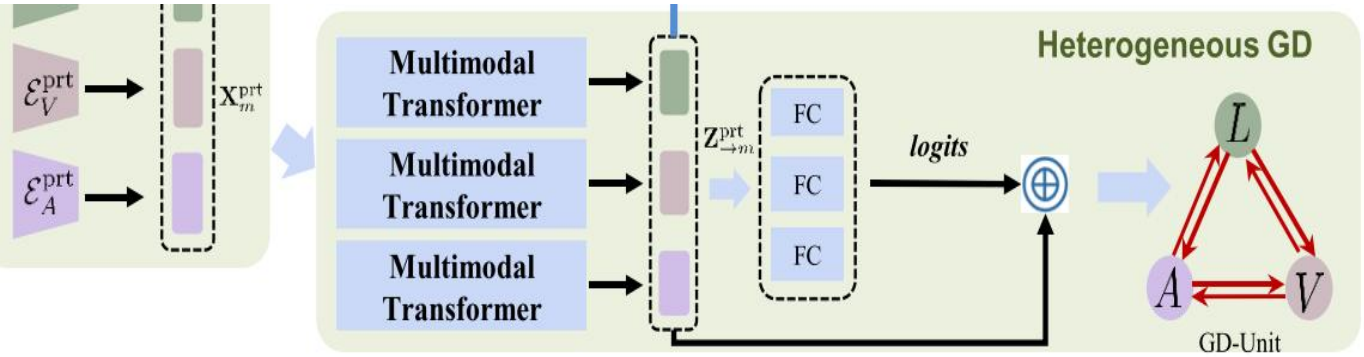
Method



$$\zeta_{:j} = \sum_{v_i \in \mathcal{N}(v_j)} w_{i \to j} \times \epsilon_{i \to j}, \tag{7}$$

$$w_{i \to j} = g([[f(\mathbf{X}_i, \theta_1), \mathbf{X}_i], [f(\mathbf{X}_j, \theta_1), \mathbf{X}_j]], \theta_2), \tag{8}$$

$$\mathcal{L}_{\text{dtl}} = \|\mathbf{W} \odot \mathbf{E}\|_1, \tag{9}$$



$$\mathbf{Z}_{L \to V}^{\text{prt}} = \text{softmax}(\frac{\mathbf{Q}_V \mathbf{K}_L^\top}{\sqrt{d}})\mathbf{V}_L, \tag{10}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{dec}} + \lambda_2 \mathcal{L}_{\text{dtl}}, \tag{11}$$

Table 1. Comparison on CMU-MOSI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	33.7	75.3	75.2
LF-LSTM		35.3	76.8	76.7
TFN [33]		32.1	73.9	73.4
LMF [14]		32.8	76.4	75.7
MFM [29]		36.2	78.1	78.1
RAVEN [30]		33.2	78.0	76.6
MCTN [26]		35.6	79.3	79.1
MulT [28]		40.0	83.0	82.8
PMR [17]		40.6	83.6	83.4
DMD (**Ours**)		**41.4**	**84.5**	**84.4**
MISA [7]*	Aligned	42.3	83.4	83.6
FDMER [32]*		44.1	84.6	84.7
DMD (**Ours**)*		**45.6**	**86.0**	**86.0**
EF-LSTM	Unaligned	31.0	73.6	74.5
LF-LSTM		33.7	77.6	77.8
RAVEN [30]		31.7	72.7	73.1
MCTN [26]		32.7	75.9	76.4
MulT [28]		39.1	81.1	81.0
PMR [17]		40.6	82.4	82.1
MICA [13]		40.8	82.6	82.7
DMD (**Ours**)		**41.9**	**83.5**	**83.5**

* means the input language features are BERT-based.

Table 2. Comparison on CMU-MOSEI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	47.4	78.2	77.9
LF-LSTM		48.8	80.6	80.6
Graph-MFN [36]		45.0	76.9	77.0
RAVEN [30]		50.0	79.1	79.5
MCTN [26]		49.6	79.8	80.6
MulT [28]		51.8	82.5	82.3
PMR [17]		52.5	83.3	82.6
DMD (**Ours**)		**53.7**	**85.0**	**84.9**
MISA [7]*	Aligned	52.2	85.5	85.3
FDMER [32]*		54.1	86.1	85.8
DMD (**Ours**)*		**54.5**	**86.6**	**86.6**
EF-LSTM	Unaligned	46.3	76.1	75.9
LF-LSTM		48.8	77.5	78.2
RAVEN [30]		45.5	75.4	75.7
MCTN [26]		48.2	79.3	79.7
MulT [28]		50.7	81.6	81.6
PMR [17]		51.8	83.1	82.8
MICA [13]		52.4	83.7	83.3
DMD (**Ours**)		**54.6**	**84.8**	**84.7**

* means the input language features are BERT-based.

Experiment

Table 3. Ablation study of the key components in DMD.

Dataset	FD	HomoGD	CA	HeteroGD	ACC_7	F1
MOSI	✓	✓	✓	✓	**41.9**	**83.5**
	✓	✓	✓	×	38.8	81.1
	✓	✓	×	✓	37.5	80.6
	✓	✓	×	×	37.2	80.8
	✓	×	×	×	34.7	79.3
	×	×	×	×	32.4	79.0
MOSEI	✓	✓	✓	✓	**54.6**	**84.7**
	✓	✓	✓	×	53.2	84.1
	✓	✓	×	✓	52.4	83.8
	✓	✓	×	×	52.4	84.3
	✓	×	×	×	51.6	82.8
	×	×	×	×	50.0	81.9

Table 4. Unimodal accuracy comparison on MOSEI dataset.

Methods	w/o FD	w/ FD
	Acc_2 (%) / F1 (%)	Acc_2 (%) / F1 (%)
L only	81.2 / 81.4	**82.7 / 82.5**
V only	58.2 / 52.2	**62.8 / 60.0**
A only	53.4 / 54.0	**64.9 / 62.5**
Mean	64.3 / 62.5	**70.1 / 68.3**
STD	12.1 / 13.4	**8.9 / 10.1**

Table 5. Ablation study of graph distillation (GD) on MulT.

Methods	CMU-MOSI			CMU-MOSEI		
	ACC_7	ACC_2	F1	ACC_7	ACC_2	F1
MulT	39.1	81.1	81.0	50.7	81.6	81.6
MulT (*w/ GD*)	39.4	82.2	82.2	51.0	82.3	82.5
DMD (**Ours**)	**41.9**	**83.5**	**83.5**	**54.6**	**84.8**	**84.7**

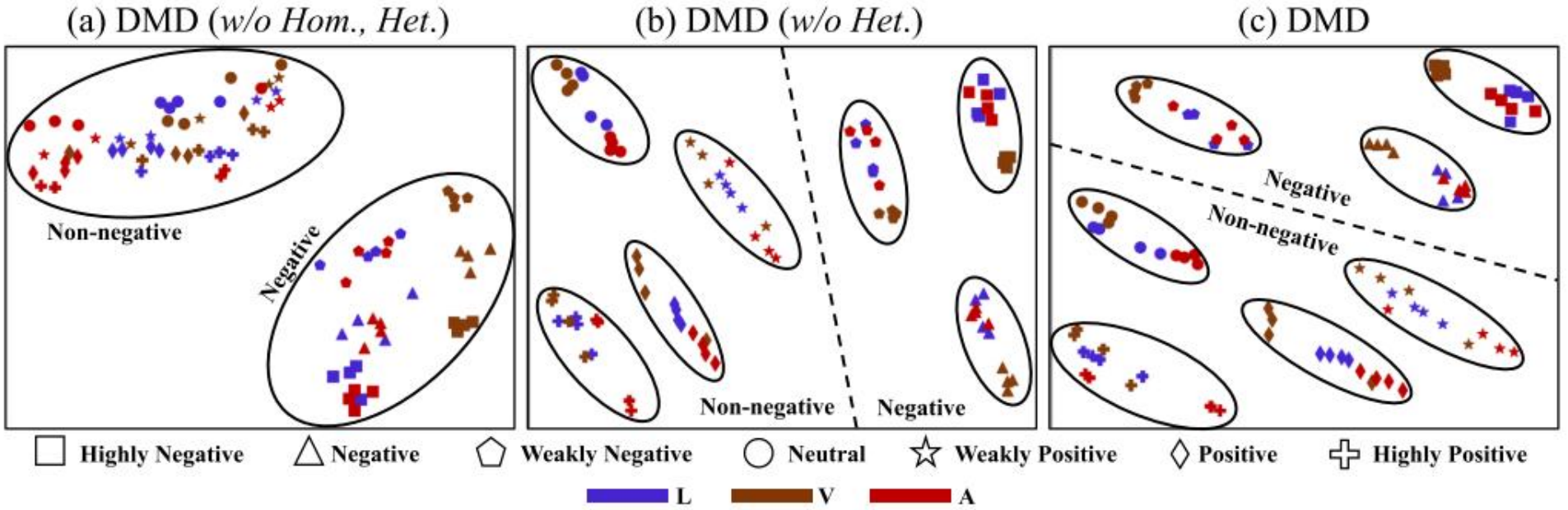


Figure 3. t-SNE visualization of decoupled homogeneous space on MOSEI. DMD shows the promising emotion category (binary or 7-class) separability in (c).

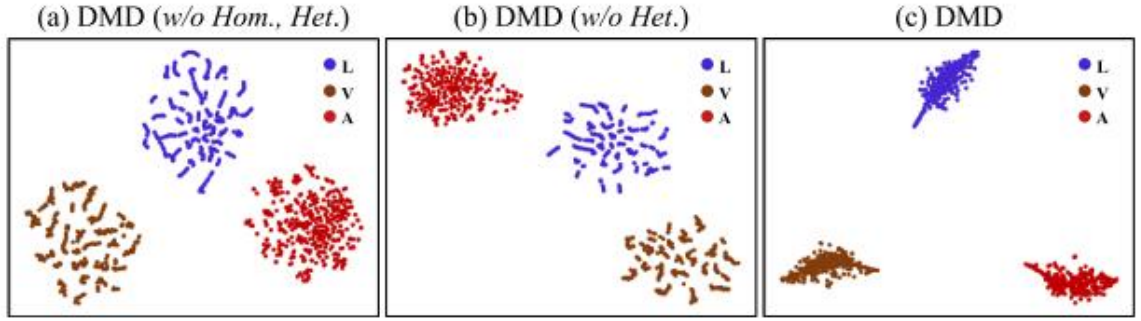
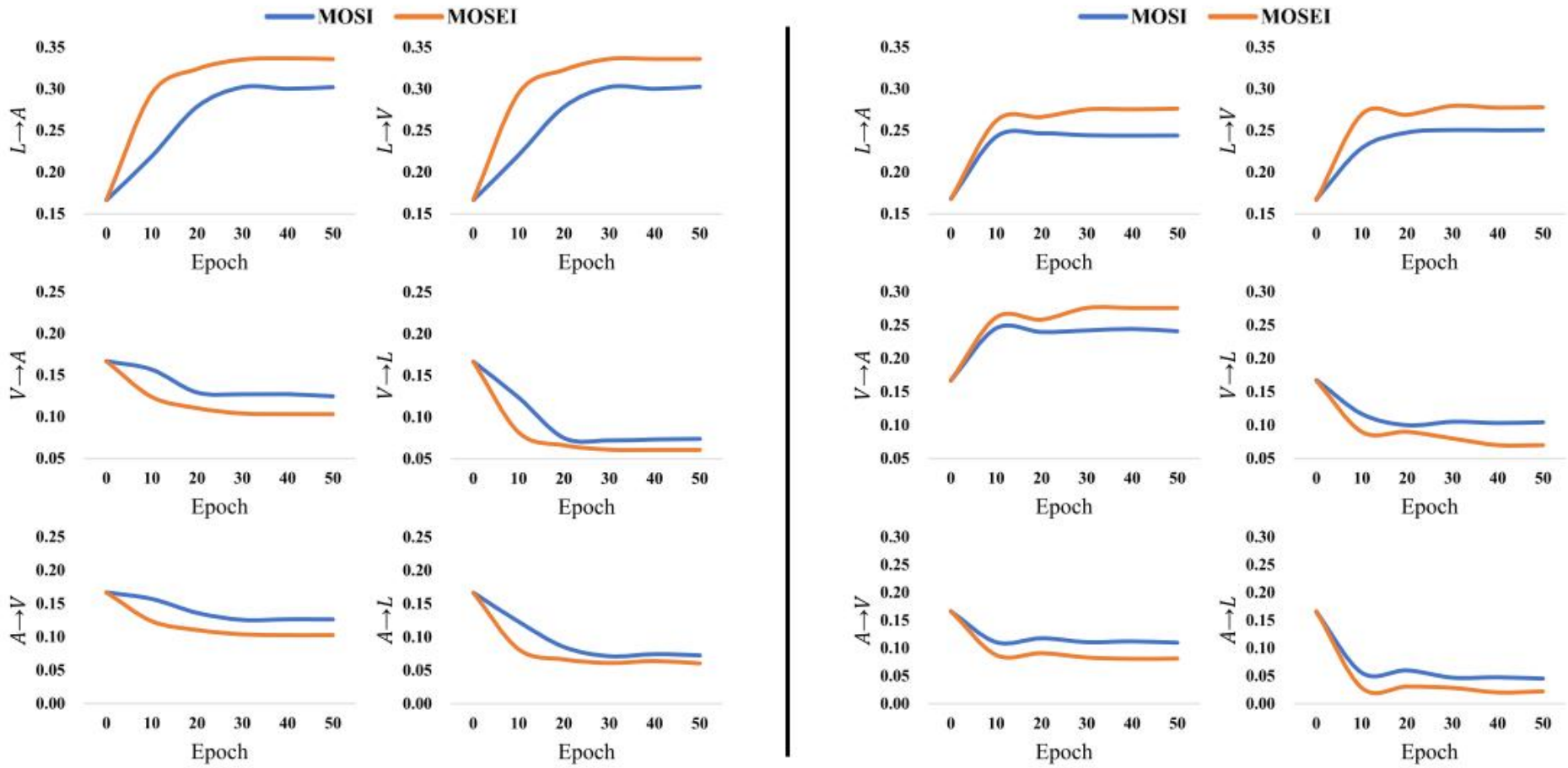


Figure 4. Visualization of the decoupled heterogeneous features on MOSEI. DMD shows the best modality separability in (c).

Experiment



(a) Graph edges in HomoGD

(b) Graph edges in HeteroGD

Figure 5. Illustration of the graph edges in HomoGD and HeteroGD. In (a), $L \rightarrow A$ and $L \rightarrow V$ are dominated because the homogeneous language features contribute most and the other modalities perform poorly. In (b), $L \rightarrow A$, $L \rightarrow V$, and $V \rightarrow A$ are dominated. $V \rightarrow A$ emerges because the *visual* modality enhanced its feature discriminability via the multimodal transformer mechanism in HeteroGD.

Thanks!