实验简述和论文图表详细分析

4 、实验简述和论文图表详细分析

4.1 数据集

我们在CMU-MOSI 和CMU-MOSEI 数据集上评估DMD。实验在单词对齐和非对齐设置下进行，以进行更全面的比较。

CMU-MOSI数据集包含2,199个短单词视频片段。CMU-MOSI中的声学和视觉特征的采样率分别为12.5和15 Hz。在样本中，有1,284个样本用作训练集，229个样本用作验证集，686个样本用作测试集。

CMU-MOSEI包含来自YouTube的22,856个电影评论视频片段（大约是CMU-MOSI的10倍大小）。声学和视觉特征的采样率分别为20和15 Hz。根据预定的协议，有16,326个样本用于训练，其余的1,871个和4,659个样本用于验证和测试。

CMU-MOSI和CMU-MOSEI中的每个样本都被标记有情感分数，范围从-3到3，包括高度负面、负面、弱负面、中性、弱正面、正面和高度正面。

按照先前的工作 [13, 17]，我们使用以下指标评估MER性能：七分类准确率（ACC7），二分类准确率（ACC2）和F1分数。

4.2 实验实施细则

在这两个数据集上，我们通过GloVe提取了单模态语言特征，并获得了300维的词特征。

为了在对齐设置下与MISA [7]和FDMER [32]进行公平比较，我们额外使用了一个BERT-base-uncased预训练模型，以获得768维的隐藏状态作为词特征。

对于视觉模态，我们使用Facet对每个视频帧进行编码，以表示35个面部动作单元的存在。

声学模态则通过COVAREP进行处理，得到74维的特征。

通过在验证集上评估MER性能，我们将λ1、λ2和γ的最佳设置设置为0.1、0.05和0.1。

我们使用PyTorch在一台具有24GB内存的RTX 3090 GPU上实现了所有实验。我们将训练批次大小设置为16，并在30个epoch直到收敛时训练DMD模型。

4.3 DMD与当前最先进的MER方法比较实验

我们将DMD与当前最先进的MER方法在相同的数据集设置下进行比较（未对齐或对齐），包括EF-LSTM、LF-LSTM、TFN [33]、LMF [14]、MFM [29]、RAVEN [30]、Graph-MFN [36]、MCTN

[26]、MulT [28]、PMR [17]、MICA [13]、MISA [7]和FDMER [32]。表1和表2分别说明了在CMU–MOSI和CMU–MOSEI数据集上的比较结果。

与基于特征解耦的MER方法[7, 29, 32]相比，我们提出的DMD取得了一致的改进，表明了内部GD–Unit的可行性，该单元能够感知各种跨模态动态。

DMD在使用多模态Transformer学习跨模态交互和进行多模态融合的方法[13, 17, 28]方面始终表现出色。原因有两个：（1）DMD同时考虑了模态无关/独有空间，并通过特征解耦减少了信息冗余。（2）DMD利用双GD–Unit在模态之间自适应地进行知识蒸馏。

4.3.1 实验图表详细分析

表一：

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.

表一为DMD与当前最先进的MER方法在CMU-MOSI数据集设置下进行比较的结果。

从表中数据可以看到：

（1）在单词对齐且未额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为41.4%，二分类准确率为84.5%，F1分数为84.4%。准确率和得分均明显高于其他各组。

（2）在单词对齐且额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为45.6%，二分类准确率为86.0%，F1分数为86.0%。准确率和得分不仅明显高于其他各组，而且明显高于未额外使用了BERT-base-uncased预训练模型组，证明了添加BERT-base-uncased预训练模型对模型的有效帮助。

（3）在单词未对齐且未额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为41.9%，二分类准确率为83.5%，F1分数为83.5%。准确率和得分均明显高于其他各组。

结果表明，我们提出的DMD在未对齐和对齐设置下的MER准确率优于其他MER方法。

表二：

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.

表一为DMD与当前最先进的MER方法在CMU-MOSEI数据集设置下进行比较的结果。

从表中数据可以看到：

（1）在单词对齐且未额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为53.7%，二分类准确率为85.0%，F1分数为84.9%。准确率和得分均明显高于其他各组。

（2）在单词对齐且额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为54.5%，二分类准确率为86.6%，F1分数为86.6%。准确率和得分不仅明显高于其他各组，而且明显高于未额外使用了BERT-base-uncased预训练模型组，证明了添加BERT-base-uncased预训练模型对模型的有效帮助。

（3）在单词未对齐且未额外使用了BERT-base-uncased预训练模型情况下，DMD组七分类准确率为54.6%，二分类准确率为84.8%，F1分数为84.7%。准确率和得分均明显高于其他各组。

结果表明，我们提出的DMD在未对齐和对齐设置下的MER准确率优于其他MER方法。

注：在CMU-MOSEI数据集上，Graph-MFN [36]的结果不理想，因为跨模态的异质性和分布差异阻碍了模态融合的学习。相比之下，DMD中的多模态特征被解耦为模态无关/独有的空间。对于后者的空间，我们使用多模态Transformer来弥合分布差异并对齐高层语义，从而减轻了吸收来自异质特征的知识的负担。

4.4 消融实验

我们评估了DMD的关键组成部分，包括特征解耦（FD）、HomoGD、跨模态注意力单元（CA）、HeteroGD的效果。结果如表3所示。

4.4.1 实验图表详细分析

表三：

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

表中打勾表示参与评估的DMD的关键组成部分，打叉表示未参与评估的DMD的关键组成部分。

从表中数据可以看到：

（1）在MOSI和MOSEI数据集下，DMD的四个关键组成部分均设置的组准确率和得分最高。在MOSI数据集下，七分类准确率41.9%，F1得分83.5%；在MOSEI数据集下，七分类准确率54.6%，F1得分84.7%。

（2）在MOSI和MOSEI数据集下，DMD的四个关键组成部分均未设置的组准确率和得分最低。在MOSI数据集下，七分类准确率32.4%，F1得分79.0%；在MOSEI数据集下，七分类准确率50.0%，F1得分81.9%。

（3）在MOSI和MOSEI数据集下，DMD设置FD部分的组准确率和得分高于未设置的组。在MOSI数据集下，七分类准确率34.7%，F1得分79.3%；在MOSEI数据集下，七分类准确率51.6%，F1得分82.8%。

（4）在MOSI和MOSEI数据集下，DMD设置FD和HomoGD部分的组准确率和得分高于未设置的组，而且相较于仅设置FD的组准确率和得分提高。在MOSI数据集下，七分类准确率37.2%，F1得分80.8%；在MOSEI数据集下，七分类准确率52.4%，F1得分84.3%。

（5）在MOSI和MOSEI数据集下，DMD设置FD和HomoGD基础上再分别设置CA或HeteroGD部分。两种设置的实验结果准确率均高于未设置的组。在MOSI数据集下，FD+HomoGD+CA组七分类准确率38.8%，F1得分81.1%；在MOSEI数据集下，七分类准确率53.2%，F1得分84.1%。在MOSI数据集下，FD+HomoGD+HeteroGD组七分类准确率37.5%，F1得分80.6%；在MOSEI数据集下，七分类准确率52.4%，F1得分83.8%。

结果表明：

（1）特征解耦（FD）显著提升了MER性能，这表明解耦和精炼的特征可以减少信息冗余并提供具有辨别力的多模态特征。为进一步证明FD的有效性，后续在MOSEI数据集上对基线模型进行了带有和不带有FD的实验，如表4所示。

（2）将FD与HomoGD结合带来进一步的好处。虽然同质特征嵌入在相同维度空间中，但模态之间仍然存在不同的辨别能力。HomoGD可以通过GD改善弱模态。

为验证这一点，我们在MOSEI数据集上进行了带有或不带有HomoGD的实验，论文仅展示结果没有做图表：ACC2的结果分别为：语言模态80.9%对比82.4%，视觉模态56.5%对比61.8%，音频模态54.4%对比64.1%。然而，进行HeteroGD而没有跨模态注意力单元会导致性能下降，这表明多模态转换器在缩小多模态分布差距方面发挥着关键作用。

（3）通过使用CA单元和HeteroGD，DMD获得了显著的改进，表明利用模态专属特征对于稳健的MER非常重要。

（4）我们设置的DMD的四个关键组成部分，包括特征解耦（FD）、HomoGD、跨模态注意力单元（CA）、HeteroGD对实验准确率有着重要的有益影响。

表四：MOSEI数据集上的单模态精度比较

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**

根据消融实验得出结论：FD显著提升了MER性能，这表明解耦和精炼的特征可以减少信息冗余并提供具有辨别力的多模态特征。为进一步证明FD的有效性，在MOSEI数据集上对基线模型进行了带有和不带有FD的实验，结果如上表所示。w/o FD表示不带FD组，w/FD表示带FD组，Mean为平均值，STD为标准差。

从表中数据可以看到：

（1）仅语言模态，带FD组的二分类准确率为82.7%，F1得分为82.5%，均高于不带FD组，其二分类准确率为81.2%，F1得分为81.4%。

（2）仅视觉模态，带FD组的二分类准确率为62.8%，F1得分为60.0%，均高于不带FD组，其二分类准确率为58.2%，F1得分为52.2%。

（3）仅听觉模态，带FD组的二分类准确率为64.9%，F1得分为62.5%，均高于不带FD组，其二分类准确率为53.4%，F1得分为54.0%。

（4）带FD的三种模态的准确率和得分的平均值为70.1%和68.3%，均高于不带FD的64.3%和62.5%。

（5）带FD的三种模态的准确率和得分的标准差为8.9%和10.1%，均低于不带FD的12.1%和13.4%。

结果表明：

（1）FD对每个单模态都带来了一致的改进。

（2）与此同时，三个模态之间的性能差距也缩小了，因为ACC2和F1的标准差都减小了。

表五：图形蒸馏（GD）在MulT上的消融研究

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**

我们将我们提出的DMD与经典的MulT [28]进行了比较以进一步研究。结果如表5所示。

其中MulT（w/ GD）表示我们在MulT上添加了一个GD单元，以便对强化的多模态特征进行自适应知识蒸馏。从本质上讲，MulT（w/ GD）和DMD之间的核心区别在于DMD包含了特征解耦。

从表中数据可以看到：

（1）我们提出的DMD的准确率和F1得分在MOSI和MOSEI数据集上均高于经典的MulT和添加了一个GD单元的MulT（w/ GD）组.

（2）对于MulT组，添加一个GD单元可以提高其准确率和F1得分。但准确率和F1得分仍达不到DMD组。

结果表明：

（1）在蒸馏之前对多模态特征进行解耦是可行和合理的。

（2）此外，添加GD单元，DMD比普通的MulT实现了更显著的改进，表明结合特征解耦和图蒸馏机制的好处。

4.4.2 解耦特征的可视化

我们在图3和图4中可视化了DMD、DMD（无HomoGD、Het.）、DMD（无Het.）的解耦合同质和异质特征，以进行定量比较。

图三：对于CMU-MOSEI数据集，我们在解耦的同质空间上进行了t-SNE可视化。DMD在图(c)中展示了有希望的情绪类别（二分类或七分类）的可分性。

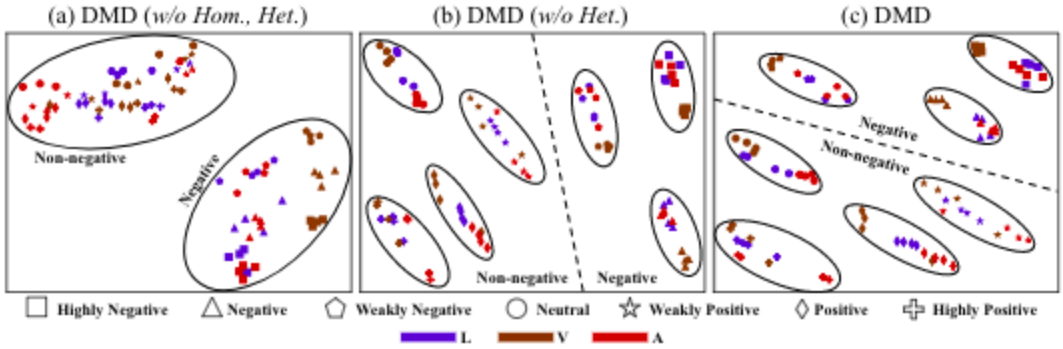


Figure 3. t-SNE visualization of decoupled homogeneous space on MOSEI. DMD shows the promising emotion category (binary or 7-class) separability in (c).

图三可视化了DMD、DMD（无HomoGD、Het.）、DMD（无Het.）的解耦合同质特征。该过程随机选择了CMU-MOSEI数据集测试集中的28个样本（每个情绪类别选择四个样本）可视化。其中 DMD (w/o Hom., Het.)表示没有HomoGD和HeteroGD的DMD。DMD（w/o Het.）表示没有HeteroGD的DMD。

从图中可以看出：

（1）对于DMD和DMD (w/o Het.)的同质多模态特征，同一情绪类别的样本自然地聚集在一起，因为它们在跨模态上具有一致性。

（2）在DMD（w/o Hom., Het.）中，具有解耦的同质特征但没有图蒸馏机制，样本在二分类的非负和负面类别上表现出基本的可分性。然而，在七分类设置下，样本无法区分，表明特征不如DMD或DMD（w/o Het.）那么具有区分性。

因此得出结论：

图蒸馏对同质多模态特征起有效帮助作用。

图三：在MOSEI上可视化解耦的异质特征。DMD在(c)中展示了最佳的模态可分性。

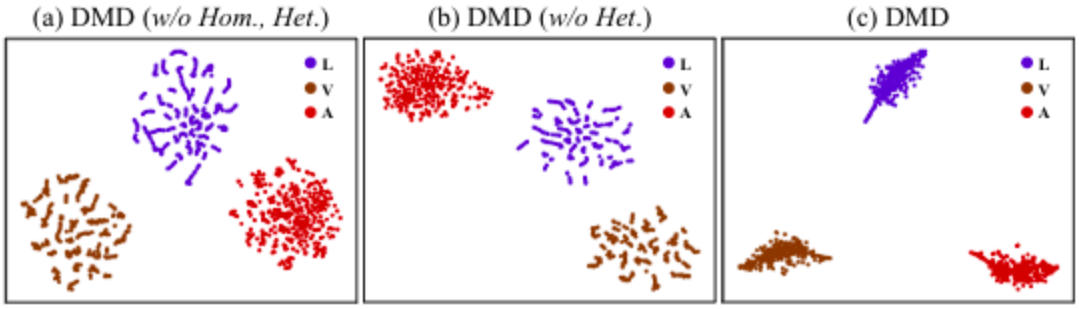


Figure 4. Visualization of the decoupled heterogeneous features on MOSEI. DMD shows the best modality separability in (c).

图四可视化了DMD、DMD（无HomoGD、Het.）、DMD（无Het.）的解耦合异质特征。该过程随机选择了CMU-MOSEI数据集测试集中的400个样本可视化。其中 DMD (w/o Hom., Het.)表示没有HomoGD和HeteroGD的DMD。DMD（w/o Het.）表示没有HeteroGD的DMD。

在异质空间中，由于其模态间的异质性，不同样本的特征应该按照模态进行聚类。如图4所示，DMD展示了最佳的特征可分性，表明模态之间的互补性得到了显著增强。DMD (w/o Hom., Het.)和DMD (w/o Het.)的特征可分性较DMD较低，表明图蒸馏对于异质多模态特征的重要性。

从图中可以看出：

（1）DMD各模态的样本距离最为接近。

（2）在DMD（w/o Het.）中，样本距离较为分散。

（3）在DMD（w/o Hom., Het.）中，样本距离最为分散。

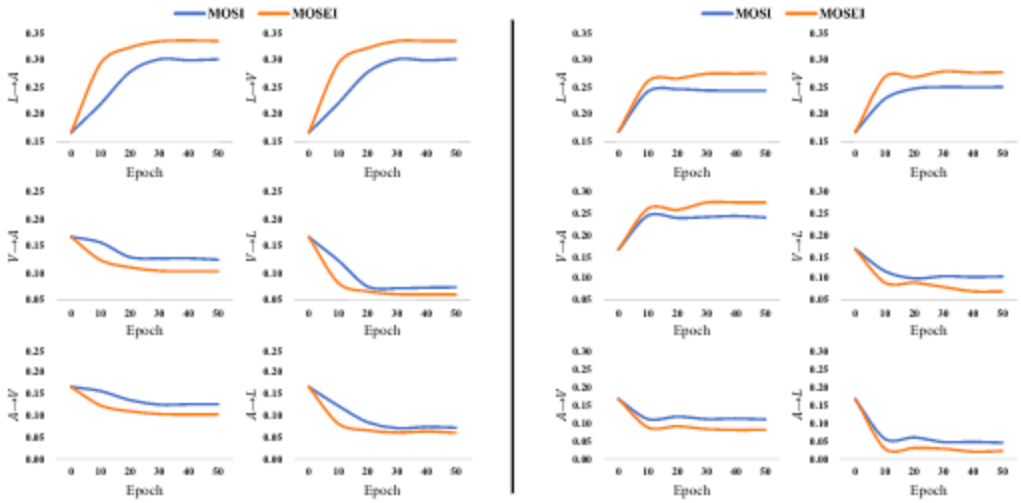
结果表明：

（1）DMD展示了最佳的特征可分性，表明模态之间的互补性得到了显著增强。

（2）在DMD（w/o Hom., Het.）中，具有解耦的同质特征但没有图蒸馏机制。论文补充，样本在二分类的非负和负面类别上表现出基本的可分性。在七分类设置下，样本无法区分，特征不如DMD或DMD（w/o Het.）那么具有区分性。

4.4.3 图形边在GD-Units中的可视化

图五： HomoGD和HeteroGD中图边缘的示意图。在(a)中，L→A和L→V占主导地位，因为同质语言特征贡献最大，而其他模态性能较差。在(b)中，L→A、L→V和V→A占主导地位，V→A出现是因为在HeteroGD中，视觉模态通过多模态变换机制增强了其特征的可分性。



(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.

在图5中，我们展示了每个GD单元中动态的边，以进行分析。每个图边对应着有向蒸馏的强度。

我们得出以下观察结果：

(1) HomoGD中的蒸馏主要由L→A和L→V主导。这是因为解耦的同质语言模态仍然发挥着最关键的作用，并以显著优势超越了视觉或声学模态。对于CMU-MOSEI数据集上的二元MER，使用解耦的同质特征，语言、视觉和声学模态分别获得80.9%、56.5%和54.4%的准确率。

(2) 对于HeteroGD，L→A、L→V和V→A占主导地位。一个有趣的现象是出现了V→A。这是合理的，因为视觉模态通过HeteroGD中的多模态变换机制增强了其特征的可分性。实际上，这三个模态分别获得了84.5%、83.8%和71.0%的准确率。

得出结论：

图边学习了有关自适应跨模态蒸馏的有意义的模式。