arXiv:2305.02814v1 [cs.MM] 4 May 2023

用于情绪识别的噪声鲁棒多模态Transformer

Yuanyuan Liu[1], Haoyu Zhang*[1,2], Yibing Zhan[4],Zijing Chen[5],Guanghao Yin[1],Lin Wei[1], 和 Zhe Chen[3]

[1]中国地质大学（武汉）计算机科学学院[2]中国地质大学（武汉）地理信息系统国家工程研究中心[1,2]{*liuyy, zhanghaoyu,ygh2,linw*}*@cug.edu.cn*[3]悉尼大学计算机科学学院[3] *zhe.chen1@sydney.edu.au*[4]京东探索研究院，京东集团[4]*zhanyibing@jd.com*[5]澳大利亚天主教大学彼得·费伯商学院[5]*zijing.chen@acu.edu.au*

摘要

多模态情感识别从视频、文本和音频等不同数据模态中识别人类情绪。然而，我们发现，这项任务很容易受到噪声信息的影响；这些信息不包含有用语义，且可能出现在多模态输入序列的不同位置。为此，我们提出了一种新范式，尝试在其流程中提取抗噪特征，并引入噪声感知学习方案，以有效提升多模态情感理解对噪声信息的鲁棒性。我们提出的新流程称为噪声鲁棒多模态Transformer（NORM-TR），主要引入了一个噪声鲁棒通用特征（NRGF）提取器，以及一个用于多模态情感识别任务的多模态融合Transformer。具体而言，我们让NRGF提取器学习提供一种通用且不易受干扰的表征，从而获得一致且有意义的语义。此外，我们应用多模态融合Transformer，将多模态输入的多模态特征（MFs）（作为键和值）基于它们与 NRGF（充当查询）的关系。因此，NRGF 中可能不敏感但有用的信息，可以通过包含更多细节的多模态特征加以补充，从而在保持对噪声鲁棒性的同时，实现更准确的情绪理解。为恰当训练 NORM-TR，我们提出的噪声感知学习方案通过增强模型应对噪声的学习能力，对常规情绪识别损失进行补充。该学习方案会在多模态输入序列中随机选择位置，向所有模态或某个特定模态添加噪声。相应地，我们引入两种对抗损失，促使 NRGF 提取器学习提取对所添加噪声不变的 NRGF，从而帮助 NORM-TR 获得更优的多模态情绪识别表现。在实践中，大量实验表明，NORM-TR 和噪声感知学习方案能够有效处理显式添加的噪声信息，以及含有隐式噪声的常规多模态序列。在多个热门多模态数据集（例如，MOSI、MOSEI、IEMOCAP 和 RML）上，NORM-TR 达到了最先进的性能，并大幅超越现有方法，

*通讯作者

这表明，多模态输入抵御噪声信息的能力对于有效的情绪识别至关重要。

1 引言

准确理解人类情绪有益于多种应用，例如多媒体分析、数字娱乐、健康监测、人机交互，等（Shen等人，2009；Beale & Pe-ter，2008；Qian等人，2019年；D'Mello & Kory，2015）。与仅使用单模态数据源的传统情绪识别相比，利用并探索不同数据源（如视觉、音频和文本）的多模态情绪识别在提升情绪理解方面表现出显著优势（Zadeh等人，2017年；Tsai、Bai等，2019年；Lv等人，2021年；Hazarika等，2020；Yuan 等人，2021年），包括快乐、愤怒、厌恶、恐惧、悲伤、中性和惊讶。

最近，大多数现有的多模态情绪识别方法主要关注多模态数据融合，包括基于张量的融合方法（Z. Liu 等人，2018年；Zadeh 等人，2017年；Sahay 等人，2020；Yuan 等人，2021年）和基于注意力的融合方法（Zhao等人，2020；Huang 等人，2020；Zhou 等人，2021年）。基于张量的融合方法旨在通过多线性函数计算，获得具有不同模态的数据的联合表示。例如，张量融合网络（Z. Liu 等人，2018年）使用笛卡尔积运算来计算不同模态之间的关系，从而获得令人满意的性能。由于笛卡尔积的计算复杂度会随着特征维度和模态数量的增加而急剧上升，如果不使用极大的模型容量，其性能很容易受到影响而下降。低秩多模态融合（Zadeh 等人，2017年）引入低秩矩阵分解运算，以降低计算成本。然而，低秩多模态融合往往会减少有用信息，导致模型性能下降。基于注意力的融合方法主要采用注意力学习机制，使多模态信息彼此交互。例如，Zhou 等人（2021年）引入注意力学习，自动计算音频和视频模态的重要性权重，从而获得有效的情绪-

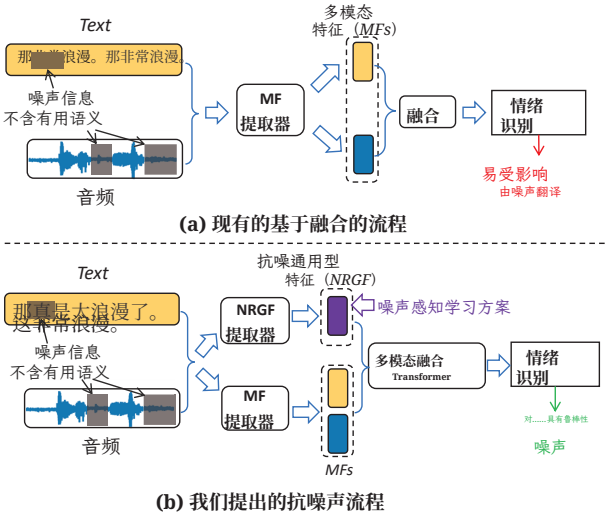


图 1：我们提出的多模态情绪识别方法，即抗噪模态变换器（NORM-TR）（如图 (b) 所示），与常见的多模态融合方法（如图 (a) 所示）的对比。对于可能包含带有无关、无用语义的噪声信息的多模态输入，现有直接融合多模态特征（多模态特征）的多模态情绪识别方法很容易受到噪声影响。作为替代，我们借助一种新颖的噪声感知学习方案，让 NORM-TR 学习提取抗噪通用特征（NRGF）。随后，通过使用多模态融合Transformer让多模态特征补充 NRGF，我们利用 NORM-TR 获得了稳健得多的多模态情绪识别结果。

相关信息。Zhao等人（2020）提出了一种新的基于注意力机制的VAANET，将空间注意力、通道注意力和时间注意力整合起来，用于音视频情感识别。总体而言，大多数注意力方法倾向于探索模态的重要性，以实现更有效的融合，却忽略了模态内部有用信息和无用信息的影响。

尽管融合策略目前已有进展，我们认为，减轻噪声信息带来的负面影响也很重要。具体而言，我们发现，在多模态序列中，可能存在大量与情感理解关联不大的信息，这些信息可以视为噪声信息。例如，

音频数据中的背景声音与对应视频中微笑的人无关。因此，对这些背景声音中的琐碎信息进行建模可能会影响多模态融合和最终的理解性能。在我们的实验中，我们可以证明，噪声信息会大幅降低情感识别准确率，这进一步表明，对噪声不敏感有助于实现准确的情绪理解。然而，据我们所知，当前关于多模态情感理解的文献对噪声信息缺乏充分研究，因此仍然只能取得次优性能。

针对上述问题，我们提出了一种新颖的抗噪多模态Transformer（NORM-TR），以应对噪声信息对多模态情绪识别造成的不利影响。NORM-TR的设计动机以及与现有融合方法的比较如图1所示。总体而言，我们让NORM-TR学习提取抗噪通用特征（NRGF），然后使用Transformer（Vaswani et al., 2017年）根据从多模态输入中提取的多模态特征（MFs）与NRGF之间的关系，将这些特征纳入其中，从而获得更能抵御噪声、更稳健且更准确的情感理解结果。具体来说，我们力求使NRGF具有通用性，并且不易受到噪声所致干扰的影响。为获得NRGF，我们采用NRGF提取器，使其学习从多模态数据中概括有意义的语义。提取出的NRGF能够提供抵御噪声信息的稳健表征，但也可能因此对准确情绪识别所需的有用细节不够敏感。因此，我们进一步引入多模态融合Transformer，以NRGF作为查询，以多模态特征（MFs）作为键和值；由此推理多模态特征（MFs）与NRGF之间的关系，使多模态特征（MFs）能够补充NRGF，从而实现稳健且准确的情绪识别预测。

为了有效训练 NORM-TR，我们首先采用常规的情绪识别损失，使其学会估计人类情绪。同时，我们进一步采用所提出的噪声感知学习方案，帮助模型增强抵御噪声的鲁棒性。我们的噪声感知学习方案会显式地向输入中添加噪声，并通过对抗损失训练 NORM-TR 中的 NRGF提取器。显式添加噪声可以明确指出输入中噪声信息出现的位置，从而帮助相关对抗损失中的判别器区分特征是否包含噪声。具体而言，我们采用两种方式添加噪声信息：1）随机地将添加的噪声信息置于所有模态中；2）在某些随机时段内，仅向特定模态添加噪声信息。针对这两种添加噪声的方式，我们分别引入了两种对抗损失。通过欺骗用于区分第一种添加噪声的判别器，我们可以使 NRGF提取器专注于多模态数据中更具通用性、能够抵御噪声的特征。同样，欺骗用于区分第二种添加噪声的判别器，可以使提取出的 NRGF 更能抵御各个特定模态中的噪声。结合情绪识别损失，我们提出的噪声感知学习方案有助于获得鲁棒的 NRGF，并使 NORM-TR 模型取得良好的多模态情感识别性能。

总而言之，本文的主要贡献可概括为：

- 我们针对多模态情感理解任务中的噪声信息开展了一项全面的新颖研究。为实现更稳健的情感理解性能，我们提出了抗噪多模态 Transformer（NORM-TR），以提取抗噪通用特征（NRGF），并显著降低噪声对多模态数据造成的负面影响。

- 基于 NRGF，我们设计了一种新颖的、基于 Transformer 的端到端多模态情感识别流程。此外，我们进一步设计了一种新颖的噪声感知学习方案，以帮助合理优化 NORM-TR。

- 实践中，我们证明 NORM-TR 能够有效获得噪声不变表示。此外，我们对多个热门数据集进行的大量实验分析也表明，NORM-TR 通过使用 NRGF 缓解噪声信息带来的不利影响，显著提高了情绪识别准确率，并取得了最先进的性能，这说明处理噪声信息至关重要。

2 相关工作

2.1 多模态情绪识别

多模态情感识别旨在根据视频、音频和文本等多种模态预测人类情绪。现有的大多数方法（Hazarika等人，2020；Zhao等人，2020；Tsai、Bai 等人，2019；Yang 等人，2022；Sun等人，2020；Tsai、Liang等人，2019）主要着眼于如何从不同模态的数据中学习并融合多模态情感表征，同时考虑不同模态之间的差异与一致性。例如，Hazarika等人（2020）提出了一种多模态表征学习方法，用于模态不变与模态特定子空间投影。此外，为了学习不同模态之间的交互信息，近期越来越多的研究聚焦于多模态融合机制，并提出了许多精细的多模态融合方法。Tsai、Liang等人（2019）提出了多模态因子分解模式（MFM），通过将模态表征分解为两组相互独立的因子，探索模态内和跨模态交互。Sun等人（2020）提出了交互典型相关网络（ICCN），该网络使用典型相关分析（CCA）对音频-文本和视频-文本模态之间的关系进行建模。Lv等人（2021）提出渐进式模态强化（PMR）方法，通过考虑所有涉及模态之间的三方交互来进行多模态融合。

尽管取得了进展，当前方法却很少关注多模态情绪识别中的噪声问题。Yuan 等人（2021年）也指出，多模态数据中包含大量噪声，例如某些模态序列中的数据缺失，这会大幅降低多模态融合方法的效果。如何有效降低噪声对多模态数据的影响，仍然是一个尚未解决的问题。

2.2 对抗学习

对抗学习广泛应用于领域自适应学习（Ganin和Lempitsky，2015；Pei等人，2018年；Y. Wang等人，2020；Z. He等人，2022年）以及跨模态检索（B. Wang等人，2017年；Li等人，2018年）等领域。近年来，为提高融合效果，对抗学习越来越多地用于多模态情绪识别中，旨在学习共同子空间表征。 Yang等人 (2022年)采用结合对抗学习的FDMER来挖掘不同模态的共性与差异性，在多模态情感识别方面取得了出色的表现。尽管取得了进展，FDMER却未考虑噪声对模型鲁棒性的副作用。为解决这一局限，我们提出了NORM-TR，并采用新颖的噪声感知对抗学习来提取抗噪通用特征（NRGF），从而大幅降低噪声信息带来的负面影响，并提高多模态融合的鲁棒性。

2.3 Transformer

Transformer 是一种基于注意力机制的构建模块，由 Vaswani 等人于 2017年提出，用于机器翻译。通过聚合整个序列中的数据，Transformer 能够学习随时间扫描的词元之间的关系，并在多种任务中取代 RNN，例如自然语言处理（Kenton 和 Toutanova，2019年；Ding 等人，2021年）、计算机视觉（Q. Zhang 等人，2022年；Y. Liu 等人，2021年），以及多模态情绪识别（Hazarika 等人，2020；Liang 等人，2020；Tsai、Bai等人，2019；Huang 等人，2020；Yuan 等人，2021年；Y. Liu 等人，2022年）。Tsai、Bai等人（2019）提出了多模态 Transformer（MulT），以解决模态数据未对齐和长距离依赖问题。Huang 等人（2020）利用 Transformer 在模型层面融合音频和视觉信息，表明模型层面的融合优于其他层次的融合策略。Yuan 等人（2021年）提出了一种基于 Transformer 的特征重建网络，以实现更稳健的多模态情绪识别。尽管取得了进展，现有的基于 Transformer 的方法主要将特定模态的特征作为查询，这可能会引入与该模态相关的不必要噪声或无关信息。

3 方法

3.1 概述

所提出的抗噪声多模态 Transformer（NORM-TR）的整体处理流程，用于

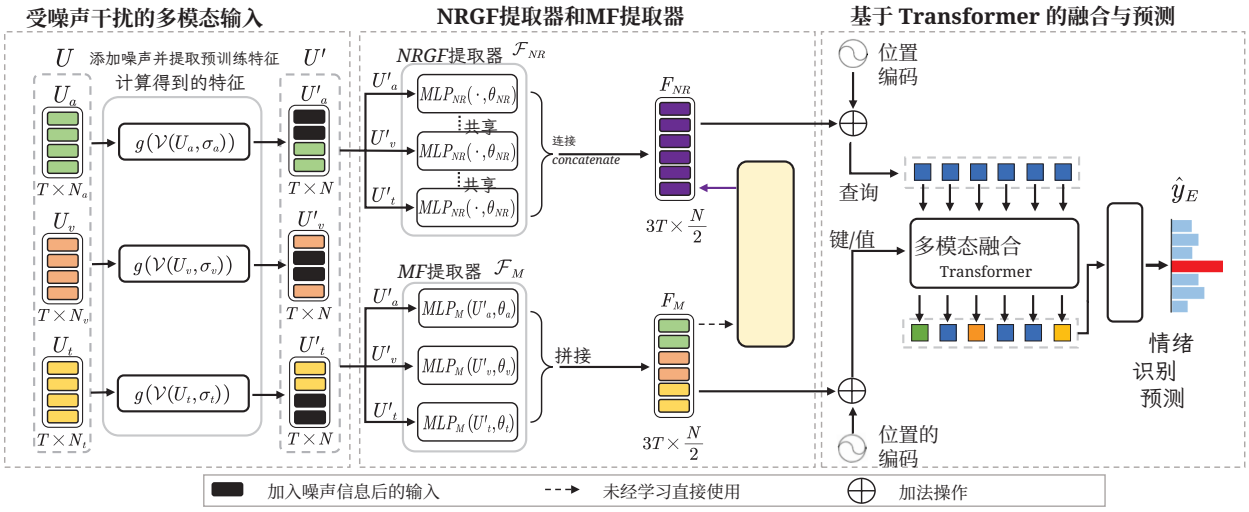


图 2：所提出的用于多模态情感识别的 NORM-TR 处理流程。对于受噪声干扰的多模态输入 U'，我们首先使用抗噪通用特征（NRGF）提取器，提取通用且对干扰不敏感的表征。随后，我们采用多模态融合 Transformer，基于更细致的多模态特征（MFs）改进 NRGF。此外，我们提出了一种新颖的噪声感知学习方案，以恰当地训练 NORM-TR，从而获得用于情绪理解的端到端抗噪模型

鲁棒情绪识别2如图所示。我们首先利用 NORM-TR 从输入中提取抗噪通用特征（NRGFs）和多模态特征（MFs）。随后，采用多模态融合 Transformer，根据 MFs 与 NRGFs 之间的关系对 MFs 进行整合，从而获得一个用于情感理解的端到端抗噪模型。为了正确训练 NORM-TR，我们提出了一种噪声感知学习方案。在该学习方案中，我们手动擦除多模态输入中的某些特定时间段的信息，以显式引入不包含任何有用语义的噪声信息。通过将噪声信息显式添加到所有多模态输入中，或添加到来自特定模态的输入数据中，我们设计了两个对抗学习目标，使 NORM-TR 能够抵抗这两种类型的附加噪声信息。

形式上，我们的 NORM-TR 采用一个 NRGF 提取器，记为 $\mathcal{F}_{NR}$，以及一个 MF提取器，记为$\mathcal{F}_{M}$，分别从其输入U'中提取详细的 NRGF 和多模态特征。然后，多模态融合Transformer，$Trans(\cdot)$，根据 NRGF 将多模态特征转换为情绪识别所需的输出 $\hat{}y_E$。因此，我们的整体流程可以描述为：

$$\hat{y}_E = Trans\Big(\mathcal{F}_{NR}(U'), \mathcal{F}_{M}(U')\Big), \tag{1}$$

其中，$\hat{}y_E$ 是情绪识别输出，而 U'是受噪声干扰的多模态输入。值得一提的是，如公式1所述，我们的 NORM-TR 也适用于未显式添加噪声的正常多模态输入。

3.2 受噪声干扰的多模态输入

关于常规的多模态情感识别，我们使用符号 U 表示给定序列的输入多模态信息。 U 可表示音频、视频、文本等输入。在本文其余部分，我们使用 U_a， U_v和 U_t 分别表示音频、视频和文本。在相关文献中（Tsai、Bai等，2019年；Hazarika等，2020；Mao等，2022年），通常使用预计算特征而非不同模态的原始数据。因此，为了公平起见，本文中的 U 表示预计算的特征向量。例如，不使用某个

对于视频，我们可以将预计算特征$U_v \in \mathbb{R}^{T\times N_v}$ 作为输入，其中 T 表示视频的长度， N_v 表示特征向量的长度。相应地，对于音频和文本模态，我们也可以分别将预计算特征 $U_a \in \mathbb{R}^{T\times N_a}$和 $U_t \in \mathbb{R}^{T\times N_t}$作为输入，其中 N_a 和 N_t 是相应特征向量的长度，而 T是每种模态的特征向量的统一长度。我们想要说明的是，在多模态情感识别领域的文献中，使用预计算特征已被广泛接受；而研究如何选择更合适的特征来描述多模态原始数据，则超出了本研究的范围。借助预计算特征$U = \{U_a, U_v, U_t\}$，我们分别使用三个全连接层来统一不同特征向量的长度。我们用符号 g 表示这些全连接层。经过 g后，得到的 $U' = \{U'_a, U'_v, U'_t\}$中的每种模态都具有 $T \times N$维，其中 N 是特征向量的统一长度。同时，我们使用 $\mathcal{V}(U, \sigma)$表示将噪声信息 σ 显式添加到输入U中的过程。因此，输入 NORM-TR 的受噪声干扰的多模态输入可表示为：

$$U' = g(\mathcal{V}(U, \sigma)). \tag{2}$$

接下来的章节将依次详细介绍我们提出的 NORM-TR，包括 NRGF提取器、MF提取器、多模态融合 Transformer 以及噪声感知学习方案。

3.3 NRGF提取器和MF提取器

3.3.1 NRGF提取器

对于多模态输入 U'，我们引入 NRGF提取器，以获得对噪声信息不敏感且具有不变性的通用特征。为此，我们只需采用一个多层感知机：

$$F_{NR} = \mathcal{F}_{NR}(U') = MLP_{NR}(U', \theta_{NR}), \tag{3}$$

其中， F_{NR} 是得到的 NRGF， MLP_{NR} 表示用于提取 NRGF 的共享参数多层感知机。实际中， MLP_{NR} 被设计为一个 2 层全连接网络，每层后接 Leaky-ReLU 激活函数（Maas等人，2013年）。使用 MLP_{NR}，我们降低了 U'的特征维度。具体而言，若考虑 3 种模态，通过式 U'2 得到的 维度为 $3T \times N$，而 F_{NR} 的维度为 $3T \times \frac{N}{2}$。事实上，我们发现，降低特征维度可以减少参数量并提高效率，同时不损失性能。

为了使 NRGF 能够有效提取对噪声不敏感的特征，我们引入了噪声感知学习方案，具体内容将在第 3.5 节中详细介绍。

3.3.2 MF提取器

除了 NRGF提取器之外，我们还引入了 MF提取器，从输入中提取并利用详细的多模态特征。这是因为我们发现，NRGF 可能会丢失一些有助于准确进行情感理解的细节。为了实现 MF提取器，我们仍然使用多层感知机，因此有：

$$F_M = \mathcal{F}_M(U') = MLP_M(U', \theta_*). \tag{4}$$

MLP_M 表示三个独立的多层感知机，用于 MF 特征提取，其中 MLP_M中的每个多层感知机架构都与 MLP_{NR}相同。$* \in \{a, v, t\}$。因此，给定 3 个输入模态， F_M 的尺寸为 $3T \times \frac{N}{2}$。需要说明的是，输入特征 U' 由 NRGF提取器和 MF提取器共享。此外，我们不会让 $\mathcal{F}_M$ 学习适应噪声引起的变化，从而使其对多模态输入中的变化更加敏感。

3.4 Transformer结构与情感识别输出

获得 NRGF 和多模态特征后，我们采用多模态融合 Transformer 来实现有效的多模态情绪识别。Transformer 能够对 NRGF 与多模态特征之间的关系进行建模，使多模态特征能够根据其与 NRGF 的相关性，更恰当地补充 NRGF。我们发现这种关系建模十分重要，因为多模态特征可能更容易受到噪声信息的影响，而直接融合

将它们与 NRGF 结合会引入噪声的负面影响（更多详情见4.7.1）。

3.4.1 多模态融合 Transformer

根据Vaswani 等 (2017年的定义，Transformer 以查询、键和值张量作为输入。随后，它将查询张量用作参考，根据查询张量与键张量之间的关系，将值张量转换为所需的输出。在转换过程中，通过多头注意力机制实现关系建模和数据融合。更多详情请参阅Tsai、Bai等(2019年)。

在本研究中，我们将提取的 NRGF 用作查询，将多模态特征用作键和值，遵循 Transformer 结构的典型形式，以实现更有效的多模态融合，并将 $Trans(\cdot)$ 实现如下：

$$Trans(\cdot) = Trans(q = F_{NR}, k/v = F_M), \quad (5)$$

其中，q、k 和 v 分别表示 Transformer 中的查询、键和值张量。具体而言，查询（即由 $\mathcal{F}_{NR}$ 提取的 NRGF）的形状为 $3T \times \frac{N}{2}$。键和值张量共用由 $\mathcal{F}_M$ 获得的同一多模态特征，其形状为 $3T \times \frac{N}{2}$。$Trans(\cdot)$ 包含八个注意力头，并在不同数据集上采用不同深度的注意力模块，例如，MOSI（Zadeh 等，2016）和 RML（Y. Wang 和 Guan，2008）数据集采用 2 层深度，而 MOSEI（Zadeh 等，2018）和 IEMOCAP（Busso 等，2008）数据集采用 4 层深度。这是因为在 MOSI 和 RML 等较小的数据集上使用更深的网络容易导致过拟合。此外，遵循 Dosovitskiy 等（2021）的做法，我们对序列时间戳进行可学习的位置编码，并将其加到 Transformer 的输入中。值得一提的是，我们尝试了更复杂的融合结构，但性能提升并不明显（更多详情见 4.7.4 ）。我们发现，较深的 Transformer 结构也容易导致数据集过拟合。

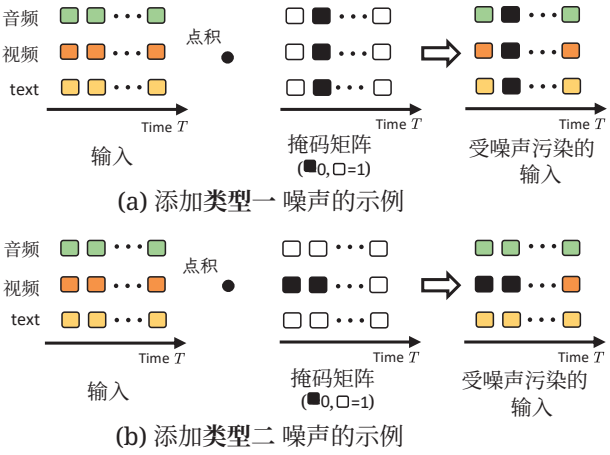
3.4.2 情绪识别输出

经过多模态融合Transformer后，我们将情绪分类层应用于Transformer的输出，得到最终的情感识别输出 $\hat{}y_E$。具体而言，经过关系建模和张量融合后，Transformer输出一个形状为$3T \times \frac{N}{2}$ 的张量。随后，我们对该张量进行平均池化操作，将其维度从 $3T \times \frac{N}{2}$ 降至 $1 \times \frac{N}{2}$，以减轻计算负担。之后，将池化后的张量输入全连接层进行情绪分类——类别。最后，我们有一个形状为 $1 \times C$的$\hat{}y_E$，其中 C 表示类别数量。$\hat{}y_E$ 的每个元素都代表一种特定的情绪，例如开心或愤怒。一般而言，在NRGF提取器和Transformer的帮助下，我们能够准确估计$\hat{}y_E$，使其受噪声信息的影响大大降低。

3.5 噪声感知学习方案

在设计 NORM-TR 时，采用恰当的学习目标来帮助模型学习抵御噪声信息至关重要。因此，我们提出一种新颖的噪声感知学习方案，用于训练 NORM-TR，使其能够抵御噪声。我们的新颖学习方案会显式添加两种噪声信息来扰乱原始多模态数据，随后引入两种相应的对抗损失，促使 NORM-TR 提供对所添加噪声信息具有不变性的 NRGF。

3.5.1 显式噪声信息



(a) 添加**类型**一 噪声的示例

(b) 添加**类型**二 噪声的示例

图3：添加两种类型噪声的示例。(a) 添加**类型**一 噪声，(b) 添加**类型**二 噪声

我们按照公式 2 所述显式添加噪声信息，因为很难准确定义哪些模式应归为噪声信息、哪些不应归为噪声信息。如果没有显式的噪声信息，我们便无法知道模型何时应对输入序列中模式的变化保持不敏感。因此，我们尝试向多模态输入中显式添加噪声信息。为此，我们随机擦除输入多模态表示中的某些时间段，以移除输入数据中任何潜在的语义。本研究主要考虑两种噪声信息：（**类型**一）所有输入多模态数据都包含随机噪声信息（见图 3(a)）；以及（**类型 2**）只有特定模态的输入数据包含随机噪声信息（见图 3(b)）。在第一种情况下进行训练，有助于 NORM-TR 总结通用且全局一致的语义；第二种情况则有助于 NORM-TR 专注于提升其抵御各特定模态噪声信息的鲁棒性。我们的实验结果验证了添加这两种噪声信息的重要性。实际操作中，我们通过屏蔽输入信息来实现公式2。例如，添加 Type-1噪声时，我们为每个模态预先计算的多模态特征向量生成一个掩码。在每个掩码中，我们随机采样一个长度在 0 到 $\frac{T}{2}$ 个时间步之间的时间窗口，然后将该时间窗口内的值设为 0，同时将此掩码中的其余值设为 1。因此，每个模态都有各自的掩码进行处理。随后，我们将这些掩码与预先计算的多模态特征向量 U相乘，从而擦除该掩码所采样时间窗口对应时段中包含的语义。此外，添加 Type-2噪声时，语义擦除操作与添加 Type-1噪声的过程类似，但掩码的生成方式不同。在这种情况下，我们只生成一个掩码，并随机采样一个包含 0 的时间窗口。我们将该掩码与随机选取的某个模态的预先计算特征相乘，从而仅向该模态添加噪声。

如图4，添加噪声信息后，我们分别使用两个对抗损失函数来定义针对两种添加噪声信息的学习目标。我们采用对抗损失函数，因为它们能够有效地使模型生成高
具有某些所需模式的维度输出，例如，专业-

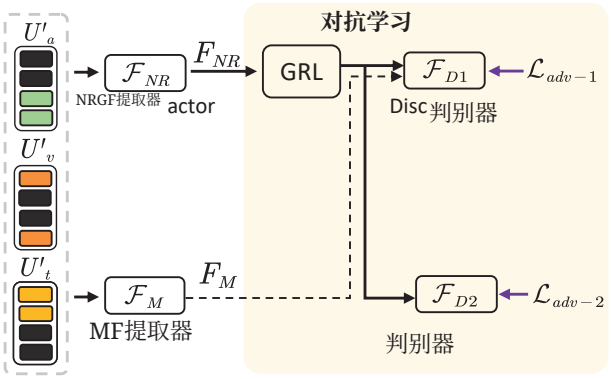


图4：噪声感知学习方案的流程

生成一幅虚假的二维图像，其图像模式与目标真实二维图像非常相似。为了实现基于对抗的学习，我们采用判别器来学习区分生成输出是否未能获得目标模式。判别器能够轻易识别生成失败的输出，而成功生成的输出则会使判别器难以区分。详细的生成对抗学习方案参见Goodfellow 等人(2014)。

3.5.2 基于对抗学习的学习方案

在此，我们尝试利用对抗损失，使 NORM-TR 学会生成不受噪声影响的 NRGF。通过对抗学习抵御添加的噪声信息所带来的影响，NRGF 提取器能够更有效地从输入数据中提取具有语义意义的特征，从而进一步提升情绪识别的鲁棒性。具体而言，我们使用符号 $\mathcal{F}_{D1}$、 $\mathcal{F}_{D2}$ 分别表示类型一和类型二添加噪声的判别器。图4 展示了噪声感知学习方案的详细对抗学习结构。

更具体地说，由于 Type-1 噪声会随机添加到所有多模态输入中，因此 NRGF提取器应从受噪声干扰的多模态输入中找出通用且普遍的模式，使提取出的 NRGF 不受 Type-1 噪声影响。为此，我们采用一个判别器 $\mathcal{F}_{D1}$ ，并使其学习判断输入特征是否包含模态-

依赖模式：如果输入到 $\mathcal{F}_{D1}$ 的是来自多模态特征 F_M 的模态特定特征，我们就让该判别器学习预测此特征代表哪种模态；如果输入到 $\mathcal{F}_{D1}$ 的是 NRGF，我们就让 NRGF提取器混淆该判别器。因此，当 NRGF提取器无法提供通用且一致的特征时，它就无法有效地混淆判别器。通过混淆判别器，提取出的 NRGF 将能够代表整个多模态输入，而非某一特定模态；后者更容易受到所添加的 Type-1 噪声影响。总之，假设 NRGF提取器的模型参数为 θ_{NR}，判别器 $\mathcal{F}_{D1}$ 的模型参数为 θ_{D1}。那么，我们的噪声感知学习方案的第一个目标可以通过针对 Type-1 添加的噪声信息的对抗损失 $\mathcal{L}_{adv-1}$ 的优化来描述：

$$\min_{\theta_{D1}}\max_{\theta_{NR}}\mathcal{L}_{adv-1}=-\frac{1}{N_b}\sum_{i=0}^{N_b}y_M^i\cdot log\,\mathcal{F}_{D1}(F_{NR}/F_M;\theta_{D1}), \tag{6}$$

其中，N_b 是训练集中的样本数，y_M^i表示指明 F_{NR} 或F_M 来自哪个模态的标签，而 F_{NR} 是根据参数 θ_{NR}提取的 NRGF。为使 NRGF 与用于识别模态相关特征的 $\mathcal{F}_{D1}$ 兼容，我们额外添加了 3 个多层感知机，分别将$3T\times\frac{N}{2}$ 的 F_{NR} 映射到与 3 种输入模态相关的特征中，其形状为 $3\times(T\cdot\frac{N}{2})$。

针对 Type-2噪声，我们也采用判别器 $\mathcal{F}_{D2}$ 来帮助训练 NRGF提取器。与 Type-1噪声不同，我们仅向一个随机模态添加噪声。针对 Type-2噪声进行学习，有助于 NORM-TR 在输入中未添加显式噪声的情况下有效运行，同时保持其抵御随机模态中现有噪声的能力。针对 Type-2噪声，我们只需让判别器 $\mathcal{F}_{D2}$ 判断某个模态是否含有噪声：如果输入到 $\mathcal{F}_{D2}$ 的内容与未添加噪声的输入相关，我们就让该判别器预测负标签；否则，就让该判别器预测正标签。因此，如果 NRGF 未能对仅添加到特定模态的噪声保持不变性，NRGF提取器也将无法混淆判别器，而 $\mathcal{F}_{D2}$ 则可以轻易通过预测正标签识别噪声。为此，假设 $\mathcal{F}_{D2}$ 的模型参数为 θ_{D2}。我们得到第二个目标，即针对 Type-2 添加的噪声信息优化对抗损失 $\mathcal{L}_{adv-2}$ 的噪声感知学习方案目标：关于

$$\min_{\theta_{D2}}\max_{\theta_{NR}}\mathcal{L}_{adv-2}=-\frac{1}{N_b}\sum_{i=0}^{N_b}y_N^i\cdot log\,\mathcal{F}_{D2}(F_{NR};\theta_{D2}), \tag{7}$$

其中，y_N^i 是指明哪个模态受到添加的噪声干扰的标签。

在实际应用中，我们使用全连接层实现这两个判别器，每个判别器均由一个全连接层构成。此外，我们采用梯度反转层（GRL）（Ganin和Lempitsky，2015）实现针对 NRGF 提取器的对抗学习，该提取器旨在迷惑这两个判别器。关于NRGF 提取器，该提取器旨在迷惑这两个判别器。

3.6 总体学习目标

总而言之，我们的方法涉及三个学习目标，包括两个对抗损失函数 $\mathcal{L}_{adv-1}$ 和$\mathcal{L}_{adv-2}$，以及一个最终的情绪学习损失 $\mathcal{L}_{er}$。考虑到不同数据集上的情绪标签存在差异，例如，RML（Y. Wang &Guan，2008）和 IEMOCAP（Busso 等人，2008）数据集上的标签是离散的，而 MOSI（Zadeh 等人，2016）和 MOSEI（Zadeh 等人，2018年）数据集上的标签是连续的。因此，我们引入交叉熵损失作为情绪学习损失 $\mathcal{L}_{er}$，用于 RML 和 IEMOCAP 数据集上的分类，并采用均方误差（MSE）作为 $\mathcal{L}_{er}$，用于 MOSI 和 MOSEI 数据集上的回归。情绪学习损失 $\mathcal{L}_{er}$ 可表示为：

$$\mathcal{L}_{er}=\begin{cases}-\frac{1}{N_b}\sum_{i=0}^{N_b}y^i\cdot\log\hat{y}_E^i & \text{for classification}\\ \frac{1}{N_b}\sum_{i=0}^{N_b}\left\|y^i-\hat{y}_E^i\right\|_2^2 & \text{for regression}\end{cases} \tag{8}$$

其中，y^i 是第 i个样本的情绪标签。$\hat{}y_E^i$ 是 NORM-TR 的预测。总目标函数$\mathcal{L}$ 是 $\mathcal{L}_{adv-1}$、$\mathcal{L}_{adv-2}$和 $\mathcal{L}_{er}$之和。数学上，$\mathcal{L}$ 可写为：

$$\mathcal{L}=\mathcal{L}_{adv-1}+\mathcal{L}_{adv-2}+\mathcal{L}_{er}. \tag{9}$$

4 实验

4.1 数据集

我们在三个三模态数据集（MOSI（Zadeh 等人，2016）、MOSEI（Zadeh 等人，2018年）和 IEMOCAP（Busso 等人，2008））以及一个双模态情感数据集（RML（Y. Wang &Guan，2008））上进行了大量实验。这些数据集涵盖了不同的语言和各种场景，如自然场景和实验室场景、对话以及单人演讲。

MOSIMOSI 数据集包含 2,199 个由视频、音频和文本模态构成的多模态序列样本。MOSI 的训练集、验证集和测试集分别包含 1,284、229 和 686 个样本。每个多模态样本都有一个范围为 -3 到 3 的统一标签。-3 和 +3分别表示强烈负面情绪和强烈正面情绪。

MOSEI MOSEI 数据集包含从 YouTube 收集的 22,851 个视频片段，这些片段具有自发的表达、头部姿态、遮挡、光照等特征。在说话人独立的设置下，该数据集被划分为 16,326 个训练样本、1,871 个验证样本和 4,659 个测试样本。每个样本都经过人工标注，情感得分范围为 -3（强烈负面）到 3（强烈正面）。

IEMOCAP IEMOCAP 包含视频、音频和文本转录，以及 12 小时的情境对话视频录制。数据被划分为五个会话，共包含 10,039 个样本和 9 个情绪类别。遵循对比方法（Tsai、Bai等、2019年；Lv等人、2021年），我们使用了四个情绪类别，即即快乐、愤怒、悲伤和中性。数据被划分为 2,717 个训练样本、798 个验证样本和 938 个测试样本。

RML RML 是一个音视频情感数据集，包含 8 个主题和 720 个视频样本，涵盖六种不同的语言（即英语、普通话、乌尔都语、旁遮普语、波斯语和意大利语）。每条数据被标注为六种情绪之一，即，愤怒、厌恶、恐惧、快乐、悲伤和惊讶。在说话人独立的交叉验证设置下，训练集和测试集的比例约为 8:2，这确保了训练集中的说话人不会出现在对应的测试集中。

4.2 实现细节

我们使用 PyTorch 实现了所提出的方法。实验在一台配备 Intel(R) Xeon(R) Gold 6240C CPU（2.60GHz、128GB 内存）和 NVIDIA GeForce RTX 3090 的 PC 上进行。关键训练参数包括初始学习率（0.0001）、用于调整学习率的余弦退火调度、小批量大小（16）以及预热。

对于序列长度设置，我们将 RML 和 IEMOCAP 数据集中的输入序列长度统一为 8（即，$T = 8$），将 MOSI 和 MOSEI 数据集中的输入序列长度统一为 50（即，$T = 50$）。视频模态的生成方式与音频和文本模态不同。具体而言，对于视频模态，我们将输入视频划分为 T 个片段，并从每个片段中随机采样一帧，以形成长度为 T 的视频序列。对于音频和文本模态，我们直接截取数据的前 T 帧作为输入序列。

4.3 预计算特征提取

视频特征 U_v**：** 对于 RML 和 IEMOCAP 数据集，遵循现有方法（Zhao等人，2020），我们采用 ResNet-18（K. He等人，2016）提取 ResNet-18 的最后一个全局平均池化输出，作为预计算的视频特征。对于 MOSI 和 MOSEI 数据集，参考现有方法（Tsai 、 Bai 等，2019年；Hazarika等，2020；Mao等，2022年），我们使用数据集提供的特征，这些特征由 OpenFace（Bal-trusaitis等人，2016）提取。

音频特征 U_a**：** 对于 RML 和 IEMOCAP，我们首先使用 Librosa 计算每个样本的对数梅尔频谱图及其一阶和二阶微分，然后采用 ResNet-18（K. He 等人，2016）提取特征。最后，我们将所有特征向量堆叠起来，获得预计算的音频特征。对于 MOSI 和 MOSEI，我们使用数据集提供的特征，这些特征由 Librosa 提取[1]。

文本特征 U_t**：** 对于 IEMOCAP，我们使用预训练的 BERT（Kenton &Toutanova，2019年）作为特征提取器，将转录后的词序列编码为预先计算好的文本特征。对于 MOSI 和 MOSEI

[1]https://librosa.org

表 1：MOSI 数据集上的比较结果

方法	七分类准确率 (↑)	二分类准确率 (↑)	F1 分数 (↑)	平均绝对误差（↓）	相关系数 (↑)
张量融合网络（Zadeh 等人，2017年）	34.9	- / 80.8	- / 80.7	0.901	0.698
低秩多模态融合（Z. Liu 等人，2018年）	33.2	- / 82.5	- / 82.4	0.917	0.695
MFM(Tsai、Liang等人，2019年)	35.4	- / 81.7	- / 81.6	0.877	0.706
ICCN (Sun等人，2020)	39.0	- / 83.0	- / 83.0	0.877	0.706
MulT (Tsai、Bai等，2019年)	40.0	- / 83.0	- / 82.8	0.871	0.698
MISA (Hazarika等，2020)	42.3	81.8 / 83.4	81.7 / 83.6	0.783	0.761
PMR (Lv等人，2021年)	40.6	- / 83.6	- / 83.4	-	-
Self-MM(Yu等人，2021年)	45.8	84.0 / 86.0	84.4 / 86.0	0.713	0.798
FDMER（Yang等人，2022年）	44.1	- / 84.6	- / 84.7	0.724	0.788
我们的 NORM-TR	**48.5**	**84.3/ 86.1**	**84.4/ 86.2**	**0.698**	**0.808**

注：对于每项评估指标，↑ 表示数值越大越好，↓ 表示数值越小越好。最佳结果以粗体标出

数据集，我们也使用了数据集提供的文本特征，这些特征由 BERT 提取。

4.4 评估指标

在 RML 数据集上，我们选用两种广泛使用的评估指标，即六分类准确率（Acc-6）和 F 分数（F1），以评估模型性能。在 IEMO-CAP 数据集上，我们遵循以往研究（Lv 等人，2021年），报告二分类准确率（Acc-2）以及各情绪类别的 F1 分数。在 MOSI 和 MOSEI 上，参考先前研究（Yuetal.，2020），我们采用六种广泛使用的评估指标，即，二分类准确率（Acc-2）、F1 分数、三分类准确率（Acc-3）、七分类准确率（Acc-7）、平均绝对误差（MAE），以及模型预测与人工判断之间的相关系数（Corr）。具体而言，遵循先前研究（Hazarika等，2020; Yuet al.，2021年），我们在 MOSI 和 MOSEI 数据集上分别采用负面/非负面和负面/正面两种方式计算二分类准确率（Acc-2）和 F1 分数。

此外，为了验证我们的方法对不同噪声强度的鲁棒性，参考先前的研究（Yuan 等人，2021年），我们引入了指标折线图下面积（AUILC），用于评估我们的方法在测试集上的噪声鲁棒性。AUILC 可表示为：

$$\text{AUILC} = \sum_{t} \frac{(x_t + x_{t+1})}{2} \cdot (r_{t+1} - r_t) \qquad (10)$$

其中，x_t 和 x_{t+1} 分别表示在掩码比例为 r_t 和 r_{t+1} 时的第 t和第 $t+1$次评估结果，分别。

4.5 总体性能

4.5.1 MOSI 数据集上的实验

表1 列出了我们提出的方法与最先进方法在 MOSI 数据集上的比较结果。如表所示，我们提出的 NORM-TR 在七分类准确率上比 Self-MM（Yu 等人，2021年）取得的次优结果高出 2.7%。与另一种基于 Transformer 的方法 FD-MER （Yuan 等人，2021年）相比，我们的方法在七分类准确率上相对提升了 9.98%。此外，我们在所有其他指标上也达到了最先进性能，尤其是在难度更高的七分类任务上。我们认为，这一显著提升得益于提取的抗噪特征能够帮助 NORM-TR 在融合过程中抑制无用的噪声信息，从而提升多模态情感识别效果。

4.5.2 MOSEI 数据集上的实验

表2 报告了我们的方法与最先进的方法在 MOSEI 数据集上的比较结果。我们的 NORM-TR 在 Acc-2、F1 分数、Acc-3、Acc-5、平均绝对误差和相关系数方面取得了显著提升。与这些基于 Transformer 的方法，即 FDMER（Yang等人，2022年）和 MulT（Tsai、Bai等，2019年）相比，我们在所有指标上都取得了相对提升，例如，提升了 0.92%

表 2：MOSEI 数据集上的对比结果

方法	七分类准确率 (↑)	二分类准确率 (↑)	F1 分数 (↑)	平均绝对误差（↓）	相关系数 (↑)
张量融合网络（Zadeh 等人，2017年）	50.2	- / 82.5	- / 82.1	0.593	0.700
低秩多模态融合 (Z. Liu 等人，2018年)	48.0	- / 82.0	- / 82.1	0.623	0.677
MFM（Tsai、Liang等人，2019年）	51.3	- / 84.4	- / 84.3	0.568	0.717
ICCN（Sun等人， 2020）	51.6	- / 84.2	- / 84.2	0.565	0.713
MuLT（Tsai、Bai等，2019年）	51.8	- / 82.5	- / 82.3	0.580	0.703
MISA（Hazarika等，2020）	52.2	83.6 / 85.5	83.8 / 85.3	0.555	0.756
PMR（Lv等人，2021年）	52.5	- / 83.3	- / 82.8	-	-
Self-MM（Yu等人，2021年）	53.5	82.8 / 85.2	82.5 / 85.3	0.530	0.765
FDMER (Yang等人，2022年)	54.1	- / 86.1	- / 85.8	0.536	0.773
我们的 NORM-TR	**54.6**	**84.3/ 86.6**	**84.5/ 86.6**	**0.529**	**0.778**

注：对于每个评估指标，↑ 表示数值越大越好，而 ↓ 表示数值越小越好。最佳结果以粗体突出显示

表3：IEMOCAP 数据集上的比较结果

方法	快乐		悲伤		愤怒		中性		平均值	
	二分类准确率	F1	二分类准确率	F1	二分类准确率	F1	二分类准确率	F1	二分类准确率	F1
早期融合长短期记忆网络	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4	69.3	67.7
晚期融合长短期记忆网络	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2	68.4	66.6
RAVEN（Y.Wang 等人，2019年）	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5	67.9	66.5
多模态循环转换网络 (Pham 等人，2019年)	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3	66.7	66.0
MulT（Tsai、Bai等，2019年）	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7	74.7	71.5
PMR（Lv等人，2021年）	86.4	83.3	78.5	75.3	75.0	71.3	63.7	60.9	75.9	72.7
ScaleVLAD（Luo等人，2021年）	86.7	85.9	84.8	84.6	86.8	86.9	72.1	72.1	82.6	82.4
我们的 NORM-TR	**87.7**	**88.5**	**86.2**	**86.4**	**88.6**	**88.6**	**74.8**	**74.3**	**84.3**	**84.5**

注：最佳结果以粗体突出显示

在七分类准确率和 F1 分数上分别提升了 0.93%。在场景更复杂的大规模数据集上取得如此出色的性能，证明了我们的 NORM-TR 能够从各种场景中提取有效的情感信息。

4.5.3 IEMOCAP数据集上的实验

表3 展示了我们的方法与最先进的方法在 IEMOCAP 数据集上的比较结果，包括 MulT（Tsai、Bai等，2019年）、PMR（Lv等人，2021年）和 ScaleVLAD（Luo等人，2021年）。结果表明，我们提出的 NORM-TR 取得了最佳性能，这证明了 NORM-TR 的优越性。与最先进的方法 ScaleVLAD 相比，我们的 NORM-TR 在平均准确率和 F1 分数上分别取得了 2.06% 和 2.55% 的相对提升。此外，在二分类准确率对应的 F1 分数上，我们在全部四个类别中均取得了最先进性能。

4.5.4 RML数据集上的实验

表4 报告了我们的方法与最先进的方法在 RML 数据集上的对比结果。与 MulT（Tsai、Bai等，2019年）取得的次优结果相比，我们的 NORM-TR 在平均准确率上相对提升了 3.23%。与Cornejo与Pedrini（2019年）提出的 Census-Transform 相比，我们的 NORM-TR 实现了更大的相对提升，达到

表4：RML上的比较结果

方法	六分类准确率	F1
Madany等人 （El-Madany等人，2016）	75.00	-
Zhang等人（S. Zhang等人，2018年）	80.36	-
Ma等人（Ma等人，2019年）	80.46	-
Cornejo等（Cornejo与Pedrini，2019年）	82.50	-
张量融合网络* (Zadeh 等人，2017年)	83.19	83.22
MulT* （Tsai、Bai等，2019年）	86.11	85.87
MISA* (Hazarika等，2020)	81.11	80.78
我们的 NORM-TR	**88.89**	**88.81**

注：最佳结果以粗体标出，* 表示该结果由作者复现。

7.75%。这表明我们的方法能够有效处理噪声信息的影响，从而提升 Transformer 的性能。

4.6 噪声数据的鲁棒性评估

为了进一步验证我们的方法对噪声数据的鲁棒性，我们使用不同比例的掩码对测试数据进行评估 r_t，其中 $r_t \in \{0, 0.1, ..., 1.0\}$。表5分别报告了我们的方法和最先进的方法在 MOSI 和 MOSEI 数据集上的 AUILC 结果。显然，在 MOSI 数据集上，我们的 NORM-TR 在几乎所有指标上都取得了更好的表现，即七分类准确率为 31.3%，二分类准确率为 68.6%，平均绝对误差为 1.093，相关系数为 0.506。在 MOSEI 数据集上，我们的方法在所有指标上也取得了显著提升。例如，与 TFR-Net（Yuan 等人，2021年）相比，我们的方法在七分类准确率上相对提升了 1.72%，在二分类准确率上提升了 2.05% / 2.41%，平均绝对误差降低了 1.90%，相关系数提升了 6.60%。这表明我们的方法在面对噪声干扰时具有很强的鲁棒性。

此外，图5展示了 MOSI 数据集上，在不同掩码比例下测试数据的指标曲线。如图所示，在各种掩码比例下，NORM-TR 在几乎所有评估指标上的表现都优于其他方法 $r_t \in \{0, 0.1, \cdots, 0.8\}$，表明我们的 NORM-TR 对噪声具有更强的鲁棒性。

4.7 消融研究与分析

4.7.1 不同组件的影响

为更好地研究所提出的 NORM-TR 中各组件的作用，表6分别报告了在 MOSI 和 MOSEI 数据集上，从 NORM-TR 框架中移除各组件后的消融结果。值得注意的是，所提出的 NORM-TR 已实现最先进性能。如表所示，移除 NRGF/MF提取器会使准确率下降至次优水平，这证明了 NRGF 和 MF提取器对于实现鲁棒的多模态情绪识别至关重要。移除多模态融合 Transformer 后，性能再次下降，这表明多模态融合Transformer 能够有效地建模 NRGF 与多模态特征之间的关系。最后，移除噪声感知学习方案后，性能显著下降，尤其是七分类准确率。这表明噪声感知学习方案有助于模型从多模态数据中学习更有用的情绪语义。此外，我们还观察到，移除 NRGF 提取器所导致的性能下降，比移除多模态特征提取器更为显著。可能的原因是，多模态特征包含更多噪声信息。

4.7.2 不同模态的影响

为探讨各模态对性能的影响，表 7 分别展示了不同模态设置在 MOSI 和 MOSEI 数据集上的消融结果。我们观察到，视频、音频和文本三者的组合取得了最佳性能，这表明我们的模型能够学习有效的多模态情绪表征，从而实现鲁棒的情绪识别。在这两个数据集上，移除文本模态后，模型性能均显著下降，这说明文本模态在多模态情绪识别中发挥着重要作用。

此外，我们尝试分别在 MOSI 数据集的各个模态中加入不同程度的 Type-2 噪声，以探讨各模态对噪声的敏感性。我们发现，在文本模态中加入噪声时，性能下降更为显著，例如，掩码比例为 50% 时，七分类准确率下降了 13%；而在音频和视频模态中加入噪声时，例如，下降了 0.6%

表 5：MOSI 和 MOSEI 数据集上的模型鲁棒性比较

方法	MOSI				MOSEI			
	七分类准确率 (↑)	二分类准确率 (↑)	平均绝对误差 (↓)	相关系数 (↑)	七分类准确率 (↑)	二分类准确率 (↑)	平均绝对误差 (↓)	相关系数 (↑)
张量融合网络 (Zadeh 等人，2017年)	-	- / 60.4	1.327	0.300	-	- / -	-	-
MulT（Tsai、Bai等，2019年）	-	- / 61.8	1.288	0.334	-	- / -	-	-
MISA（Hazarika等，2020）	-	- / 63.2	1.209	0.403	-	- / -	-	-
TFR-Net(Yuan 等人，2021年)	-	- /**69.0**	1.155	0.467	-	- / -	-	-
TFN* (Zadeh 等人，2017年)	22.0	62.6 / 62.6	1.281	0.342	45.1	70.8 / 72.6	0.726	0.409
MulT* (Tsai、Bai等, 2019年)	22.7	63.5 / 64.0	1.265	0.339	46.4	74.5 / 75.5	0.692	0.504
MISA* (Hazarika等,2020)	27.0	65.4 / 65.4	1.181	0.412	44.5	73.3 / 74.5	0.720	0.421
TFR-Net* (Yuan 等人，2021年)	26.8	67.8 / 68.2	1.175	0.445	46.5	73.3 / 74.7	0.686	0.515
我们的 NORM-TR	**31.3**	**68.6** /68.6	**1.093**	**0.506**	**47.3**	**74.8/ 76.5**	**0.673**	**0.549**

注：对于每个评估指标，↑ 表示数值越大越好，↓ 表示数值越小越好。最佳结果以粗体突出显示，* 表示该结果由作者复现

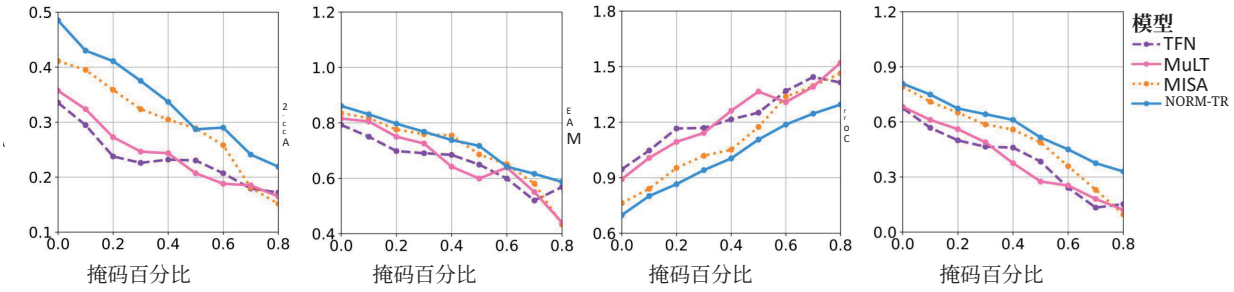


图 5：可视化 MOSI 数据集上不同掩码百分比下测试数据的指标曲线

掩码比例为 50% 时，七分类准确率分别下降了 0.6% 和 1.1%。这表明，与其他模态相比，文本模态对噪声更为敏感。

4.7.3 Transformer 中不同的查询、键和值设置

表 8分别展示了 Transformer 在 MOSI 和 MOSEI 数据集上采用不同查询、键和值设置时的实验结果。我们发现，当使用 NRGF 作为查询、MFs 作为键和值时，模型取得了最佳性能。这表明，我们的 NORM-TR 能够为 Transformer 提取更加通用且有用的查询，从而帮助实现更好的情绪识别性能。

4.7.4 Transformer 中超参数设置的影响

图 6 （a）展示了 Transformer 架构中注意力块数量对 RML 数据集上情绪识别准确率的影响。从结果来看，当深度设为 2 时，准确率达到最高值 88.89%。此外，我们还观察到，不同的 Transformer 深度仅导致性能出现较小变化，这表明该参数对我们的方法影响较小。

此外，我们尝试在 RML 数据集上使用更复杂的 Transformer 模型代替我们的多模态融合Transformer进行融合，例如，1）使用类似于 MulT 的 Transformer 对（Tsai、Bai等，2019年）；2）将 NRGF 和多模态特征拼接起来，并使用更深层的 ViT（Dosovitskiy 等，2021年）进行融合。我们发现，这些复杂模型并未提升性能，即，在 RML 数据集上获得了 88.75% 和 88.19% 的六分类准确率

表6：所提出的 NORM-TR 的消融研究。分别在 MOSI 和 MOSEI 数据集上，分析移除 NORM-TR 中不同组件的影响

方法	MOSI		MOSEI	
	七分类准确率 (↑)	平均绝对误差 (↓)	七分类准确率 (↑)	平均绝对误差 (↓)
NORM-TR	**48.5**	**0.698**	**54.6**	**0.528**
不使用 NRGF提取器	46.8	0.726	52.3	0.547
不使用 MF提取器	48.1	0.706	53.6	0.534
无多模态融合Transformer	45.6	0.734	53.8	0.536
无噪声感知学习方案	46.9	0.721	53.2	0.544

注意：对于每个评估指标，↑ 表示数值越大越好，而 ↓ 表示数值越小越好。最佳结果以粗体标出

表 7：不同模态的影响

方法	MOSI		MOSEI	
	七分类准确率 (↑)	平均绝对误差 (↓)	七分类准确率 (↑)	平均绝对误差 (↓)
NORM-TR	**48.5**	**0.698**	**54.6**	**0.528**
不含音频	43.6	0.761	50.6	0.655
无视频	43.3	0.765	51.8	0.592
无文本	18.1	1.410	41.2	0.831

注：对于每个评估指标，↑ 表示数值越大越好，↓ 表示数值越小越好。最佳结果以粗体突出显示。

表 8：Transformer 中不同查询、键和值设置的影响

Q K &V	MOSI		MOSEI	
	七分类准确率 (↑)	平均绝对误差 (↓)	七分类准确率 (↑)	平均绝对误差 (↓)
多模态特征 NRGF	47.1	0.706	53.2	0.539
NRGF 多模态特征	**48.5**	**0.698**	**54.6**	**0.529**

注：对于每项评估指标，↑ 表示越大越好，而 ↓ 表示越小越好。最佳结果以粗体标出

分别通过 MulT 和 ViT 在数据集上实现。与此同时，这些复杂模型需要更多参数进行训练，并产生额外的计算成本（约为 1 MACs）。我们提出的简单而有效的多模态融合Transformer能够利用包含更多细节的多模态特征，补充 NRGF 中可能不敏感但有用的信息，从而实现更准确的情感理解（六分类准确率达到 88.89%）。

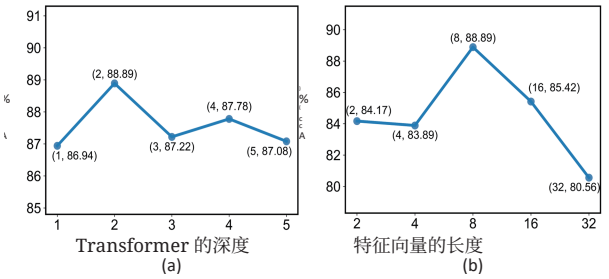
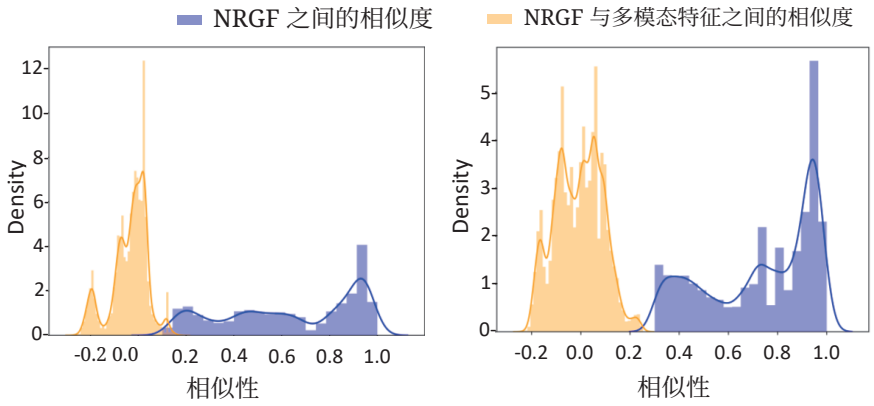


图 6：Transformer 中重要参数设置对 RML 数据集的影响。(a) Transformer 中注意力块的影响，以及 (b) 每个序列采样帧数的影响

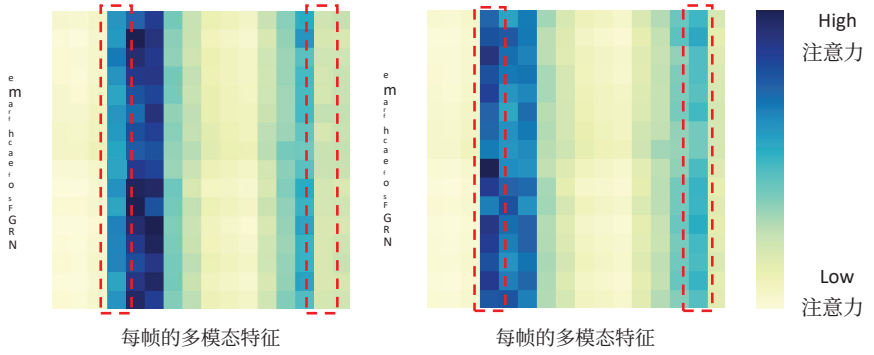
4.7.5 输入特征向量长度的影响

在图 6 （b）中，我们展示了受输入特征向量长度影响的准确率曲线，即从原始序列中采样的帧数。 如图所示，当我们将长度设置为 8 时，准确率达到最高值 88.89%。因此，我们将输入特征向量的长度设置为 8。此外，每个-即，从原始序列中采样的帧数。如图所示，当我们将长度设置为 8 时，准确率达到最高值 88.89 %。因此，我们将输入特征向量的长度设置为 8。此外，每个-



(a) 无噪声感知学习方案　　(b) 采用噪声感知学习方案

图 7：在 MOSI 数据集上，分别对使用和未使用所提出的噪声感知学习方案时的 NRGF 和多模态特征分布进行可视化。(a) 未使用噪声感知学习方案时，MOSI 数据集上的相似度分布；(b) 使用噪声感知学习方案时，MOSI 数据集上的相似度分布



(a) 含噪声　　(b) 无噪声

图 8：在 RML 数据集上，对随机选取样本在有无随机掩码时由多模态融合Transformer学习到的注意力权重进行可视化。(a) 存在掩码噪声时的注意力权重；(b) 不存在掩码噪声时的注意力权重。注：颜色越深表示学习到的注意力权重越高，红色虚线框表示序列中随机选取的两帧对应的注意力分布

当长度设置为 32 时，性能急剧下降，这表明从数据中采样过多的帧往往会引入额外的噪声，导致结果不佳。

4.7.6 抗噪多模态特征分布的可视化

在图 7中，我们对提取出的抗噪通用特征（NRGF）的相似度分布进行了可视化以及 MOSI 数据集上的多模态特征（MFs），分别对应使用和不使用所提出的噪声感知学习方案的情况。我们采用核密度估计（KDE）和余弦相似度来描述两类特征的相似度分布。如图 7 （b）所示，在采用噪声感知学习方案时，大多数 NRGF 彼此相似（相似度接近 1），而大多数 NRGF 与 MFs 彼此不同（相似度接近 0），这表明 NRGF 更具通用性

且与噪声保持一致，而 MFs 包含更多特定于模态的特征。相反，在不采用噪声感知学习方案时，NRGF 之间的相似度显著提高（见图7 （a）），这表明它们保留了来自不同模态的更多噪声信息。

4.7.7 Transformer 学习到的注意力权重可视化

图 8 (a) 和 (b) 分别展示了在 RML 数据集上，Transformer 针对随机选取的样本在随机掩码和无随机掩码情况下学习到的注意力权重矩阵。实际处理中，对于音频和视频数据，我们将掩码帧的预计算特征向量 U_a， U_v 设为 0。由于 NRGF 和多模态特征的形状为$2T \times \frac{N}{2}$，其中 $T = 8$，因此权重矩阵的大小为 16×16。矩阵中的每个方格表示 Transformer 学习到的 NRGF 与多模态特征对应帧之间的注意力权重分数。红色虚线框显示了矩阵中随机选取的两个帧的注意力分布。与未掩码帧的注意力权重（图 8(a) 中的方格颜色较浅）相比，掩码帧的注意力权重显著降低（图 8(b) 中的方格颜色较深）。这表明，Transformer 倾向于将未掩码帧的信息作为重点，通过提取 NRGF 作为查询、多模态特征作为键和值，而不是关注带有掩码的噪声帧，从而抑制噪声信息带来的负面影响。

5 结论

本文提出了一种用于多模态情绪识别的新型噪声鲁棒多模态Transformer（NORM-TR）方法。NORM-TR由噪声鲁棒通用特征（NRGF）提取器、多模态特征（MF）提取器以及用于融合NRGF和多模态特征的多模态融合Transformer组成，从而显著降低噪声对多模态数据的负面影响。为此，本文进一步设计了一种新型噪声感知学习方案，以帮助适当优化NORM-TR，从而获得噪声不变情感表示。在包括MOSI、MOSEI、IEMOCAP和RML在内的多个多模态数据集上进行的大量实验，结果表明，我们的方法优于其他方法，证明了处理噪声信息的重要性以及我们方法的有效性。尽管我们的方法效果显著，但我们发现，该方法尚未解决情感标签不一致和缺失的问题。未来，我们将把更先进的半监督或自监督学习机制引入该方法，从无标签数据中学习，从而实现更稳健的情感理解。

参考文献

Baltrusaitis, T., Robinson, P., & Morency, L.（2016）。Openface：一款开源面部行为分析工具包。收录于*2016*年 *IEEE* 冬季计算机视觉应用会议（*WACV*）（第 1–10 页）。IEEE计算机学会。

Beale, R., & Peter, C.（2008）。情感和情绪在人机交互中的作用。收录于人机交互中的情感与情绪（第 4868 卷，第 1–11 页）。施普林格。

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S.（2008）。IEMOCAP：交互式情绪双人动作捕捉数据库。语言资源与评估，*42*（4），335–359。

Cornejo, J. Y. R., & Pedrini, H.（2019年）。基于人口普查变换的混合深度卷积神经网络视听情绪识别。收录于 *2019*年 *IEEE* 国际系统、人类与控制论会议（第 3396–3402 页）。IEEE。

Ding, L., Wang, L., Liu, X., Wong, D. F., Tao, D., & Tu, Z. (2021 年). 理解并改进非自回归翻译中的词汇选择。载于国际学习表征会议（*ICLR*）。

D'Mello, S. K., &Kory, J. M. (2015). 多模态情感检测系统综述与荟萃分析。*ACM*计算调查（*CsUR*）,*47*(3), 1–36.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N.（2021）。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于*ICLR*。

El-Madany, N. E., He, Y., & Guan, L.（2016）。通过多集局部保持典型相关分析实现多视图情绪识别。载于*IEEE* 电路与系统国际研讨会（*ISCAS*）（第590–593 页）。IEEE。

Ganin, Y., & Lempitsky, V. S. (2015). 通过反向传播进行无监督领域适应。载于 国际机器学习会议（第 37 卷，第 1180–1189 页）。JMLR。

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . .. Bengio, Y. (2014). 生成对抗网络。载于 神经信息处理系统进展（第 27 卷，第 2672–2680 页）。

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA：用于多模态情感分析的模态不变与特定表示。载于 第 *28* 届 *ACM*国际多媒体会议论文集（第 1122–1131 页）。ACM。

He, K., Zhang, X., Ren, S., & Sun, J. (2016). 用于图像识别的深度残差学习。载于 *2016 IEEE* 计算机视觉与模式识别会议（*CVPR*）（第 770–778 页）。

He, Z., Zhang, L., Gao, X., & Zhang, D. (2022年). 面向非受限目标检测的多对抗 Faster R-CNN 与范式教师。*International Journal of Computer Vision*, 1–21。

Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020). 用于连续情绪识别的多模态 Transformer 融合。载于*2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）（第 3507–3511 页）。IEEE。

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019年). BERT：用于语言理解的深度双向 Transformer 预训练。载于*NAACL-HLT* 会议论文集（第 4171–4186 页）。

Li, C., Deng, C., Li, N., Liu, W., Gao, X., & Tao, D.（2018年）。用于跨模态检索的自监督对抗哈希网络。载于*2018 IEEE/CVF*计算机视觉与模式识别会议（第4242–4251页）。

Liang, J., Li, R., & Jin, Q.（2020）。采用跨模态分布匹配的半监督多模态情绪识别。载于第二十八届*ACM*国际多媒体会议论文集（第2852–2861页）。ACM。

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., & Shan, S. (2022年). MAFW：一个大规模、多模态复合情感数据库，用于野外动态面部表情识别。载于（第 24–32 页）。ACM。

Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., & Zhan, Y.（2021年）。用于鲁棒的基于视频的面部表情识别的表情片段变换器。相关系数，*abs/2109.08409*。

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L.（2018年）。利用模态特定因子实现高效的低秩多模态融合。在第*56*届计算语言学协会年度会议论文集，*ACL 2018*，澳大利亚墨尔本，*2018*年*7*月*15*日至*20*日，第*1*卷：长篇论文（第2247–2256页）。计算语言学协会。

Luo, H., Ji, L., Huang, Y., Wang, B., Ji, S., & Li, T.（2021年）。ScaleVLAD：通过对局部描述子进行多尺度融合改进多模态情感分析。*arXiv*预印本 *arXiv:2112.01368*。

Lv, F., Chen, X., Huang, Y., Duan, L., & Lin, G. (2021年). 面向未对齐多模态序列的人类多模态情绪识别的渐进式模态强化。在 *2021*年 *ieee/cvf* 计算机视觉与模式识别会议 *(cvpr)*（第2554–2562页）。计算机视觉基金会 / IEEE。

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Kosir, A. (2019年). 视听情绪融合（AVEF）：一种深度高效的加权方法。信息融合，*46*，184–192。

Maas, A. L., Hannun, A. Y., Ng, A. Y., 等（2013年）。整流非线性提升了神经网络声学模型的性能。载于*ICML* 论文集（第30卷，第3页）。

Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., & Gao, K.（2022年）。M-SENA：用于多模态情感分析的集成平台。载于计算语言学协会第*60*届年会论文集（第204–213页）。计算语言学协会。

Pei, Z., Cao, Z., Long, M., & Wang, J.（2018年）。多重对抗域自适应。载于*AAAI*人工智能大会论文集（第32卷，第3934–3941页）。AAAIPress。

Pham, H., Liang, P. P., Manzini, T., Morency, L., & P´oczos, B.（2019年）。翻译中发现：通过模态间循环翻译学习鲁棒的联合表示。载于人工智能 *AAAI* 会议论文集 （第 33 卷，第 6892–6899 页）。AAAI出版社。Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L.（2019年）。EARS：基于混合信息融合的情绪感知推荐系统。信息融合，*46*，141–146。Sahay, S., Okur, E., Kumar, S. H., & Nachman, L.（2020）。用于多模态序列的低秩融合 Transformer。*arXiv* 预印本 *arXiv:2007.02038*。Shen, L., Wang, M., & Shen, R.（2009）。情感化电子学习：利用“情绪”数据改进泛在学习环境中的学习。教育技术与社会期刊，*12*（2），176–189。Sun, Z., Sarma, P. K., Sethares, W. A., & Liang, Y.（2020）。通过深度典型相关分析学习文本、音频和视频之间的关系，以进行多模态语言分析。载于人工智能 *AAAI* 会议论文集 （第 34 卷，第 8992–8999 页）。AAAI出版社。Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R.（2019年）。用于未对齐多模态语言序列的多模态 Transformer。载于第 *57* 届计算语言学协会会议论文集，*ACL 2019*，意大利佛罗伦萨，*2019* 年 *7* 月 *28* 日至 *8* 月 *2* 日，第 *1* 卷：长篇论文（第 6558–6569 页）。计算语言学协会。

Tsai, Y. H., Liang, P. P., Zadeh, A., Morency, L., & Salakhutdinov, R.（2019年）。学习因子化多模态表示。载于*ICLR*。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., J ones, L., Gomez, A. N., . . . Polosukhin, I.（2017年）。注意力就是你所需要的一切。载于神经信息处理系统进展（第30卷，第5998–6008页）。

Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T.（2017年）。对抗式跨模态检索。载于第*25*届*ACM*多媒体国际会议论文集（第154–162页）。ACM。

Wang, Y., & Guan, L. (2008). 从视听信号中识别人类情绪状态。*IEEE*多媒体汇刊, *10*(5), 936–946。Wang, Y., Herranz, L., & van de Weijer, J. (2020). 混合匹配网络：用于零样本图像到图像转换的跨模态对齐。国际计算机视觉期刊, *128*(12), 2849–2872。Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. (2019). 词语会产生偏移：利用非语言行为动态调整词语表示。载于*AAAI*人工智能大会论文集（第33卷，第7216–7223页）。AAAI出版社。Yang, D., Huang, S., Kuang, H., Du, Y., & Zhang, L. (2022). 面向多模态情绪识别的解耦表示学习。载于第*30*届*ACM*国际多媒体会议论文集 （第1642–1651页）。ACM。Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., . . . Yang, K. (2020). CH-SIMS：一个带有模态细粒度标注的中文多模态情感分析数据集。载于第*58*届计算语言学协会年会论文集 （第3718–3727页）。计算语言学协会。Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). 通过自监督多任务学习为多模态情感分析学习模态特定表示。载于*AAAI*人工智能大会论文集（第35卷，第10790–10797页）。

Yuan, Z., Li, W., Xu, H., & Yu, W. (2021年). 用于鲁棒多模态情感分析的基于 Transformer 的特征重建网络. 载于 第*29*届*ACM*国际多媒体会议论文集（第4400–4407页）. ACM.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.（2017）。用于多模态情感分析的张量融合网络。在自然语言处理中的经验方法*2017*年会议论文集（第1103–1114页）。计算语言学协会。

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L.（2018）。开放环境中的多模态语言分析：CMU-MOSEI数据集与可解释的动态融合图。在第*56*届计算语言学协会年会论文集（第2236–2246页）。

Zadeh, A., Zellers, R., Pincus, E., & Morency, L.（2016）。视频中的多模态情感强度分析：面部动作与口头信息。*IEEE*智能系统，*31*（6），82–88。Zhang, Q., Xu, Y., Zhang, J., & Tao, D.（2022）。Vitaev2：通过探索归纳偏置推进视觉Transformer，用于图像识别及其他任务。相关系数，*abs/2202.10108*。Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q.（2018）。利用混合深度模型学习用于视听情绪识别的情感特征。*IEEE*电路与系统视频技术汇刊，*28*（10），3030–3043。Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., . . . Keutzer, K.（2020）。用于用户生成视频情绪识别的端到端视听注意力网络。在*AAAI*人工智能会议论文集（第34卷，第303–311页）。AAAI出版社。Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q., & Lee, C.（2021）。在注意力网络中使用自适应多级因子分解双线性池化进行视听情绪识别的信息融合。*IEEE/ACM*音频、语音与语言处理汇刊，*29*，2617–2629。