AGSP-DSA：一种采用动态语义对齐的自适应图信号处理框架，用于实现稳健的多模态融合

KV Karthikeya、Dr. Ashok Kumar Das、Dr. Shantanu Pal、Dr. Vivekananda Bhat K 和 Dr. Arun Sekar - Rajasekaran

[1]

摘要——**本文提出一种带动态语义对齐的自适应图信号处理（AGSP- DSA）框架，用于对包括文本、音频和图像在内的异构来源进行稳健的多模态数据融合。该方法采用双图构建来学习模态内和模态间关系，通过谱图滤波增强信息丰富的信号，并利用多尺度图卷积网络（GCN）进行有效的节点嵌入。语义感知注意力机制使各模态能够根据上下文相关性动态地为上下文作出贡献。在包括 CMU-MOSEI、AVE 和 MM-IMDB 在内的三个基准数据集上的实验结果表明，AGSP-DSA 达到了当前最先进水平。具体而言，该方法在 CMU-MOSEI 上取得了 95.3% 的准确率、0.936 的 F1分数和 0.924 的平均精度均值，准确率较 MM-GNN 提高了 2.6 个百分点。在 AVE 上，该方法取得了 93.4% 的准确率和 0.911 的 F1分数；在 MM-IMDB 上取得了 91.8% 的准确率和 0.886 的 F1分数，展现出在模态缺失设置下良好的泛化能力和鲁棒性。这些发现验证了 AGSP-DSA 在推动情感分析、事件识别和多媒体分类中的多模态学习方面的有效性。**

索引词—**自适应图信号处理、图卷积网络、多模态融合、语义注意力、情感分析、事件识别、多媒体分类。**

I. 引言

The 文本、图像和音频等异构数据源日益普及，极大丰富了多媒体内容，为情感分析、事件识别和跨模态检索等应用中的智能分析带来了新的可能。深度学习驱动的多媒体系统利用多种类型的信息，依赖于同时处理来自不同模态的数据，以改进其决策[1], [2]。

传统上，早期融合方法通过汇集特征信息，或后期融合方法通过组合分类决策，都难以处理不同数据源之间的复杂关系。大多数此类方法都假设特征始终以相同方式发挥作用，不会对上下文含义或所收集数据类型的变化作出反应[3]。此外，只有少数方法认识到比较不同样本的相关性，因为在部分数据缺失时，这有助于更有效地引导模型。

基于图的学习被引入作为应对这些挑战的一种方案，因为它能够呈现数据中存在的结构和上下文[4]。图信号处理使我们能够处理各个图节点上的信号，并利用其谱特性实现可靠的滤波和网络学习。因此，可以利用图为每个样本融合来自不同来源的信息，并通过连接节点将相似样本关联起来[5]。

本文提出采用自适应图信号处理（AGSP），将图信号处理与图卷积网络相结合，以实现对来自不同来源的数据进行稳健且灵活的融合。现有方法要么侧重于模态，要么侧重于图的结构。AGSP通过同时获取各模态的信号及图的细节，改善特征表示。本研究的主要贡献如下。

- 我们提出采用完整的AGSP框架融合多模态信号，因为该框架结合谱图滤波和深层图卷积网络进行数据表示与学习。
- 我们的方法能够动态平衡各模态的影响，同时保留不同样本之间的联系。
- 我们在多个基准数据集（MM-IMDb、VGGSound 和 MELD）上，对情感分析、事件识别和跨模态检索这三项任务进行了 AGSP 框架的测试。该框架取得了良好结果，表现优于已知的融合方法。
- 我们的实验和深入研究证明，该模型能够稳健应对来自不同来源的信息缺失。

在本文其余部分中，第二节将讨论多模态融合和基于图的学习相关研究，并介绍 AGSP 的方法以及实验方案。第四节将列出评估标准和系统性能结果。最后，

[1]本研究未获得公共、商业或非营利部门资助机构的任何专项拨款。

（通讯作者：K. V. Karthikeya。）

First A KV Karthikeya 任职于印度海得拉巴的 AT&T 首席信息办公室（电子邮件：kvkarthikeya02@gmail.com）

第五节将总结本文工作，并详细说明后续研究方向。

II. 文献综述

多模态融合有助于在多媒体系统中整合图像、文本和声音。早期融合和后期融合效率较高，但可能难以捕捉不同类型数据之间的深层关系。张量融合和基于注意力的融合方法日益受到重视，这些方法有助于筛选出对融合最重要的特征。尽管近期的深度多模态融合模型使用变换器和交叉注意力，在共享层中整合不同模态，但当某个输入缺失或输入的重要性发生变化时，这些模型的表现并不理想。因此，有必要探索图融合等自适应方案。

作者提出的方法[6] 采用跨模态注意力与语义图嵌入来解决多标签分类问题。利用图结构，服务器端能够纳入客户端标签之间的特殊关系，从而实现更优的跨模态依赖共享。将基于图的知识与基于注意力的特征对齐相结合，可提升复杂分类数据集上的表现。Xi等人[7], 提出使用图注意力网络层，在跨模态匹配过程中关联图像与文本的含义。该方法构建跨模态图，并运用注意力机制找出不同模态输入之间共享的含义。他们的研究显著推动了图像和文本检索的发展，表明图神经网络能够有效匹配不同类型的语义信息。

Zhang等人提出了TCTFusion，这是一种跨模态变换器，能够自适应地融合可见光和红外光谱中的图像 [8]。该模型采用专用变换器和一个融合变换器，以确保各类特征得到均衡融合。这种方法在图像融合方面取得了领先成果，证实了多分支变换器适用于多传感器数据。Hu和Yamamura [9] 提出了全局-局部融合神经网络方法，用于从多种类型数据的不同上下文中提取特征，以进行多模态情感分析。网络中的注意力机制使其能够调整局部活动和全局语义的重要性，从而以强大而灵活的方式整合不同类型的信息。标准情感数据集上的测试表明，该方法的准确率表现出色。

Liu等人 [10] 提出了一种分层注意力融合网络，通过在时间和空间维度上运用注意力机制，对多个层级的特征进行划分和学习，以助力视频情绪识别。该网络能够从视觉信息和音频的组合中提取情绪。与传统的长短期记忆网络和卷积神经网络相比，这些网络在处理长程视频特征方面表现更佳。He及其合作者 [11] 开发了一种多模态互注意力框架，专注于复杂情境下的情感分析。借助这一框架，他们的模型能够学会在任务的不同阶段关注各种模态。这种设计增强了模型对模糊或含噪数据的鲁棒性，因此适用于情感预测的实际应用。

在论文中，Yu等人 [12] 提出了多模态Transformer和多视图视觉表征，以提升图像描述的性能。他们利用图像的多个视图，并通过特征嵌入和注意力将其与文本关联起来。因此，生成过程更加流畅可靠，使该模型在生成准确且内容丰富的描述时优于标准方法。Tahir等人 [13] 深入研究了图信号处理（GSP），并阐述了其在图像分析、传感器网络和语音处理中的数据分析应用。论文重点讨论了图构建、谱域处理以及应对计算挑战。他们的工作有助于推动GSP在结构化数据和多模态数据类型中的应用。

在 Yao 等人的 [14] 研究中，他们提出了一种自适应同质性图学习方法，用于高光谱图像聚类。该方法通过在训练过程中改变图的构建方式，解决光照不同和区域相似带来的问题。使用自适应滤波器后，聚类结果更加充分，模型也能适用于不同情境。Liu 等人 [15] 提出了一种音频-视觉元音图注意力网络，用于自动评估汉语构音障碍。该系统将两种类型的特征结合起来，构建了一个侧重关注特定音素的图结构。将模型结果与专家评估进行比较后发现，图神经网络可能有助于医学语音诊断。

Wu 等人 [16] 汇总了有关图神经网络（GNN）的信息，并介绍了 GCN、GAT 和 GraphSAGE 等几种重要模型，同时说明了此类网络在自然语言处理、计算机视觉和其他多模态任务中的应用。文中还提到了这种方法面临的问题，例如过平滑和可扩展性，并建议未来的研究致力于学习动态且可调整的图数据。任何想了解图上的深度学习的人都应参考这篇文章。

三、 Proposed Methodology

本节提出一种用于融合多种类型数据的新方法，称为带动态语义对齐的自适应图信号处理（AGSP-DSA）。不同于一次只考察特征交互，我们的方法通过两阶段流水线 [17] 同时处理模态内信息和模态间信息。其关键创新在于结合图谱滤波与动态语义图对齐，从而在两种模态之间选取相关特征，同时不改变各张图像的风格。该处理流程包含五个主要模块：从多个来源进行特征编码、构建两张图、对齐两者的频谱以进行比较、生成节点

嵌入（通过多尺度图卷积网络），以及使用语义感知注意力融合结果。

*A. AGSP*框架概述

首先，使用多个基础编码器处理原始数据，从图像、文本和音频中生成特征向量。随后，利用这些特征向量构建多模态图，其中每个节点表示一个数据实例，边表示两个实例之间的关联程度 [18]。之后，使用谱滤波器处理图信号，以去除噪声并增强有价值模式的重要性。接着，使用多层图卷积网络进一步处理这些信号。最后，选择采用注意力或门控的方法来融合特征，以进行处理流程中的下一步预测。AGSP-DSA流程的第一步是接收来自多个输入模态的数据样本xi（m），这些输入模态记为Rdm，并由i 和m 表示（例如，m=1 表示图像，m=2 表示音频，m=3 表示文本）。每个模态特定特征都经过一个线性投影层，以映射到与其他特征相同的潜在空间中R^d

$$\{\tilde{x}\}_i^{\{(m)\}} = W^{\{(m)\}\{x\}_i^{\{(m)\}}} + b(m) \qquad (1)$$

其中，$W(m)$ 是 $Rd\times dm$ 中的可学习权重矩阵，而 $b(m)$ 是偏置项。这些表示被输入到两个构建图的部分，即模态内图 $G(m)$ 和模态间语义图 Gcm，以帮助收集局部细节并实现不同组之间的语义共享 [19]。

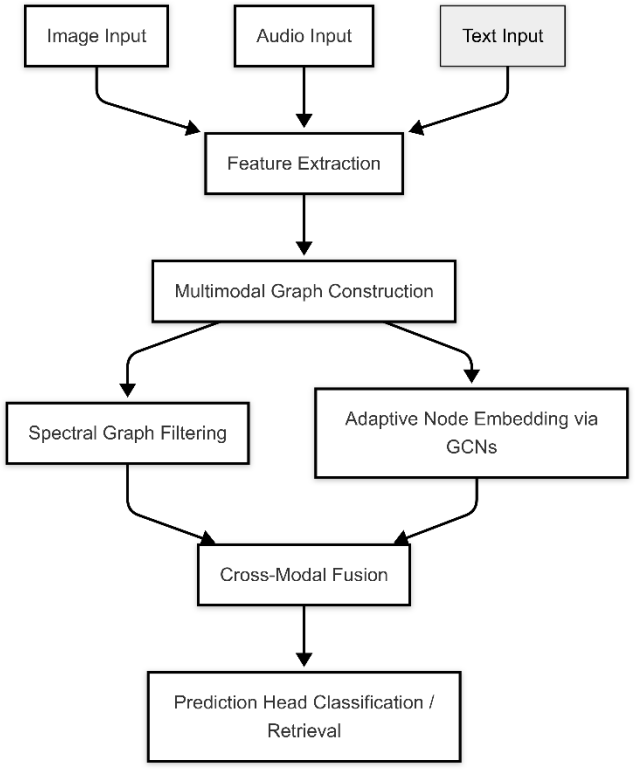


图1。所提出的 AGSP-DSA 框架的架构。

如图1所示，所提出的用于多模态数据融合的AGSP-DSA框架以完整架构的形式进行了说明。该过程首先分别接收图像、音频和文本，随后将其输入专门的特征提取器，例如用于图像的 ResNet、用于音频的卷积神经网络以及用于文本的 BERT，以获得高维嵌入。这些嵌入被发送到下一阶段，在该阶段构建双图，包括一个用于表示每种数据类型之间相似性的图，以及另一个用于展示不同数据类型之间关系的图 [20]。图构建完成后，输入被送入称为谱图滤波和基于图卷积网络的自适应节点嵌入（GCN）的两个分支中进行处理。图神经网络通过对图的谱进行滤波来去除噪声并增强所需信号，同时利用其邻域信息微调节点表示。每个分支的结果通过跨模态融合模块，利用与输入语义相关的注意力机制进行融合。融合完成后，结果被发送至预测头，用于分类或检索访问。该框架能够有效跟踪不同媒体中各个部分及其语义之间的联系，因此能够很好地适应缺失数据，并适用于深度学习分析。

B. 双图构建

该框架设计了两个图，用于描述各模态内部的连接以及模态之间的关联。第一部分针对每种输入类型（图像、音频、文本）构建一个图 $G(m)$，其中每个节点表示一个数据实例，而边则通过计算余弦相似度 [21] 来表示两个实例之间的相似程度。第二部分是一个名为 Gcm 的语义图，它根据共享标签或通过比较预训练编码器得到的表示来衡量相似度，从而连接任意输入类型中含义相同的样本。得益于这一构建方式，各模态的结构得以保留，并且一种表示中的有用语义信息能够在训练过程中传播到另一种表示中。为了保留各模态内部及模态之间的结构，我们构建了两个图。关于模态特定图的详情

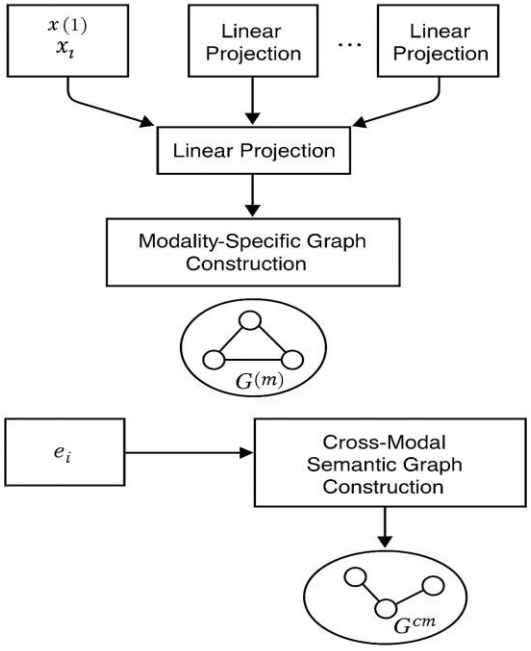
$G(m)$=($V(m)$,$E(m)$) 通过计算同一类别样本之间的余弦相似度构建。

$$A_{ij}^{(m)} = \frac{\tilde{x}_i^{(m)} . \tilde{x}_j^{(m)}}{\|\tilde{x}_i^{(m)}\|_2 \|\tilde{x}_j^{(m)}\|_2} \cdot I[\| \tilde{x}_i^{(m)} - \tilde{x}_j^{(m)} \|_2^2 < \epsilon] \qquad (2)$$

在此部分，$I[\cdot]$ 确保只有由ϵ定义的、彼此配对且具有足够相似度的项目才会成为图的一部分。在跨模态图Gcm 中使用共同标签，通过高斯核计算语义相似度，并根据这些嵌入构建邻接矩阵。

$A_{ij}^{cm} = exp(-\frac{\|ei-ej\|_2^2}{\sigma^2})$ (3)

此处， σ表示一个可调整的带宽。借助这种表示，我们的模型能够在每个模态内部以及不同模态之间传递特征，同时识别所有模态共有的细微关系，以及某一[22]特有的关系。



.

图2. 双图构建

图2展示了双图的构建过程，这是 AGSP-DSA 方法学习鲁棒多模态特征的关键。该任务首先按照顺序获取特征向量 x1(1)、x1(2)、……，并将其映射自图像、声音和文本等不同媒体来源。通过可学习的线性投影层，将所有模态的特征投影到同一潜在空间中，使每个模态具有相同数量的维度。随后，利用预测得到的特征构建两种不同类型的图。第一种是表示每个模态路径的图 $G(m)$。

此类方法通过在给定阈值内使用余弦相似度或欧几里得距离比较相似性，将来自同类信息的数据点进行分组。这种结构确保模型不会丢失每个模态的局部结构 [23]。与此同时，第二个组件构建语义图 Gcm，其中节点为数据实例，边则基于标签嵌入 ei 表示不同模态之间的语义相似性，即在特定上下文中的理解。类似 $G(m)$ 和 Gcm 这样的结构至关重要，因为它们支持在后续阶段使用谱滤波器和图卷积，确保每种数据模态在含义上既保持一致又实现对齐。

C. 谱图对齐

谱图滤波有助于增强特征，因为它既应用于特定模态图，也应用于跨模态图。进行谱滤波时，图信号会被投影到其拉普拉斯特征基上，并在重新组合信号之前，将 g(Λ) 应用于系数。因此，模型可以聚焦于重要信号，并削弱频谱中的干扰噪声。为提高这种方法的效率，我们使用切比雪夫多项式近似滤波函数，因此无需进行特征分解。图构建完成后，我们使用图信号处理（GSP）提升节点特征的质量。根据邻接矩阵 A，以下公式定义了归一化图拉普拉斯矩阵：

p

$L = I - D^{-1/2}AD^{-1/2}$ (4)

D 是一个对角度矩阵，且对于所有 i，Di=∑$jAij$。让我们对拉普拉斯矩阵中的每一项同时乘以并除以某个值，以简化表达式。

$L = U\Lambda U\top$ (5)

在这种情况下，U 表示特征向量矩阵，而Λ是一个主对角线上存有特征值的矩阵。其定义为对信号 x 应用谱图滤波器。

$xfiltered = Ug(\Lambda)U\top x$ (6)

为了以更简单的方式解决这个问题，我们用部分切比雪夫多项式序列近似谱滤波器$g(\Lambda)$

$g(\Lambda) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\Lambda}), T0(x) = 1, T1(x) = x, Tk + 1(x) = 2xTk(x) - Tk - 1(x)$ (7)

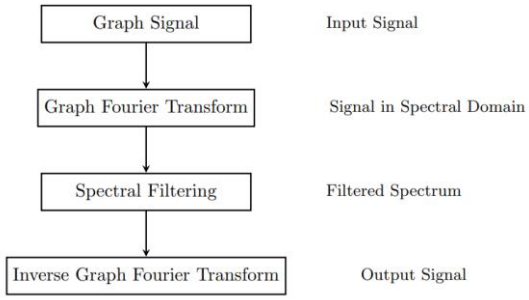


图3. 谱图对齐过程

图3展示了AGSP-DSA框架如何执行谱图对齐过程。这使得系统能够有效地滤波并改进基于各种信息源生成的图信号。首先，处理流程获取一个图信号$x \in RN$ ，其中每个条目对应于早期图构建中的一个节点。初始步骤是使用图傅里叶变换（GFT），借助图拉普拉斯矩阵的特征向量 U 将信号投影到谱域。

D. 基于图卷积网络的多尺度节点嵌入

采用图卷积网络（GCN）的多尺度架构继续处理图信号。每个 GCN 层都会整合来自邻居及远处区域的信息，以检测数据中近距离和远距离的连接。经过滤波后，信号被输入多层 GCN，通过累加附近节点的输入来优化嵌入。每一层都会应用 GCN 更新规则。l 是：

$$H^{(l+1)} = \sigma(\sum_{k=0}^{K}(\tilde{A})^k H^{(l)} W_k^{(l)}) \quad (8)$$

在这种情况下，$\tilde{A}=D^{-1/2}AD^{-1/2}$ 是归一化邻接矩阵，而 $\sigma(\cdot)$ 表示一种非线性函数（例如 ReLU）。通过这种设置，GCN 不仅聚合直接邻居的连接，也聚合更远的 K 跳连接。对 $G(m)$ 和 Gcm 分别执行相同的处理，然后将结果拼接起来。

$$H_i = Concat(H_i^{ntra}, H_i^{nter}) \quad (9)$$

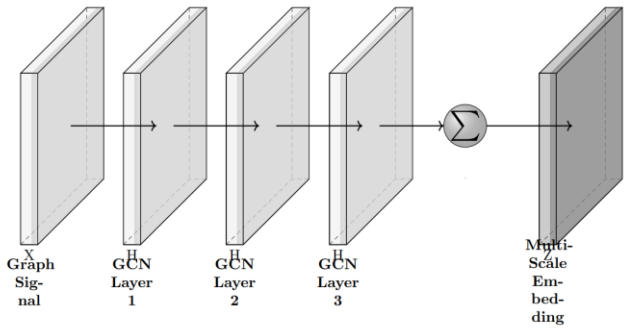


图 4. 基于图卷积网络的多尺度节点嵌入

图 4 展示了 AGSP-DSA 系统中使用图卷积网络进行多尺度节点嵌入的过程。图信号被输入处理流程，每个节点都会将多模态数据转换为有用的表示。图卷积网络的每一层都会接收信号、聚合邻居节点的信息，并随着图卷积网络层数的增加而扩展到更多邻居（例如，第一层覆盖 1 跳邻居，第二层覆盖 2 跳邻居）。借助这些图卷积网络层，模型能够提取各种数据输入中相似和不同方面之间的关系。所有输出通过求和（或拼接）操作合并，从而生成一个包含多个分辨率语义信息的统一嵌入。因此，模型能够学习重要特征，并在分类或查找数据中相似对象等后续任务中表现出色。

E. 语义感知注意力融合

为整合不同类型的信息，我们加入了一种注意力方法，使各模态中的细节获得不同程度的相关性 [24]。对于每个节点 i，我们将 $hi(m)$ 作为该节点在模态 m中的嵌入，并将 si 作为该节点的语义锚点（可以是类别标签或语义质心）。模态 m）的注意力权重为：g

$$\alpha_i^{(m)} = \frac{exp(h_i^{(m)\top} W_a si)}{\sum_{m'=1}^{m} exp(h_i^{(m')\top} W_a si)} \quad (10)$$

在此hi Wa sa 是一个可训练的权重矩阵。最终结果如下所示。

$$z_i = \sum_{m=1}^{M} \alpha i(m) hi(m) \quad (11)$$

为提升系统性能，我们让门控方法控制每个模态对结果的影响程度。

$$gi = \sigma(Wgzi + bg), zifused = gi \odot zi \quad (12)$$

这里，$\sigma(\cdot)$ 表示 Sigmoid函数，而“⊙”表示逐元素乘法。之后，任务专用头接收融合后的向量 z^{ifused}，用于分类、检索或回归 [25]。

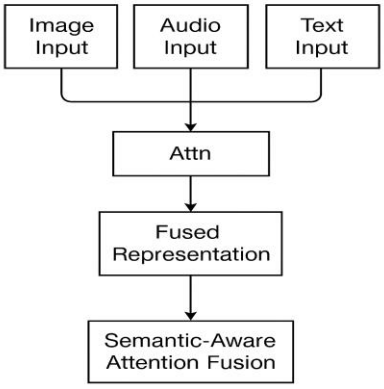


图 5. 语义感知注意力融合

图 5 描述了 AGSP-DSA框架中所使用的语义感知注意力融合模块。该图展示了图像、音频和文本模块的信息如何通过注意力机制进行整合。所有模态都会生成一个特征向量，并将其输入注意力单元；该单元利用标签或特定场景描述符中的语义信息，为每种模态赋予一个反映其重要性的注意力分数。随后，这些特征被整合为一个统一表示，同时保留原始数据的信息。因此，模型能够学会关注有助于改进多模态推理和辅助任务的重要模态。

F. 计算复杂度分析

为了评估 AGSP-DSA 框架与基线算法相比的效率，我们对融合过程主要组件的理论复杂度进行了评估，如表 1 所示。

表 I：计算复杂度分析

方法d	图构建on	谱滤波	GCN嵌入g	注意力n融合	总计复杂度y
早期融合	-	-	O(n)	-	O(n)
MMT	-	-	O(n2d)	O(nd2)	O(n2d)
MM-GNN	O(n2d)	-	O(k·n·d2)	-	O(n2d + k·n·d2)
AGSP-DSA	O(n2d)(双图)	O(K·n·d)(切比雪夫特征滤波器)	O(L·n·d2)(多规模图卷积网络)	O(n·d2)	O(n2d + K·n·d + L·n·d2)

IV. 结ESULTS 与讨论

所提出的 AGSP-DSA 在知名的多模态数据集 CMU-MOSEI、AVE 和 MM-IMDB 上进行了验证。每种模态均采用最新的特征提取模型：文本使用 BERT，图像使用 ResNet-50，音频使用 VGG。对于构建图和训练过程，神经网络的学习率设为 0.001，每次训练 64 个样本，并采用具有 3 层和 ReLU 激活函数的图卷积网络，如表2所示。

表II – AGSP-DSA框架中使用的仿真参数

参数	值	描述

学习率	0.001	优化器的步长 梯度下降
优化器	Adam	自适应 时刻 估计优化器
批量大小	64	每次的样本数 梯度更新
数量 of 训练轮次	100	总训练 训练周期 在数据集上
数量 of GCN层	3	图 卷积 用于特征的层 精炼
隐藏 单元 每个 GCN 层	128	维度 of 中间 node 嵌入
丢弃率	0.3	正则化 to 防止过拟合
切比雪夫 多项式 阶数 (K)	3	谱滤波器阶数 近似
融合策略	语义 感知的 注意力	自适应 模态 加权 基于 on 注意力
模态 Used	文本、图像、 音频	多模态数据流
输入 图 Type	Dual 图 （内部/跨模态）	分离 图 for 模态特定 and 共享表示

表 III – AGSP-DSA 与现有方法的比较评估

数据集	方法	准确率 (%)	F1分数	MAP
CMU-MOSEI	早期融合	88.2	0.873	0.861
	MMT	90.1	0.888	0.877
	MM-GNN	92.7	0.906	0.894
	AGSP-DSA	95.3	0.936	0.924
AVE	后期融合	84.6	0.832	0.821
	MMT	86.9	0.853	0.847
	MM-GNN	89.2	0.877	0.863
	AGSP-DSA	93.4	0.911	0.898
MM-IMDB	早期融合	80.1	0.778	0.765
	MM-GNN	83.5	0.802	0.790
	AGSP-DSA	91.8	0.886	0.871

在 CMU-MOSEI、AVE 和 MM-IMDB 这三个基准多模态数据集上，所提出的 AGSP-DSA（带动态语义对齐的自适应图信号处理）框架表现更优。与经典融合方法早期融合（88.2 %）和 MMT（90.1%），以及最先进的基于图的方法 MM-GNN（92.7%）相比，AGSP-DSA 在 CMU-MOSEI 数据集上取得了最高的准确率 95.3%。此外，AGSP-DSA 的 F1分数最高，达到 0.936，平均精度均值达到 0.924，表明它能够以更有效的方式对跨模态语义关联进行建模。在 AVE 数据集上，该模型仍表现出色，准确率为 93.4%，F1分数为 0.911，平均精度均值为 0.898，显著优于后期融合（84.6%）和 MM-GNN（89.2%）。同样，在 MM-IMDB 数据集上，AGSP-DSA 取得了 91.8% 的优异准确率，显著优于早期融合（80.1%）和 MM-

图神经网络（83.5%），同时 F1分数和平均精度均值分别达到 0.886 和 0.871，表现更佳。

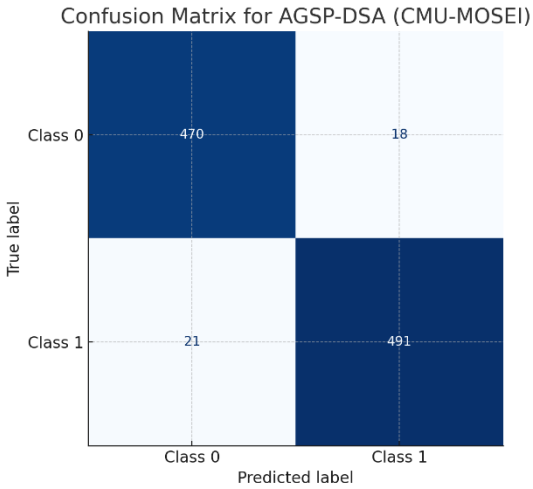


图6：混淆矩阵

图6展示了混淆矩阵，说明 AGSP-DSA 模型在 CMU-MOSEI 数据集的二元分类任务中的分类结果。如矩阵所示，470 个类别0样本和 491 个类别1样本被正确分类，表明敏感度和特异度较高。仅有 18 个类别0样本被误判为类别1，另有 21 个类别1样本被误判为类别0，因此错误率非常低。这些表现说明，基于图的语义对齐和自适应特征融合策略在 AGSP-DSA 模型中具有较强的鲁棒性，并且对不同类别具有较高的精确率和召回率。如此高的分类保真度证明了该模型适用于实际的多模态情感分析任务。

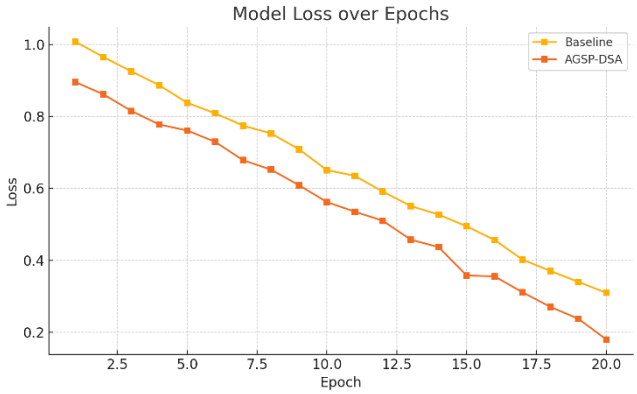


图7：基线模型与 AGSP-DSA 模型的损失收敛分析比较

图7展示了基线模型和所提出的 AGSP-DSA 模型在 20 个训练轮次中的损失下降情况。AGSP-DSA 在整个训练过程中也始终具有较低的损失值，形成了更高效、波动更小的学习过程。值得注意的是，AGSP-DSA 模型的初始损失较低，且下降趋势更陡，这意味着其收敛更快、优化效果更好。这证实了 AGSP-DSA 框架相比基线方法具有更高的训练效率和泛化能力。

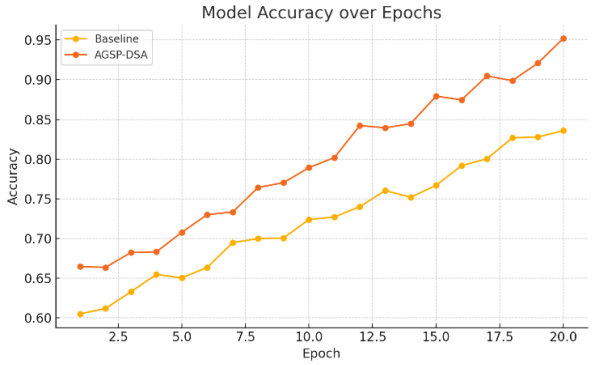


图8：基线模型与 AGSP-DSA 模型在各训练轮次中的准确率变化比较

图8展示了基线模型和 AGSP-DSA 模型在 20 个训练轮次中的准确率变化。无论是提升速率还是最终准确率，AGSP-DSA 模型都显著优于基线模型（~95% 对比 ~84%）。AGSP-DSA 的准确率曲线增长更快，且呈现稳定上升趋势，这表明其泛化能力更强、学习效率更高。因此，AGSP-DSA 能够有效学习复杂的多模态关系，并在整个训练过程中提升预测结果。

我们在每个数据集上使用随机种子进行了 5 次独立运行，以检验 AGSP-DSA 的可复现性和泛化能力。表 III 展示了结果，数值为均值 +/- 标准差。

表 III – 5 次运行的性能统计

数据集	模型	准确率（%）	F1分数	mAP
CMU-MOSEI	AGSP-DSA	95.3 ± 0.41	0.936 ± 0.015	0.924 ± 0.019
AVE	AGSP-DSA	93.4 ± 0.38	0.911 ± 0.014	0.898 ± 0.017
MM-IMDB	AGSP-DSA	91.8 ± 0.45	0.886 ± 0.020	0.871 ± 0.022

多次运行时离散程度较小，证明了 AGSP-DSA框架的稳健性。该模型在不同初始化下均保持稳定，且不会过拟合，即使在多模态数据缺失的情况下也是如此。

V. 结ONCLUSION

本文提出了 AGSP-DSA，这是一种自适应图信号处理模型，能够通过动态语义对齐稳健地融合多模态数据。针对异构模态集成的所有挑战，

该模型采用双图构建机制加以解决，同时捕获模态间和模态内关系。模型通过谱图滤波提升信号质量，并借助多尺度图卷积网络嵌入层，使全局和局部信息得以有效传播。语义感知注意力机制根据上下文以自适应方式平衡各模态的贡献，从而对融合表示进行上下文敏感的调节。在CMU-MOSEI、AVE 和 MM-IMDB 等大量基准数据集上的实验表明，与现有方法相比，AGSP-DSA 是一种更优的融合策略。该框架达到当前最先进的性能，在准确率、F1分数和平均精度均值（mAP）方面均有所提升，同时在模态缺失时仍具备良好的鲁棒性。消融研究验证了各个架构组件的价值，其他实验则表明该模型具有良好的训练稳定性和计算效率。

该方案是一种可扩展、准确且自适应的方法，可用于情感分析、事件识别和多媒体分类等多模态任务。后续工作将致力于将AGSP-DSA框架扩展至少样本学习、数据时序融合，以及采用基于Transformer的语义编码器进行实时推理。

References

[1] T. Baltrušaitis、C. Ahuja 和 L. -P. Morency，《多模态机器学习：综述与分类法》，发表于《IEEE模式分析与机器智能汇刊》，第 41 卷，第 2 期，第 423-443 页，2019年2月1日，doi: 10.1109/TPAMI.2018.2798607。[2] Atrey、Pradeep、Hossain, M.、El Saddik, Abdulmotaleb 和 Kankanhalli, Mohan。（2010）。多媒体分析中的多模态融合：综述。《多媒体系统》，16，345-379。10.1007/s00530-010-0182-0。[3] D. I. Shuman、S. K. Narang、P. Frossard、A. Ortega 和 P. Vandergheynst，《图信号处理这一新兴领域：将高维数据分析扩展至网络和其他不规则域》，发表于《IEEE信号处理杂志》，第 30 卷，第 3 期，第 83-98 页，2013年5月，doi: 10.1109/MSP.2012.2235192。[4] Georgiou、Efthymios、Papaioannou、Charilaos 和 Potamianos、Alexandros。（2019）。用于情感分析的深度层次融合。1646-1650。10.21437/Interspeech.2019-3243。[5] Kipf、Thomas 和 Welling、Max。（2016）。基于图卷积网络的半监督分类。10.48550/arXiv.1609.02907。[6] You、Renchun、Zhiyao、Guo、Cui、Lei、Long、Xiang、Bao、Yingze 和 Wen、Shilei。（2020）。用于多标签分类的语义图嵌入跨模态注意力。AAAI人工智能会议论文集。34。12709-12716。10.1609/aaai.v34i07.6964。[7] Xi、Xiaocong、Chow、Chee Onn、Chuah、Joon Huang 和 Kanesan、Jeevan。（2025）。基于图注意力网络的图文匹配跨模态语义关系增强。IEEE Access。PP。1-1。10.1109/ACCESS.2025.3549781。[8] Zhang、L.; Jiang、Y.; Yang、W.; Liu、B. TCTFusion：用于自适应红外与可见光图像融合的三分支跨模态 Transformer。*Electronics* **2025**，*14*，731。https://doi.org/10.3390/electronics14040731。[9] Yamamura M. 用于多模态情感分析的全局局部融合神经网络。应用科学。2022；12(17):8453。https://doi.org/10.3390/app12178453。[10] Liu、Xiaodong、Li、Songyang 和 Wang、Miao。（2021）。用于视频情绪识别的层次注意力多模态融合网络。《计算智能与神经科学》。2021。1-11。10.1155/2021/5585041。[11] He、Lijun、Wang、Ziqing、Wang、Liejun 和 Li、Fan。（2023）。适用于复杂情境的基于多模态互注意力的情感分析框架。《IEEE视频技术电路与系统汇刊》。PP。1-1。10.1109/TCSVT.2023.3276075。[12] Yu、Jun、Li、Jing、Yu、Zhou 和 Huang、Qingming。（2019）。采用多视图视觉表征的图像描述多模态Transformer。《IEEE视频技术电路与系统汇刊》。PP。1-1。10.1109/TCSVT.2019.2947482。[13] Tahir、Sohaib、Mohsin、Muhammad、Hassan、Muhammad、Yousif、Muhammad、Mannan、Abdul 和 Wattoo、Waqas Ahmad。（2022）。图信号处理的挑战与应用。《国际电气工程杂志》。[14] Yao、Ding、Zhang、Zhili、Kang、Weijie、Yang、Aitao、Zhao、Junyang、Feng、Jie、Hong、Danfeng 和 Zheng、Qinghe。（2025）。自适应同质性聚类：采用自适应滤波器的高光谱图像结构同质性图学习。《IEEE地球科学与遥感汇刊》。63。1-13。10.1109/TGRS.2025.3556276。[15] Liu、Xiaokang、Du、Xiaoxia、Liu、Juan、Su、Rongfeng、Ng、Manwa、Zhang、Yumei、Yang、Yudong、Zhao、Shaofeng、Wang、Lan 和 Yan、Nan。（2025）。使用音视频元音图注意力网络自动评估汉语构音障碍。《IEEE音频、语音与语言处理汇刊》。PP。1-13。10.1109/TASLPRO.2025.3546562。[16] Wu、Zonghan、Pan、Shirui、Chen、Fengwen、Long、Guodong、Zhang、Chengqi 和 Yu、Philip。（2019）。图神经网络综合综述。10.48550/arXiv.1901.00596。[17] Vu、Huy-The、Nguyen、Minh-Tien、Nguyen、Van-Chien、Pham、Minh-Hieu、Nguyen、Van-Quyet 和 Nguyen、Van-Hau。（2022）。用于多标签文本分类的标签代表性图卷积网络。《应用智能》。53。1-16。10.1007/s10489-022-04106-x。[18] Zhou、Haibing、Qian、Zhong、Li、Peifeng 和 Zhu、Qiaoming。（2024）。用于谣言检测的跨模态交互图注意力网络。1-8。10.1109/IJCNN60899.2024.10650542。[19] Zhang J、Wu G、Bi X、Cui Y. 基于动态图对比学习与特征融合的视频摘要生成网络。*Electronics*。2024；13(11):2039。https://doi.org/10.3390/electronics13112039。[20] Desmarais、Jaida、Klassen、Riordan、Patel、Eira 和 Chaudhry、Taryn。（2023）。DMFLC：基于深度多模态特征融合与低秩表示的短视频分类。10.21203/rs.3.rs-2662848/v1。[21] Kipf、Thomas 和 Welling、Max。（2016）。基于图卷积网络的半监督分类。10.48550/arXiv.1609.02907。[22] Shuman、David、Narang、Sunil K.、Frossard、Pascal、Ortega、Antonio 和 Vandergheynst、Pierre。（2012）。图信号处理这一新兴领域：将高维数据分析扩展至网络和其他不规则域。《IEEE信号处理杂志》。30。10.1109/MSP.2012.2235192。[23] Z. Wu、S. Pan、F. Chen、G. Long、C. Zhang 和 P. S. Yu，《图神经网络综合综述》，发表于《IEEE神经网络与学习系统汇刊》，第 32 卷，第 1 期，第 4-24 页，2021年1月，doi: 10.1109/TNNLS.2020.2978386。[24] Yu、Qiankun、Wang、XueKui、Wang、Dong、Chu、Xu、Liu、Bing 和 Liu、Peng。（2022）。BiTMulV：基于双向解码的Transformer与多视图视觉表征。10.1007/978-3-031-18907-4_57。[25] Yan、Kun、Zhang、Chenbin、Hou、Jun、Wang、Ping、Bouraoui、Zied、Jameel、Shoaib 和 Schockaert、Steven。（2022）。利用词向量引导注意力为多标签少样本图像分类推断原型。AAAI人工智能会议论文集。36。2991-2999。10.1609/aaai.v36i3.20205。