多模态表征学习：关于演进、预训练及其应用的综述

MUHAMMAD ARSLAN MANZOOR，Mohamed bin Zayed University of Artificial Intelligence，阿联酋SARAH ALBARRI，Mohamed bin Zayed University of Artificial Intelligence，阿联酋ZITING XIAN，中山大学，中国ZAIQIAO MENG，格拉斯哥大学，英国PRESLAV NAKOV，Mohamed bin Zayed University of Artificial Intelligence，阿联酋SHANGSONG LIANG*，Mohamed bin Zayed University of Artificial Intelligence，阿联酋

arXiv:2302.00389v2 [cs.AI] 1 Mar 2024

多模态表征学习是一种学习嵌入来自不同模态的信息及其相关性的技术，已在各种应用中取得显著成功，例如视觉问答（VQA）、视觉推理自然语言（NLVR）和视觉语言检索（VLR）。在这些应用中，不同模态之间的跨模态交互和互补信息对于高级模型最佳地执行任何多模态任务至关重要，例如理解、识别、检索或生成。研究人员提出了多种方法来解决这些任务。Transformer 架构的不同变体在多种模态上表现出色。本综述全面介绍了深度学习多模态架构的演进与发展，探讨其如何处理文本、视觉和音频特征，以应对各种跨模态任务和现代多模态任务。本文总结了（*i*）近期面向特定任务的深度学习方法，（*ii*）预训练类型和多模态预训练目标，（*iii*）从最先进的预训练多模态方法到统一架构，以及（*iv*）多模态任务类别和可进一步设计以改善多模态学习的潜在改进方向。此外，我们为新研究人员准备了数据集部分，涵盖大多数用于预训练和微调的基准数据集。最后，本文探讨了主要挑战、研究空白和潜在研究主题。与本综述相关且持续更新的论文列表维护于 https://github.com/marslanm/multimodality-representation-learning。

CCS 概念：• **计算方法** → **自然语言处理**；**神经网络**。

其他关键词与短语：多模态性、表征学习、预训练模型、多模态方法、多模态应用

ACM 引用格式：
Muhammad Arslan Manzoor、SarahAlBarri、Ziting Xian、Zaiqiao Meng、Preslav Nakov 和 Shangsong Liang。2023。多模态表征学习：演进、预训练及其应用综述。载于 。 ACM，纽约，美国，32 页。
https://doi.org/XXXXXXX.XXXXXXX

1 引言

多模态系统利用两种或更多输入模态（如音频、文本、图像或视频）来生成输出模态，而输出模态可能与输入模态不同。跨模态系统是多模态系统的一个子类，它利用一种模态的信息来提升另一种模态的性能。例如，多模态系统会使用图像和文本模态来评估情境并执行任务，而跨模态系统则会使用图像模态来输出文本模态 [1, 2]。视听语音识别（AVSR） [3], 检测

*Shangsong Liang 是本文的通讯作者。

©2023 计算机协会。稿件已提交至 ACM

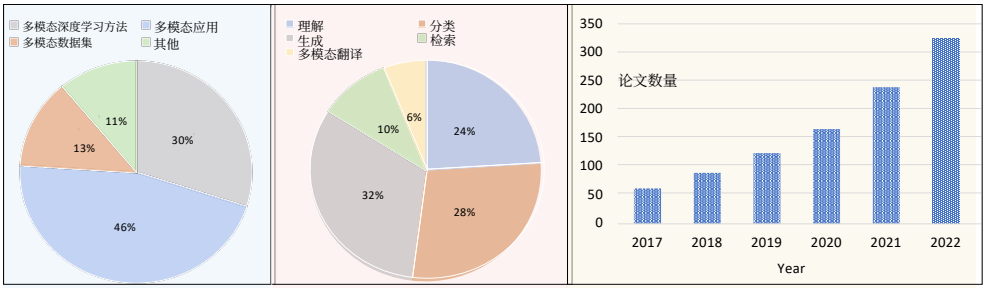

图 1. 左侧饼图展示了本综述所涵盖各章节论文的占比。中间的饼图展示了各类多模态应用论文的占比。最右侧的图展示了过去六年谷歌学术上基于深度学习的多模态论文数量增长情况。

模因中的宣传 [4], 和视觉问答（VQA） [5] 都是多模态系统的示例。多模态表示学习技术通过分层处理原始异构数据，缩小不同模态之间的异质性差距 [6]。来自不同模态的异构特征以上下文信息的形式提供了额外的语义 [6]。因此，可以通过多种模态学习互补信息。例如，视觉模态可以通过提供唇部运动信息来帮助语音识别 [7] ，这在 AVSR 中尤为如此。近期先进的深度学习方法变体通过将不同模态映射到标准表示空间，解决了经典的多模态挑战（相关性、翻译、对齐、融合）。

过去，许多特定任务深度学习策略显著提升了各类多模态任务的性能 [8]。近来，自然语言处理（NLP）和计算机视觉（CV）中的预训练与微调方法备受关注，这主要得益于其语义丰富的表示，以及大规模公开模型的可用性 [9]。本研究总结了多模态表示中特定任务方法的流程。我们讨论了预训练类型以及使预训练模型能够稳健应对各种多模态或跨模态下游任务所需的前置任务。我们指出，大多数预训练方法都采用 Transformer，从而催生了统一架构。这些模型能够处理多种下游任务中的所有模态，减少了对单独进行特定任务微调的需求，进而降低了计算复杂度和处理时间 [10]。此外，本综述全面介绍了下游任务评估、多模态基准、广泛的多模态应用，以及通过融合视觉和音频模态而得到增强的自然语言处理任务，包括情感分析、文档理解、假新闻检测、检索、翻译和其他推理应用。图1 展示了本综述所收录深度学习多模态论文的类别占比。条形图展示了谷歌学术上深度学习多模态方法逐年的发展和可用情况。

2 背景——多模态架构的演进与进展

多模态研究的概念最初受到视听语音识别（AVSR）领域的启发 [11],，在该领域中，视觉与听觉的良好对齐整合是一项基本要求。AVSR研究取得的令人鼓舞的成果，促使研究界将这一方法拓展至其他多模态任务，例如内容索引与检索 [12], 、视频摘要 [13] 和镜头边界检测 [14]。21世纪初，研究人员创建了AMI会议 [15] 和 SEMAINE [16] 语料库，旨在理解人类多模态行为以及人际互动的动态过程。这些基础数据集不仅促成了

视听情感挑战赛（AVEC）的创立 [17],，还推动其后来拓展至抑郁症和焦虑症的自动检测领域 [18]。随着时间推移，多模态方法不断进步，使语音、视觉和文本等不同模态得以同时处理 [2, 19]。最新类别涵盖了来自多媒体来源的数据，其复杂度超越了两种或三种模态。这些复杂的多模态方法促使研究人员采用预训练和微调等更有效的策略，应对相关挑战。事实证明，与传统的特定任务框架相比，这些方法能够取得更出色的性能。先进方法尤其擅长处理复杂的多模态任务，表现优于以往的模型。

多模态方法面临的首要障碍，在于理解多种模态的不同编码方案之间的关系，这一问题通常被称为表征学习中的异质性差距 [20]。多模态学习系统中的融合模块通过关联不同模态之间的相似语义，有助于缩小异质性差距；事实证明，这一策略能够提升大多数任务的性能 [21]。近期研究表明，由于具备强大的学习能力，深度学习方法能够有效地理解和处理表征 [22]。深度学习的主要优势之一，是能够以通用学习的方式分层学习表征，而无需额外的结构信息。这些因素使深度学习成为多模态表示学习中极具适应性的方法。

早期的特定任务多模态方法经过训练，仅用于解决单一的多模态任务。图 2 **（a）**概括性地展示了特定任务多模态方法的处理流程。例如，在 VQA 中，视觉编码器通常基于某种卷积神经网络（CNN）的变体 [23, 24], , 用于提取视觉特征。同时，文本特征由文本编码模块进行编码，这些模块通常基于词袋模型 [25], 序列到序列模型 [26], , 并可能采用注意力机制 [27]。随后，编码后的表征被传入多模态融合模块（该模块在多模态学习中得到了广泛研究），该模块融合不同模态的特征并建立其间的交互，从而生成统一表征 [28]。这些多模态融合方法逐渐演进为：（*i*）简单融合，通过拼接或逐元素求和、求积来融合特征 [29],（*ii*）模态间注意力，对齐不同模态的特征，并构建更具信息价值的联合表征 [30],（*iii*）模态内注意力，通过在模态内部构建图形表征来开展关系推理 [31],（*iv*）*Transformer*，通过同时关注其他模态以及同一模态中的相关区域，利用上述技术 [32]。由于 Transformer 能够学习并建立不同模态之间复杂的交互 [33, 34],因此，它们被视为通用多模态预训练方法的结构基础模块。

随着大规模数据集和计算资源的日益丰富，预训练范式已在文本 [35], 视觉 [36], 和语音 [37]等多个领域展现出高效性能。这推动了人们对多模态预训练的兴趣，使得能够通过半监督、自监督或无监督方法处理原始数据（包括有标签和无标签数据）。如图 2**(b)**所示，每种模态都由专用编码器处理，以生成嵌入。随后，这些嵌入会传递给专门针对多模态前置任务训练的融合模块。专用编码器生成内容丰富、具有上下文信息的嵌入，使融合模块能够学习不同模态之间更稳健的关联。进而，随着模型逐渐擅长处理复杂的多模态数据，其在下游任务中的表现也可能更加准确。在多模态学习的微调阶段，作为灵活骨干网络的预训练模型会使用面向目标任务的特定多模态数据集进一步训练。通过微调，模型会调整其学习到的表征，以更好地适应特定任务数据集中不同模态之间的关系和交互。因此，预训练所获得的丰富通用知识得以转化为任务特定洞见，从而提升模型在

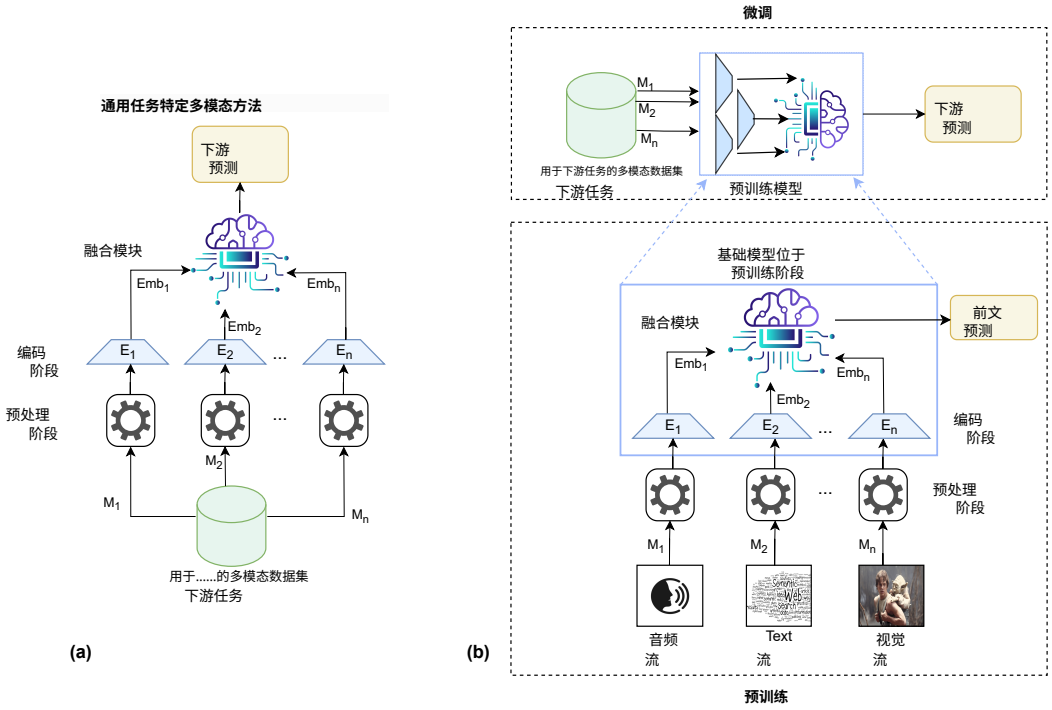

图 2. **(a)**特定任务多模态方法是在为专用下游任务构建的基准数据集上训练的。经过预处理后，每个编码器（E_i）接收一种模态（M_i）的数据，并生成嵌入。融合模块负责交互不同模态的特征，并通过预测下游任务进行训练。**(b)** 在预训练中，来自网络或数据集的原始数据以任意模态（M_i）的形式经过预处理，并在编码阶段传递给专用编码器（M_i）。编码器生成的嵌入由融合模块整合，并通过预测前置任务形成有意义的统一表征。预训练阶段蓝色边框的框表示基础模型，该模型在微调阶段可作为预训练模型直接使用。用于下游任务的多模态数据集可以利用这一通用预训练模型进行下游预测。

下游任务中的表现。图 2 清晰展示了多模态学习中的两种关键方法：特定任务方法和预训练-微调范式。二者的主要区别在于输入数据的类型：特定任务方法利用专门的基准数据集，而预训练可以使用任意类型的数据。特定任务方法直接针对下游任务进行训练，而预训练则面向前置任务，生成可适配多种下游任务的通用预训练模型，从而节省分别训练各个模型所需的计算时间并降低复杂度。

2.1 与以往综述的比较

大量综述对多模态性进行了深入研究，重点关注多模态学习技术、预训练技术、特定模型拓扑结构和特定任务应用。表示、相关性、融合、翻译和协同学习是多模态学习的重要组成部分， [2, 6, 7, 38, 39]对这些内容进行了综述。 [2],建立了多模态学习组件的分类体系。此外， [7] 从计算机视觉领域出发，对这些组件提出了独到的见解。 [6] 和 [39] 的工作分别仅研究了表示

表1. 不同综述工作的内容比较。考察的内容包括是否涉及预训练概念、多模态应用、自然语言处理侧重（文本和音频形式）、演变视角、统一架构、基准数据集和活跃代码仓库

Ref	Year	预训练	应用	自然语言处理侧重	演进	统一架构	数据集	活跃代码仓库
[43]	2017	✗	✓	✗	✗	✗	✓	✗
[2]	2018	✗	✓	✓	✓	✓	✗	✗
[6]	2019	✗	✗	✗	✓	✓	✗	✗
[1]	2021	✗	✓	✓	✓	✗	✓	✗
[44]	2021	✓	✓	✓	✗	✗	✓	✗
[41]	2022	✓	✓	✗	✗	✓	✓	✗
[7]	2022	✗	✓	✓	✗	✓	✓	✗
[42]	2022	✓	✓	✓	✗	✗	✓	✗
[28]	2022	✓	✓	✗	✗	✓	✓	✗
[40]	2023	✓	✓	✗	✗	✗	✓	✗
我们的工作	2023	✓	✓	✓	✓	✓	✓	✓

和融合，并分别介绍了它们的数学框架、架构、挑战和前景。尽管多模态学习组件对于理解和处理不同模态至关重要，但这些综述关注范围有限，未能全面呈现多模态性的整体图景。

视觉-语言预训练目标、策略、架构和数据集已在 [28, 40–42]中得到广泛研究。多模态预训练综述 [41] 聚焦于图文任务，因为其回顾了原始图像和文本的编码方法，以及用于交互的架构。虽然所有预训练综述都涉及下游任务，但对于最先进的框架，其细节程度仍然有限，而且对多模态领域内多样化应用的覆盖也不充分。研究 [43], [1]和 [44]考察了多模态深度学习模型，这些模型按照所涉及的模态 [1] 以及 [44]中的代表性任务进行介绍。 [43]的工作介绍了寻找最优多模态架构和跨模态正则化的方法。各种研究从特定应用的角度评估了多模态学习，例如视觉问答 [45],生物医学应用 [46], 事件检测 [47], 情感分析 [48, 49], 和目标检测 [50]。

我们的工作拓展了现有研究，更全面地探讨了特定任务方法和多模态预训练进展。我们详细阐述了预训练任务对于大规模预训练多模态模型的重要性，以及这些模型在各类下游任务中的鲁棒性。此外，我们还汇总了预训练基准数据集，介绍其规模、所含模态及目标任务。本综述还考察了统一架构的发展，并对多种最先进模型进行了比较分析。我们将探讨范围扩展至广泛的应用领域，包括视觉、语言、音频、医疗保健和自然语言处理特定任务。表 1 对比了我们的工作与其他综述，着重展示其对预训练概念、下游应用、自然语言处理应用、演化视角、统一架构、数据集以及活跃代码仓库可用性的涵盖情况。这种广泛的视角有助于全面把握多模态学习，并通过揭示高效的多模态处理技术，为未来研究提供指引。

3 多模态深度学习方法

本节探讨多模态架构，并将其分为特定任务架构和预训练-微调架构（图2）。第3.1 小节回顾特定任务方法，这些方法近年来逐渐发展为大规模预训练方法。第3.2 小节概述预训练过程，包括用于多模态任务的前置任务和最先进（SOTA）框架。第3.3小节比较最先进方法的结果，而3.4小节列出了文中使用的首字母缩略词。

3.1 多模态任务特定方法

多模态表示学习通过引入互补特征，提升了深度学习（DL）模型的鲁棒性。本节介绍特定任务多模态方法，这些方法通常可分为基于编码器-解码器、基于注意力和基于强化学习的模型。

3.1.1 基于编码器-解码器的模型。 编码器-解码器模型（如级联 CNN-RNN 和 RNN-RNN）通过语义丰富的潜在表示，将输入数据转换为输出。Hu 和 Wu [51] 提出了一种级联循环神经网络（CRNN），用于学习图像-文本交互。该网络以 VGG16 网络作为编码器，并以堆叠门控循环单元（SGRU）作为解码器。Ji 等人 [52] 提出了一种全局增强变换（GET）网络，采用 Faster-RCNN 和 ResNet-101 作为编码器。GET 网络在 MS-COCO 数据集上的表现优于 CRNN [53]，取得了更高的 BLEU 分数。

自编码器 [54], 是一种采用无监督学习的编码器-解码器模型形式，用于构建潜在表示以重构原始数据点 [54]。在 [55], 一种多模态架构中，自编码器用于处理音频和视频模态，并将其输入传统的编码器-解码器架构。

3.1.2 基于图的模型。 在多模态学习中，促进不同模态特征交互的融合模块已成为关键且核心的组成部分 [56]。特定任务方法广泛采用基于图的方法来执行融合任务 [28]。图神经网络（GNN）在处理多模态数据时表现出卓越的鲁棒性，能够利用其他模态的信息，巧妙地弥补某一模态中不完整或有噪声的输入 [57]。在多模态任务中，GNN 能够通过迭代式邻域聚合，同时捕获局部和全局上下文 [58]。用于神经机器翻译（NMT）的基于图的多模态融合编码器 [59] 在 Multi30K 数据集上的表现更优。该方法构建了包含两类节点的图：句子中的所有词语均作为文本节点，而由 Stanford Parser 识别出的句中名词则由视觉定位工具包检测为视觉节点。此外，方法还采用了两类边：用于连接同一模态中节点的模态内边，以及用于连接不同模态中对应节点的模态间边，从而捕获多模态数据的语义关系。这些图经过多个基于图的多模态融合层处理，逐步促进语义交互并学习节点表示，最终为解码器生成基于注意力的上下文向量。同样，Gao 等人 [60]通过设计多模态图神经网络（MM-GNN），以比预训练词嵌入更好的方式解决了场景文本歧义。VQA-GNN [61] 将图像级信息和概念知识整合到统一的多模态语义图中，以实现联合推理。该模型的有效性体现在其在 VCR 任务中取得了排名靠前的表现，超越了先前的模型；值得注意的是，它还首次为视觉问答任务中的视觉和文本知识提供了跨域可解释性。

3.1.3 基于注意力的模型。注意力机制通过检查信息流、特征和资源来改进模型学习。这些模型通过将注意力分配给关键信息并过滤所有不必要或无关的刺激，帮助网络处理长期依赖关系。Jiang 等人近期的一项工作 [62] 在多门控注意力网络（MGAN）流程中采用了注意力权重门（AWG）模块和自门控（SG）模块来实现注意力概念。此外，MGAN 还提取并利用了网络中的对象内关系 [62]。注意力引导的多模态相关性（AMC）中也探索了另一种基于注意力的变体 [63]。注意力机制应用于模态，而非向量中的语义上下文；每种模态的重要性会根据系统所需的查询而变化

。AMC 的应用使用了搜索日志：用户输入图像和文本进行搜索，系统则确定哪种模态更有价值。

3.1.4 基于强化学习的模型。 强化学习模型通过环境感知智能体与其环境之间的试错交互进行学习，在利用已有知识与探索新动作之间寻求平衡。该算法会“强化”能够带来高奖励的动作。与监督学习不同，它不需要标注数据。这一概念被融入深度学习中 [64],采用分层强化学习（HCL）框架，其中包含管理者、工作者和评论者等角色。管理者为工作者设定目标，工作者通过执行动作来实现目标，而评论者则评估目标是否完成。HCL 在将视频模态转换为文本方面优于基线模型。

随着多模态框架的复杂度不断提高，研究者对更多深度学习架构进行了改进和增强，例如生成对抗网络（GAN）和概率图模型（PGM） [65, 66]。然而，由于每项特定任务都需要重新训练，本节提到的方法在时间和计算资源效率方面面临挑战。认识到这些挑战后，许多研究者如今已将关注点转向预训练框架。这些框架旨在解决并克服特定任务方法固有的局限，减少针对不同任务反复重新训练的需求，从而为未来的多模态表示学习提供更高效的途径。

3.2 多模态预训练方法

本节将详细讨论预训练框架，重点介绍多模态预训练，包括预训练类型、前置任务和最先进的方法。早期，自监督学习（SSL）推动了语言任务的预训练 [27] ，之后被用于视觉任务并取得了高效的成果[67]。这一概念随后扩展到多模态场景。在大规模标注或未标注数据集上进行预训练，并在特定任务数据集上进行微调，已成为适用于各个领域的现代范式 [68]。

3.2.1 预训练类型。 Kalyan 等人 [46] 提出了研究人员用于大规模设计和训练模型的不同预训练方法。这些方法包括：（1）从头预训练（PTS），即使用大量未标注文本，通过随机初始化语言模型的所有层，从头训练模型，例如 ELECTRA [69], 、RoBERTa [70] 和 BERT[35]。（2）持续预训练（CPT），使用现有预训练模型初始化模型。这种方法可以减少所需的计算资源。特定领域任务使用这种方法来缓解目标词汇表较短所带来的挑战。BioBERT [71], 、HateBERT [72] 和 InfoXLM[73] 都是 CPT 的例子。（3）同步预训练（SPT），用于 [74]。该方法从头训练时同时使用特定领域文本和通用文本。（4）任务自适应预训练（TAPT），用于 [75] ，只需要少量数据。（5）知识继承预训练（KIPT） [76]采用知识蒸馏技术。

3.2.2 预训练任务。 在预训练阶段，通过解决预定义任务或前置任务来学习语言表示。这些任务基于自监督学习，且具有足够的挑战性，能够利用训练信号提升模型的鲁棒性，其中包括：(1) 因果语言建模（CLM）是一种单向建模方法，利用上下文预测下一个词，GPT-1 [77] 首次采用了该方法。(2) 掩码语言模型（MLM）通过考虑双向上下文来增强表示，并且只使用 15% 的词元， [35]首次采用了该方法。(3) 替换词元检测（RTD）通过将被替换的词元分类为原始词元或非原始词元，缓解了 MLM 在预训练中的挑战（监督信号较少） [69]。(4) 打乱词元检测（STD）通过利用判别式任务（检测打乱的词元） [78]，缩小了预训练与微调之间的差异。其他预训练任务包括

随机令牌替换（RTS）、下一句预测（NSP）、交换语言模型（SLM）、句子顺序预测（SOP），这些任务已成为多模态预训练的基线前置任务。

3.2.2.1 多模态预训练任务。多模态预训练任务具有诸多优势，包括增强表征学习、跨模态迁移、提升泛化能力、有效的多模态融合，以及提高多模态应用在广泛任务中的表现。这些任务包括：(1) 跨模态掩码语言模型（CMML），与 BERT 中的掩码语言模型类似；此外，预训练模型还会考虑上下文视觉特征和未掩码词元。这是一种有效的视觉语言预训练方法，可帮助模型更好地对齐并考量不同模态之间的关系 [41]。(2) 跨模态掩码区域预测（MRP） [79]：与前述目标类似，MRP 将感兴趣区域置零进行掩码。上下文视觉特征和文本推理用于预测这些掩码。(3) 图像-文本匹配（ITM）或视觉语言匹配（VLM） [80]：这一目标通过匹配和对齐，考量图像和文本特征之间的粗粒度关系。它在较高层面衡量不同模态之间的相关性。(4) 跨模态对比损失（CMCL）：这一目标通过拉近相关图像-文本对、推远不相关图像-文本对，学习通用的视觉和语言特征。不同模态嵌入之间最高的相似度分数有助于做出判断。类似的多模态预训练任务还包括掩码对象分类（MOC） [79], 短语-区域对齐（PRA） [81], 词语-区域对齐（WRA） [82], 视频-字幕匹配（VSM）[83], 掩码帧建模（MFM） [83], 帧顺序建模（FOM） [83], 视觉翻译语言模型（VTLM）[84], 等等。

预训练方法已广泛应用于自然语言处理下游任务，并取得了卓越表现。研究人员提出了 BERT [35], 、ALBERT [85], 、RoBERTa [70], 和 T5 [86], 等预训练模型，以学习通用文本表征并提升下游任务的性能。类似地，预训练方法也被用于减少基于视觉的下游任务所需的时间和计算量 [87, 88]。与此同时，预训练范式在多模态数据领域不断取得进展，提升了从跨模态到完全多模态的各类任务的性能。下一节将介绍近期为充分利用丰富多模态数据而构建的架构。

3.2.3 基于*Transformer*的架构。近期大规模预训练方面的研究激励研究界推动多模态任务的发展 [89]。基于大规模图像-文本对进行联合预训练的先进方法发展迅速，其表现优于特定任务架构。例如，在 ViLBERT [9], 中，作者对 BERT 进行预训练，使其分别处理文本和视觉输入，并通过协同注意力层维持两者之间的关系。在 Conceptual Captions数据集 [90], 上进行预训练后，该模型被迁移至主流语言-视觉任务，包括视觉问答、视觉常识推理和基于图注的图像检索，并且取得了稳定的提升，优于特定任务架构。 [91] 提出了统一形式的视觉语言预训练（VLP），其在图像描述和视觉问答任务上的表现超越了此前的最先进水平。该模型采用共享编码器-解码器架构，并在大规模图像-文本对上以无监督方式进行预训练。他们在 VQA 2.0、COCO Captions 和 Flickr30k [92] Captions 上评估了该模型。与先前工作一致，Li 等人 [93] 提出了一种预训练方法，将对象标签用作锚点，以增强与配对文本的对齐学习。

除了上述方法外，近期研究还从场景图中提取结构化知识[95], ，以辅助学习文本与视觉的联合表示，将其作为语义相关联的特征。作者在预训练阶段新提出了场景图预测任务，包括物体预测（OP）、属性预测（AP）和关系预测（RP）。所提出的模型能够学习语义对齐的联合表示。研究人员在五项跨模态任务上评估了该模型的有效性，该模型在 VCR 任务上表现领先，提升了

表 2. 基于 Transformer 的预训练方法、基准数据集、任务及架构设置。

模型	基准数据集	预训练任务	架构比较
VLP [91]	VQA2.0, COCO 图像描述, Flickr30k	图像分类、视觉问答	12层，768隐藏维度，12个注意力头，1.1亿个参数。
VirTex [94]	COCO Captions	图像分类，属性预测	视觉主干网络：ResNet-50；文本的头部：两个单向Transformer
ViLBERT [9]	视觉问答、视觉常识推理、RefCOCO、信息检索	视觉问答、视觉常识推理、图文生成与检索、组合图像检索	文本编码器：BERT-base；视觉骨干网络：Faster R-CNN（ResNet 101）
ERNIE-ViL [95]	VCR、RefCOCO+、VQA、信息检索-Flickr30K、QR-Flickr30K	SGP、VCR、VQA、图文生成与检索、信息检索、TR	文本编码器：BERT-base；视觉骨干网络：Faster R-CNN (ResNet 101)
OSCAR [93]	COCO, CC, SBU 图像描述, flicker30k, GQA	视觉问答、图文检索、图像分类、名词短语生成、GQA、NLVR2	基础版：12层， 768隐藏维度，12个注意力头，110M 个参数；大型版：24层，1024-隐藏维度，16个注意力头，3.4亿个参数
Vokenizer [96]	SST-2。QNLI, QQP, MNLI、SQuAD v1.1、SQuAD v2.0、SWAG 平均值	GLUE、SQuAD、SWAG, GLUE	文本编码器：BERT；视觉编码器：ResNeXt-101-32x8d
AV-HuBERT [3]	LRS3、VoxCeleb2	ASR	混合式 ResNet-Transformer 架构-ture
CLIP [89]	COCO, 视觉 Genome, YFCC100M	图像 分类, IC, OCR, AR, Geo-定位	文本编码器：GPT-2、GPT-3；视觉编码器：ViT-B/32、ViT-B/16、ViT-L/14
BLIP [97]	No-Caps、COCO、Flickr30k	图文检索、图像分类、视觉问答、视觉对话、自然语言视觉推理	文本编码器：BERTbase；图像编-码器：ViT-B/16、ViT-L/16
BLIP-2 [98]	No-Caps, COCO, Flickr30k, VQAv2、OK-VQA、GQA	指令微调 零样本图像到文本 生成、视觉问答、图像分类、图文检索	文本编码器：Q-Former；视觉编码器：FlanT5、ViT-L/14、ViT-g/14

3.7%。Desai 等人 [94] 提出了 VirTex，这是一种数据高效的监督预训练替代方案，将图像描述用作视觉任务的监督信号。他们从头开始对模型进行联合预训练，该模型由用于视觉和语言模态的卷积与 Transformer 主干网络组成，旨在为图像生成描述。模型采用 ResNet-50 作为视觉主干网络，计算输入图像的视觉特征，随后将这些特征传递给文本头，为相应的图像生成描述。训练完成后，文本头会被丢弃，因为其主要目的是利用学习到的特征完成后续视觉任务。此外，Shi 等人 [3] 提出了视听隐藏单元 BERT（AV-HuBERT），并在由语音和视觉特征组成的 LRS3 基准上从头开始训练。混合架构 ResNet-transformer 充当所提模型的骨干网络。AV-HuBERT 同时从音频和视觉流中提取语音和语言特征，以生成潜在表示。该表示可用于预训练基线，例如 HuBERT [99]，其性能优于在单一模态上预训练的模型。表 2 根据基准数据集、任务和架构设置，对所讨论的方法进行了比较。

Word2Vec [100], GPT [77], ELMO [101], BERT 等基于 SSL 的框架在自然语言理解方面表现良好，但没有考虑视觉世界中的接地信息，正如 Bender 等人 [102]和 Bisk 等人 [103]所指出的那样。Tan 等人 [96], 设计的预训练过程称为 Vokenization，它将语言标记及其对应的视觉信息（称为“视觉词元”）作为监督输入。作者

利用小型图像描述数据集训练 Vokenizer，进而为大型语料库生成语言视觉词元。视觉监督模型在 GLUE、SWAG 和 SQuAD 基准测试上的表现优于基于 SSL 的预训练方法。然而，使用监督方法训练的模型往往严重依赖标注数据，这可能限制其通用性和可用性。因此，为了克服这些挑战，自监督多模态学习方法应运而生，例如 CLIP [89]。CLIP（对比式语言-图像预训练）在 WIT（WebImageText）上进行预训练 [89], ，该数据集包含从网络收集的图像-文本对。与标准视觉模型不同，CLIP 在训练阶段联合训练图像编码器和文本编码器。CLIP 的性能在 30 多个视觉和视觉-语言数据集上进行基准测试，包括 COCO [53], Visual Genome [104],和 YFCC100M [105]等。BLIP（引导式语言图像预训练） [97],是一种统一的视觉-语言框架，通过对图像描述进行知识蒸馏来提升性能。它在多种任务上取得了最先进的结果，包括图像-文本检索、图像描述、视觉问答，甚至在文本到视频检索和视频问答的零样本性能方面也表现出色。BLIP 在 No-Caps 上进行了广泛评估 [106], COCO [53],和 Flickr30k [92]。BLIP-2 [98],作为一种基于全面预训练策略改进的视觉-语言模型，在众多视觉-语言任务中取得了最先进的性能。BLIP-2 展现出广泛的零样本图像到文本能力，包括视觉知识推理、视觉常识推理、视觉对话和个性化图像到文本生成。与其前身一样，BLIP-2 使用相同的评估数据集，并继续采用描述与筛选（CapFilt）方法。

3.2.4 统一架构。 统一架构旨在接收不同模态作为输入，并在多个任务上训练模型，以减少任务特定参数，从而构建通用架构。Li 等人 [107] 提出了适用于仅文本和纯视觉任务的统一模型。作者致力于创建一种适用于任何模态、可针对任何下游任务进行微调的单一基于Transformer的预训练模型。该基础模型是在未配对图像和文本上联合预训练的Transformer。预训练采用从教师预训练模型进行知识蒸馏，以实现更好的联合训练，并使用梯度掩码来平衡参数更新。该模型采用共享Transformer，通过任务特定分类器为不同任务编码所有模态。其目标是构建更通用的模型，以减少任务特定和模态特定参数，而大部分计算都在共享Transformer模块中进行。

UNITER [82] 是一种在四个大型基准数据集上进行的预训练方法。通过学习两种模态的联合嵌入来完成异构下游任务。MLM、MRM（掩码区域建模）、ITM（图像-文本匹配）和WRA（词-区域对齐）被用作预训练任务。此外，该预训练任务通过条件掩码实现全局图像-文本对齐。最优传输是作者为WRA任务提出的第二个概念，旨在提升图像与词语之间的对齐效果。Hu 等人 [108]提出了一种基于多模态学习、可同时进行多任务处理的统一Transformer。所提出的架构为每种模态使用专用编码器，并为每项任务使用共享解码器。DETR [87] 用于视觉特征编码，BERT [35] 则用于文本特征编码。Akbari 等人[109] 采用对比学习对多模态数据进行学习，训练能够同时处理音频、文本和视频的Transformer编码器。

Wang 等人 [110] 提出了 One For All（OFA）方法，该方法采用基于统一词汇表（适用于所有类型的模态）的序列到序列框架，统一任务与模态。OFA 将不同模态的数据表示在统一空间中，并将图像和文本离散化，以构建统一的输出词汇表。他们提出了统一模型应具备的三项特性，以支持任意模态下的多任务处理：（1）任务无关：利用人工设计的基于指令的学习来实现这一特性。（2）模态无关：基于单一 Transformer 的

表 3. 作者采用的统一架构、基准数据集、任务及架构设置。

模型	基准数据集	任务	架构比较
UniT [108]	COCO、QQP、 VQAv2 QNLI、MNLI, SST-2 ，SNKI-VE	视觉问答、目标检测、视觉蕴含	2.01 亿参数，由 ResNet-50 + BERT
ViT-BERT [107]	MNLI、QQP、QNLI、SST-2、RTE CIFAR、ImageNet、Flowers、Pet	仅视觉、文本-only	12 层 Transformer，768隐藏维度，3072 多层感知机
UNITER [82]	视觉问答、Flickr30k、 自然语言视觉推理、Ref-COCO	信息检索、文本检索 视觉问答、关系抽取	12–24 层，768–1024 隐藏层，12–16 个头，86–303M 个参数。
VATT [109]	AudioSet, HowTo UCF101, HMDB51, MSR-VTT	视觉到音频检索、音频事件分类、文本到视频检索	基于 Transformer 的变体；1.55亿至4.15亿参数。
OFA [110]	SST-2、RTE、MRPC、QQP、QNLI, MNLI、SNLI-VE	NLU、NLG、图像分类	基于 Transformer 的变体；3300万至9.4亿参数。
InstructBLIP [98]	No-Caps、Flickr30k、GQA、VSR、IconQA、TextVQA、VisDial、HM、VizWiz、SciQA 图像、MSVD QA、MSRVTT QA、iVQA	IC, VR, 图像描述与推理复杂度, 视频问答、图像问题生成, 图像问答，LLaVa-Instruct-150k	变体 of BLIP-2, 图像编码器：ViT-g/14，大语言模型：FlanT5-XL (3B)、FlanT5-XXL（11B）、Vicuna-7B 和 Vicuna-13B

架构采用全局共享的多模态词汇，使其不受模态限制。（3）任务全面性：通过在多种单模态和多模态任务上进行预训练，实现任务全面性。Transformer 是编码器-解码器统一网络的骨干网络，用于预训练、微调和零样本任务。该网络兼顾多模态性与多任务处理，从而更好地泛化到未见任务。OFA 在多模态任务上达到了最先进水平，在单模态任务上的表现也优于其他知名预训练模型。与 GPT [77] 和 BART [111], 类似，它使用 BPE（字节对编码）将词语序列拆分为子词序列，并将其嵌入为特征。

InstructBLIP [112] 是一种通用视觉-语言模型，在众多视觉-语言任务上实现了最先进的零样本性能。InstructBLIP 是一种视觉-语言指令微调框架，由图像编码器、大语言模型（LLM）和 Q-Former [98]组成。在其评估中，InstructBLIP 使用了两个大语言模型：FlanT5[113] 和 Vicuna [114]。评估指标涵盖 28 个数据集上的 11 项任务，包括图像描述、视频推理、视觉对话问答、知识支撑的图像问答、视频问答、图像描述阅读理解、图像问题生成、图像分类，以及 LLaVa-Instruct-150k；后者是一个基准测试，涵盖视觉对话、复杂推理和详细图像描述。

3.3 讨论

大多数多模态方法都使用 Transformer 作为骨干网络。由于自注意力在较低层阶段侧重于全局上下文学习，这些架构在大多数视觉任务上的表现优于卷积神经网络。不过，大规模多模态架构的泛化能力很大程度上取决于预训练目标、学习技术以及本综述所述基准数据集的性质。

VLP [91] 被认为是在联合视觉-语言多模态任务（尤其是生成和理解任务）上首个达到最先进水平的方法。同样，VirTex 在多模态任务上展现了最先进的性能 [94]。ViLBERT [9] 在 VQA 和 RefCOCO+上以巨大优势达到最先进水平。AV-HuBERT [3] 在基准测试（纯音频语音识别）中采用视听表征，使词错误率相较于最先进水平降低了 40%。ERNIE-ViL [95] 在 5 项跨模态下游任务上达到最先进水平，并在 VCR排行榜上名列第一，绝对提升了 3.7%。OSCAR [93] 在六项成熟的视觉-语言理解任务上创下了新的最先进水平

和生成任务。Vokenizer [96] 提出了视觉监督语言模型，在多项语言任务上持续取得提升。在类似的基准测试上进行微调可带来出色的结果，但需要更多参数、时间和资源。许多研究人员正通过提出统一架构来应对这些挑战，使同一骨干网络能够处理多种模态并支持多任务处理。此类通用模型总结于表3 中，可为不同的下游任务共享参数，并避免针对每项任务进行大量预训练或微调。尽管统一架构并非针对特定任务进行专门设计，但仍能取得有竞争力的结果。

研究发现，在预训练阶段，如果考虑表示之间的强对齐和对应关系，以相关方式综合不同模态进行预训练能够提升性能。在多种模态上训练的模型在单模态任务上也表现良好，例如 ViT-BERT [107] 在纯视觉任务上超越了 ViT。然而，由于用于微调的数据量较少，尤其是在 RTE 数据集上，ViT-BERT 未能超越 BERT 在仅文本任务上的表现。同样，基线方法的联合训练结果优于跨模态微调，因为后者面临预训练目标与下游任务不匹配的挑战。ViT-BERT 在仅文本和纯视觉任务上分别取得了 83% 和 89% 的平均分数，在两类任务上的平均分为 86%，是一个适中的选择（且比 UniT 使用更少的参数）。UniT [108]作为通用模型，在所有任务上使用共享参数，取得了颇具潜力的结果。然而，UniT 未能超越 VisualBERT 或 BERT，后两者仅针对特定任务进行训练。较小的批大小和较高的学习率分别会导致性能下降和发散问题。

UNITER-large [82]模型在所有基准数据集上都取得了相当的性能，但消耗了 303M 个参数。UNITER-base 模型也表现出色，但在 VQA 上的表现除外，其参数量为 86M。VATT [109] 在音频事件识别和视频动作识别任务的所有指标上均优于基于 CNN 的方法。此外，它在文本到视频检索方面也取得了有竞争力的结果。作者提出了 DropToken 策略，以降低训练的计算复杂度。DropToken 是一种采样技术，可作为维度缩减和分辨率缩减的替代方法，以应对冗余挑战。此外，基于噪声对比估计和多实例学习的多模态对比学习，在投影到共享空间后用于对齐视频-音频对和视频-文本对。OFA [110] 在 VQA 和 SNLI-VE 上超越了 UNITER、UNIMO 及其他最先进水平的方法。值得注意的是，它在 RTE 和 RefCOC0+ 上取得了最高分，超过了特定任务架构和统一架构。具体而言，OFA 的结果表明，大规模多模态预训练方法在理解和生成任务上可与自然语言预训练的最先进水平方法相媲美。此外，OFA 在域外任务（例如单句分类和句对分类）上优于 Uni-Perceiver。

InstructBLIP [112] 提出了一种面向通用视觉语言模型的新型指令微调框架。其提出的方法在一系列基准数据集上取得了最先进的性能，并采用了严格的评估方案。此外，InstructBLIP 展现出作为先进初始模型的潜力，可用于下游任务的微调。

3.4 首字母缩略词

本节列出了表格中所有的首字母缩略词。**AEC**：音频事件分类，**Auton. Driving**：自动驾驶，**AR**：动作识别，**AVSS**：视听语音合成，**CIR**：文本到图像检索，**CMR**：跨模态检索，**ED**：事件检测，**EL**：实体标注，**E QA**：具身问答，**GR**：手势识别，**GRE**：指代表达生成，**IC**：图像描述，**ICRC**：图像描述阅读理解，**IE**：信息抽取，**IQG**：图像问题生成，**IR**：图像检索，**IS**：室内场景分割，**ITR**：图像到文本检索，**LaVa-Instruct-150k**：包含视觉对话、复杂推理和详细图像描述，**MML**：多模态学习，**MMSA**：多模态情感分析，**NLG**：自然语言生成，**NLU**：

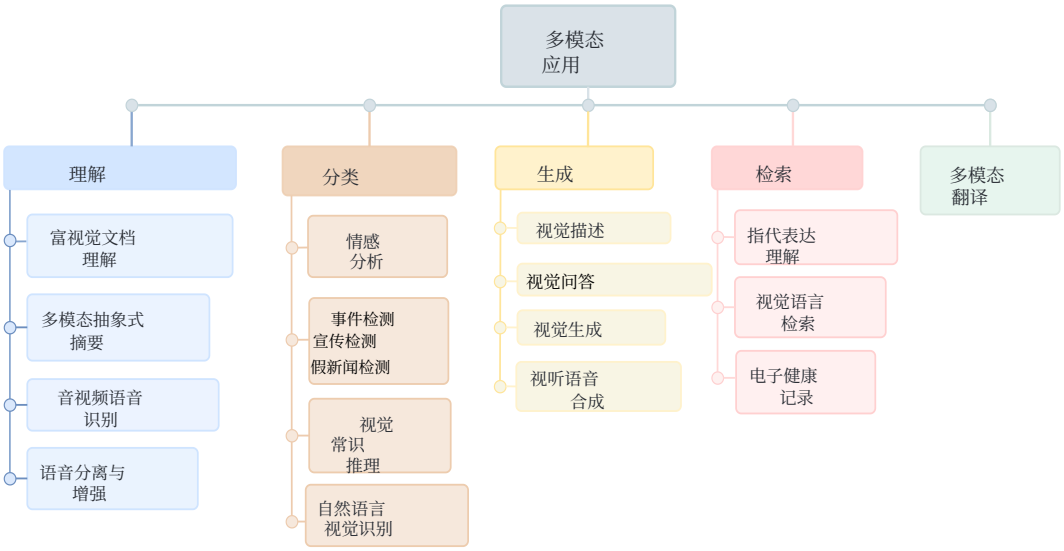

图 3. 第 4 节多模态应用的分类体系。

自然语言理解，**NLVR**：视觉推理自然语言，**OCR**：光学字符识别，**OD**：目标检测，**RE**：指代表达，**SSS**：声源分离，**TD**：文本检测，**TR**：文本检索，**TVR**：文本到视频检索，**VAC**：视频动作识别，**VAR**：视觉到音频检索，**VC**：视频描述，**VCR**：视觉常识推理，**VD**：视觉对话，**VE**：视觉蕴含，**VG**：视觉生成，**VLP**：视觉语言预训练，**VQA**：视觉问答，**VR**：视频推理，**VU**：视觉理解。

4 多模态应用

本节将详细介绍深度学习架构增强的多模态应用类别，如图 3 所示。多模态任务分为五大类：理解、分类、检索、生成和翻译。针对每种多模态应用，本文均讨论其基准测试、评估指标、描述以及性能最佳架构的比较。

4.1 理解

理解文本、语音和视觉是基于人工智能的系统最重要且首要的任务。从不同媒介（文本、图像、语音）中提取相关信息、识别并辨认实体，是其子任务。通过理解获得的信息能够直接或间接地惠及许多下游任务。

4.1.1 视觉丰富文档理解。理解包含结构化信息的富视觉文档对各种应用都很重要 [115]。与传统信息抽取方法不同，视觉丰富文档理解会同时考虑视觉布局和文本。LayoutLMv2 [116]在预训练阶段通过整合并对齐布局、文本和视觉信息，对能够从不同模态中学习的 Transformer 进行了预训练。LayoutMv2 使用超过 100 亿份文档对模型进行预训练，使其在文档理解任务上的表现超越了最先进水平。最近，Li 等人 [117] 提出了 StrucTexT，通过以自监督方式预训练 Transformer，以更少的参数实现了高分。公开可用的基准测试

下游任务数据集包括 SROIE [118] 和 FUNSD [119]。LayoutLMv2 在 SROIE 和 FUNSD 上分别取得了 97% 和 84% 的 F1 分数。StrucTexT 在 SROIE 和 FUNSD 上分别取得了 96.88% 和 85.68% 的 F1 分数。

Gu 等人 [120] 发布了 LayoutLM2 [116] 的改进版本 XYLayoutLM。他们提出了增强型 XY 切割，以捕获作为布局信息的阅读顺序，而这一信息此前一直被忽略。此外，输入序列的可变长度通过 DCPE（膨胀条件位置编码——受 CPE [121] 启发）进行处理，该方法分别为文本输入和视觉输入生成一维和二维特征，用于卷积。该模型在 XFUN [116] 基准测试的 SER（语义实体识别）和 RE（关系抽取）任务上成功超越了其他方法。

4.1.2 多模态抽象式摘要。 该任务从互联网中获取大量多模态内容（视频、图像及相应文本），提取关键信息并生成摘要 [122]。Palaskar 等人[123] 使用融合不同模态信息的多源序列到序列模型和层次注意力生成摘要。同样，有研究提出了多阶段融合网络，以建立不同源模态之间的交互。最近，有一种方法被提出 [124]，利用视觉模态与生成式预训练语言模型（GPLMs）生成摘要。该方法在视觉引导的 GPLMs 中插入基于注意力的附加层，以保持视觉信息的融合和文本生成。该模型在 How2[125] 数据集上进行了评估，该数据集包含时长为 2000 小时、每段配有两到三句话的教学视频。所提出模型的视觉接地变体超越了基线方法，在该基准测试上取得了最高的 ROGUE-1、ROGUE-2 和 ROGUE-L 分数。

4.1.3 视听语音识别。视听语音识别（AVSR）是最早的多模态研究领域之一，推动研究界同时利用听觉和视觉特征来理解语音。视觉特征和语音在理解语音方面发挥着至关重要的作用，尤其是在噪声环境下，并有助于为语音障碍患者提供支持。近期的大多数模型 wav2vec2.0 [37], HuBERT [99]和 De-CoAR2.0[126] 都仅使用音频进行语音识别。通过利用自监督学习，唇部运动提供了监督信号，有助于提高识别效果。

近期发布的视听隐藏单元 BERT（AV-HuBERT） [3] 是一种自监督表征学习框架。该方法借助唇读和自动语音识别，学习视听语音表征。所提出的模型使用来自 433 小时数据中的 30 小时标注数据，在 LRS3 [127] 基准测试上取得了 32.5% 的词错误率，并以 26.9% 的词错误率超越了此前在 31K 小时数据上训练的最先进水平 [128]。作者还指出，该表征在纯音频任务上进行语音识别时，词错误率降低了 40%。通过利用 AV-HuBERT 的上下文表征，该模型在纯视觉模态下也表现出色。

4.1.4 语音分离与增强。语音增强（SE）是指在声学噪声环境中提取目标说话人语音信号的过程。SE 可改善语音质量（声音）和语音可懂度（语言内容）。对语音中多个目标进行估计称为语音分离或声源分离 [129]。过去，这些任务通过信号处理中的统计和数学准则来处理 [130]。如今，监督学习、多模态方法和融合技术的演进，能够在嘈杂环境中获得说话人不受干扰的视觉特征，并结合声学特征。

音视频句子提取（AV-SE）和音视频句子分离（AV-SS）系统会统计说话人的数量并追踪他们的面部。检测与跟踪算法用于获取人脸的高维视觉帧。通过基于主成分分析（PCA）的主动外观模型降低维度[131]。Zhou 等人 [132] 使用外观注意力模块实现声源分离，该模块利用单帧视频中的类别信息。作者通过关联外观来优化模型

嵌入和特征图。声源的位置通过嵌入与特征图的标量积确定。MUSIC 数据集包含由不同乐器演奏的 YouTube 视频，并带有画外噪声 [133]。该模型使用标准化评估指标在 MUSIC 数据集上进行评估，即信号失真比（SDR）、信号干扰比（SIR）和信号伪影比（SAR）。基于注意力的方法表现优于 Resnet-18 和 Resnet-50，达到 SDR 10.74、SIR 17.29 和 SAR 13.04。

4.2 分类

分类是按照既定标准将样本系统地归入某个群组或类别。视觉、文本和语音领域的多项任务都属于分类范畴。本研究讨论的主要多模态分类任务包括情感分析、假新闻检测和事件检测。

4.2.1 情感分析。 情感分析是自然语言处理中的核心任务，旨在提取并分类针对特定实体的评论、感受、姿态和行为。情感分析能够理解人们的观点，从而支持多个领域中的高效决策。以往研究已广泛分析文本表征，例如 [134], [135],和 [136]。然而，由于社交媒体和互联网的发展，情感分析任务已从文本模态转向其他模态。Chen 等人 [137] 设计了深度学习架构，通过设计由浅层融合和聚合两个部分组成的组件式方法，从复杂的多模态数据中提取情感。浅层融合组件利用注意力机制从不同领域提取上下文信息，而聚合部分则实现情感词感知融合。所提出的架构在 CMU-MOSI、CMU-MOSEI 和 YouTube 数据集等多模态数据集上取得最高分，优于其他方法。

4.2.2 事件检测。通过社交媒体上的内容来检测事件、趋势或情境时，纳入来自不同模态的数据 [138]能够提升效率并拓宽信息来源。大量数据由此产生，对人类的生活、财产和心理产生重大影响 [139]。事件检测可应用于其他现实场景，例如应急管理、灾害检测和主题检测 [47]。尽管来自不同视角的数据能够提升性能，但也增加了方法应对冗余和异质特征时面临的挑战。除 CrisisMMD [140], 、CrisisNLP 和 CrisisLex 外，目前尚无足够大型的数据集可用于灾害检测中的深度学习架构。针对交通事件检测，Chen等人 [141] 通过整合经过筛选的交通相关推文与传感器数据，创建了一个多模态数据集。作者使用卷积神经网络、循环神经网络和多模态生成对抗网络（多模态 GAN）模型，分别取得了 84%、83% 和 87% 的 F1 分数。

4.2.3 检测模因中的宣传。宣传是一种影响人们心理的传播形式，会促使人们对某个实体持有特定观点或采取某些行动。由于社交媒体的使用率很高，宣传已成为一个社会和政治问题。包含视觉和文本内容的模因正被广泛用作互联网传播媒介，以引发这一问题。

Dimitrov 等人 [4] 将这一问题视为多模态多标签任务，通过检测模因中的宣传手法来推广宣传。他们发布了一个新的数据集，包含 950 个模因（同时含有文本和视觉内容），并标注了 22 种不同的宣传手法。以往的数据集在文档级别 [142], 、句子级别和片段级别 [143]研究宣传。最先进水平的模型开展了实验。实验针对两类预训练模型的不同变体进行：（i）单模态预训练；（ii）采用多模态目标进行预训练。在第一种变体中，BERT 和 ResNet-152 分别进行训练，并通过多模态双向变换器（MMBT）以早期、中期或后期融合的方式进行融合。ViLBERT 和 Visual BERT 分别在 Conceptual Captions 和 MS COCO 上进行多模态预训练。ViLBERT 和

Visual BERT 的结果优于基线和其他变体。结果分析表明，基于 Transformer 的多模态方法切实可行，且能取得有效结果。

4.2.4 假新闻检测。假新闻可以利用多媒体在社交媒体上引发恐慌。文本和图像内容会曲解事实并操纵人类心理，从而导致假新闻迅速传播 [144]。多模态架构可以通过查找不同模态之间的不匹配来检测假新闻 [145]。

文献 [146] 针对未见过和新出现事件的假新闻检测问题，采用对抗神经网络予以解决。该网络利用事件判别器移除事件特定特征，并保留事件和模态之间的共享特征。Khattar 等人 [147] 提出了 SpotFake，该方法利用 BERT 学习文本特征，并使用 VGG-19（在 ImageNet 上预训练）学习视觉特征。他们采用简单的拼接技术来组合从不同模态中提取的特征。类似地， [147] 提出了多模态变分自编码器（MVAE），由编码器、解码器和假新闻检测器模块组成。变分自编码器通过优化观测数据边际似然的下界，利用概率潜变量模型。借助从双模态变分自编码器中获得的多模态表征，假新闻检测器将多模态帖子分类为真实或虚假。

近期，Wang 等人 [148] 提出了用于检测虚假新闻的细粒度多模态融合网络（FMFN）。首先，卷积神经网络提取视觉特征，并使用 RoBERTa [70] 获取词语的上下文嵌入。随后，在视觉特征和文本特征之间使用注意力机制，以增强二者的相关性并融合特征。最后，采用二元分类器对融合后的特征进行检测。该模型仅针对同时包含文本和图像的推文，在微博数据集[144] 上取得了 88% 的准确率。

4.2.5 视觉常识推理。视觉常识推理通过将认知、视觉落地和语言推理与识别任务相结合，实现稳健的视觉理解。Zellers 等人 [149]介绍了 VCR 数据集，其中包含 29 万个多项选择问答（MCQA）问题。作者提出了从识别到认知网络，该网络考虑了认知层面视觉理解中的视觉落地和上下文信息。该模型超越了此前在 VQA [25]上开发的最先进水平。Song 等人 [150] 提出了知识增强视觉与语言 BERT，利用了外部常识知识。与基于 BERT 的方法和先前的 VCR 模型相比，该方法取得了小幅提升。

4.2.6 用于视觉识别的自然语言。NLVR（自然语言视觉识别）任务用于检验给定的一段文本描述与多张图像之间关系的一致性。以往的预训练方法[82, 151, 152]将 NLVR 视为二元分类器。视觉语言导航是一项类似的任务，智能体通过遵循语言指令在真实世界的动态环境中穿行。Zhu 等人提出了基于自监督的辅助推理导航（AuxRN） [153] ，用于从周围环境中观察语义信息，以实现视觉语言导航。换言之，模型会从环境中学习隐式信息、估计导航路线并预测下一个位置。

4.3 生成

在多模态生成任务中，输入模态提供的数据经过处理后，会生成另一种模态，或以改写的形式输出为相同模态。生成类别涵盖的任务高度依赖前面提到的理解和分类任务。图像描述任务会先对输入图像中的物体进行分类，然后生成文本。同样，文本到图像生成任务会理解输入文本，并据此生成目标图像。

4.3.1 视觉描述。视觉描述是将图像或视频模态转换为相关文本模态的任务。输入是图像像素，输出则是语法和语义上都有意义的文本。根据图像生成图像描述，弥合了图像与文本模态中低层和高层语义特征之间的差距。图像描述领域一些最先进的数据集包括 MS-COCO [53] 和 Flickr30k [92]。基准和传统的视觉描述流程由图像编码器和文本解码器组成。近期研究利用了 BERT 的能力，例如 [154]。BERT 模型 [35] 最初使用大型文本语料库进行训练，随后采用右侧掩码预训练技术，对融合后的文本与视觉模态嵌入进行微调。Vinyals 等人 [155] 提出了一种基于神经网络的端到端系统，并结合了用于视觉与语言模型的 SOFA 子网络。本文介绍的方法显著超越了以往方法的性能。2017年，Chen 等人 [156] 改进了卷积神经网络，并基于卷积神经网络能够从图像中提取的空间、通道和多层特征，提出了 SCA-CNN。SCA-CNN 能够在卷积神经网络中整合空间与通道注意力，并在多层特征图中动态调节句子生成的上下文。与以往研究不同，Rennie 等人 [157] 使用强化学习来优化图像描述系统。他们提出，使用指标 SCST（自批评序列训练）优化 CIDEr 非常有效。

4.3.2 视觉问答。视觉问答（VQA）旨在通过从视觉线索中检索信息来回答语言问题 [45]。VQA 将书面问题与高维视觉图像或视频中的信息结合起来。问题范围从简单的是非题到基于知识的问题和开放式问题不等；视觉内容则可以是简单的草图或图像，也可以是视频。此外，VQA 融合了自然语言处理和计算机视觉领域的多项功能，例如语言理解、关系抽取、属性与物体类别分类、计数、知识库以及常识推理 [1]。基准 VQA 分别将问题文本和通过循环神经网络与卷积神经网络（RNN 和 CNN）获得的视觉嵌入映射到共同的向量空间。将嵌入表示映射到共享向量空间，使 VQA 能够处理开放式自由形式的问题。在相关文献中，VQA 深度学习技术分为联合嵌入模型 [158], 、注意力机制 [159],、组合模型 [160], 、图神经网络（GNN） [60] 和知识库模型 [161]。已有综述论文将 VQA 作为独立研究领域进行探讨，例如 [45] 和 [5]。除综述外，还有卡内基梅隆大学（CMU）的基准测试 [162] 和 [163]。基准数据集包括 VAQ v1. 0 [5], 、VAQ-X [164], 和 VAQ-CP [165]。

4.3.3 视觉生成。 视觉生成也称为文本到图像生成，通常使用输入文本生成图像。视觉生成是图像描述的反向过程。2016年，Reed 等人 [166] 首次提出了深度卷积生成对抗网络（GAN），根据文本描述合成图像。该训练模型基于深度卷积生成对抗网络（DC-GAN），但与传统生成对抗网络不同的是，判别器 D 的输入中加入了真实图像与虚假文本描述对。作者训练了一个卷积神经网络（CNN），利用生成器 G 生成的图像预测风格。预测出的风格可用于 G 的构成。本文使用的数据集包括：加州理工学院-加州大学圣迭戈分校鸟类数据集 [167]、牛津102类花卉数据集和 MS COCO [53] 数据集。Xu 等人 [168] 提出了注意力生成对抗网络（AttnGAN），用于细粒度文本到图像生成。AttnGAN 由注意力生成网络和深度注意力多模态相似度模型（DAMSM）组成，重点关注相关词语，以描绘特定的图像子区域。DAMSM 通过细粒度图像到文本匹配损失来增强生成器训练。综合评估结果表明，AttnGAN 的表现显著优于以往的生成对抗网络模型。该方法在 CUB [167]和 COCO [53] 数据集上进行了评估。

4.3.4 视听语音合成。Prajwal 等人 [129] 提出了一种能够在存在噪声的情况下根据唇部运动合成语音的模型。该作者发布了基准测试 Lip2Wav 数据集，因为先前的工作 [169] 是在

小规模且受限的基于词汇表的数据集上进行评估的 [127, 170, 171]。该数据集旨在使模型能够根据不受限制的唇部动作合成语音。

高质量语音由 [172]受 Tacotron [173]启发的解码器生成，该解码器根据文本输入生成梅尔频谱图。它以先前表示中编码的人脸嵌入为条件，在唇动转语音任务的所有客观指标上均优于先前工作 [174–176] ，尤其是在 TIMIT [171] 数据集上。

4.4 检索

检索系统将用户输入作为查询，并结合上下文信息提供最相关的结果。多媒体增强的演进体现为成熟的跨模态任务。

4.4.1 指代表达理解。在指代表达理解（REC）应用中，使用一种表达来指称并定位图像中的目标对象。指代表达具有明确针对性，会详细描述对象属性及其与周围环境的关系。因此，需要考虑视觉属性、关系和上下文信息。在 [50], REC中，监督式深度学习架构分为单阶段或两阶段。两阶段流程中，第一阶段根据图像生成并列出候选对象；第二阶段对指代表达进行编码，并计算候选对象与编码后的指代表达之间的匹配分数。在单阶段流程中，图像和语言特征经过拼接后输入模型，通过一个更快的步骤完成处理。用于REC的部分任务无关的视觉与语言应用预训练模型包括 VL-BERT [177] 和 ViLBERT [9]。

4.4.2 视觉语言检索。构建信息检索（IR）应用需要经过索引、查询构建、检索和评估几个步骤。在索引和查询构建过程中，文档和用户界面查询分别通过其特征进行表示 [178]。随后，检索系统将这两种表示映射起来，以检索或提取所需的有用信息。检索任务的性能根据召回率和精确率进行评估。在多模态信息检索中，系统会搜索包含不同模态的文档，例如文本、图像、视频，或生理信号和图像。使用多种模态能够丰富信息检索过程。电子健康记录就是多模态信息检索的一个应用实例。

4.4.3 电子健康记录。患者健康记录包含多种模态，例如分类数据、文本、磁共振成像扫描等图像，或心电图（ECG）等信号 [178]。信息检索系统能够从不同模态中提取信息，用于报告、呈现和/或预测患者的健康状况。例如，监督式深度患者表征学习框架（SDPRL）利用不同模态的信息来学习患者表示[179]。SDPRL 基于 MIMIC-III 基准数据集构建并进行测试。Chen 等人 [180] 针对重度抑郁障碍（MDD）检测问题，提出了一种新颖的、基于图神经网络（GNN）的多模态融合策略，称为“模态共享与模态特定图神经网络”。该方法考虑了心理生理模态之间的异质性/同质性以及模态间/模态内特征，并采用重建网络和注意力机制来获得紧凑的多模态表示。同样，多模态图神经网络框架也被用于利用基因表达、拷贝数变异和临床数据等多模态数据预测癌症生存期 [181]。

4.5 多模态翻译（MMT）

由于图像中缺乏足够的特征，视觉模态被认为对机器翻译的帮助有限。Caglan 等人 [182]提出，当源句缺乏语言上下文时，视觉信息有助于机器翻译（MT）。作者采用颜色退化技术，

剥夺、实体掩码和渐进式掩码来退化源句。此外，研究还输入无关图像（违反语义兼容性）以评估该方法对视觉信息的敏感度。该方法使用基于 ResNet-50 卷积神经网络的编码器 [24] 提取视觉特征。受 [183] 启发的用于神经机器翻译的多模态注意力（NMT）从文本和视觉特征中获得上下文向量。训练使用了 Multi30K 和 flicker30 数据集。多模态机器翻译（MMT）的表现优于 NMT。作者提出了一种用于神经机器翻译（NMT）的新型编码器 [60]，该编码器采用基于图的方法来组合多种模态。该方法利用不同数据类型之间细致的语义联系，增强组合表示的学习。

Su 等人 [184] 提出了一种利用跨模态语义表示交互的方法，其灵感来自问答和命名实体识别领域。他们提出了两种基于注意力的模型：一种利用双向注意力机制，从文本与视觉特征的交互中进行学习；另一种采用协同注意力机制，根据文本表征生成上下文向量，随后利用该向量结合更多交互信息更新文本上下文向量。这些模型在扩展后的 Multi30K 数据集上进行了评估，无论是否经过预训练，其性能均优于以往的基线方法。

5 数据集

本节总结了多模态模型预训练、微调和评估所用的基准数据集。据我们所知，我们涵盖了所有包含图像、文本、视频和音频模态的基准数据集。本节旨在为新入门的研究人员提供可直接使用的信息，包括基准测试的性质、各基准测试中的样本数量、对不同任务的适用性，以及表 4 所示的对比信息。MS COCO 数据集 [53] 用于图像识别、图像检测、图像分割和图像描述。数据集中的每张图像都配有五种不同的描述。该数据集包含 91 种易于识别的物体类别，其中有 30 多万张图像和 250 万个标注实例。

Flickr30k [92] 收集自 Flickr，并附有人工标注者提供的 5 条图像描述。Flickr8k [185]是最小版本，同样收集自 Flickr，因此更易于快速训练模型。FUNSD [119] 是一个小型数据集，包含 199 份真实且经过完整标注的扫描表单。FUNSD 包含 31485 个词语、9707 个语义实体和 5304 个关系，旨在提取并构建表单中的文本内容。SROIE [118] 包含一个由 1000 张完整扫描的收据图像及其标注组成的数据集，用于扫描收据光学字符识别和关键信息提取竞赛（SROIE）。卡内基梅隆大学多模态观点、情感极性与情绪强度（CMU-MOSEI）数据集 [186] 是迄今为止规模最大、用于多模态情感分析和情绪识别的数据集。2018年，CMU-MOSEI 数据集创建完成，包含 23,453 个经过标注的视频片段。这些片段来自 1,000 个不同的 YouTube 视频，涵盖 250 个不同主题。

视觉问答 [25] 是一个大型数据集，包含超过 250,000 张图像，每张图像至少有 3 个问题，每个问题有 10 个真实答案。VQA v1 提供了 6,141,630 个真实答案和 1,842,489 个合理答案，使模型更难正确回答问题。VAQ-CP（先验变化下的视觉问答）是 VQA v1 和 VQA v2 数据集的划分版本。CrisisMMD [140] 是一个规模庞大的自然灾害多模态数据集，采集自 Twitter，涵盖 2017 年世界各地发生的地震、飓风、野火和洪水。它包含三类标注。第一类是“有信息价值”或“无信息价值”，用于判断内容是否有助于人道主义援助。第二类是“人道主义类别”，例如基础设施和公用设施损坏、车辆损坏等。最后一类是“损害严重程度评估”，用于描述损害的严重程度。

MIMIC（重症监护医疗信息集市）[187] 是由计算生理学实验室开发并向公众开放的数据集，包含与数千例重症监护病房入院病例相关的去标识化健康数据。目前，该数据集有三个版本：MIMIC-II、MIMIC-III 和 MIMIC-IV。MIMIC-III 包含 2001 年至 2012 年间入住重症监护病房的 53,423 名成年人的信息，例如患者的性别、身高等基本信息，以及血常规、肝功能等医院检查数据和用药信息。

表 4. 本研究中各任务所用基准数据集的信息。任务列列出了相应论文中的任务。

数据集	模态	数量	任务	数据类别
MS COCO [53]	文本 + 图像	328k	图像识别、图像分类	多种
Conceptual Captions [90]	文本 +图像	3.3M	视觉问答、视觉常识推理、信息检索、图像分类	多种
Flickr30k [92]	文本 + 图像	31k	信息检索、跨模态检索、图像分类	多种
Flickr8K [185]	文本 + 图像	8k	信息检索、跨模态检索、图像分类	多种
FUNSD [119]	文本 +图像	199	文本检测、光学字符识别、实体链接	扫描表单
SROIE [118]	文本 +图像	1k	光学字符识别，信息抽取	扫描图像
CMU-MOSEI [186]	文本 + 视频	23,453	MMSA	YouTube 视频
视觉问答 [25]	文本 + 图像	250k	VQA	多个
CrisisMMD [140]	文本 + 图像	16,097	图像分类，事件检测	灾害推文
MIMIC [187]	MML 数据	53,423	IR	患者信息
Fashion-200K [188]	文本 + 图像	200k	图像分类，信息检索	服装
纽约大学深度数据集 v1 [189],v2 [190]	红绿蓝 + 深度	4GB，90GB	IS，图像分类	室内场景
SKIG [74]	文本 +图像	2160	GR	手势
GoodNews [191]	文本 + 图像	466,000	IC	News
MSR-VTT [192]	文本 + 视频	200k	VU	商业视频
MSVD-QA [193]	文本 + 视频	2089	视觉推理、视觉问答、视频描述	多种
TGIF-QA [194]	文本 + 图像（gif）	103,919	VQA	多种
EQA-v1 [195]	文本 + 3D 环境。	9,000	EQA、视觉问答	多种
VideoNavQA [196]	文本 + 3D 环境。	28	EQA、视觉问答	多种
TDIUC [197]	文本 + 图像	1,654,167	VQA	多种
nuScenes [198]	图像 +传感器数据	1.4M	信息检索、自动驾驶	驾驶信息
CUB-200 [167]	文本 + 图像	11,788	VG	鸟类
牛津102种花卉 [199]	文本 +图像	4,080-26,316	VG	Flowers
视觉常识推理 [149]	文本 + 图像	263k	VU	电影场景
如何2 [125]	音频 +文本 + 视频	2000 小时	MML	教学视频
Lip2Wav [129]	音频 + 视频	120 小时	AVSS	说话人脸视频
MUSIC [133]	音频 + 视频	685	SSS	音乐视频
MOSI [200]	音频 +文本 + 视频	3702	多模态分析	YouTube 视频
GRID [201]	音频 + 视频	33,000	VSR	语音

Fashion200K [188] 包含 20 万张时尚图像，每张图像都附有简洁的类属性商品描述。MIT-Stata 中心 [202] 是一个多模态数据集，包含视觉（立体视觉和 RGB-D）、激光及本体感知数据。该数据集容量超过 2.3 TB，包含 38 小时的数据和 42 千米的记录。该数据集还包含机器人在每个时刻的真实位置估计。这是用于机器人建图和计算机视觉研究的重要数据集。NYU-Depth 数据集由微软 Kinect 的红绿蓝和深度摄像机录制的视频序列构成。NYU-Depth v1 [189] 提供约 4GB 的标注数据和约 90GB 的原始数据。NYU-Depth v1 包含 64 个不同的室内场景和 7 种场景类型。NYU-Depth v2 [190]包含 464 个不同的室内场景和 26 种场景类型。

Shefeld Kinect Gesture（SKIG） [74] 是一个手势数据集，包含 2160 个手势序列。这些手势可分为 10 类：顺时针画圆、逆时针画三角形、上下、左右、挥手、“Z”形、交叉、招手、转身和轻拍。这些序列在 3 种不同背景和 2 种光照条件下录制，因而具有多样性。GoodNews [191] 是一个大型数据集，包含 466,000 张图像及其图像描述、标题和文本文章。然而，与 MSCOCO 或 Flicker8k 等数据集不同，GoodNews 为每张图像只提供一条真实描述。由专业期刊撰稿人撰写的 GoodNews 图像描述平均长度较长，高于通用图像描述数据集，这意味着这些描述更加详尽。

MSR-VTT [192]是一个大规模视频描述数据集，包含 1 万个网络视频片段，共计 38.7 小时，以及 20 万个视频片段-句子对。MSR-VTT 根据某商业视频搜索引擎的 257 个热门查询创建。MSR-VTT 中的每个片段都标注了约 20 条自然语言句子。该数据集用于视频理解。微软研究院视频描述语料库（MSVD）数据集 MSVD-QA 包含对 2089 个短视频片段（通常不足 10 秒）的 12.2 万条描述。MSVD 数据集包含多种语言的描述，如英语、印地语、罗马尼亚语、斯洛文尼亚语等。MSVD-QA 是用于视频检索、视觉问答和视频描述的基准测试。

TGIF-QA [194]是一个大规模数据集，包含从 56,720 个动画 GIF 中收集的 103,919 个问答对。这些 GIF 来自 TGIF 数据集。TGIF 数据集基于 GIF 数据，因为 GIF 格式简洁且叙事连贯。TGIF-QA 可用于视觉问答研究。EQA（具身问答）v1.0[195] 是一个包含来自 774 个环境的 9,000 个问题的数据集。该数据集中的视觉问题和答案以 House3D 为基础。EQA-v1 包含有关位置、颜色和地点介词的问题。

VideoNavQA [196]也是一个用于研究 EQA 任务的数据集。VideoNavQA 包含 28 个问题，分属 8 个类别，共有 70 个可能的答案。VideoNavQA 中问题的复杂度远高于类似任务；这类任务使用生成方法，从视频中提取真实信息来生成问题。任务导向图像理解挑战赛 (TDIUC) [197] 是一个包含 167,437 张图像和 1,654,167 组问答对的数据集。TDIUC 将 VQA 划分为 12 个组成任务，从而更容易衡量和比较视觉问答算法的性能。12 种不同的问题类型按照这些任务进行分组。

nuScenes [198] 是一个面向自动驾驶的大规模公开数据集，包含三维物体标注。它也是一个多模态数据集。nuScenes 提供了 140 万张相机图像、来自 4 座城市（波士顿、匹兹堡、拉斯维加斯和新加坡）的 1500 小时驾驶数据，以及公开发布的 150 小时传感器数据（5 台激光雷达、8 台相机、惯性测量单元、全球定位系统）、详细的地图信息、针对 23 个物体类别人工标注的 140 万个三维边界框等。nuScenes 可用于智能体研究。nuImages 是一个包含图像级二维标注的大规模自动驾驶数据集。它包含 9.3 万个时长各为 6 秒的视频片段、9.3 万张已标注图像和 110 万张未标注图像。加州理工学院—加州大学圣迭戈分校鸟类数据集 (CUB-200) [167] 包含 200 种不同鸟类物种的图像，共计 11,788 张。该数据集中的每张图像都带有标注，包括鸟类周围的边界框、鸟类的基本分割，以及各种属性的标签。牛津102类花卉数据集包含来自 102 个鸟类类别的图像，每个类别包含 40 至 258 张图像。这些图像呈现出丰富多样的尺寸、姿势和光照条件。

视觉常识推理（VCR）[149] 包含从 11 万个电影场景中提取的超过 21.2 万（训练集）、2.6 万（验证集）和 2.5 万（测试集）个问题、答案和理由。它被广泛用于认知层面的视觉理解。How2[125] 是一个多模态数据集，包含带有英文字幕和众包葡萄牙语翻译的教学视频。How2 涵盖了丰富多样的主题，共有 8 万个片段（约 2,000 小时）。Lip2Wav[129] 数据集包含 5 位说话人的 120 小时说话人脸视频。该数据集中，每位说话人约有 20 小时的自然语音，词汇量均超过 5,000 个词语。MUSIC（乐器组合的多模态来源）数据集[133]包含 685 个未经剪辑的器乐独奏与二重奏视频。该数据集涵盖 11 种乐器类别：手风琴，

原声吉他、大提琴、单簧管、二胡、长笛、萨克斯管、小号、大号、小提琴和木琴。MOSI[200]（多模态意见级情感强度）数据集包含 93 个视频和 3,702 个视频片段。该数据集不仅提供情感标注，还提供人工手势标注。GRID[201]由 54 位说话人的视频录制内容构成，每位说话人有 100 条话语，共计 33,000 条话语。

6 未来展望

本节介绍研究人员从以往研究中观察到，并认为可采用和考虑的技术，以取得更好的结果并降低计算成本。此外，尽管这些方法在相关研究中表现出良好的效果，但近期研究尚未对其进行充分考虑或深入探索。

构建统一模型。 以往研究证明了 Transformer 在自然语言处理、计算机视觉和多模态任务中的有效性。近期，研究人员提出了一种统一架构，以单智能体Transformer作为骨干网络，处理多种模态并完成多项任务 [203]。尽管该架构在下游任务上的表现令人瞩目，但在部分任务上仍无法超越特定任务模型 [204]。对于零样本学习，性能很大程度上取决于指令。经过优化的指令可以带来更令人满意的结果。研究人员正努力解决模型对提示和参数的细微变化表现出的敏感性这一挑战。

利用GNN进行多模态预训练。 Ektefai等人 [205] 提出了多模态图学习（MGL），用于处理多模态数据输入，并为各种下游任务生成通用的输出表示。该方法首先识别不同数据模态中的相关实体，并将其映射到共享命名空间中，随后结合这些不同模态。该过程利用消息传递模块，根据模态内和模态间的邻接矩阵学习节点表示。最后，通过聚合过程生成适用于下游任务的节点、子图或图级表示。

预训练策略。 为多模态任务开发一组最优的预训练目标值得探索。前置任务会直接影响下游任务的表现。选择预训练目标时，需要考虑模态、数据集、网络架构和下游任务。多模态方法有望采用高效的分阶段预训练，正如 Bao等人[33] 在单一模态中成功采用的做法。

利用多语言特征增强预训练。 由于人们对纯英语多模态基准数据集进行了大量实验，多模态系统的多语言预训练尚未得到充分研究。UC2 [84], M3P [206], 和MURAL [207]通过为多语言或翻译后的数据添加单独的编码器，利用多语言特征。基于 ALBEF [209] 的 CCLM [208] 提出了一种方法，在零样本跨语言设置中超越了最先进水平。总之，多语言视角对于增强多模态预训练框架的扩展效果至关重要 [210]。

提示调优。 提示调优在 GPT-2 和 GPT-3 之后受到关注。这项技术将隐含信息转换为问题或描述性文本，从而有助于利用语言模型的文本概率。采用预训练目标（掩码语言模型）的提示学习可以应对微调面临的两项挑战：i）通过避免为每个下游任务设置不同的参数，降低计算复杂度；ii）缩小预训练任务与下游任务的表征学习之间的差距 [211]。目前，多模态方法尚未探索这项技术。从技术上讲，提示调优或许能将统一架构转变为更规范、更通用的版本。

掩码阶段。 不同模态之间的交互面临对齐挑战。掩码是一种有效策略，可通过生成文本和图像来解决对齐问题。掩码可应用于不同阶段，例如输入级和任务级；但根据 Zhuge 等人， [212], 嵌入级掩码是对齐不同模态表示最有效的阶段。如果我们专注于音频、视频和文本模态之间的最佳对齐，就能学习到细粒度表示。

知识注入。为了让模型充分掌握视觉、语言和音频信息，有必要将外部知识融入多模态表示中，正如 Chen 等人 [213] 将说明性知识用于视觉与语言任务。我们需要更智能的架构，将知识库特征融入多模态预训练和下游任务。与不利用知识蒸馏的其他变体相比，知识蒸馏通过提供有效的学习表示，提升了 ViT-BERT 在所有任务上的表现。在 SST-2 和 IN-1K 上，准确率分别提高了 10% 和 5%。因此，在设计预训练目标和认知架构时，必须考虑常识、情境知识、层次结构、结构信息或网络信息。

考虑预期评估。深度学习架构需要大量计算资源来进行实验。当前的评估方法会在大量实验后提供有关模型有效性的信息。对于大规模多模态方法，必须制定评估策略，在较早阶段检查模型 [42]，以验证模型是否与下游任务兼容，然后再投入计算成本。

多模态架构的加速与扩展。除了知识蒸馏之外，[214], 剪枝和量化仍有待探索，以压缩和加速模型，并提升跨模态推理速度。受大型版预训练语言模型性能的启发，人们通过大规模训练和复杂架构，努力在多模态任务上取得成功。然而，要以更高的熟练度应对大规模多模态场景，仍需要成熟的基准数据集和模型 [215, 216]。

包括音频流。近期专门针对预训练的多模态综述 [41, 42] 尚未考虑音频流。音频与文本一样，语义信息丰富，还能提供有关说话人的情感和补充信息；在存在多位说话人的情况下，也能提供边界数据。利用音频进行预训练对于统一架构极为重要，因为这使模型能够执行包含音频流的下游任务[217]。然而，随着模态数量增加，对齐和对应关系方面的挑战也会加剧。

7 结论

本综述探讨了深度多模态学习及其架构在有效处理高级多模态任务方面所发挥的作用。我们首先介绍了基于深度学习、面向特定任务的多模态架构，这些架构以编码器-解码器、注意力和强化学习为基础。随后，我们回顾了硬件、大规模计算资源及预训练方法方面的进展。预训练和微调能够缓解多模态及跨模态任务中的各种挑战，包括多任务处理。因此，我们讨论了预训练的类型和任务，这些内容有助于提升训练过程的有效性；还对近期产生重大影响的最先进水平的基于 Transformer 的预训练方法进行了比较总结。我们进一步指出，研究界致力于大规模预训练，借助以 Transformer 为骨干网络的统一通用架构，在参数相同的情况下执行更复杂的任务。我们还介绍了面向不同多模态应用的最先进水平的多模态架构，以及它们在基准数据集上的表现。这些应用展现了多模态系统的实用性和现实价值。多模态的潜力巨大，但迄今为止，其能力仅在英语环境下得到探索。未来可能的研究方向之一，是构建其他语言的多模态数据集，并开发多语言框架。

参考文献

[1] JabeenSummaira、Xi Li、Amin Muhammad Shoib、Songyuan Li 和 Jabbar Abdul。《多模态深度学习的最新进展与趋势：综述》。*arXiv*预印本 *arXiv:2105.11087*，2021年。[2] Tadas Baltrušaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。《多模态机器学习：综述与分类体系》。*IEEE*模式分析与机器智能汇刊，41(2):423–443，2018年。

[3] Bowen Shi、Wei-Ning Hsu、Kushal Lakhotia 和 Abdelrahman Mohamed。通过掩码多模态聚类预测学习视听语音表征。载于《国际学习表征会议论文集》，2022年。[4] Dimitar Dimitrov、Bishr Bin Ali、Shaden Shaar、Firoj Alam、Fabrizio Silvestri、Hamed Firooz、Preslav Nakov 和 Giovanni Da San Martino。检测模因中的宣传手法。载于《计算语言学协会第59届年会暨自然语言处理国际联合会议第11届会议论文集》（第1卷：长篇论文），第6603–6617页，2021年。[5] Stanislaw Antol、Aishwarya Agrawal、Jiasen Lu、Margaret Mitchell、Dhruv Batra、C Lawrence Zitnick 和 Devi Parikh。VQA：视觉问答。载于《IEEE国际计算机视觉会议论文集》，第2425–2433页，2015年。[6] Wenzhong Guo、Jianwen Wang 和 Shiping Wang。深度多模态表征学习：综述。IEEE Access，7:63373–63394，2019年。[7] Khaled Bayoudh、Raja Knani、Fayçal Hamdaoui 和 Abdellatif Mtibaa。计算机视觉深度多模态学习综述：进展、趋势、应用与数据集。The Visual Computer，第1–32页，2021年。[8] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。VisualBERT：一种简洁且性能出色的视觉与语言基线方法。arXiv预印本 arXiv:1908.03557，2019年。[9] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练。神经信息处理系统进展，32，2019年。[10] Jiasen Lu、Vedanuj Goswami、Marcus Rohrbach、Devi Parikh 和 Stefan Lee。12-in-1：多任务视觉与语言表征学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10437–10446页，2020年。[11] Harry McGurk 和 John MacDonald。听见嘴唇，看见声音。《自然》，264(5588):746–748，1976年。[12] Pradeep K Atrey、M Anwar Hossain、Abdulmotaleb El Saddik 和 Mohan S Kankanhalli。用于多媒体分析的多模态融合：综述。Multimedia Systems，16:345–379，2010年。[13] Georgios Evangelopoulos、Athanasia Zlatintsi、Alexandros Potamianos、Petros Maragos、Konstantinos Rapantzikos、Georgios Skoumas 和 Yannis Avrithis。基于听觉、视觉和文本注意力的多模态显著性与融合，用于电影摘要。IEEE多媒体汇刊，15(7):1553–1568，2013年。[14] Rainer W Lienhart。自动镜头边界检测算法比较。载于《图像与视频数据库的存储与检索 VII》，第3656卷，第290–301页。SPIE，1998年。[15] Jean Carletta、Simone Ashby、Sebastien Bourban、Mike Flynn、Mael Guillemot、Thomas Hain、Jaroslav Kadlec、Vasilis Karaiskos、Wessel Kraaij、Melissa Kronenthal、Guillaume Lathoud、Mike Lincoln、Masson Agnes Lisowska、Iain McCowan、Wilfried Post、Dennis Reidsma 和 Pierre D. Wellner。AMI会议语料库：预告。载于《多模态交互机器学习：第二届国际研讨会，MLMI 2005，爱丁堡，英国，2005年7月11日至13日，修订精选论文2》，第28–39页。施普林格，2006年。[16] Gary McKeown、Michel F Valstar、Roderick Cowie 和 Maja Pantic。带有情感色彩的角色互动SEMAINE语料库。载于《2010年IEEE多媒体与博览会国际会议论文集》，第1079–1084页。IEEE，2010年。[17] Björn Schuller、Michel Valstar、Florian Eyben、Gary McKeown、Roddy Cowie 和 Maja Pantic。AVEC 2011——首届国际视听情感挑战赛。载于《情感计算与智能交互：第四届国际会议，ACII 2011，美国田纳西州孟菲斯，2011年10月9日至12日，论文集，第二部分》，第415–424页。施普林格，2011年。[18] Michel Valstar、Björn Schuller、Kirsty Smith、Timur Almaev、Florian Eyben、Jarek Krajewski、Roddy Cowie 和 Maja Pantic。AVEC 2014：三维情感与抑郁识别挑战赛。载于《第4届国际视听情感挑战赛研讨会论文集》，第3–10页，2014年。[19] Guanzheng Chen、Fangyu Liu、Zaiqiao Meng 和 Shangsong Liang。重新审视参数高效微调：我们真的已经实现目标了吗？载于《2022年自然语言处理实证方法会议论文集》，第2612–2626页，2022年。[20] Nikhil Rasiwasia、Jose Costa Pereira、Emanuele Coviello、Gabriel Doyle、Gert RG Lanckriet、Roger Levy 和 Nuno Vasconcelos。一种跨模态多媒体检索新方法。载于《第18届ACM多媒体国际会议论文集》，第251–260页，2010年。[21] Amirhossein Habibian、Thomas Mensink 和 Cees GM Snoek。样例稀缺时，Video2vec嵌入仍能识别事件。IEEE模式分析与机器智能汇刊，39(10):2089–2103，2016年。[22] Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton。深度学习。《自然》，521(7553):436–444，2015年。[23] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。国际学习表征会议，2015年。[24] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《IEEE计算机视觉与模式识别会议论文集》，第770–778页，2016年。[25] Stanislaw Antol、Aishwarya Agrawal、Jiasen Lu、Margaret Mitchell、Dhruv Batra、C Lawrence Zitnick 和 Devi Parikh。VQA：视觉问答。载于《IEEE国际计算机视觉会议论文集》，第2425–2433页，2015年。[26] Sepp Hochreiter 和 Jürgen Schmidhuber。长短期记忆。Neural Computation，9(8):1735–1780，1997年。[27] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力即所需的一切。神经信息处理系统进展，30，2017年。[28] Zhe Gan、Linjie Li、Chunyuan Li、Lijuan Wang、Zicheng Liu 和 Jianfeng Gao。视觉语言预训练：基础、最新进展与未来趋势。Foundations and Trends® in Computer Graphics and Vision，14(3–4):163–352，2022年。[29] Allan Jabri、Armand Joulin 和 Laurens van der Maaten。重新审视视觉问答基线方法。载于《欧洲计算机视觉会议论文集》，第727–739页。施普林格，2016年。

[30] Duy-Kien Nguyen 和 Takayuki Okatani。通过密集对称协同注意力改进视觉问答中的视觉与语言表征融合。载于 IEEE计算机视觉与模式识别会议论文集，第6087–6096页，2018年。[31] Ronghang Hu、Anna Rohrbach、Trevor Darrell 和 Kate Saenko。用于关系推理的语言条件图网络。载于 IEEE/CVF国际计算机视觉会议论文集，第10294–10303页，2019年。[32] Peng Gao、Zhengkai Jiang、Haoxuan You、Pan Lu、Steven CH Hoi、Xiaogang Wang 和 Hongsheng Li。通过模态内与模态间注意力流实现视觉问答的动态融合。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第6639–6648页，2019年。[33] Hangbo Bao、Wenhui Wang、Li Dong、Qiang Liu、Owais Khan Mohammed、Kriti Aggarwal、Subhojit Som、Songhao Piao 和 Furu Wei。VLMo：采用混合模态专家的统一视觉语言预训练。神经信息处理系统进展，35：32897–32912，2022年。[34] Wonjae Kim、Bokyung Son 和 Ildoo Kim。ViLT：无需卷积或区域监督的视觉语言变换器。载于国际机器学习会议论文集，第5583–5594页。PMLR，2021年。[35] Jacob Devlin、Ming-Wei Chang、Kenton 和 Lee Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。载于 NAACL-HLT 论文集，第4171–4186页，2019年。[36] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。使用深度卷积神经网络进行 ImageNet 分类。ACM通讯，60(6)：84–90，2017年。[37] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：语音表征自监督学习框架。神经信息处理系统进展，33：12449–12460，2020年。[38] Anil Rahate、Rahee Walambe、Sheela Ramanna 和 Ketan Kotecha。多模态协同学习：挑战、应用、数据集、最新进展与未来方向。信息融合，81：203–239，2022年。[39] Jing Gao、Peng Li、Zhikui Chen 和 Jianing Zhang。多模态数据融合深度学习综述。Neural Computation，32(5)：829–864，2020年。[40] Fei-Long Chen、Du-Zhen Zhang、Ming-Lun Han、Xiu-Yi Chen、Jing Shi、Shuang Xu 和 Bo Xu。VLP：视觉语言预训练综述。机器智能研究，20(1)：38–56，2023。[41] Yifan Du、Zikang Liu、Junyi Li 和 Wayne Xin Zhao。视觉语言预训练模型综述。载于第三十一届国际人工智能联合会议（IJCAI-22）综述专题论文集，2022年。[42] Siqu Long、Feiqi Cao、Soyeon Caren Han 和 Haiqin Yang。视觉语言预训练模型：综述。载于 Luc De Raedt 编，《第三十一届国际人工智能联合会议论文集》，IJCAI 2022，第5530–5537页。ijcai.org，2022年。[43] Dhanesh Ramachandram 和 Graham W Taylor。深度多模态学习：最新进展与趋势综述。IEEE信号处理杂志，34(6)：96–108，2017年。[44] Aditya Mogadala、Marimuthu Kalimuthu 和 Dietrich Klakow。视觉与语言研究融合趋势：任务、数据集与方法综述。人工智能研究杂志，71：1183–1317，2021年9月。[45] Qi Wu、Damien Teney、Peng Wang、Chunhua Shen、Anthony Dick 和 Anton Van Den Hengel。视觉问答：方法与数据集综述。计算机视觉与图像理解，163：21–40，2017年。[46] Katikapalli Subramanyam Kalyan、Ajit Rajasekharan 和 Sivanesan Sangeetha。AMMU：基于变换器的生物医学预训练语言模型综述。生物医学信息学杂志，126：103982，2022年。[47] Kejing Xiao、Zhaopeng Qian 和 Biao Qin。多模态事件检测与演进的数据表征综述。《应用科学》，12(4)：2204，2022年。[48] Lukas Stappen、Alice Baird、Lea Schumann 和 Schuller Bjorn。汽车评论多模态情感分析（MUSE-CAR）数据集：收集、洞见与改进。IEEE情感计算汇刊，2021年。[49] Ganesh Chandrasekaran、Tu N Nguyen 和 Jude Hemanth D。面向社交媒体应用的多模态情感分析：全面综述。Wiley跨学科评论：数据挖掘与知识发现，11(5)：e1415，2021年。[50] Yanyuan Qiao、Chaorui Deng 和 Qi Wu。指代表达理解：方法与数据集综述。IEEE多媒体汇刊，23：4426–4440，2021年。[51] Jie Wu 和 Haifeng Hu。用于图像描述生成的级联循环神经网络。电子学快报，53(25)：1642–1643，2017年。[52] Jiayi Ji、Yunpeng Luo、Xiaoshuai Sun、Fuhai Chen、Gen Luo、Yongjian Wu、Yue Gao 和 Rongrong Ji。利用变换器网络中的层内与层间全局表征改进图像描述。载于 AAAI人工智能会议论文集，第35卷，第1655–1663页，2021年。[53] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于《计算机视觉—ECCV 2014：第13届欧洲会议，瑞士苏黎世，2014年9月6日至12日，论文集，第V部分13》，第740–755页。施普林格，2014年。[54] Jiquan Ngiam、Aditya Khosla、Mingyu Kim、Juhan Nam、Honglak Lee 和 Andrew Y Ng。多模态深度学习。载于第28届国际机器学习会议（ICML-11）论文集，第689–696页，2011年。[55] Pascal Vincent、Hugo Larochelle、Yoshua Bengio 和 Pierre-Antoine Manzagol。使用去噪自编码器提取并组合鲁棒特征。载于第25届国际机器学习会议论文集，第1096–1103页，2008年。[56] Xiangru Zhu、Zhixu Li、Xiaodan Wang、Xueyao Jiang、Penglei Sun、Xuwu Wang、Yanghua Xiao 和 Nicholas Jing Yuan。多模态知识图谱构建与应用：综述。IEEE知识与数据工程汇刊，2022年。

[57] Yujia Li、Richard Zemel、Marc Brockschmidt 和 Daniel Tarlow。门控图序列神经网络。载于国际学习表征会议论文集，2016年。[58] Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。stat，1050(20)：10–48550，2017年。[59] Yongjing Yin、Fandong Meng、Jinsong Su、Chulun Zhou、Zhengyuan Yang、Jie Zhou 和 Jiebo Luo。用于神经机器翻译的新型基于图的多模态融合编码器。载于计算语言学协会第58届年会论文集，第3025–3035页，2020年。[60] Difei Gao、Ke Li、Ruiping Wang、Shiguang Shan 和 Xilin Chen。用于视觉与场景文本联合推理的多模态图神经网络。载于IEEE/CVF计算机视觉与模式识别会议论文集，第12746–12756页，2020年。[61] Yanan Wang、Michihiro Yasunaga、Hongyu Ren、Shinya Wada 和 Jure Leskovec。VQA-GNN：利用多模态语义图进行视觉问答推理。arXiv预印本 arXiv:2205.11501，2022年。[62] Weitao Jiang、Xiying Li、Haifeng Hu、Qiang Lu 和 Bohong Liu。用于图像描述的多门控注意力网络。IEEE Access，9：69700–69709，2021年。[63] Kan Chen、Trung Bui、Chen Fang、Zhaowen Wang 和 Ram Nevatia。AMC：注意力引导的图像搜索多模态相关性学习。载于 IEEE计算机视觉与模式识别会议论文集，第2644–2652页，2017年。[64] Xin Wang、Wenhu Chen、Jiawei Wu、Yuan-Fang Wang 和 William Yang Wang。通过分层强化学习生成视频描述。载于 IEEE计算机视觉与模式识别会议论文集，第4213–4222页，2018年。[65] Jinyuan Fang、Shangsong Liang、Zaiqiao Meng 和 Qiang Zhang。用于关系学习的图卷积核高斯过程。载于第27届ACM SIGKDD知识发现与数据挖掘会议论文集，第353–363页，2021年。[66] Guanzheng Chen、Jinyuan Fang、Zaiqiao Meng、Qiang Zhang 和 Shangsong Liang。基于贝叶斯高斯过程网络的多关系图表示学习。载于 AAAI人工智能会议论文集，第36卷，第5530–5538页，2022年。[67] Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jianfeng Gao。VinVL：重新审视视觉语言模型中的视觉表征。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第5579–5588页，2021年。[68] Junyang Lin、Rui Men、An Yang、Chang Zhou、Yichang Zhang、Peng Wang、Jingren Zhou、Jie Tang 和 Hongxia Yang。M6：用于统一预训练的多模态到多模态多任务巨型变换器。载于第27届ACM SIGKDD知识发现与数据挖掘会议论文集，第3251–3261页，2021年。[69] Kevin Clark、Minh-Thang Luong、Quoc V. Le 和 Christopher D. Manning。ELECTRA：将文本编码器作为判别器而非生成器进行预训练。载于国际学习表征会议论文集。OpenReview.net，2020年。[70] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种鲁棒优化的BERT预训练方法。arXiv预印本 arXiv:1907.11692，2019年。[71] Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型。生物信息学，36(4)：1234–1240，2020年。[72] Tommaso Caselli、Valerio Basile、Jelena Mitrović 和 Michael Granitzer。HateBERT：针对英语辱骂性语言检测重新训练BERT。载于第5届在线辱骂与伤害研讨会（WOAH 2021）论文集，第17–25页，2021年。[73] Zewen Chi、Li Dong、Furu Wei、Nan Yang、Saksham Singhal、Wenhui Wang、Xia Song、Xian-Ling Mao、He-Yan Huang 和 Ming Zhou。InfoXLM：跨语言语言模型预训练的信息论框架。载于2021年计算语言学协会北美分会：人类语言技术会议论文集，第3576–3588页，2021年。[74] Shoya Wada、Toshihiro Takeda、Shiro Manabe、Shozo Konishi、Jun Kamohara 和 Yasushi Matsumura。用于医学BERT本地化并增强生物医学BERT的预训练技术。arXiv预印本 arXiv:2005.07202，2020年。[75] Suchin Gururangan、Ana Marasović、Swabha Swayamdipta、Kyle Lo、Iz Beltagy、Doug Downey 和 Noah A Smith。不要停止预训练：使语言模型适应领域与任务。载于计算语言学协会第58届年会论文集，第8342–8360页，2020年。[76] Yujia Qin、Yankai Lin、Jing Yi、Jiajie Zhang、Xu Han、Zhengyan Zhang、Yusheng Su、Zhiyuan Liu、Peng Li、Maosong Sun 和 Jie Zhou。预训练语言模型的知识继承。载于2022年计算语言学协会北美分会：人类语言技术会议论文集，第3921–3937页，2022年。[77] Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。通过生成式预训练改进语言理解。不列颠哥伦比亚大学，2018年。[78] Subhadarshi Panda、Anjali Agrawal、Jeewon Ha 和 Benjamin Bloch。通过乱序词元检测改进预训练RoBERTa。载于2021年计算语言学协会北美分会学生研究研讨会论文集，第88–93页，2021年。[79] Xunlin Zhan、Yangxin Wu、Xiao Dong、Yunchao Wei、Minlong Lu、Yichi Zhang、Hang Xu 和 Xiaodan Liang。Product1M：通过跨模态预训练实现弱监督实例级商品检索。载于 IEEE/CVF国际计算机视觉会议论文集，第11782–11791页，2021年。[80] Junyang Lin、An Yang、Yichang Zhang、Jie Liu、Jingren Zhou 和 Hongxia Yang。InterBERT：用于多模态预训练的视觉与语言交互。arXiv预印本 arXiv:2003.13198，2020年。

[81] Yongfei Liu、Chenfei Wu、Shao-Yen Tseng、Vasudev Lal、Xuming He 和 Nan Duan。Kd-vlp：通过目标知识蒸馏改进端到端视觉语言预训练。载于《计算语言学协会：NAACL 2022论文集》，第1589–1600页，2022年。[82] Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng 和 Jingjing Liu。Uniter：通用图像文本表征学习。载于《欧洲计算机视觉会议》，第104–120页。施普林格，2020年。[83] Linjie Li、Yen-Chun Chen、Yu Cheng、Zhe Gan、Licheng Yu 和 Jingjing Liu。Hero：用于视频语言全方位表征预训练的层次编码器。载于《2020年自然语言处理实证方法会议论文集》（EMNLP），第2046–2065页，2020年。+ Mingyang Zhou、Luowei Zhou、Shuohang Wang、Yu Cheng、Linjie Li、Zhou Yu 和 Jingjing Liu。Uc2：通用跨语言跨模态视觉语言预训练。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4155–4165页，2021年。[84] Zhenzhong Lan、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Soricut。Albert：用于语言表征自监督学习的轻量级BERT。载于《国际学习表征会议论文集》，2019年。[85] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。探索采用统一文本到文本变换器进行迁移学习的极限。《机器学习研究期刊》，21(1)：5485–5551，2020年。[86] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。使用Transformer进行端到端目标检测。载于《欧洲计算机视觉会议》，第213–229页。施普林格，2020年。[87] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。可变形DETR：用于端到端目标检测的可变形Transformer。载于《国际学习表征会议》，2021年。[88] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移的视觉模型。载于《国际机器学习会议》，第8748–8763页。PMLR，2021年。[89] Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。Conceptual Captions：一个经过清洗和上位词化、用于自动图像描述的图像替代文本数据集。载于《计算语言学协会第56届年会论文集》（第一卷：长篇论文），第2556–2565页，2018年。[90] Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason Corso 和 Jianfeng Gao。用于图像描述和视觉问答的统一视觉语言预训练。载于《AAAI人工智能会议论文集》，第34卷，第13041–13049页，2020年。[91] Bryan A Plummer、Liwei Wang、Chris M Cervantes、Juan C Caicedo、Julia Hockenmaier 和 Svetlana Lazebnik。Flickr30k实体：收集区域到短语对应关系，构建更丰富的图像到句子模型。载于《IEEE国际计算机视觉会议论文集》，第2641–2649页，2015年。[92] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei、Yejin Choi 和 Jianfeng Gao。OSCAR：面向视觉-语言任务的目标语义对齐预训练。载于《欧洲计算机视觉会议》，第121–137页。施普林格，2020年。[93] Karan Desai 和 Justin Johnson。VirTex：从文本注释中学习视觉表征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第11162–11173页，2021年。[94] Fei Yu、Jiji Tang、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。ERNIE-ViL：通过场景图实现知识增强的视觉语言表征。载于《AAAI人工智能会议论文集》，第35卷，第3208–3216页，2021年。[95] Hao Tan 和 Mohit Bansal。Vokenization：通过上下文化视觉锚定监督改进语言理解。载于《2020年自然语言处理实证方法会议论文集》（EMNLP），第2066–2080页，2020年。[96] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：通过引导式语言-图像预训练实现统一的视觉-语言理解与生成。载于Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），《第39届国际机器学习会议论文集》，第162卷，《机器学习研究论文集》，第12888–12900页。PMLR，2022年7月17–23日。[97] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型进行引导式语言-图像预训练。arXiv预印本 arXiv:2301.12597，2023年。[98] Wei-Ning Hsu、Yao-Hung Hubert Tsai、Benjamin Bolte、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：糟糕的教师能为自动语音识别预训练带来多大帮助？载于《ICASSP 2021—2021年IEEE国际声学、语音与信号处理会议论文集》（ICASSP），第6533–6537页。IEEE，2021年。[99] Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。词语和短语的分布式表征及其组合性。《神经信息处理系统进展》，第26卷，2013年。[100] Matt Gardner、Joel Grus、Mark Neumann、Oyvind Tafjord、Pradeep Dasigi、Nelson F Liu、Matthew E Peters、Michael Schmitz 和 Luke Zettlemoyer。AllenNLP：深度语义自然语言处理平台。载于《NLP开源软件研讨会论文集》（NLP-OSS），第1–6页，2018年。[101] Emily M Bender 和 Alexander Koller。迈向自然语言理解：数据时代的意义、形式与理解。载于《计算语言学协会第58届年会论文集》，第5185–5198页，2020年。[102] Yonatan Bisk、Ari Holtzman、Jesse Thomason、Jacob Andreas、Yoshua Bengio、Joyce Chai、Mirella Lapata、Angeliki Lazaridou、Jonathan May、Aleksandr Nisnevich、Nicoloas Pinto 和 Joseph P. Turian。经验奠定语言基础。载于《2020年自然语言处理实证方法会议论文集》（EMNLP），第8718–8735页，2020年。[103] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A. Shamma、Michael S. Bernstein 和 Li Fei-Fei。Visual Genome：利用众包密集图像标注连接语言与视觉。《国际

计算机视觉期刊》，123：32–73，2017年。[105] Bart Thomee、David A Shamma、Gerald Friedland、Benjamin Elizalde、Karl Ni、Douglas Poland、Damian Borth 和 Li-Jia Li。YFCC100M：多媒体研究的新数据。《ACM通讯》，59(2)：64–73，2016年。[106] Harsh Agrawal、Karan Desai、Yufei Wang、Xinlei Chen、Rishabh Jain、Mark Johnson、Dhruv Batra、Devi Parikh、Stefan Lee 和 Peter Anderson。NoCaps：大规模新颖目标描述。载于《IEEE/CVF国际计算机视觉会议论文集》，第8948–8957页，2019年。[107] Qing Li、Boqing Gong、Yin Cui、Dan Kondratyuk、Xianzhi Du、Ming-Hsuan Yang 和 Matthew Brown。迈向统一基础模型：在未配对图像和文本上联合预训练Transformer。arXiv预印本 arXiv:2112.07074，2021年。[108] Ronghang Hu 和 Amanpreet Singh。UniT：使用统一Transformer进行多模态多任务学习。载于《IEEE/CVF国际计算机视觉会议论文集》，第1439–1449页，2021年。[109] Hassan Akbari、Liangzhe Yuan、Rui Qian、Wei-Hong Chuang、Shih-Fu Chang、Yin Cui 和 Boqing Gong。VATT：从原始视频、音频和文本中进行多模态自监督学习的Transformer。《神经信息处理系统进展》，第34卷，2021年。[110] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou 和 Hongxia Yang。OFA：通过简单的序列到序列学习框架统一架构、任务和模态。载于《国际机器学习会议》，第23318–23340页。PMLR，2022年。[111] Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于《计算语言学协会第58届年会论文集》，第7871–7880页，2020年。[112] Wenliang Dai、Junnan Li、Dongxu Li、Anthony Meng Huat Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale Fung 和 Steven Hoi。InstructBLIP：通过指令调优迈向通用视觉语言模型。arXiv预印本 arXiv:2305.06500，2023年。[113] Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma、Albert Webson、Shixiang Shane Gu、Zhuyun Dai、Mirac Suzgun、Xinyun Chen、Aakanksha Chowdhery、Dasha Valter、Sharan Narang、Gaurav Mishra、Adams Wei Yu、Vincent Zhao、Yanping Huang、Andrew M. Dai、Hongkun Yu、Slav Petrov、Ed Huai-hsin Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny Zhou、Quoc V. Le 和 Jason Wei。扩展指令微调语言模型。arXiv预印本 arXiv:2210.11416，2022年。[114] Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica 和 Eric P. Xing。Vicuna：一款开源聊天机器人，其表现令人惊叹，在 GPT-4 测试中达到 90%* 的 ChatGPT 质量，2023 年 3 月。[115] Jinbae Im、Moonki Kim、Hoyeop Lee、Hyunsouk Cho 和 Sehee Chung。自监督多模态观点摘要。载于《计算语言学协会第59届年会暨自然语言处理国际联合会议第11届会议论文集》（第一卷：长篇论文），2021年8月。[116] Yang Xu、Yiheng Xu、Tengchao Lv、Lei Cui、Furu Wei、Guoxin Wang、Yijuan Lu、Dinei Florencio、Cha Zhang、Wanxiang Che、Min Zhang 和 Lidong Zhou。LayoutLMv2：面向视觉丰富文档理解的多模态预训练。载于《计算语言学协会第59届年会暨自然语言处理国际联合会议第11届会议论文集》（第一卷：长篇论文），第2579–2591页，2021年。[117] Yulin Li、Yuxi Qian、Yuechen Yu、Xiameng Qin、Chengquan Zhang、Yan Liu、Kun Yao、Junyu Han、Jingtuo Liu 和 Errui Ding。StrucTexT：使用多模态Transformer进行结构化文本理解。载于《第29届ACM多媒体国际会议论文集》，第1912–1920页，2021年。[118] Zheng Huang、Kai Chen、Jianhua He、Xiang Bai、Dimosthenis Karatzas、Shijian Lu 和 CV Jawahar。ICDAR2019扫描收据OCR与信息抽取竞赛。载于《2019年国际文档分析与识别会议》（ICDAR），第1516–1520页。IEEE，2019年。[119] Guillaume Jaume、Hazim Kemal Ekenel 和 Jean-Philippe Thiran。FUNSD：用于噪声扫描文档表单理解的数据集。载于《2019年国际文档分析与识别研讨会》（ICDARW），第2卷，第1–6页。IEEE，2019年。[120] Zhangxuan Gu、Changhua Meng、Ke Wang、Jun Lan、Weiqiang Wang、Ming Gu 和 Liqing Zhang。XYLayoutLM：面向视觉丰富文档理解的布局感知多模态网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4583–4592页，2022年。[121] Xiangxiang Chu、Bo Zhang、Zhi Tian、Xiaolin Wei 和 Huaxia Xia。视觉Transformer真的需要显式位置编码吗？arXiv预印本 arXiv:2102.10882，3(8)，2021年。[122] Nayu Liu、Xian Sun、Hongfeng Yu、Wenkai Zhang 和 Guangluan Xu。采用遗忘门的多阶段融合，用于开放领域视频的多模态摘要。载于《2020年自然语言处理实证方法会议论文集》（EMNLP），第1834–1845页，2020年。[123] Shruti Palaskar、Jindřich Libovický、Spandana Gella 和 Florian Metze。面向How2视频的多模态抽象式摘要。载于《计算语言学协会第57届年会论文集》，第6587–6596页，2019年。[124] Tiezheng Yu、Wenliang Dai、Zihan Liu 和 Pascale Ngan Fung。用于多模态抽象式摘要的视觉引导生成式预训练语言模型。载于《2021年自然语言处理实证方法会议论文集》，2021年。[125] Ramon Sanabria、Ozan Caglayan、Shruti Palaskar、Desmond Elliott、Loïc Barrault、Lucia Specia 和 Florian Metze。How2：用于多模态语言理解的大规模数据集。载于《神经信息处理会议》，2018年。[126] Shaoshi Ling 和 Yuzong Liu。DecoAR 2.0：通过向量量化获得深度上下文化声学表示。arXiv预印本 arXiv:2012.06659，2020年。[127] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman。LRS3-TED：用于视觉语音识别的大规模数据集。arXiv预印本 arXiv:1809.00496，2018年。

[128] Takaki Makino、Hank Liao、Yannis Assael、Brendan Shillingford、Basilio Garcia、Otavio Braga 和 Olivier Siohan。用于视听语音识别的循环神经网络转导器。载于《2019年IEEE自动语音识别与理解研讨会（ASRU）论文集》，第905–912页。IEEE，2019年。[129] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。学习个体说话风格，以实现准确的唇语到语音合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第13796–13805页，2020年。[130] Robert Rehr 和 Timo Gerkmann。论超高斯语音先验对基于机器学习的语音增强的重要性。《IEEE/ACM音频、语音与语言处理汇刊》，26(2)：357–366，2017年。[131] Timothy F. Cootes、Gareth J. Edwards 和 Christopher J. Taylor。主动外观模型。《IEEE模式分析与机器智能汇刊》，23(6)：681–685，2001年。[132] Lingyu Zhu 和 Esa Rahtu。利用类别信息进行单帧视觉声源分离。载于《2021年第九届欧洲视觉信息处理研讨会（EUVIP）论文集》，第1–6页。IEEE，2021年。[133] Hang Zhao、Chuang Gan、Andrew Rouditchenko、Carl Vondrick、Josh McDermott 和 Antonio Torralba。像素之声。载于《欧洲计算机视觉会议（ECCV）论文集》，第570–586页，2018年。[134] Erin Hea-Jin Kim、Yoo Kyung Jeong、Yuyoung Kim、Keun Young Kang 和 Min Song。Twitter和新闻中的埃博拉病毒：基于主题的内容与情感分析。《信息科学杂志》，42(6)：763–781，2016年。[135] Jose Camacho-Collados 和 Mohammad Taher Pilehvar。论文本预处理在神经网络架构中的作用：文本分类与情感分析评估研究。载于《2018年EMNLP BlackboxNLP研讨会：自然语言处理神经网络的分析与解释论文集》，第40–46页，2018年。[136] Benjamin Wood、Owain Williams、Vijaya Nagarajan 和 Gary Sacks。加工食品制造商用以扩大和巩固其权力的市场策略：系统综述与文献分析。《全球化与健康》，17(1)：1–23，2021年。[137] Minping Chen 和 Xia Li。Swafn：用于多模态情感分析的情感词感知融合网络。载于《第28届国际计算语言学会议论文集》，第1067–1077页，2020年。[138] Linmei Hu、Bin Zhang、Lei Hou 和 Juanzi Li。新闻流中的自适应在线事件检测。《知识型系统》，138：105–112，2017年。[139] Nilani Algiriyage、Raj Prasanna、Kristin Stock、Emma EH Doyle 和 David Johnston。用于灾害应对的多源多模态数据与深度学习：系统综述。《SN计算机科学》，3(1)：1–29，2022年。[140] Firoj Alam、Ferda Ofli 和 Muhammad Imran。CrisisMMD：来自自然灾害的多模态Twitter数据集。载于《第十二届AAAI国际网络与社交媒体会议论文集》，2018年。[141] Qi Chen、Wei Wang、Kaizhu Huang、Suparna De 和 Frans Coenen。用于智慧城市交通事件检测的多模态生成对抗网络。《应用中的专家系统》，177：114939，2021年。[142] Alberto Barrón-Cedeno、Israa Jaradat、Giovanni Da San Martino 和 Preslav Nakov。Proppy：根据新闻的宣传性内容对新闻进行分类。《信息处理与管理》，56(5)：1849–1864，2019年。[143] Giovanni Da San Martino、Seunghak Yu、Alberto Barrón-Cedeno、Rostislav Petrov 和 Preslav Nakov。新闻文章中宣传内容的细粒度分析。载于《2019年自然语言处理实证方法会议与第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》，第5636–5646页，2019年。[144] Zhiwei Jin、Juan Cao、Han Guo、Yongdong Zhang 和 Jiebo Luo。利用循环神经网络进行多模态融合，以检测微博中的谣言。载于《第25届ACM多媒体国际会议论文集》，第795–816页，2017年。[145] Xinyi Zhou、Jindi Wu 和 Reza Zafarani。：相似性感知的多模态假新闻检测。载于《知识发现与数据挖掘进展：第24届亚太地区会议（PAKDD 2020），新加坡，2020年5月11–14日，论文集，第二部分》，第354–367页。施普林格，2020年。[146] Yaqing Wang、Fenglong Ma、Zhiwei Jin、Ye Yuan、Guangxu Xun、Kishlay Jha、Lu Su 和 Jing Gao。Eann：用于多模态假新闻检测的事件对抗神经网络。载于《第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第849–857页，2018年。[147] Dhruv Khattar、Jaipal Singh Goud、Manish Gupta 和 Vasudeva Varma。Mvae：用于假新闻检测的多模态变分自编码器。载于《万维网大会论文集》，第2915–2921页，2019年。[148] Jingzi Wang、Hongyan Mao 和 Hongwei Li。Fmfn：用于假新闻检测的细粒度多模态融合网络。《应用科学》，12(3)：1093，2022年。[149] Rowan Zellers、Yonatan Bisk、Ali Farhadi 和 Yejin Choi。从识别到认知：视觉常识推理。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6720–6731页，2019年。[150] Dandan Song、Siyi Ma、Zhanchen Sun、Sicheng Yang 和 Lejian Liao。Kvl-bert：用于视觉常识推理的知识增强型视觉与语言BERT。《知识型系统》，230：107408，2021年。[151] Hao Tan 和 Mohit Bansal。Lxmert：从Transformer中学习跨模态编码器表征。载于《2019年自然语言处理实证方法会议与第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》，第5100–5111页，2019年。[152] Zhicheng Huang、Zhaoyang Zeng、Bei Liu、Dongmei Fu 和 Jianlong Fu。Pixel-bert：通过深度多模态Transformer对齐图像像素与文本。arXiv预印本 arXiv:2004.00849，2020年。[153] Fengda Zhu、Yi Zhu、Xiaojun Chang 和 Xiaodan Liang。通过自监督辅助推理任务进行视觉语言导航。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10012–10022页，2020年。

[154] Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason Corso 和 Jianfeng Gao。用于图像描述和视觉问答的统一视觉语言预训练。载于《AAAI人工智能会议论文集》，第34卷，第13041–13049页，2020年。[155] Oriol Vinyals、Alexander Toshev、Samy Bengio 和 Dumitru Erhan。展示与讲述：神经网络图像描述生成器。载于《IEEE计算机视觉与模式识别会议论文集》，第3156–3164页，2015年。[156] Long Chen、Hanwang Zhang、Jun Xiao、Liqiang Nie、Jian Shao、Wei Liu 和 Tat-Seng Chua。Sca-cnn：用于图像描述的卷积网络中的空间与通道注意力。载于《IEEE计算机视觉与模式识别会议论文集》，第5659–5667页，2017年。[157] Steven J Rennie、Etienne Marcheret、Youssef Mroueh、Jerret Ross 和 Vaibhava Goel。用于图像描述的自批评序列训练。载于《IEEE计算机视觉与模式识别会议论文集》，第7008–7024页，2017年。[158] Christel Chappuis、Sylvain Lobry、Benjamin Alexander Kellenberger、Bertrand Le Saux 和 Devis Tuia。如何为遥感视觉问答找到合适的图文嵌入？载于《欧洲机器学习会议（ECML）研讨会》，编号CONF，2021年。[159] Tanzila Rahman、Shih-Han Chou、Leonid Sigal 和 Giuseppe Carenini。改进视觉问答中的注意力机制。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1653–1662页，2021年。[160] Sanjay Subramanian、Sameer Singh 和 Matt Gardner。分析视觉问答中的组合性。《神经信息处理系统进展》，第7卷，2019年。[161] Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。Ok-vqa：需要外部知识的视觉问答基准测试。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第3195–3204页，2019年。[162] Paul Pu Liang、Yiwei Lyu、Xiang Fan、Zetian Wu、Yun Cheng、Jason Wu、Leslie Yufan Chen、Peter Wu、Michelle A Lee、Yuke Zhu、Ruslan Salakhutdinov 和 Louis-Philippe Morency。MultiBench：多模态表征学习的多尺度基准测试。载于《第三十五届神经信息处理系统会议数据集与基准测试专题（第一轮）》。[163] Xingjian Shi、Jonas Mueller、Nick Erickson、Mu Li 和 Alex Smola。针对含文本字段的表格数据对多模态自动化机器学习进行基准测试。载于《第三十五届神经信息处理系统会议数据集与基准测试专题（第二轮）》。[164] Dong Huk Park、Lisa Anne Hendricks、Zeynep Akata、Bernt Schiele、Trevor Darrell 和 Marcus Rohrbach。注意力解释：为决策提供依据并指出证据。arXiv预印本 arXiv:1612.04757，2016年。[165] Aishwarya Agrawal、Dhruv Batra、Devi Parikh 和 Aniruddha Kembhavi。不要只凭假设；观察并作答：克服视觉问答中的先验偏见。载于《IEEE计算机视觉与模式识别会议论文集》，第4971–4980页，2018年。[166] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。生成对抗式文本到图像合成。载于《国际机器学习会议论文集》，第1060–1069页。PMLR，2016年。[167] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。加州理工学院-加州大学圣迭戈分校鸟类200-2011数据集。加州理工学院，2011年。[168] Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成。载于《IEEE计算机视觉与模式识别会议论文集》，第1316–1324页，2018年。[169] Leyuan Qu、Cornelius Weber 和 Stefan Wermter。LipSound：用于唇读的神经网络梅尔频谱图重建。载于《INTERSPEECH》，第2768–2772页，2019年。[170] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman。深度视听语音识别。《IEEE模式分析与机器智能汇刊》，44(12)：8717–8727，2018年。[171] Naomi Harte 和 Eoin Gillen。TCD-TIMIT：连续语音的视听语料库。《IEEE多媒体汇刊》，17(5)：603–615，2015年。[172] Wei Ping、Kainan Peng、Andrew Gibiansky、Sercan O Arik、Ajay Kannan、Sharan Narang、Jonathan Raiman 和 John Miller。Deep Voice 3：利用卷积序列学习扩展文本转语音。载于《国际学习表征会议论文集》，2018年。[173] Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan、Rif A. Saurous、Yannis Agiomyrgiannakis 和 Yonghui Wu。通过以梅尔频谱图预测结果为条件的WaveNet实现自然文本转语音合成。载于《2018年IEEE国际声学、语音与信号处理会议（ICASSP）》，第4779–4783页。IEEE，2018年。[174] Ariel Ephrat 和 Shmuel Peleg。Vid2Speech：从无声视频重建语音。载于《2017年IEEE国际声学、语音与信号处理会议（ICASSP）》，第5095–5099页。IEEE，2017年。[175] Hassan Akbari、Himani Arora、Liangliang Cao 和 Nima Mesgarani。Lip2AudSpec：从无声唇动视频重建语音。载于《2018年IEEE国际声学、语音与信号处理会议（ICASSP）》，第2516–2520页。IEEE，2018年。[176] Konstantinos Vougioukas、Pingchuan Ma、Stavros Petridis 和 Maja Pantic。利用生成对抗网络进行视频驱动的语音重建。《INTERSPEECH 2019论文集》，第4125–4129页，2019年。[177] Weijie Su、Xizhou Zhu、Yue Cao、Bin Li、Lewei Lu、Furu Wei 和 Jifeng Dai。VL-BERT：通用视觉语言表征的预训练。载于《国际学习表征会议》，2020年。[178] Wei-Hung Weng 和 Peter Szolovits。电子健康记录的表征学习。arXiv预印本 arXiv:1909.09248，2019年。[179] Xianli Zhang、Buyue Qian、Yang Li、Yang Liu、Xi Chen、Chong Guan 和 Chen Li。从多模态电子健康记录中学习稳健的患者表征：一种监督式深度学习方法。载于《2021年SIAM国际数据挖掘会议（SDM）论文集》，第585–593页。SIAM，2021年。

[180] Tao Chen、Richang Hong、Yanrong Guo、Shijie Hao 和 Bin Hu。MS^2-gnn：探索用于抑郁症检测的基于图神经网络的多模态融合网络。《IEEE控制论汇刊》，2022年。[181] Jianliang Gao、Tengfei Lyu、Fan Xiong、Jianxin Wang、Weimao Ke 和 Zhao Li。利用多模态图神经网络预测癌症患者的生存情况。《IEEE/ACM计算生物学与生物信息学汇刊》，19(2):699–709，2021年。[182] O Caglayan、P Madhyastha、L Specia 和 L Barrault。探究多模态机器翻译对视觉上下文的需求。载于：2019年计算语言学协会北美分会年会论文集：人类语言技术，第1卷（长短篇论文），第4159–4170页。计算语言学协会，2019年。[183] Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。通过联合学习对齐与翻译实现神经机器翻译。载于 Yoshua Bengio 和 Yann LeCun（编），国际学习表征会议论文集，2015年。[184] Jinsong Su、Jinchang Chen、Hui Jiang、Chulun Zhou、Huan Lin、Yubin Ge、Qingqiang Wu 和 Yongxuan Lai。具有深层语义交互的多模态神经机器翻译。《信息科学》，554:47–60，2021年。[185] Micah Hodosh、Peter Young 和 Julia Hockenmaier。将图像描述构建为排序任务：数据、模型和评估指标。《人工智能研究期刊》，47:853–899，2013年。[186] Amir Zadeh 和 Paul Pu。真实场景中的多模态语言分析：CMU-MOSEI 数据集与可解释动态融合图。载于：计算语言学协会第56届年会论文集（长篇论文），2018年。[187] Alistair EW Johnson、Tom J Pollard、Lu Shen、Li-wei H Lehman、Mengling Feng、Mohammad Ghassemi、Benjamin Moody、Peter Szolovits、Leo Anthony Celi 和 Roger G Mark。MIMIC-III：一个免费开放的重症监护数据库。《科学数据》，3(1):1–9，2016年。[188] Xintong Han。Fashion 200K 基准。https://github.com/xthan/fashion-200k，2017年。 [在线；访问日期 2017]。[189] N. Silberman 和 R. Fergus。使用结构光传感器进行室内场景分割。载于：国际计算机视觉会议——3D 表示与识别研讨会论文集，2011年。[190] Nathan Silberman、Derek Hoiem、Pushmeet Kohli 和 Rob Fergus。利用 RGBD 图像进行室内场景分割和支撑关系推断。载于：第12届欧洲计算机视觉会议论文集（第V部分），第746–760页，2012年。[191] Ali Furkan Biten、Lluis Gomez、Marçal Rusinol 和 Dimosthenis Karatzas。好消息，大家！用于新闻图像的上下文驱动实体感知图像描述。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第12466–12475页，2019年。[192] Jun Xu、Tao Mei、Ting Yao 和 Yong Rui。MSR-VTT：一个用于衔接视频与语言的大型视频描述数据集。载于：IEEE计算机视觉与模式识别会议论文集，第5288–5296页，2016年。[193] Dejing Xu、Zhou Zhao、Jun Xiao、Fei Wu、Hanwang Zhang、Xiangnan He 和 Yueting Zhuang。通过对外观和运动进行逐步精炼的注意力实现视频问答。载于：第25届ACM多媒体国际会议论文集，第1645–1653页，2017年。[194] Yunseok Jang、Yale Song、Youngjae Yu、Youngjin Kim 和 Gunhee Kim。TGIF-QA：探索视觉问答中的时空推理。载于：IEEE计算机视觉与模式识别会议论文集，第2758–2766页，2017年。[195] Licheng Yu、Xinlei Chen、Georgia Gkioxari、Mohit Bansal、Tamara L Berg 和 Dhruv Batra。多目标具身问答。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第6309–6318页，2019年。[196] Catalina Cangea、Eugene Belilovsky、Pietro Liò 和 Aaron C. Courville。VideoNavQA：弥合视觉问答与具身问答之间的差距。载于《视觉基础交互与语言》（ViGIL），NeurIPS 2019 研讨会，2019年。[197] Kushal Kafle 和 Christopher Kanan。视觉问答算法分析。载于：IEEE国际计算机视觉会议论文集，第1965–1973页，2017年。[198] Holger Caesar、Varun Bankiti、Alex H Lang、Sourabh Vora、Venice Erin Liong、Qiang Xu、Anush Krishnan、Yu Pan、Giancarlo Baldan 和 Oscar Beijbom。nuScenes：一个用于自动驾驶的多模态数据集。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第11621–11631页，2020年。[199] Maria-Elena Nilsback 和 Andrew Zisserman。面向大量类别的自动花卉分类。载于：2008年第六届印度计算机视觉、图形学与图像处理会议论文集，第722–729页。IEEE，2008年。[200] Amir Zadeh、Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency。MOSI：在线评论视频中情感强度和主观性分析的多模态语料库。arXiv 预印本 arXiv:1606.06259，2016年。[201] Yuanhang Zhang、Shuang Yang、Jingyun Xiao、Shiguang Shan 和 Xilin Chen。我们能否读懂唇语之外的语音？重新思考深度视觉语音识别中的感兴趣区域选择。载于：2020年第15届IEEE自动人脸与手势识别国际会议（FG 2020），第356–363页。IEEE，2020年。[202] Maurice Fallon、Hordur Johannsson、Michael Kaess 和 John J Leonard。MIT Stata 中心数据集。《国际机器人研究期刊》，32(14):1695–1699，2013年。[203] Alexei Baevski、Wei-Ning Hsu、Qiantong Xu、Arun Babu、Jiatao Gu 和 Michael Auli。Data2vec：一种用于语音、视觉和语言自监督学习的通用框架。载于：国际机器学习会议论文集，第1298–1312页。PMLR，2022年。[204] Amanpreet Singh、Ronghang Hu、Vedanuj Goswami、Guillaume Couairon、Wojciech Galuba、Marcus Rohrbach 和 Douwe Kiela。FLAVA：一种联合缩放的语言与视觉对齐模型。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第15638–15650页，2022年。[205] Yasha Ektefaie、George Dasoulas、Ayush Noori、Maha Farhat 和 Marinka Zitnik。基于图的多模态学习。《自然·机器智能》，第1–11页，2023年。

[206] Minheng Ni、Haoyang Huang、Lin Su、Edward Cui、Taroon Bharti、Lijuan Wang、Dongdong Zhang 和 Nan Duan。M3P：通过多任务多语言多模态预训练学习通用表征。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第3977–3986页，2021年。[207] Aashi Jain、Mandy Guo、Krishna Srinivasan、Ting Chen、Sneha Kudugunta、Chao Jia、Yinfei Yang 和 Jason Baldridge。MURAL：跨语言多模态多任务检索。*arXiv* 预印本 *arXiv:2109.05125*，2021年。[208] Yan Zeng、Wangchunshu Zhou、Ao Luo 和 Xinsong Zhang。跨视图语言建模：迈向统一的跨语言跨模态预训练。*arXiv* 预印本 *arXiv:2206.00621*，2022年。[209] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。先对齐再融合：通过动量蒸馏进行视觉与语言表征学习。神经信息处理系统进展，34:9694–9705，2021年。[210] Xi Chen、Xiao Wang、Soravit Changpinyo、AJ Piergiovanni、Piotr Padlewski、Daniel Salz、Sebastian Goodman、Adam Grycner、Basil Mustafa、Lucas Beyer、Alexander Kolesnikov、Joan Puigcerver、Nan Ding、Keran Rong、Hassan Akbari、Gaurav Mishra、Linting Xue、Ashish V. Thapliyal、James Bradbury、Weicheng Kuo、Mojtaba Seyedhosseini、Chao Jia、Burcu Karagol Ayan、Carlos Riquelme、Andreas Steiner、Anelia Angelova、Xiaohua Zhai、Neil Houlsby 和 Radu Soricut。Pali：一种联合缩放的多语言语言-图像模型。*arXiv* 预印本 *arXiv:2209.06794*，2022年。[211] Maria Tsimpoukelli、Jacob L Menick、Serkan Cabi、SM Eslami、Oriol Vinyals 和 Felix Hill。利用冻结语言模型进行多模态少样本学习。神经信息处理系统进展，34:200–212，2021年。[212] Mingchen Zhuge、Dehong Gao、Deng-Ping Fan、Linbo Jin、Ben Chen、Haoming Zhou、Minghui Qiu 和 Ling Shao。Kaleido-bert：时尚领域的视觉语言预训练。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第12647–12657页，2021年。[213] Fei-Long Chen、Du-Zhen Zhang、Ming-Lun Han、Xiu-Yi Chen、Jing Shi、Shuang Xu 和 Bo Xu。VLP：视觉语言预训练综述。机器智能研究，20(1):38–56，2023年。[214] Zhiyuan Fang、Jianfeng Wang、Xiaowei Hu、Lijuan Wang、Yezhou Yang 和 Zicheng Liu。通过知识蒸馏压缩视觉-语言模型。载于 *IEEE/CVF*国际计算机视觉会议，第1428–1438页，2021年。[215] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉-语言表征学习。载于 国际机器学习会议，第4904–4916页。PMLR，2021年。[216] Nanyi Fei、Zhiwu Lu、Yizhao Gao、Guoxing Yang、Yuqi Huo、Jingyuan Wen、Haoyu Lu、Ruihua Song、Xin Gao、Tao Xiang、Haoran Sun 和 Jiling Wen。借助多模态基础模型迈向通用人工智能。《自然-通讯》，13(1):3094，2022年。[217] Daniel Michelsanti、Zheng-Hua Tan、Shi-Xiong Zhang、Yong Xu、Meng Yu、Dong Yu 和 Jesper Jensen。基于深度学习的视听语音增强与分离综述。*IEEE/ACM*音频、语音与语言处理汇刊，29:1368–1396，2021年。