跨媒体智能关联分析与语义理解理论与技术研究进展

于俊清[1)], 王鑫[2)], 况琨[3)], 刘偲[4)], 张新峰[5)], 宋子恺[1)]

[1)] (华中科技大学计算机科学与技术学院 武汉 430074)
[2)] (清华大学计算机科学与技术系 北京 100084)
[3)] (浙江大学计算机科学与技术学院 杭州 310027)
[4)] (北京航空航天大学计算机学院 北京 100191)
[5)] (中国科学院大学计算机科学与技术学院 北京 100049)
(yjqing@hust.edu.cn)

摘 要: 深入分析了跨媒体智能关联分析与语义理解理论技术的最新研究进展, 包括多模态数据的统一表达、知识引导的数据融合、跨媒体关联分析、基于知识图谱的跨媒体表征技术以及面向多模态的智能应用. 其中, 多模态数据的统一表达是对跨媒体信息进行分析推理的先决条件, 利用多模态信息间的语义一致性剔除冗余信息, 通过跨模态相互转化来实现跨媒体信息统一表达, 学习更全面的特征表示; 跨媒体关联分析立足于图像语言、视频语言以及音视频语言的跨模态关联分析与理解技术, 旨在弥合视觉、听觉以及语言之间的语义鸿沟, 充分建立不同模态间的语义关联; 基于知识图谱的跨媒体表征技术通过引入跨媒体的知识图谱, 从跨媒体知识图谱构建、跨媒体知识图谱嵌入以及跨媒体知识推理3个方面展开研究, 增强跨媒体数据表征的可靠性, 并提升后续推理任务的分析效率和准确性; 随着跨模态分析技术的快速发展, 面向多模态的智能应用得到了更多的技术支撑, 依据智能应用所需要的领域知识, 选取了多模态视觉问答, 多模式视频摘要、多模式视觉模式挖掘、多模式推荐、跨模态智能推理和跨模态医学图像预测等跨模态应用实例, 梳理了其在多模态数据融合以及跨媒体分析推理方面的研究进展.

关键词: 跨媒体信息统一表达; 知识引导的数据融合; 跨媒体关联分析; 跨媒体知识图谱; 跨媒体分析与推理; 多模态智能应用
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19296

Advances in Theory and Technology of Cross-Media Intelligent Association Analysis

Yu Junqing[1)], Wang Xin[2)], Kuang Kun[3)], Liu Si[4)], Zhang Xinfeng[5)], and Song Zikai[1)]

[1)] (*School of Computer of Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074)
[2)] (*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
[3)] (*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310027)
[4)] (*School of Computer Science and Engineering, Beihang University, Beijing* 100191)
[5)] (*School of Computer of Science and Technology, University of Chinese Academy of Sciences, Beijing* 100049)

Abstract: This paper provides an analysis of the latest research trends of theories and technologies in cross-media intelligent correlation analysis and semantic understanding. The main content of this report includes a unified representation of cross-media information, knowledge-guided data fusion, cross-media correlation analysis, cross-media knowledge graph, and intelligent applications for multi-modal. Unified representations are preconditions for analyzing and inference about multi-modal information. The semantic con-

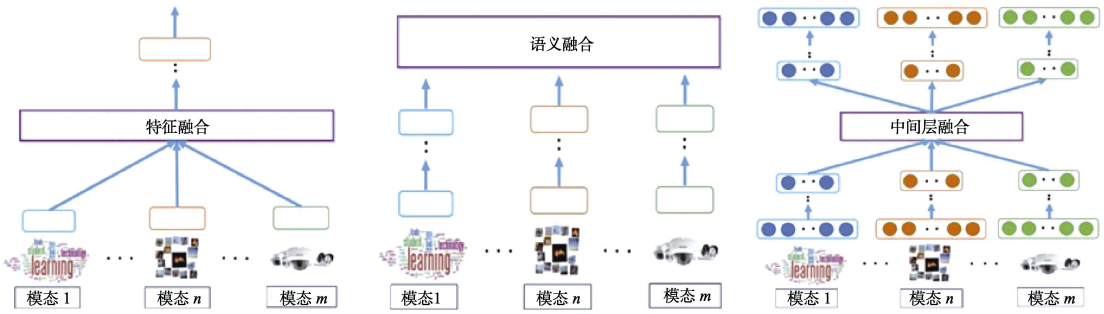
收稿日期: 2021-08-10; 修回日期: 2022-05-28. **基金项目**: 国家自然科学基金(62050110, 62006207, 61876177, U20A20184). **于俊清**(1975—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为智能媒体计算、网络安全; **王鑫**(1988—), 男, 博士, 助理研究员, CCF 会员, 主要研究方向为媒体大数据分析、机器学习、多媒体智能; **况琨**(1992—), 男, 副教授, 博士生导师, CCF 会员, 主要研究方向为因果推理、人工智能、因果指导的可信机器学习; **刘偲**(1985—), 女, 博士, 副教授, 博士生导师, CCF 会员, 主要研究方向为视觉和语言、目标跟踪; **张新峰**(1983—), 男, 博士, 助理教授, 博士生导师, CCF 会员, 主要研究方向为视频编码、特征压缩、视频图像质量评价、视频图像处理; **宋子恺**(1993—), 男, 博士研究生, 主要研究方向为目标跟踪.

sistency between multi-modal information is utilized to eliminate redundant information and achieve unified representation through cross-modal interconversion to learn more comprehensive feature representation. The cross-media association analysis focuses on image-language, video-language, and audio-video-language, aiming to bridge the semantic gap between visual, auditory, language, and fully establish the semantic association between different modalities. By introducing the construction of cross-media knowledge graph, cross-media knowledge graph construction, cross-media knowledge graph embedding, and cross-media knowledge inference, the cross-media representation based on knowledge graph enhances the reliability and improves the efficiency and accuracy of subsequent inference tasks. With the rapid development of cross-modal analysis, intelligent applications for multi-modal are supported by more technologies. According to the required domain knowledge, this paper selects cross-modal applications such as multi-modal visual question answering, multi-modal video summarization, multi-modal visual pattern mining, multi-modal recommendation, cross-modal intelligent inference, and cross-modal medical image prediction, their research progress is compared and reviewed in terms of multi-modal fusion and cross-media inference.

Key words: unified representation of cross-media information; knowledge-guided data fusion; cross-media correlation analysis; cross-media knowledge graph; cross-media analysis and inference; multi-modal intelligent applications

认知科学的前沿发现[1]，人类能够融合多个感官的反馈来感知周围的环境．人类获取到的信息已经从一种媒体形式逐渐转变为文本、图像、视频、音频等数据结合在一起的跨媒体数据．如何有效地处理这些跨媒体信息成为目前亟待解决的问题．本文深入分析了跨媒体智能关联分析与语义理解理论与技术最新的进展，从国际研究现状和国内研究现状介绍了多模态数据的统一表达、知识引导的数据融合、跨媒体关联分析、基于知识图谱的跨媒体表征技术，以及面向多模态的智能应用．

多模态数据的统一表达和知识引导的数据融合是对跨媒体信息进行分析推理的先决条件．要对跨媒体信息进行分析与推理，首先要利用多模态信息间的语义一致性，剔除模态间的冗余信息，通过跨模态相互转化来实现跨媒体信息统一表达．多模态数据融合通过融合来自不同模态的信息使得多模态分析的方法能够优于原来仅利用单个模态信息的方法．多模态数据融合的方法可以分为传统的数据融合方法和深度学习的数据融合方法，图 1 所示为多模态的数据融合方法．传统的数据融合方法有特征融合和语义融合 2 类，深度网络的数据融合方法为中间层融合．其中，如图 1a 所示，特征融合是对来自不同模型的特征进行拼接．如图 1b 所示，语义融合是在语义层面对多模态数据进行融合,能够保证语义融合后的可解释性，但是不能充分地利用多模态数据的全部信息．伴随着深度神经网络的成功，图 1c 所示在中间层融合不同模态的隐藏空间信息方法，以数据驱动的方式学习不同模态的相关表示，可以充分地利用多模态数据．



a. 特征融合　　b. 语义融合　　c. 中间层融合

图 1　3 种融合方法示意图[2]

跨媒体关联分析与理解技术旨在弥合模态间的语义鸿沟，充分建立不同模态间的语义关联. 跨媒体关联分析重点介绍了图像语言、视频语言以及音视频语言的跨模态关联分析与理解技术. 图像与自然语言的跨媒体关联分析与理解由指代表达理解、指代表达分割以及短语定位 3 个主要的研究方向组成. 视频与自然语言的跨媒体关联分析与理解旨在基于自然语言描述从视频中提取出感兴趣的信息，主要聚焦于由语言指导的对视频的事件定位、目标定位和目标分割 3 个方向. 视频与音频的跨媒体关联分析与理解主要聚焦在视听定位以及视听关联学习.

基于知识图谱的跨媒体表征技术，通过引入构建跨媒体知识图谱，增强跨媒体数据表征的可靠性，并提升后续推理任务的准确性. 为实现基于知识图谱的跨媒体表征学习，主要从跨媒体知识图谱构建、跨媒体知识图谱嵌入以及跨媒体知识推理 3 个方面开展研究. 在跨媒体知识图谱构建过程中，需要处理以特定数据结构展示和存储的结构化数据，借助知识图谱补全、实体识别与链接以及实体关系学习等技术来保证跨媒体知识图谱的完备性. 基于跨模态知识图谱的嵌入技术能够更好地建模跨模态数据中广泛存在的多样性，有效地提升跨模态数据的表征能力.

随着跨模态分析技术的快速发展，面向多模态的智能应用得到了更多技术支撑. 为了清晰地阐述现有知识引导的跨模态数据融合思想，以及跨媒体信息的统一表达，本文依据所需要的领域知识，选取了多模态视觉问答、多模式视频摘要、多模式视觉模式挖掘、多模式推荐、跨模态智能推理和跨模态医学图像预测等跨模态应用实例，讨论了其在知识引导的多模态数据融合和跨媒体分析与推理方面的研究方向.

1　多模态数据的统一表达

1.1　国际研究现状

跨媒体数据承载着不同种类的信息，需要利用多模态信息间的语义一致性，剔除模态间的冗余，来学习更全面的特征表示. 多模态数据的统一表达主要利用不同模态之间的语义相似性来学习和构建不同模态数据之间的共同投影子空间，并在子空间上实现统一表达.

多模态数据的统一表示主要包含联合表征和协同表征 2 大研究方向[3]. 其中，联合表征尝试构建一个公共空间，将数据映射到该空间，使用常见的距离度量直接计算异构数据对象之间的相似性，以减少来自不同模态的数据之间的异构差距. 第 1 个经典模型是典型相关性分析(canonical correlation analysis, CCA)[4]，通过最大化成对异构数据之间的相关性来学习公共空间，并通过线性函数执行投影. 在 CCA 的基础上，有许多后续的研究，如核典型相关性分析(kernel canonical correlation analysis, KCCA)[5]在 CCA 的基础上使用较低阶的核函数提高模型的描述能力；Andrew 等[6]使用深度学习技术扩展 CCA，将传统的 CCA 拓展为深度典型相关分析(deep canonical correlation analysis, DCCA). Srivastava 等[7]提出了深度玻尔兹曼机，对不同模态分别学习底层表示，利用高层语义对不同模态融合建立关联. Karpathy 等[8]提出深度视觉语义对齐(deep visual-semantic alignments, DVSA)算法，使用区域卷积神经网络(region convolutional neural network, R-CNN)检测和编码图像区域. Castrejón 等[9]提出了正则化跨模态深度学习网络，旨在模态差异很大的情况下学习不同模态的共同表示.

知识引导的数据融合方法主要有贝叶斯推理、师生网络和强化学习 3 类. 其中，贝叶斯理论[10]是统计学中非常流行的工具. 贝叶斯推理[11-13]的目的是通过将一些“先验”知识编码到模型，以模拟人类的推理能力. 因此，利用贝叶斯先验知识融合领域知识是知识引导的多模态融合的一个很好的选择. 师生网络[14]最初是在训练有素的网络(教师网络)的指导下，提出对复杂的深层模型(学生网络)进行压缩. 师生网络还被应用于图像集[15]，RGB 图像和深度图[16]以及视频集[17]之间的信息与知识传递. 强化学习[18-20]旨在采取适当的行动，在某些情况下使奖励最大化. 在过去的几十年里，强化学习一直是一个成熟的机器学习课题，特别是在机器人领域[21-22]. 利用领域知识来指导强化框架中的奖励和反馈是处理知识引导的多模态融合的一种较有前途的方法.

1.2　国内研究现状

国内研究多模态数据统一表达的主要方法是基于 CCA 来构建投影子空间. Liang 等[23]通过无监督聚类算法，将学习不同模态之间的群组相关性来构建投影子空间. Song 等[24]和 Hua 等[25]通过构造非参数映射和线性交叉模态投影，拓展了传统典型 CCA. Xia 等[26]引入了 DCCA 框架，将随机梯度方法中 Levinberg-Marquardt 和高斯-牛顿方法的

优势互补, 该框架下多模态数据统一表达陷入局部最优的问题. 协同表征学习是多模态数据统一表征的另一类方法. 在协同表征学习方面, Wu 等[27]构建了图像-文本和文本-图像双向匹配关系, 通过双向匹配之间的语义相似性来计算不同模态数据之间的双线性相似度; 在构建文本与图像之间的统一表征过程中, 通过学习不同文本与多层次图像特征交互, 多模态卷积神经网络(multimoda convolutional neural network, Multimodal-CNN)[28]模型实现了多模态统一表征; 面向多标签数据集, Huang 等[29]提出通过构建多区域多标签的神经网络来学习不同模态之间的语义特征; 基于注意力机制学习框架, Wang 等[30]提出全局注意力机制来学习不同文档之间单词和句子、词语之间的语义一致性, 实现跨模态数据的统一表征; 基于对抗生成网络框架, Wu 等[31]在离线训练中引入生成对抗方式, 实现了跨模态数据对齐和跨模态语义一致性.

1.3 国内外进展对比与总结

跨媒体信息统一表达领域国外起步较早, 如在文本-图像跨模态检索任务中引入 CCA 方法, 通过最大化不同模态特征到子空间投影向量的相关度, 学习文本和图像不同模态数据的共同子空间. 国内学者后续则尝试了为具有更多媒体类型的场景开发数据集和方法, 如北京大学多媒体信息处理研究室彭宇新教授课题组采集并发布的 XMedia 数据集是第一个包含 5 种媒体类型(文本、图像、视频、音频和 3D 模型)的数据集. 在该数据集上尝试(如在统一框架中)对数据集上 5 种媒体类型的图正则化联合建模相关性和语义信息等[32]方法. 总体而言, 在跨媒体信息统一表达方面国内外进展情况较为接近.

2 跨媒体关联分析

跨媒体关联分析从图像与自然语言、视频与自然语言以及音频与视频 3 个方面介绍了跨模态关联分析与理解技术.

2.1 国际研究现状

2.1.1 图像与自然语言的跨媒体关联分析

图像与自然语言的跨媒体关联分析与理解主要由指代表达理解、指代表达分割以及短语定位等方向组成.

指代表达的任务要求是根据给定语言表达, 在图像的区域提案集合中选出最相关的 1 个. 这些模型需要先使用目标检测器提取区域提案, 再计算指代表达和每个目标之间的相似性. 基于 YOLOv3[33]提出了单阶段指代表达理解方法[34] (fast and accurate one-stage approach, FAOA), 该方法使用 Darknet[35]和 BERT[36]分别提取语言特征和视觉图像金字塔特征用于预测目标物体的边界框坐标. Sun 等[37]提出了基于深度强化学习的迭代收缩算法, 通过顺序观察收缩过程, 模型的迭代推理过程可以更好地得到解释.

对于指代表达分割任务, Huang 等[38]提出了渐进式跨模态理解(cross-modal progressive comprehension, CMPC)模型. 此前的方法通常在视觉和语言模态之间进行隐式的特征交互与融合, 无法有效地感知指代表达中的关键词语, 因此, 2 种模态的特征对齐程度较低, 限制了目标物体的定位准确度. 而渐进式跨模态理解模型先在视觉和语言模态搜索可能的实体, 再使用指代表达中的关系词定位目标物体, 从而准确定位所指代的目标. 印度海得拉巴研究所[39]提出了综合多模态交互方法, 该方法主要由一个联合推理(joint reasoning module, JRM)模块和一个跨模态多层融合(cross-modal multi-level fusion, CMMLF)模块组成, 能够对多种关系同时实现跨模态联合推理, 从而消除解析单个关系时可能引入的歧义.

在短语定位方面, 一种基于弱监督对比学习方法[40]创新地使用语言模型指导的词替换机制, 在已有指代表达的基础上构建大量难负样本. 该方法证明了通过最大化图像和短语之间的跨模态交互信息, 学习词语-区域注意力来进行短语定位. Wang 等[41]针对弱监督短语定位任务提出了多模态对齐框架(multimodal alignment framework, MAF). 该框架使用细粒度视觉特征和视觉感知的语言特征对短语-物体关联程度进行建模, 然后使用对比学习方法提高模型性能. 总体来说, 上述方法从不同角度出发, 实现了短语定位任务的弱监督训练, 避免了大规模标注短语定位数据集的成本, 对于该领域的落地应用具有推动作用.

2.1.2 视频与自然语言的跨媒体关联分析

视频与自然语言的跨媒体关联分析与理解由语言指导的事件定位、目标定位以及目标分割组成. 输入一段视频和自然语言描述, 事件定位需要确定该语言描述的事件在该视频中的起止时间点, 目标定位或分割任务则需要确定该语言描述的目标在视频帧图像中的边界框或者像素级别的二值掩码.

对于语言指导的视频事件定位任务, 已有方

法主要是提取若干个预选视频子片段，然后基于各个视频子片段与自然语言的融合特征来选择某个子片段，并对其起止时间进行更精细地调整. Rodriguez-Opazo 等[42]提出了一种端到端且无需预选子片段的方法，该方法通过基于注意力机制的动态转化器将语言信息迁移到视觉信息域中，在训练中引入新的损失函数来指导模型自适应地注意到视频中与语言最相关的部分. Mun 等[43]提出了利用文本查询中的语义短语来进行更细粒度的交互的方法，除了采用整句自然语言描述的全局特征和视频的视觉特征的交互外，还引入了句子中的重要语义实体与视频对应片段的交互. Varol 等[44]将 Transformer 模型[45]引入该任务中，实现了手语视频与其文本含义的对齐.

对于由语言指导的视频目标定位任务，大多数方法是直接应用基于视频帧图像的目标定位方法. Sadhu 等[46]引入了多帧视频中目标之间的关系信息，通过具有相对位置编码的自注意力机制对多对象关系建模来消除歧义，从而保证自然语言描述定位目标的准确性和唯一性. Yang 等[47]针对视频时空结构复杂和缺乏细粒度标注的特点，提出了弱监督框架，通过时间定位模块对查询对象和视频帧之间的潜在关系进行建模，从而定位到关键帧以学习上下文感知的目标特征表示.

由语言指导的视频目标分割任务中，McIntosh 等[48]首次将动态路由胶囊网络[49]引入了该任务当中，首先通过胶囊网络对视频和文本输入进行编码来提供更有效的特征表示，然后通过视觉文本路由机制实现视频和文本胶囊的融合，多个胶囊按协议路由程序来学习不同实体之间的关系，实现文本描述对象的像素级定位. Seo 等[50]提出了联合图像指代分割和视频目标分割的框架，通过指代自然语言和先前帧中的分割掩码预测来估计当前帧中的对象掩码，以此迭代逐帧处理视频，直至所有帧中的掩码预测均收敛.

2.1.3　音频与视频的跨媒体关联分析

音频与视频的跨媒体关联分析与理解包括视听定位和视听关联学习. 视听定位的目标是根据输入音频，对视频中的音频信号进行定位，其难点在于需要分离不同物体发出的特定声音，并在视觉环境中定位每个声音信号. 在视听定位任务中，目标说话人的视觉信息可以辅助声音分离，如嘴唇的动作、音调和空间位置等. Gu 等[51]通过基于因子化注意力的融合方法获得每个模态的语义信息，利用说话人的信息辅助定位. Zhu 等[52]则通过引入外观注意模块来增加额外的信息，以分离不同的语义表征. 基于视频中运动信息的重要性，Zhao 等[53]提出了端到端的深稠密轨迹网络来学习视频中的运动信息，实现了音视频分离.

视听关联学习的重点是发现音频和视觉模态之间的语义关系. 针对视听关联学习任务，Surís 等[54]提出了新的联合嵌入模型，将 2 种模态映射到一个联合嵌入空间，直接计算两者之间的欧几里得距离. Nagrani 等[55]提出了跨模态的自监督方法来学习视频中嵌入的音视频信息，大大降低了网络的复杂度，并设计了新的课程学习时间表来进行样本选择，以进一步提高网络的性能.

2.2　国内研究现状

2.2.1　图像与自然语言的跨媒体关联分析

对于指代表达理解任务，Liao 等[56]首次在精度不下降的条件下实现了实时推理，极大地推动了指代表达理解任务的研究边界；Yang 等[57]学者从结构化推理的角度出发，提出了基于图的指代表达理解模型. 考虑到指代表达理解与指代表达分割具有高度相关性，Luo 等[58]学者提出了多任务协作网络(multi-task collaborative network, MCN). 具体来说，指代表达分割可以帮助指代表达理解更准确地对齐视觉和语言特征，而指代表达理解则可以帮助指代表达分割定位目标. 该方法首次实现了上述任务的联合学习，为指代表达理解与指代表达分割任务提供了新的解决思路. 在短语定位方面，Mu 等[59]提出了解耦的介入式图网络(graph learning framework for phrase grounding, DIGN)，将场景图上下文中的不同图案加入分布式表示. 该方法细化了跨模态上下文信息的理解粒度，打破了模型理解复杂信息的限制.

2.2.2　视频与自然语言的跨媒体关联分析

对于视频的事件定位任务，国内的研究人员也提出了许多新方法. Liu 等[60]引入了图网络，通过自模态交互图建立视频帧或者句子单词之间的关联，利用跨模态交互图建模句子和视频之间的相关实例，两者构成的联合图能够更有效地捕捉 2 种模态之间的交互. Zeng 等[61]针对训练过程中带注释的起始结束帧和其他帧的不平衡问题，提出了一种新的密集回归网络，预测每帧到语言所描述的视频片段的开始或结束帧的距离. Wang 等[62]和 He 等[63]提出了使用强化学习来解决该任务的方法.

对于视频目标定位任务，Zhang 等[64]引入图网络来增强模型的推理能力，该方法基于视频每个帧图像中的空间子图和跨帧的时间子图来建立时

空区域图，然后将文本线索合并到图中进行多步跨模态图推理. Wang 等[65]针对该任务缺乏细粒度标注的特点提出了一种弱监督的方法，关注了视频中不同对象区域之间的时空相关性，通过自注意机制捕获多对象特征之间的潜在时空相关性.

对于由语言指导的视频目标分割任务，动态卷积良好的感受野自适应性较好地匹配了分割任务的细粒度要求. Wang 等[66]提出了一种包含上下文信息的动态卷积网络，根据自然语言和上下文特征生成特定区域的卷积核. Wang 等[67]还针对之前工作通常只用语言来指导视觉特征而忽略了语言描述多样性这一问题，提出了非对称的交叉引导注意力网络. 为了准确地表示目标对象，Ning 等[68]针对此特性在视觉语言特征融合中引入了基于极坐标的相对位置编码，以提取自然语言中隐含的相对位置关系.

2.2.3 音频与视频的跨媒体关联分析

对于在视觉背景下对声音进行定位的视听定位任务中，国内研究人员也提出了许多方法. 受人类听觉系统选择性地接受信息的启发，Sun 等[69]提出了基于超材料的单麦克风监听系统(metamaterial-based single-sensor listening system, MSLS)来定位和分离 3D 空间中的固定声音信号. Lu 等[70]提出的模型由多个视频流和一个音频流组成，将来自不同流的特征连接成一个联合的音频-视频特征表示；同时，提出了一个视听匹配网络来建立语音和人类嘴唇运动之间的对应关系，从而使得模型获得更好的效果. 在利用音频与视觉信息的跨媒体分析理解的视听关联学习任务中，国内最近的研究引入了注意机制来突出音频或视频表征中包含的一些重要信息. Zhou 等[71]通过多模态注意机制来融合不同模态的特征. Zhang 等[72]提出了一种分解双线性池化，嵌入注意力机制来学习各个模态的特征.

2.3 国内外进展对比与总结

在指代表达理解领域中，国内外的进展基本一致. 在单阶段指代表达理解模型中，国内率先实现了实时高精度检测，极大地推动了研究前沿. 在指代表达分割方面，国内进展显著超过了国际进展. 大多数有影响力的工作均出自国内研究机构，这说明国内团队已经在该领域取得了一定的领先地位. 与此同时，Zhou 等[73]率先对中文指代表达分割做出了尝试. 短语定位领域，国内研究稍落后于国际水平. 由于短语定位数据集的标注成本较高，国际研究团队近年针对弱监督学习发表了一些工作，而国内则主要着眼于模型结构和训练策略的创新.

针对视频与自然语言的跨媒体关联分析与理解的研究，一方面，集中在跨模态特征的提取与交互机制方面，国内外的研究人员引入了众多新的网络结构来探究更好的跨模态特征提取方式，例如，尝试了包含上下文信息的动态卷积网络[66]，引入了动态路由胶囊网络[48]. 国内外的研究人员也尝试更充分地利用视觉和语言的信息来进行更细粒度的特征交互，例如，尝试引入图像与语言中的空间相对位置信息之间的对应[67]，尝试引入图像与语言中定位目标的参考实体之间的对应[43]. 国内外的研究人员还在探索更好的跨模态交互机制，尝试了将 Transformer 结构和图神经网络进行跨模态的交互[44,60,64]. 另一方面，国内外研究人员将弱监督学习模式在多种任务上均进行了尝试，并取得了一定的成功[47,65].

在视听定位任务上，国内研究人员在其具体的任务上基于任务本身的性质提出了新颖的解决方法. 例如，Sun 等[69]基于人类听觉系统设计了基于超材料的单麦克风监听系统. Lu 等[70]建立了语音和人类嘴唇运动之间的对应关系. 在视听关联学习任务上，国内的研究人员引入了新的网络结构和各式的注意力机制来提升模型的性能；国外的研究人员则更多地探索了音频和视觉特征的表示学习与特征对齐.

3 基于知识图谱的跨媒体表征

3.1 国际研究现状

3.1.1 跨媒体知识图谱构建

2012 年，谷歌提出了知识图谱这一概念，有效地刻画了异构实体及关系. 目前，知识图谱主要分为 2 类：(1) 通用知识图谱，通常用于辅助问答系统、推荐系统和信息检索系统；(2) 领域知识图谱，通常用于解决医疗、教育和金融等特定领域的具体问题. 在跨媒体知识图谱构建过程中，获取到的原始知识数据的结构可以划分为以下 3 类：(1) 以特定数据结构展示和存储的结构化数据，如逗号分隔值(comma-separated values, CSV), Java 脚本对象简谱(JavaScript Object Notation, JSON)等；(2) 基于特定存储方式的半结构化数据，如超文本标记语言(hyper text markup language, HTML)等；(3) 不具备结构的非结构化数据，例如，不含超链接的文本、缺乏标注的网络视频和图片等. 为了从上述

类型数据中构建知识完备的跨媒体知识图谱，通常需要采用知识图谱补全、实体识别与链接以及实体关系学习的技术.

(1) 知识图谱补全

知识图谱补全主要分为面向三元组的单步补全方法和面向路径的多步补全方法. 在单步补全中，基于嵌入的补全方法聚焦于学习实体和关系在低维空间中的嵌入: 得分函数计算每个新的实体替换原有三元组头、尾实体后的得分，得分最高的若干个实体生成新的三元组，并加入知识图谱中实现补全. 国际上的经典方法包括 TransE[74], HolE[75]以及关系图卷积网络(relational graph convolutional network, R-GCN)[76]. 多步补全方法主要是针对知识图谱中存在的路径信息，期望获得非直接邻接实体之间的关系，从而有效地提升知识图谱对事件的推理能力. Gardner 等[77]在路径排序算法的基础上引入了向量空间相似性的启发式算法，缓解了路径排序算法中的特征稀疏问题. 在基于规则的推理方面，Abboud 等[78]提出了一个空间平移嵌入模型(box embedding model, BoxE)，解决了将知识嵌入到潜在空间的理论局限性，BoxE 可以从丰富的规则语言类中捕获和注入规则. Sen 等[79]证明了并非所有基于规则的知识库补全模型都是相同的，并提出了混合关系和混合路径 2 种不同的学习方法.

(2) 实体识别与链接

命名实体识别是指识别数据中的命名性实体，并将其划分到指定类别的任务[80]. 命名实体的神经结构条件随机场(neural architectures for named entity conditional random field, NN-CRF)[81]是一种典型的堆叠式神经网络架构，包含数个长短记忆网络(long short-term memory, LSTM)层和条件随机场(conditional random field, CRF)层以学习每个词位置处的词向量及最佳的分类标签. 为了减少实体分类过程中的标签噪声，PLE[82]提出了一种带有异构图的局部标签嵌入模型，以表示实体名、文本特征、实体类型及其关系. 为了应对类型数量的增长，Ma 等[83]提出了一种带有层级信息的原型驱动的标签嵌入方法，以解决细粒度零次学习命名实体分类. 实体链接主要解决具有不同实体名的同一实体的对齐问题，是指将数据中实体名指向其所代表的真实实体的任务，通常也被称为实体消歧[84]. Ganea 等[85]提出了局部内容窗口上的注意力机制神经网络和可微的消息传递机制以分别学习实体嵌入和推断模糊实体. Le 等[86]将实体间的关系视为隐变量，提出一种带有逐关系和逐实体名归一化的端到端神经架构. Adjali 等[87]率先实现了基于多模态数据的实体链接算法，构建了 Twitter 跨媒体知识图谱.

(3) 实体关系学习

实体关系学习的目标是从非结构化数据中提取出未知的关系事实三元组，用以自动构建大规模知识图谱，又称为关系抽取. 由于缺少标注好的关系数据，研究者常用远程监督(即弱监督)方法对知识和非结构化数据进行启发式对齐，构建大量的训练数据. 该方法的主要假设: 给定已有知识库中的一个事实三元组，若某条外部数据中同时包含了其中两个实体，则该数据一定程度上表达了该关系事实，即可对其进行自动标注. 深度神经网络已经成为实体关系学习的主流方法. Nguyen 等[88]使用带有多尺寸卷积核的多窗口卷积神经网络将已有的关系分类方法拓展至关系抽取任务. Miwa 等[89]基于依赖树堆叠序列和树结构的 LSTM 以实现关系抽取. 在此基础上，多种深度学习范式的引入进一步提高了实体关系学习的准确率. 在注意力机制方面，Soares 等[90]提出使用 Transformer 的预训练关系表征实现实体关系学习. 在图卷积网络方面，上下文图神经网络(contextualized graph convolutional network, C-GCN)[91]是一种建立在语句依存树的上下文图神经网络.

3.1.2　跨模态知识图谱嵌入

在跨模态场景下，不同模态的实体特征通常呈现异质性. 将不同模态的实体映射到统一的表示空间，是跨模态表征学习的常见做法. 该问题通常被称为嵌入.

(1) 面向文本的跨模态知识图谱嵌入

该任务通过文本主题分类器使语言模型适应特定领域或任务，以提高语音识别或机器翻译系统性能. 在常见的文本分类任务中，所利用的特征通常是局部的词汇特征，未能充分地利用词汇所在句子的结构信息. 在情感分类或空位填充等更复杂的问题中，需要引入上下文信息，基于语法特征需在评估阶段运行代价昂贵的解析模型. Wang 等[92]研究了从大型知识图谱和文本语料库中推理新关系事实的嵌入方法，在嵌入过程中兼顾知识图谱中实体关系以及文本语料库单词，同时定义了描述知识和文本一致性的概率模型. Marin 等[93]基于跨模态知识图谱探索了一种利用短语模式特征实现文本分类的方法，将每个句子拆分成多个短语，分别和知识图谱中的实体或者关系做匹配.

(2) 面向图像的跨模态知识图谱嵌入

对于包含图像的多模态表征问题，Pezeshkpour 等[94]提出基于不同神经编码器的跨模态数据多峰知识库嵌入方法. 该方法通过不同的神经编码器对不同模态知识进行建模，并将神经编码器与关系模型相结合，引入补全模型从知识图谱中生成对应的属性信息，对跨模态知识的嵌入进行补全. Zhu 等[95]提出了一个基于知识库的框架处理各种视觉查询，而无需为新任务训练新的分类器，将大型马尔可夫随机场(Markov random field, MRF)转换为知识表示形式，并结合视觉信息、文本、结构化数据以及三者之间的多种关系. 与其他针对标准识别和检索任务的专用模型相比，此系统在回答更丰富的多模态视觉查询方面表现出更大的灵活性.

(3) 面向语音对话系统的跨模态知识图谱嵌入

在面向任务的对话系统(spoken dialogue system, SDS)中，跟踪用户目标问题受到了广泛关注. Ma 等[96]提出了推理知识图谱，将现有的大规模语义知识图谱映射到 MRF，创建基于口语对话系统的用户目标跟踪模型. 由于语义知识图谱包含实体和其他多模态知识，首先在原始知识图谱的实体和关系上引入势因子，将语义知识图谱转换成 MRF 因子图. 对于知识图谱中的每个实体，根据 MRF 因子图得到其与用户目标的符合程度. 选取符合程度高的实体，生成证据实体，计算其条件概率以表示符合用户问题的程度. 最后，对所有实体进行排序，选取条件概率最高的实体，即推理系统对输入对话的回答.

(4) 面向复杂网络的跨模态知识图谱嵌入

随着在线社交网络的快速发展，了解用户行为和网络动态成为社交网络挖掘中重要且具有挑战性的问题. Yang 等[97]定义了面向社交的知识图谱嵌入学习问题. 对于给定的一个社交网络，包括对应的知识图谱以及用户在社交网络上发布的文本和视觉等多模态信息，研究者旨在将每个社交网络用户链接到给定的概念知识. 社交知识图谱嵌入学习在用户建模、推荐和搜索中均具有潜在的应用. 研究者通过形式化以上社交知识图谱嵌入学习的问题，提出了一种新颖的多模态贝叶斯嵌入模型 GenVector，使用共享的潜在嵌入空间对多模态知识进行建模.

3.2 国内研究现状

3.2.1 跨媒体知识图谱构建

跨媒体知识图谱构建方面，国内尚处于起步阶段. Richpedia[98]以从搜索引擎中得到的图像和检索词作为跨模态知识图谱的核心内容，侧重于跨模态数据的分类任务. 该图谱构建者认为已有跨模态知识图谱虽然融合了多模态知识，但均是基于已有的文本知识图谱进行构建，未考虑图像的多样性，图像未作为单独的知识实体存在.

在单步补全方法中，国内研究者在 TransE 的基础上发展出 TransR[99]和 TransH[100]等方法. 这类改进聚焦于低维嵌入空间和评分函数的选择，采用了表征能力更强的嵌入空间和评分函数以捕获更为复杂的结构信息. Guan 等[101]引入共享网络结构和自适应加权的损失函数，在不同空间中分别获得头尾实体和关系的隐式特征. 在多步补全方法中，随机游走算法受限于知识图谱的图结构，难以发现未直接连接的路径. 为解决此问题，Neelakantan 等[102]将路径递归地分解成为若干个关系的嵌入，Chen 等[103]则将其分解为路径发现和路径推理 2 个子步骤.

在命名实体识别中，国内研究者提出了多种创新框架. Xia 等[104]提出的集成框架包含多种粒度的实体位置检测，以及适用于嵌套和非重叠命名实体的基于注意力机制的实体分类. Hu 等[105]通过多任务学习框架区分多标记实体和单标记实体. 最近，Li 等[106]通过参考注释准则构造查询问题，将普通和嵌套命名实体识别任务统一为机器阅读理解框架. Zhao 等[107]使用局部类别和全局三元组知识以强化统一表征. 在实体链接领域，国内研究者致力于对知识及内容上下文进行联合建模. 例如，Fang 等[108]首先建立了联合特征学习与链接模型，而 Cao[109]等提出一种多原型实体名嵌入模块，学习包含同一实体名的多个含义的嵌入，据此判断实体名所属的实体. Fang 等[110]将任务转化为序列学习问题，并提出一种强化学习方法，从全局角度对实体进行链接. Chen 等[111]则在模型中引入了潜在的实体类型信息，以减少对不同类型实体的误链接.

国内研究者在深度实体关系学习领域做出了开创性工作. Zeng 等[112]将相对于实体的距离作为特征输入卷积神经网络进行关系分类，并进一步提出分片卷积神经网络(piecewise convolutional neural networks, PCNN)[113]以更好地捕捉实体对间的结构信息. 在此基础上，Jiang 等[114]将其扩展至多标签学习，通过跨语句池化算子进行特征选择. 循环神经网络也被引入该领域，例如，Xu 等[115]使用了最短依赖通道的长短记忆网络(shortest de-

pendency path long short-term memory, SDP-LSTM), Cai 等[116]则通过双通道双向 LSTM 和 CNN 同时捕捉序列依赖关系和局部语义信息.

国内研究者在基于规则的推理方面也做出了一定的成果. Lin 等[117]提出了一种规则增强迭代互补的方法，其由规则学习、嵌入学习器以及三重鉴别器组成，这种迭代过程丰富了知识图谱的语义，提高了规则学习的完整性. Liang 等[118]提出了基于自下而上的知识图谱完备性规则的模型(bottom-up rule learning for knowledge graph completion, HRER)，其在现有关系规则挖掘方法的基础上修改了度量指标. 新的度量 Horn 规则可靠性(Horn rule reliability, HRR)在过滤 Horn 规则方面比传统的置信度更有效. 此外，基于嵌入的方法和基于逻辑规则的方法之间的差异，HRER 提出了实体规则. 实体规则在一定程度上弥补了 Horn 规则的有限表达.

3.2.2　跨模态知识图谱嵌入

在跨模态知识图谱嵌入方面，国内研究者取得了一系列的相关研究成果. Nian 等[119]提出了一种跨模态知识表示学习方法，可以有效地从网页中挖掘结构化的文本和视觉关系. 同时，将多模态知识图谱应用在了谣言检测任务中[120]，从外部的多模态知识图谱中获取文本背后的丰富知识信息，辅助高度精炼的文本语义表达. 在后续的研究中，提出了多模态知识层次注意网络(multi-modal knowledge-aware hierarchical attention network, MKHAN)[121]和多模态多关系特征聚合网络(multi-modal knowledge-aware hierarchical attention network, MMRFAN)[122]来对医疗多模态知识图谱中实体和关系进行建模. Chen 等[123]设计了一种模态知识嵌入方法 MMEA，以解决多模态知识图谱中的实体对齐问题. Xie 等[124]提出了体现图像的知识表示学习(image-embodied knowledge representation learning, IKRL)，学习事实关系三元组和图像特征，IKRL 框架图如图 2 所示. 该方法使用自动编解码器构造所有实体图像的特征表示，通过注意力机制在实体空间中构造图像表示，同时仍然学习基于结构的实体表示. Sun 等[125]提出了多模态知识图谱注意网络(multi-modal knowledge graphs attention network, MKGAT)，其分为嵌入模块和推荐模块. 嵌入层主要包括多模态图谱实体编码器和多模态图谱注意力层，得到了比基于单一模态知识图谱的方法更好的推荐效果.

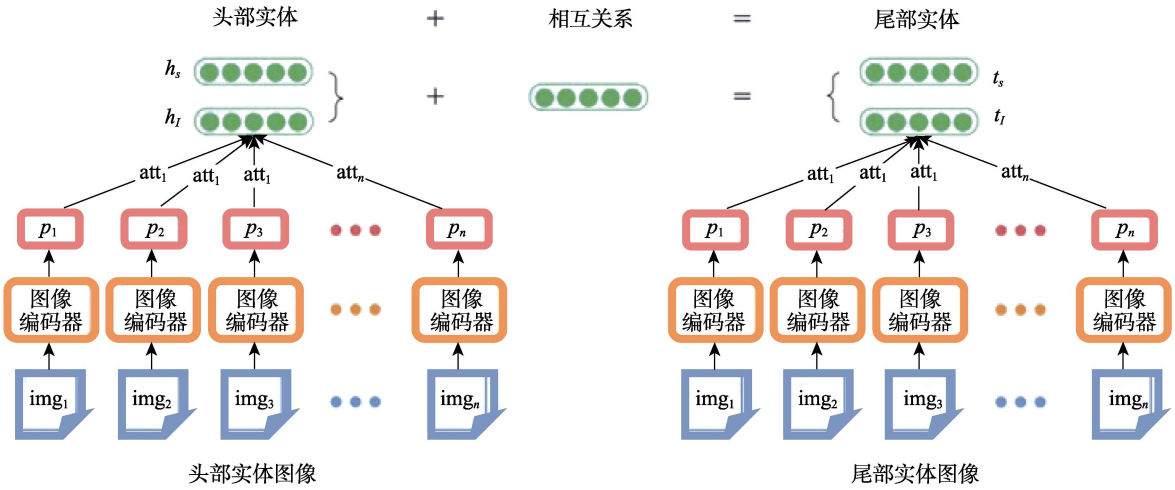


图 2　IKRL 框架图[124]

3.3　国内外进展对比与总结

跨媒体知识图谱构建方面，国外知识库构建起步较早，诞生了多个成熟的英文知识库. 国内知识图谱的构建起步较晚，但是发展迅速，尤其在领域知识图谱方面呈现百花齐放的态势：在电商领域，阿里巴巴和京东等电商平台均构建了其商品知识图谱；医疗领域，中国中医科学院研发了中医药知识服务平台，集成了多个领域知识图谱[126]；教育领域，百度发布了基础教育图谱. 而跨媒体知识图谱的构建成为近年来知识图谱相关研究的热点，国内外研究处于并驾齐驱的势态. 国内提出的多模态知识图谱 Richpedia[98]，克服了国外多模态知识图谱的短板，并提出了基于规则的构建方法，提高了知识图谱中图片实体和关系的质量.

在跨媒体知识图谱嵌入方面，多模态知识图谱中最常见的模态包括概念知识(关系知识)、文本知识以及图像视觉知识等. 国内外的多模态知识图谱嵌入研究通常是文本知识和图像视觉知识辅

助概念知识的建模. 国外的研究集中实体信息补全等任务; 而国内的研究主要集中在多个知识图谱之间的实体对齐、基于知识图谱的应用等方面.

4 面向多模态的智能应用

4.1 国际研究现状

4.1.1 多模态视觉问答

给定一幅图像和一个相关的文本问题, 视觉问答(visual question answerin, VQA)系统应该基于图像正确地回答问题, 使得 VQA 本质上是跨模态的. VQA 不仅对视觉和文本模式的衔接要求非常高, 对从目标识别和定位到高级推理和常识知识学习等多方面的能力要求也非常高. 传统的 VQA 方法通过端到端的方式训练一个神经网络, 使用(图像, 问题, 答案)三元组作为监督, 建立从给定图像和问题输入到一个候选答案的映射. 其核心思想是学习图像和问题的统一嵌入. 输入图像将通过预训练用于图像分类的卷积神经网络来获得图像表示. 音乐视频摘要的工作流程如图 3 所示.

在视觉世界, 人类有能力专注于特定的区域而不是整个场景. 在此启发下, 注意力机制[127]被广泛地应用于解决“看哪里”的问题. 注意力的核心思想是通过对相关区域的内容和边信息之间的相互作用进行建模, 使神经网络能够了解要关注的区域. Yang 等[128]提出了一种堆叠式注意网络(stacked attention networks, SAN), 利用文本问题的语义特征作为查询, 通过多层结构搜索相关的视觉区域. Lu 等[129]提出了一个层次共注意模型, 通过对图像进行问题引导注意和对问题进行图像引导注意, 该方法未采用简单的元素生成或串联, 而是采用双线性池模型及其变体, 通过计算 2 个向量的外积来实现两个向量中元素之间的交互, 进而获得了巨大成功. Fukui 等[130]提出了一种多模态紧凑双线性池化(multimodal compact bilinear, MCB)算法, 采用基于采样的计算和投影方法来降低维数, 同时保持完全双线性池的性能. Yu 等[131]利用部分矩阵分解技巧, 提出了多模态分解双线性(multimodal factorized bilinear , MFB)池化, 以提高收敛速度. 通过将低秩矩阵约束与 Tucker 分解相结合.

4.1.2 多模态视频摘要

视频摘要的目标是生成一个包含一部分视频片段的短视频摘要. 相关研究人员提出了大量的单峰方法来生成高质量的视频摘要, 其中无监督方法[132-135]通常采用人工设计的视觉准则和监督方法从视频中提取帧或镜头[136-137], 倾向于直接利用人工编辑的摘要示例来学习视频摘要模和挖掘视频摘要的特定视觉模式. 除了视觉特征外, 视频还与来自其他形式的丰富信息相结合, 如音频信号、文本描述等. 所有形式信息彼此对齐或互补, 在不同方面反映视频内容. 基于这一思想, 提出了多种多模态视频摘要方法.

传统的多模态视频摘要方法主要是对电影或音乐视频进行摘要, 从视频中检测与合成低水平的视觉、音频和文本线索, 以评估不同视频部分的显著性、代表性或质量, 然后提取这些信息的一部分, 生成最终的视频摘要. Xu 等[138]提出了基于视听文本分析和对齐的音乐视频摘要方法. 由图 3 可知, 先将音乐视频分为一个音乐曲目和一个视频曲目. 对于音乐轨迹, 基于音乐结构分析检测出合唱; 对于视频轨迹, 将视频镜头分割, 并分为近距离人脸镜头和非人脸镜头, 从这些镜头中提取歌词, 检测重复次数最多的歌词, 根据检测到的结果生成音乐视频摘要. Pan 等[139]通过编码文本和场景信息, 将故事镜头链接为图形的徽标, 提出了一种多模式面向故事的视频摘要模型.

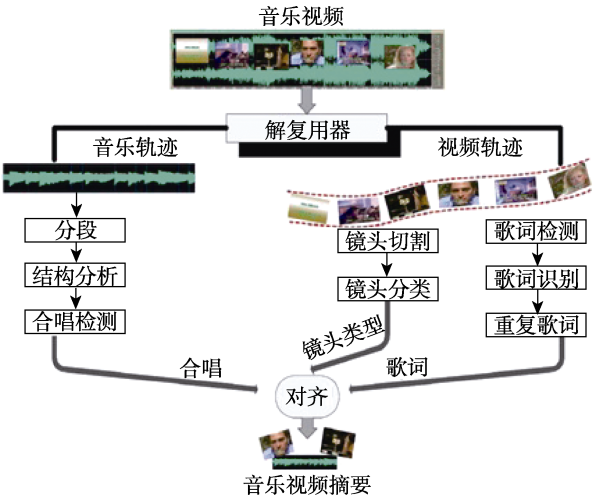


图 3 音乐视频摘要的工作流程[138]

与传统的离线视频不同, 在线网络视频中充斥着标签、标题、描述等各种辅助信息, 承载着丰富的领域知识. 几种多模式视频摘要方法是将 web 视频与其领域知识联系起来生成视频摘要. Song 等[140]观察到与标题相关的图像可以作为重要主题视觉概念的代理. 基于标题的视频摘要如图 4 所示, 利用视频标题通过图像搜索引擎检索 web 图像, 并开发一种共同原型分析技术, 学习视频和 web 图像之间共享的规范视觉概念.

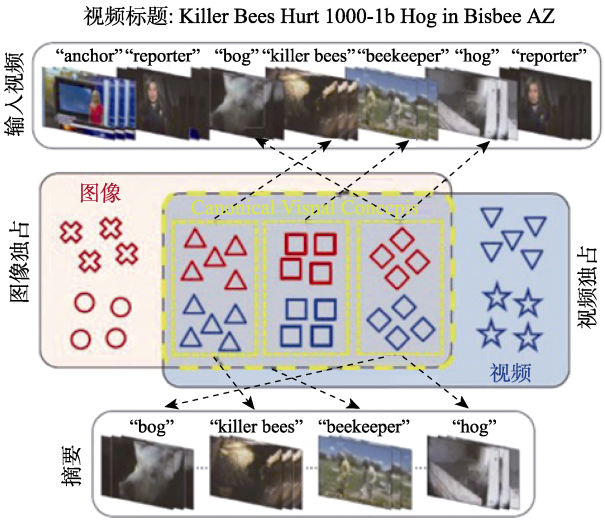


图 4　基于标题的视频摘要的图示[140]

4.1.3　多模态视觉模式挖掘

知识库是实体、属性及其相互关系的集合. 可视化知识库构建的前期工作只探索了有限的基本对象和场景关系. 一些多模态模式挖掘方法[141-145]旨在半自动地构造一个高级的"事件"模式，这种模式能够扩展纯文本模式构造方法. 并利用一个大型无约束的弱监督图像对语料库发现事件的视觉特征，并自动命名这些视觉成分.

为了挖掘用于构建知识库的与事件相关的多模态模式，寻找高质量的多模态视觉模式，本文定义了代表性和区别性 2 个准则. 其中，代表性意味着发现的模式应该是该类别中常见的，而区别性是指从一个类别中发现的模式不应该在其他类别中发现. 端到端多模态模式发现和生成管道如图 5[141]所示，它可以将发现的关联规则转换为多模态视觉模式.

多模态模式挖掘方法可以作为填补文本分析和视觉分析之间空白的桥梁. Zhang 等[143]和 Lu 等[144]利用多模态视觉模式挖掘框架[146]改进自然语言处理领域中的知识和事件抽取问题. Zhang 等[145]利用端到端的多模态模式发现和生成管道[141]进一步提升了知识和事件抽取问题的准确率. 与传统的纯文本事件抽取方法相比，多模态方法引入了从视觉领域发现的领域知识，并取得了显著的效果.

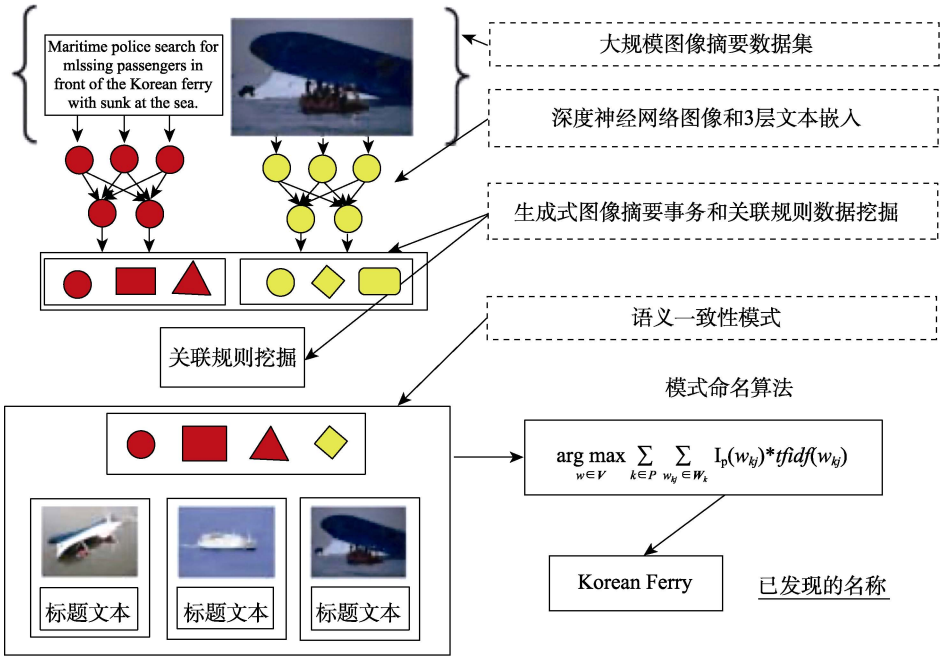


图 5　端到端多模态模式发现和生成管道[141]

4.1.4　多模态推荐

随着各种在线社交网络和多媒体网站的爆炸式增长，人们现在已经习惯于同时使用不同的媒体来满足其多样化的信息需求[147]. 跨模式信息共同反映了每个人的兴趣和偏好. 传递关联跨模态信息对于智能地为人们服务具有重要意义[148].

现有的多模式推荐工作可以从 2 个角度进行分组，即根据关联知识进行分类和根据整个模型结构进行分类.

国外的工作主要集中在根据关联知识进行分类. Yan 等[149]提出了一种基于潜在属性稀疏编码的主题关联框架，验证了在跨网络的协作问题中，利用用户协作实现异构知识关联的有效性. Lee 等[150]提出的稀疏编码算法经过多次变换后，可以有效地解决这一问题. van der Maaten 等[151]提出了一个嵌入和映射框架(embedding and mapping cross-

domain recommendation, EMCDR), 首先通过矩阵分解得到不同平台上的用户表示, 然后通过线性映射或多层感知器(multilayer perceptron, MLP)进行映射. 采用线性转移和 MLP 作为映射函数的 EMCDR 框架如图 6 所示.

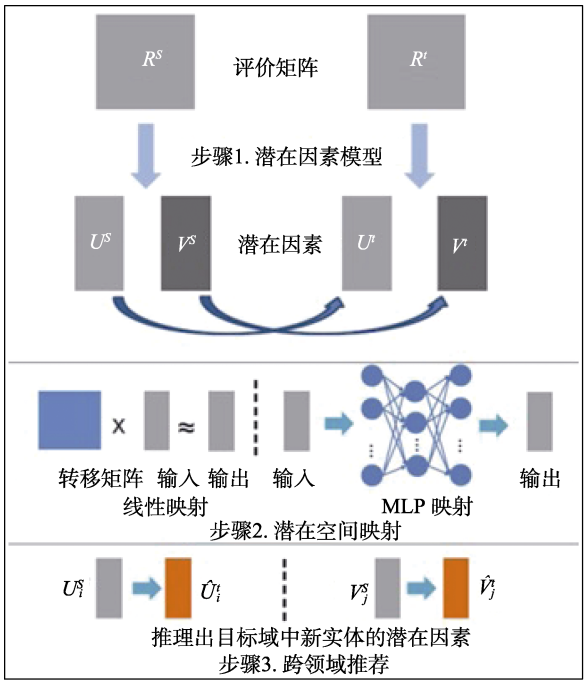


图 6 采用线性转移和 MLP 作为映射函数的 EMCDR 框架[151]

4.1.5 跨模态智能推理

跨模态智能推理的核心思想是通过刻画不同模态之间的模式, 并利用其中一种或多种模态的信息来推理与其相关的其他模态的结果. 跨模态智能推理的主流方法可以归纳为以下 4 类[152]: 子空间学习方法、深度学习方法、哈希表征方法和主题模型方法. 其中, 子空间学习方法假设不同模态的数据和特征可以被投影到一个潜在的共同子空间, 在该空间上可以刻画不同模态特征之间的相似性. Mahadevan 等[153]考虑不同模态之间的几何结构, 提出了通过学习多模态低维表征来构建不同模态之间的子空间. 基于深度学习方法的跨模态智能推理的思想类似于协同表征, 利用深度学习模型提取不同模态的特征, 并约束不同模态的表征之间的相似度. 例如, Ngiam 等[154]针对视频和语音等多模态数据, 训练了一个深度神经网络来学习跨模态融合以及共享表示方法. 基于哈希表征方法的跨模态智能推理主要是通过学习不同模态的哈希变换表征, 将不同模态的特征值映射到二值的汉明空间. 例如, Kumar 等[155]在传统单模态哈希表征方法的基础上进行拓展, 提出了多模态数据的哈希表征方法. 基于主题模型方法的跨模态智能推理的核心思想是将不同模态的数据和特征投影到一个共同的隐主题空间. 例如, Blei 等[156]提出了多模态分层概率混合模型, 通过得到跨模态数据的潜在狄里克雷分配, 将潜在主题空间作为不同模态共同的潜变量来模拟模态间的联合分布.

4.1.6 跨模态医学图像预测

跨模态医学图像预测是另一大研究领域, 跨模态医学图像预测是由一种模态图像预测另一种模态图像[157]. 常用医学模态包括计算机体层摄影(computed tomography, CT)、磁共振成像(magnetic resonance imaging, MRI)、正电子发射计算机断层显像(positron emission computed tomography, PET)、超声成像以及各类成像模态等[158]. 其中, CT 的密度分辨力较高, 成像速度快, 但具有电离辐射; MRI 图像中诊断信息丰富, 适合软组织成像, 但在某些环境下空间分辨率不及 CT, 体内有金属物品者也不宜做 MRI; PET 检查灵敏度高, 也更安全, 但价格较贵, 成本偏高; 超声成像成本相对低廉, 但是分辨率相对较低. 几种常用的医学模态各有特点, 通过跨模态医学预测能够更好地辅助医生提高诊断效率. Burgos 等[159]提出利用多图集信息生成图像的算法, 使用局部图像相似性度量将 MRI 衍生的患者形态与 MRI/CT 对数据库进行局部匹配, 并基于形态相似性赋予图集不同的权重大小. Wolterink 等[160]为了克服成对图像之间错位导致的合成图像错误, 选择直接采用不成对且未校准的 MRI-CT 图像数据来训练一个生成对抗网络(generative adversarial networ, GAN), 将原本用于合成自然图像的 CycleGAN 模型[161]应用到 CT 预测上. 2015 年, Bahrami 等[162]提出利用多级典型相关分析(canonical correlation analysis, CCA)和组稀疏性作为分层框架, 以 3T MRI 来进行 7T MRI 重构, 主要使用基于实例[163]和组稀疏的方法, 并通过将低分辨率的 3T MRI 图像与高分辨率的 7T MRI 图像映射到 CCA 空间中的方法[164], 提升两者的相关性, 尝试恢复更丰富的结构细节. 2016 年, 其又提出基于 CNN 的 7T MRI 图像构建方法[165], 通过直接学习低/高分辨率图像之间的端到端映射来完成高分辨率图像的重建[166-167].

4.2 国内研究现状

4.2.1 多模态视觉问答

与国外研究的侧重点不同, 国内研究主要集中在结合领域知识上. 正确回答视觉问题可能需

要从常识到专家领域知识的额外信息，这远远超出了训练数据集所能提供的范围. 因此，将从其他来源检索到的有用领域知识整合到 VQA 系统中是很有吸引力的. Li 等[168]提出了知识融合的动态记忆网络(knowledge dynamic memory networks, KDMN)框架，它可将海量领域知识转化为语义空间来回答视觉问题.

4.2.2　多模态视频摘要

在具有各种辅助信息的在线视频的多模式视频摘要领域里, Wang 等[169]提出了基于标记定位和关键镜头挖掘的事件驱动视频摘要方法，将与每个视频相关联的标记定位到其快照中，在获得镜头相对于所有标签的相关性得分之后，估计每个镜头相对于事件查询的相关性得分. 通过探索关键子事件的重复发生特征，识别出一组具有较高关联度的关键镜头. Yuan 等[170]提出了语义嵌入模型，利用从在线视频的侧面信息获得的领域知识生成视频摘要.

4.2.3　多模态推荐

从关联知识的角度研究多模态模型时，有一组遵循以用户为中心的方法，主要关注重叠用户的跨模态信息. 一个直接的解决方案是将交叉模态关联作为一个线性传递问题，并寻求一个基于回归的显式传递矩阵. TLRec[171]利用重叠的用户和项目作为跨越不同媒体的桥梁，对潜在向量引入平滑约束和正则化. Jiang 等[148]等通过提出 XPTrans 模型引入了一个对齐的跨模式用户行为相似性约束，该模型利用少量重叠人群来优化桥接不同媒体.

在按整个结构分组方面，一些工作在查阅有关整个结构的现有文献时，设计了一组方法来构建一个统一的框架[148,171]. 其中，前 2 个工作使用基于矩阵分解的技术，后 3 个工作使用基于概率模型的策略. 另外一些工作采用两步程序[149,172-173]，将来自不同媒体的用户表示在其潜在空间中，然后将这些表示联合起来.

上述方法的核心思想相同，即所有跨模态信息均是一致的. 然而，少数文献[173-174]发现了跨媒体关联表示过程中存在的数据不一致现象，并试图通过数据选择来解决此问题. Lu 等[174]等发现选择媒体一致的辅助数据对于跨模式协同过滤非常重要，提出了基于经验预测误差和方差的一致性评价准则，将该准则引入 boosting 框架中，实现了知识的选择性转移. Yan 等[173]将用户分为 3 组，提出了预定义的微观用户特定度量，以自适应地加权数据，同时集成不同媒体上的异构数据.

4.2.4　跨模态智能推理

子空间方法中, Feng 等[175]设计了一种跨图像-文本的多模态检索框架，该检索框架利用一致自编码器构建的多层编码网络分别提取不同模态的编码，并通过不同模态编码的最小化距离保持模态之间的对应关系. 在基于哈希变换的跨模态方法上，Ding 等[176]提出了一种集合矩阵分解的哈希方法，通过对模型的集合矩阵分解，从一个样本的不同模态中学习统一的哈希编码实现多模态推理. Wang 等[177]通过将正交正则化方法应用在加权参数上，提出了基于哈希编码的正交深度神经网络，来学习更加紧凑的多模态表示. Song[178]将不同模态的数据转换到公共汉明空间，提出了模态间的哈希模型来计算不同模态的相关性. 在字典学习框架下，Wu 等[179]通过联合不同模态的字典，学习不同模态的数据的稀疏编码方式，设计了稀疏多模态哈希算法. 基于深度学习的方法, Wang 等[180]利用基于深度学习的映射机制，设计了堆叠式自动编码器来学习各个模态的编码，以捕获来自异构源数据的模态内和模态间的语义相关性. 在主题模型方面，Liao[181]开发了复合的非参数贝叶斯多模态主题模型，用于计算任一模态内部和不同模态之间的相关性，通过在高斯过程中引入监督响应变量，提升了相关性结构学习的灵活性. Wang 等[182]通过构建联合的多模态概率图模型，提出了多模态多主题增强的建模方法，来捕获跨模态的语义信息，发现不同模态之间一致的语义主题. Huang 等[183]通过引入门限视觉语义嵌入方法来处理小样本的图像-文本匹配问题，多个视觉语义嵌入模块相互融合可以完成无监督的跨模态检索任务.

4.2.5　跨模态医学图像预测

跨模态医学图像预测领域国内也进行了一定的研究. Nie 等[184]利用 3D 全卷积神经网络(fully convolutional network, FCN)来学习从 MRI 图像到 CT 图像的端到端非线性映射，与传统 CNN 架构相比，FCN 生成结构化输出，可以更好地保留预测 CT 图像中的邻域信息，避免切片与切片间出现不连续的问题. Shin 等[185]利用迁移学习将从自然图像数据集预训练的 CNN 模型微调到医学图像任务. Han 等[186]提出了一种用于合成 CT 生成的新型深度卷积神经网络方法，其在测试时准确性和计算速度方面均优于基于图集的方法. Bi 等[187]通过多通道生成对抗网络合成 PET 数据，以解决之前方法合成 PET 图像存在的低分辨率和低信噪比等

问题.

4.3 国内外进展对比与总结

多模态视觉问答是在计算机视觉与自然语言处理领域均发展到一定程度的产物，随着近年来深度学习领域突破性的进展才逐渐走入研究者的视野. 传统的 VQA 方法是基于端到端的方式训练神经网络，这些工作大部分由国外的研究者完成. 在较为新兴的 VQA 方法，如结合领域知识方面，清华大学的研究者[168]提出了知识融合的动态记忆网络框架.

深度学习为多模态视频摘要提供了新的生命力，国内外越来越多的研究者开始注意到这一问题. 在这方面虽然国内起步稍晚，但国内研究者进行了许多的应用和优化，并取得了相当的研究成果. 例如，合肥工业大学[169]与清华大学[170]的研究团队分别在具有各种辅助信息在线视频的多模式视频摘要领域，提出了基于标记定位和关键镜头挖掘的事件驱动视频摘要方法与深层语义嵌入模型.

在多模态视觉模式挖掘研究中，国外具有较大的先发优势，大部分相关的研究均是在国外进行的. 我国只有香港城市大学[146]的研究团队提出了一种多模态视觉模式挖掘框架，跨模态的知识融合较之单纯的文本知识有大幅提升. 现有的多模态视觉模式挖掘，在传统方法的基础上引入了视觉领域发现的领域知识，进而跨模态地传递知识，抽取和挖掘模式，取得了较好的成效.

多模态推荐工作主要分为根据关联知识进行分类和根据整个模型结构进行分类. 这是唯一国内比国外占有优势的应用方向. 国外的研究主要集中在前者，而国内的研究中 2 类均有涉猎. 针对推荐算法用户重叠的情况，清华大学[148]的研究团队提出利用 XPTrans 模型引入了对齐的跨模式用户行为相似性约束. 中国科学院[173]的研究团队针对不同媒体上的异构数据，提出了预定义的微观用户特定度量，从而自适应地加权数据.

在跨模态检索方面，近年来，基于 BERT 的预训练模型也被越来越多地应用于跨模态检索任务，国内外的研究在这方面均有涉及. BERT 相关的方法主要包括单流模型和双流模型. 比较具有代表性的单流模型包括 VideoBERT[188], Visual-BERT[189]和 VLBERT[190]等，主要是将视频数据输入预训练好的语言模型中. 双流模型(如 ViLBERT[191])则是使用 2 个 BERT 分别处理视觉输入和文本输入，然后在后续层进行信息传递，完成跨模态特征之间的相互优化. 国内还有许多研究者在寻找可迁移/零样本的跨模态检索方法[192-193]，现有的深度跨模态检索方法通常需要大量标记数据进行训练才能实现高性能. 然而，手动注释多模态数据既耗时又昂贵，而且同一数据不同模态的标签也可能存在不同. 该思路聚焦于如何将有价值的知识从现有的带注释数据转移到新数据，从先前标记的源域中迁移知识，以提高对未标记的目标域的检索性能，有效利用具有不同标签空间的未标记和有标记的多模态数据.

5 发展趋势与展望

本文总结了跨媒体分析技术，并讨论了多模态数据的统一表达和知识引导的数据融合、跨媒体关联分析、基于知识图谱的跨媒体表征技术以及面向多模态的智能应用. 然而，多媒体的跨模态智能关联分析与语义理解仍面临着巨大的挑战. 未来多模态研究方向有以下几个方向.

5.1 跨模态集体智能

集体智慧的概念最初来源于昆虫学家惠勒的观察. 从表面上看，独立的个体可以非常紧密地协同工作，使其看起来像一个单一的有机体. 在人类社会中，考虑到单个个体做出的决定与大多数人做出的决定相比往往不准确，集体智慧是一种共享的智慧，也是一种汇集意见并将其转化为决策过程的过程. 所有这些现象或事例均证实了集体智慧可以产生更强大的“超有机体”或拥有更多智慧的大脑. 基于丰富的跨模态信息，可认为集体智能可以用于人类的规划，这是人类所共有的独特而复杂的特征.

5.2 多模态联合学习

随着多模态联合学习技术的发展，对大规模标签数据集的依赖性问题逐渐显露. 无监督学习可以利用大量廉价的数据进行学习，如何改进无监督多模态联合学习模型的效果可能是未来的研究热点. 零样本学习也可以在一定程度上减轻数据集标注的压力，目前零样本学习模型大多基于两个同构模态建立，未来可以尝试引入更多模态，或者使用异构模态来提升零样本学习模型的预测能力. 在多模态特征融合方面，还有众多可以探索的方向. 例如，如何有效地学习包含不精确数据、不正确数据和冗余数据的融合特征，如何平衡模态交互的丰富性与算法的复杂性以及如何解决融合模型泛化能力不足的问题. 同时，对多模态信息之间关系的通用建模方法也是一大难题.

5.3　跨媒体关联分析

图像和自然语言模态的关联分析与理解对众多跨模态应用具有重要作用. 在视频与自然语言的跨媒体关联分析与理解相关方向，更加准确地理解自然语言描述和视频图像内容是目前研究的重点. 未来更高效地建模跨模态交互的机制将成为研究的重点，此外，由于视频本身的数据密集性与标注稀疏性的矛盾，弱监督学习将成为更具有实际价值的范式，如何利用有限的标注训练出更具泛化性能的模型将是未来研究的重点. 音频与视觉的跨媒体关联分析与理解相关方向是活跃的研究领域，如何在拥有大规模但缺少监督信息的数据情况下，充分利用音频和视频模态的对应性来切实提升多模态任务的性能将是重要且具有实际意义的研究方向.

5.4　多媒体特征提取和表达

传统的技术和方法均主要注重算法和模型的高效性. 基于海量数据，许多深度学习算法与模型被提出用于充分挖掘数据中的关联知识，高效地提取特征和学习表征. 但是，这些方法均未能考虑数据偏差的问题，数据偏差会导致数据中存在虚假关联. 相对于虚假关联来说，因果关联更有效且可解释. 因果推理是甄别数据关联中虚假关联和因果关联的关键技术. 如何利用因果推理技术来实现多媒体特征提取和表达的无偏性、稳定性和可解释性是亟待解决的核心问题.

6　结　语

本文对跨媒体智能关联与语义理解理论技术进行了全面而深入地研究，主要从多模态数据的统一表达、知识引导的数据融合、跨媒体关联分析、基于知识图谱的跨媒体表征技术以及面向多模态的智能应用这 5 个方面展开论述，梳理了其在多模态数据融合以及跨媒体分析推理方面的研究进展，深入地分析了跨媒体智能理论与技术最新的国际研究现状以及国内研究现状，并对国内外进展开展了详细地对比分析. 最后，依据现有理论和技术的现状，本文介绍了人工智能新时代的未来研究方向，指出了具有前景的方向，并对未来跨媒体领域的发展趋势和研究方向进行了展望.

参考文献(References):

[1] McGurk H, MacDonald J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746-748

[2] Zhu W W, Wang X, Li H Z. Multi-modal deep analysis for multimedia[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3740-3764

[3] Wang Shuhui, Yan Xu, Huang Qingming. overview of research on cross-media analysis and reasoning technology[J]. Computer Science, 2021, 48(3): 79-86(in Chinese)
(王树徽，闫旭，黄庆明. 跨媒体分析与推理技术研究综述[J]. 计算机科学, 2021, 48(3): 79-86)

[4] Rasiwasia N, Pereira J C, Coviello E, *et al*. A new approach to cross-modal multimedia retrieval[C] //Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 251-260

[5] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis[M]. Cambridge: Cambridge University Press, 2004

[6] Andrew G, Arora R, Bilmes J, *et al*. Deep canonical correlation analysis[C] //Proceedings of the 30th International Conference on Machine Learning. New York: ACM Press, 2013: 1247-1255

[7] Srivastava N, Salakhutdinov R. Multimodal learning with deep boltzmann machines[C] //Proceedings of the 25th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2012: 2222-2230

[8] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3128-3137

[9] Castrejón L, Aytar Y, Vondrick C, *et al*. Learning aligned cross-modal representations from weakly aligned data[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2940-2949

[10] Bernardo J M, Smith A F M. Bayesian theory[M]. New York: John Wiley & Sons, 2009

[11] Dempster A P. A generalization of Bayesian inference[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1968, 30(2): 205-232

[12] Jensen F V. An introduction to Bayesian networks[M]. London: UCL Press, 1996

[13] Box G E, Tiao G C. Bayesian inference in statistical analysis[M]. New York: John Wiley & Sons, 2011

[14] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[OL]. [2021-08-10]. https://arxiv.org/pdf/1503.02531

[15] Luo Z L, Hsieh J T, Jiang L, *et al*. Graph distillation for action detection with privileged modalities[C] //Proceedings of the European Conference on Computer Vision. Heidelberg Springer, 2018: 174-192

[16] Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2827-2836

[17] Zhang C R, Peng Y X. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification[OL]. [2021-08-10]. https://arxiv.org/pdf/1804.10069

[18] Sutton R S, Barto A G. Introduction to reinforcement

learning[M]. Cambridge: MIT Press, 1998

[19] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285

[20] Chen L L, Lu K, Rajeswaran A, *et al*. Decision transformer: reinforcement learning via sequence modeling[J]. Advances in Neural Information Processing Systems, 2021, 34: 15084-15097

[21] Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: a survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274

[22] Ibarz J, Tan J, Finn C, *et al*. How to train your robot with deep reinforcement learning: lessons we have learned[J]. The International Journal of Robotics Research, 2021, 40(4/5): 698-721

[23] Liang J, He R, Sun Z N, *et al*. Group-invariant cross-modal subspace learning[C] //Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2016: 1739-1745

[24] Song G L, Wang S H, Huang Q M, *et al*. Multimodal similarity Gaussian process latent variable model[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4168-4181

[25] Hua Y, Wang S H, Liu S Y, *et al*. Cross-modal correlation learning by adaptive hierarchical semantic aggregation[J]. IEEE Transactions on Multimedia, 2016, 18(6): 1201-1216

[26] Xia D L, Miao L, Fan A W. A cross-modal multimedia retrieval method using depth correlation mining in big data environment[J]. Multimedia Tools and Applications, 2020, 79(1): 1339-1354

[27] Wu Y L, Wang S H, Huang Q M. Online asymmetric similarity learning for cross-modal retrieval[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3984-3993

[28] Ma L, Lu Z D, Shang L F, *et al*. Multimodal convolutional neural networks for matching image and sentence[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2623-2631

[29] Huang Y, Wu Q, Wang W, *et al*. Image and sentence matching via semantic concepts and order learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(3): 636-650

[30] Wang S H, Chen Y Y, Zhuo J B, *et al*. Joint global and co-attentive representation learning for image-sentence retrieval[C] //Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM Press, 2018: 1398-1406

[31] Wu Y L, Wang S H, Song G L, *et al*. Augmented adversarial training for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2021, 23: 559-571

[32] Peng Y X, Zhai X H, Zhao Y Z, *et al*. Semi-supervised cross-media feature learning with unified patch graph regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(3): 583-596

[33] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2021-08-10]. https://arxiv.org/pdf/1804.02767

[34] Yang Z Y, Gong B Q, Wang L W, *et al*. A fast and accurate one-stage approach to visual grounding[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4682-4692

[35] Redmon J. Darknet: open source neural networks in C[OL]. [2021-08-10]. https://pjreddie.com/darknet/

[36] Devlin J, Chang M W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding[C] //Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4171-4186

[37] Sun M J, Xiao J M, Lim E G. Iterative shrinking for referring expression grounding using deep reinforcement learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 14055-14064

[38] Huang S F, Hui T R, Liu S, *et al*. Referring image segmentation via cross-modal progressive comprehension[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10485-10494

[39] Jain K, Gandhi V. Comprehensive multi-modal interactions for referring image segmentation[OL]. [2021-08-10]. https://arxiv.org/pdf/2104.10412

[40] Gupta T, Vahdat A, Chechik G, *et al*. Contrastive learning for weakly supervised phrase grounding[C] //Proceedings of the European Conference on Computer Vision. Heidelberg Springer, 2020: 752-768

[41] Wang Q, Tan H, Shen S, *et al*. MAF: multimodal alignment framework for weakly-supervised phrase grounding[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 2030-2038

[42] Rodriguez-Opazo C, Marrese-Taylor E, Saleh F S, *et al*. Proposal-free temporal moment localization of a natural-language query in video using guided attention[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2020: 2453-2462

[43] Mun J, Cho M, Han B. Local-global video-text interactions for temporal grounding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10807-10816

[44] Varol G, Momeni L, Albanie S, *et al*. Read and attend: temporal localisation in sign language videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 16852-16861

[45] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010

[46] Sadhu A, Chen K, Nevatia R. Video object grounding using semantic roles in language description[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10414-10424

[47] Yang X, Liu X L, Jian M, *et al*. Weakly-supervised video object grounding by exploring spatio-temporal contexts[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1939-1947

[48] McIntosh B, Duarte K, Rawat Y S, *et al*. Visual-textual capsule routing for text-based video segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9939-9948

[49] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 3859-3869

[50] Seo S, Lee J Y, Han B. Urvos: unified referring video object segmentation network with a large-scale benchmark[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 208-223

[51] Gu R Z, Zhang S X, Xu Y, *et al*. Multi-modal multi-channel target speech separation[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 530-541

[52] Zhu L Y, Rahtu E. Separating sounds from a single image[OL]. [2021-08-10]. https://arxiv.org/pdf/2007.07984

[53] Zhao H, Gan C, Ma W C, *et al*. The sound of motions[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1735-1744

[54] Surís D, Duarte A, Salvador A, *et al*. Cross-modal embeddings for video and audio retrieval[C] //Proceedings of the European Conference on Computer Vision Workshops. Heidelberg: Springer, 2018: 711-716

[55] Nagrani A, Albanie S, Zisserman A. Learnable PINs: cross-modal embeddings for person identity[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 73-89

[56] Liao Y, Liu S, Li G B, *et al*. A real-time cross-modality correlation filtering method for referring expression comprehension[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10877-10886

[57] Yang S B, Li G B, Yu Y Z. Graph-structured referring expression reasoning in the wild[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9949-9958

[58] Luo G, Zhou Y Y, Sun X S, *et al*. Multi-task collaborative network for joint referring expression comprehension and segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10031-10040

[59] Mu Z S, Tang S L, Tan J, *et al*. Disentangled motif-aware graph learning for phrase grounding[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 13587-13594

[60] Liu D Z, Qu X Y, Liu X Y, *et al*. Jointly cross-and self-modal graph attention network for query-based moment localization[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 4070-4078

[61] Zeng R H, Xu H M, Huang W B, *et al*. Dense regression network for video grounding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10284-10293

[62] Wang W N, Huang Y, Wang L. Language-driven temporal activity localization: a semantic matching reinforcement learning model[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 334-343

[63] He D L, Zhao X, Huang J Z, *et al*. Read, watch, and move: reinforcement learning for temporally grounding natural language descriptions in videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8393-8400

[64] Zhang Z, Zhao Z, Zhao Y, *et al*. Where does it exist: spatio-temporal video grounding for multi-form sentences[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10665-10674

[65] Wang M G, Cui D, Wu L F, *et al*. Weakly-supervised video object localization with attentive spatio-temporal correlation[J]. Pattern Recognition Letters, 2021, 145: 232-239

[66] Wang H, Deng C, Ma F, *et al*. Context modulated dynamic networks for actor and action video segmentation with language queries[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12152-12159

[67] Wang H, Deng C, Yan J C, *et al*. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3938-3947

[68] Ning K, Xie L X, Wu F, *et al*. Polar relative positional encoding for video-language segmentation[C] //Proceedings of International Joint Conferences on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2020: 948-954

[69] Sun X C, Jia H, Zhang Z, *et al*. Sound localization and separation in three-dimensional space using a single microphone with a metamaterial enclosure[OL]. [2021-08-10]. https://arxiv.org/pdf/1908.08160

[70] Lu R, Duan Z Y, Zhang C S. Listen and look: audio-visual matching assisted speech source separation[J]. IEEE Signal Processing Letters, 2018, 25(9): 1315-1319

[71] Zhou P, Yang W W, Chen W, *et al*. Modality attention for end-to-end audio-visual speech recognition[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2019: 6565-6569

[72] Zhang Y Y, Wang Z R, Du J. Deep fusion: an attention guided factorized bilinear pooling for audio-video emotion recognition[C] //Proceedings of the 2019 International Joint Conference on Neural Networks. Los Alamitos: IEEE Computer Society Press, 2019: 1-8

[73] Zhou Q L, Hui T R, Wang R, *et al*. Attentive excitation and aggregation for bilingual referring image segmentation[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(2): Article No.26

[74] Bordes A, Usunier N, Garcia-Durán A, *et al*. Translating

embeddings for modeling multi-relational data[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2013: 2787-2795

[75] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 1955-1961

[76] Schlichtkrull M, Kipf T N, Bloem P, *et al*. Modeling relational data with graph convolutional networks[C] //Proceedings of the European Semantic Web Conference. Heidelberg: Springer, 2018: 593-607

[77] Gardner M, Talukdar P P, Krishnamurthy J, *et al*. Incorporating vector space similarity in random walk inference over knowledge bases[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 397-406

[78] Abboud R, Ceylan I, Lukasiewicz T, *et al*. Boxe: a box embedding model for knowledge base completion[J]. Advances in Neural Information Processing Systems, 2020, 33: 9649-9661

[79] Sen P, Carvalho B W, Abdelaziz I, *et al*. Combining rules and embeddings via neuro-symbolic AI for knowledge base completion[OL]. [2021-08-10]. https://arxiv.org/pdf/2109.09566

[80] Chinchor N, Robinson P. MUC-7 named entity task definition[J]. Proceedings of the 7th Conference on Message Understanding, 1998, 29: 1-21

[81] Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition[OL]. [2021-08-10]. https://arxiv.org/pdf/1603.01360

[82] Ren X, He W Q, Qu M, *et al*. Label noise reduction in entity typing by heterogeneous partial-label embedding[C] //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1825-1834

[83] Ma Y, Cambria E, Gao S. Label embedding for zero-shot fine-grained named entity typing[C] //Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. New York: ACM Press, 2016: 171-180

[84] Ji H, Grishman R, Dang H T, *et al*. Overview of the TAC 2010 knowledge base population track[C] //Proceedings of the Third Text Analysis Conference. 2010, 3(2): 3-3

[85] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2017: 2608-2618

[86] Le P, Titov I. Improving entity linking by modeling latent relations between mentions[C] //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 1595-1604

[87] Adjali O, Besançon R, Ferret O, *et al*. Multimodal entity linking for tweets[C] //Proceedings of the European Conference on Information Retrieval. Heidelberg: Springer, 2020: 463-478

[88] Nguyen T H, Grishman R. Relation extraction: perspective from convolutional neural networks[C] //Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Stroudsburg: ACL, 2015: 39-48

[89] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures[C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 1105-1116

[90] Soares L B, FitzGerald N, Ling J, *et al*. Matching the blanks: distributional similarity for relation learning[C] //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 2895-2905

[91] Zhang Y H, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction[C] //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 2205-2215

[92] Wang Z, Zhang J W, Feng J L, *et al*. Knowledge graph and text jointly embedding[C] //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1591-1601

[93] Marin A, Holenstein R, Sarikaya R, *et al*. Learning phrase patterns for text classification using a knowledge graph and unlabeled data[C] //Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore: ISCA, 2014: 253-257

[94] Pezeshkpour P, Chen L, Singh S. Embedding multimodal relational data for knowledge base completion[C] //Proceedings of the Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 3208-3218

[95] Zhu Y K, Zhang C, Ré C, *et al*. Building a large-scale multimodal knowledge base system for answering visual queries[OL]. [2021-08-10]. https://arxiv.org/pdf/1507.05670

[96] Ma Y, Crook P A, Sarikaya R, *et al*. Knowledge graph inference for spoken dialog systems[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2015: 5346-5350

[97] Yang Z L, Tang J, Cohen W. Multi-modal Bayesian embeddings for learning social knowledge graphs[C] //Proceedings of the 25th International Joint Conference on Artificial Intelligence. Amsterdam: Elsevier, 2016: 2287-2293

[98] Wang M, Qi G, Wang H F, *et al*. Richpedia: a comprehensive multi-modal knowledge graph[C] //Proceedings of the Joint International Semantic Technology Conference. Heidelberg: Springer, 2019: 130-145

[99] Lin Y K, Liu Z Y, Sun M S, *et al*. Learning entity and relation embeddings for knowledge graph completion[J]. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, 29(1): 2181-2187

[100] Wang Z, Zhang J W, Feng J L, *et al*. Knowledge graph embedding by translating on hyperplanes[C] //Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 1112-1119

[101] Guan S P, Jin X L, Wang Y Z, *et al*. Shared embedding based neural networks for knowledge graph completion[C] //Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2018: 247-256

[102] Neelakantan A, Roth B, McCallum A. Compositional vector space models for knowledge base completion[C] //Proceedings

of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2015: 156-166

[103] Chen W H, Xiong W H, Yan X F, *et al*. Variational knowledge graph reasoning[C] //Proceedings of the NAACL-HLT. Stroudsburg: ACL, 2018: 1823-1832

[104] Xia C Y, Zhang C W, Yang T, *et al*. Multi-grained named entity recognition[C] //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1430-1440

[105] Hu A, Dou Z, Nie J Y, *et al*. Leveraging multi-token entities in document-level named entity recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7961-7968

[106] Li X Y, Feng J R, Meng Y X, *et al*. A unified MRC framework for named entity recognition[C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 5849-5859

[107] Zhao Y, Zhang A X, Xie R B, *et al*. Connecting embeddings for knowledge graph entity typing[C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 6419-6428

[108] Fang W, Zhang J W, Wang D L, *et al*. Entity disambiguation by knowledge and text jointly embedding[C] //Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg: ACL, 2016: 260-269

[109] Cao Y X, Huang L F, Ji H, *et al*. Bridge text and knowledge by learning multi-prototype entity mention embedding[C] //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1623-1633

[110] Fang Z, Cao Y N, Li Q, *et al*. Joint entity linking with deep reinforcement learning[C] //Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 438-447

[111] Chen S, Wang J P, Jiang F, *et al*. Improving entity linking by modeling latent entity type information[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7529-7537

[112] Zeng D J, Liu K, Lai S W, *et al*. Relation classification via convolutional deep neural network[C] //Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. New York: ACM Press, 2014: 2335-2344

[113] Zeng D J, Liu K, Chen Y B, *et al*. Distant supervision for relation extraction via piecewise convolutional neural networks[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1753-1762

[114] Jiang X, Wang Q, Li P, *et al*. Relation extraction with multi-instance multi-label convolutional neural networks[C] //Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. New York: ACM Press, 2016: 1471-1480

[115] Xu Y, Mou L L, Li G, *et al*. Classifying relations via long short term memory networks along shortest dependency paths[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1785-1794

[116] Cai R, Zhang X D, Wang H F. Bidirectional recurrent convolutional neural network for relation classification[C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 756-765

[117] Lin Q K, Liu J, Pan Y D, *et al*. Rule-enhanced iterative complementation for knowledge graph reasoning[J]. Information Sciences, 2021, 575: 66-79

[118] Liang Z W, Yang J N, Liu H, *et al*. HRER: a new bottom-up rule learning for knowledge graph completion[J]. Electronics, 2022, 11(6): 908

[119] Nian F D, Bao B K, Li T, *et al*. Multi-modal knowledge representation learning via webly-supervised relationships mining[C] //Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 411-419

[120] Zhang H W, Fang Q, Qian S S, *et al*. Multi-modal knowledge-aware event memory network for social media rumor detection[C] //Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 1942-1951

[121] Zhang Y Y, Qian S S, Fang Q, *et al*. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering[C] //Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 1089-1097

[122] Zhang Y Y, Fang Q, Qian S S, *et al*. Multi-modal multi-relational feature aggregation network for medical knowledge representation learning[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 3956-3965

[123] Chen L Y, Li Z, Wang Y J, *et al*. MMEA: entity alignment for multi-modal knowledge graph[C] //Proceedings of International Conference on Knowledge Science, Engineering and Management. Heidelberg: Springer, 2020: 134-147

[124] Xie R B, Liu Z Y, Luan H B, *et al*. Image-embodied knowledge representation learning[OL]. [2021-08-10]. https://arxiv.org/pdf/1609.07028

[125] Sun R, Cao X Z, Zhao Y, *et al*. Multi-modal knowledge graphs for recommender systems[C] //Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020: 1405-1414

[126] Wu Zhaohui, Jiang Xiaohong, Chen Huajun. Knowledge service: the development trend of Chinese medicine informatization in the era of big data[J]. Chinese Journal of Library and Information Science for Traditional Chinese Medicine, 2013, 37(2): 2-5(in Chinese)
(吴朝晖, 姜晓红, 陈华钧. 知识服务: 大数据时代下的中医药信息化发展趋势[J]. 中国中医药图书情报杂志, 2013, 37(2): 2-5)

[127] Xu K, Ba J L, Kiros R, *et al*. Show, attend and tell: neural image caption generation with visual attention[C] //Proceedings of the 32nd International Conference on Machine Learning. New York: ACM Press, 2015: 2048-2057

[128] Yang Z C, He X D, Gao J F, *et al*. Stacked attention networks for image question answering[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 21-29

[129] Lu J S, Yang J W, Batra D, *et al*. Hierarchical question-image co-attention for visual question answering[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 289-297

[130] Fukui A, Park D H, Yang D, *et al*. Multimodal compact bilinear pooling for visual question answering and visual grounding [OL]. [2021-08-10]. https://arxiv.org/pdf/1606.01847

[131] Yu Z, Yu J, Fan J P, *et al*. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 1839-1848

[132] Cong Y, Yuan J S, Luo J B. Towards scalable summarization of consumer videos via sparse dictionary selection[J]. IEEE Transactions on Multimedia, 2012, 14(1): 66-75

[133] Lu S Y, Wang Z Y, Mei T, *et al*. A bag-of-importance model with locality-constrained coding based feature learning for video summarization[J]. IEEE Transactions on Multimedia, 2014, 16(6): 1497-1509

[134] Lee Y J, Ghosh J, Grauman K. Discovering important people and objects for egocentric video summarization[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 1346-1353

[135] Yao T, Mei T, Rui Y. Highlight detection with pairwise deep ranking for first-person video summarization[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 982-990

[136] Zhang K, Chao W L, Sha F, *et al*. Video summarization with long short-term memory[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 766-782

[137] Sharghi A, Gong B Q, Shah M. Query-focused extractive video summarization[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 3-19

[138] Xu C S, Shao X, Maddage N C, *et al*. Automatic music video summarization based on audio-visual-text analysis and alignment[C] //Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005: 361-368

[139] Pan J Y, Yang H, Faloutsos C. MMSS: multi-modal story-oriented video summarization[C] //Proceedings of the 4th IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2004: 491-494

[140] Song Y L, Vallmitjana J, Stent A, *et al*. TVSum: summarizing web videos using titles[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 5179-5187

[141] Li H Z, Ellis J G, Ji H, *et al*. Event specific multimodal pattern mining for knowledge base construction[C] //Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 821-830

[142] Li H Z, Ellis J G, Zhang L, *et al*. Patternnet: visual pattern mining with deep neural network[C] //Proceedings of the ACM on International Conference on Multimedia Retrieval. New York: ACM Press, 2018: 291-299

[143] Zhang T T, Li H Z, Ji H, *et al*. Cross-document event coreference resolution based on cross-media features[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 201-206

[144] Lu D, Voss C, Tao F B, *et al*. Cross-media event extraction and recommendation[C] //Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Stroudsburg: ACL, 2016: 72-76

[145] Zhang T T, Whitehead S, Zhang H W, *et al*. Improving event extraction via multimodal integration[C] //Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 270-278

[146] Zhang W, Li H Z, Ngo C W, *et al*. Scalable visual instance mining with threads of features[C] //Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 297-306

[147] Chen T, Kaafar M A, Friedman A, *et al*. Is more always merrier? A deep dive into online social footprints[C] //Proceedings of the ACM Workshop on Workshop on Online Social Networks. New York: ACM Press, 2012: 67-72

[148] Jiang M, Cui P, Yuan N J, *et al*. Little is much: bridging cross-platform behaviors through overlapped crowds[C] //Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 13-19

[149] Yan M, Sang J T, Xu C S. Mining cross-network association for YouTube video promotion[C] //Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 557-566

[150] Lee H, Battle A, Raina R, *et al*. Efficient sparse coding algorithms[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 801-808

[151] van Der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

[152] Ou Weihua, Liu Bin, Zhou Yonghui, *et al*. Survey on the cross-modal retrieval research[J]. Journal of Guizhou Normal University: Natural Sciences, 2018, 36(2): 114-120(in Chinese) (欧卫华, 刘彬, 周永辉, 等. 跨模态检索研究综述[J]. 贵州师范大学学报: 自然科学版, 2018, 36(2): 114-120)

[153] Mahadevan V, Wong C W, Pereira J C, *et al*. Maximum covariance unfolding: manifold learning for bimodal data[C] //Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2011: 918-926

[154] Ngiam J, Khosla A, KimM, *et al*. Multimodal deep learning[C] //Proceedings of the 28th International Conference on Machine Learning. New York: ACM Press, 2011: 689-696

[155] Kumar S, Udupa R. Learning hash functions for cross-view similarity search[C] //Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence. Amsterdam: Elsevier, 2011: 1360-1365

[156] Blei D M, Jordan M I. Modeling annotated data[C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003: 127-134

[157] Zhow Pei, Chen Houjin, Yu Zekuan, *et al*. Review of cross-modality medical image prediction[J]. Acta Electronica

Sinica, 2019, 47(1): 220-226(in Chinese)
(周沛, 陈后金, 于泽宽, 等. 跨模态医学图像预测综述[J]. 电子学报, 2019, 47(1): 220-226)

[158] Shi Lixing, Zhang Jiwu. Optical molecular imaging and its application[J]. Chinese Journal of Medical Imaging Technology. 2008, 24(12): 2024-2026(in Chinese)
(石立兴, 张继武. 光学分子影像学及其应用[J]. 中国医学影像技术, 2008, 24(12): 2024-2026)

[159] Burgos N, Cardoso M J, Thielemans K, *et al*. Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies[J]. IEEE Transactions on Medical Imaging, 2014, 33(12): 2332-2341

[160] Wolterink J M, Dinkla A M, Savenije M H F, *et al*. Deep MR to CT synthesis using unpaired data[C] //Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging. Heidelberg: Springer, 2017: 14-23

[161] Zhu J Y, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2242-2251

[162] Bahrami K, Shi F, Zong X P, *et al*. Hierarchical reconstruction of 7T-like images from 3T MRI using multi-level CCA and group sparsity[C] //Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 659-666

[163] Shilling R Z, Robbie T Q, Bailloeul T, *et al*. A super-resolution framework for 3-D high-resolution and high-contrast imaging using 2-D multislice MRI[J]. IEEE Transactions on Medical Imaging, 2009, 28(5): 633-644

[164] Huang H, He H T, Fan X, *et al*. Super-resolution of human face image using canonical correlation analysis[J]. Pattern Recognition, 2010, 43(7): 2532-2543

[165] Bahrami K, Shi F, Rekik I, *et al*. Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features[J]. Deep Learning and Data Labeling for Medical Applications, 2016: 39-47

[166] Dong C, Loy C C, He K M, *et al*. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307

[167] Kulkarni K, Lohit S, Turaga P, *et al*. ReconNet: non-iterative reconstruction of images from compressively sensed measurements[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 449-458

[168] Li G H, Su H, Zhu W W. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks[OL]. [2021-08-10]. https://arxiv.org/pdf/1712.00733

[169] Wang M, Hong R C, Li G D, *et al*. Event driven web video summarization by tag localization and key-shot identification[J]. IEEE Transactions on Multimedia, 2012, 14(4): 975-985

[170] Yuan Y T, Mei T, Cui P, *et al*. Video summarization by learning deep side semantic embedding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(1): 226-237

[171] Chen L H, Zheng J B, Gao M, *et al*. TLRec: transfer learning for cross-domain recommendation[C] //Proceedings of IEEE International Conference on Big Knowledge. Los Alamitos: IEEE Computer Society Press, 2017: 167-172

[172] Man T, Shen H W, Jin X L, *et al*. Cross-domain recommendation: an Embedding and mapping approach[C] //Proceedings of the 26th International Joint Conferences on Artificial Intelligence. New York: ACM Press, 2017: 2464-2470

[173] Yan M, Sang J T, Xu C S. Unified YouTube video recommendation via cross-network collaboration[C] //Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. New York: ACM Press, 2015: 19-26

[174] Lu Z Q, Zhong E H, Zhao L L, *et al*. Selective transfer learning for cross domain recommendation[C] //Proceedings of the SIAM International Conference on Data Mining. Bethesda: SIAM, 2013: 641-649

[175] Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder[C] //Proceedings of the 22nd ACM international conference on Multimedia. New York: ACM Press, 2014: 7-16

[176] Ding G G, Guo Y C, Zhou J L. Collective matrix factorization hashing for multimodal data[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. Los Alamitos: IEEE Computer Society Press, 2014: 2083-2080

[177] Wang D X, Cui P, Ou M D, *et al*. Learning compact hash codes for multimodal representations using orthogonal deep structure[J]. IEEE Transactions on Multimedia, 2015, 17(9): 1404-1416

[178] Song J K, Yang Y, Yang Y, *et al*. Inter-media hashing for large-scale retrieval from heterogeneous data sources[C] //Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2013: 785-796

[179] Wu F, Yu Z, Yang Y, *et al*. Sparse multi-modal hashing[J]. IEEE Transactions on Multimedia, 2014, 16(2): 427-439

[180] Wang W, Ooi B C, Yang X Y, *et al*. Effective multi-modal retrieval based on stacked auto-encoders[J]. Proceedings of the VLDB Endowment, 2014, 7(8): 649-660

[181] Liao R J, Zhu J, Qin Z C. Nonparametric Bayesian upstream supervised multi-modal topic models[C] //Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2014: 493-502

[182] Wang Y F, Wu F, Song J, *et al*. Multi-modal mutual topic reinforce modeling for cross-media retrieval[C] //Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 307-316

[183] Huang Y, Long Y, Wang L. Few-shot image and sentence matching via gated visual-semantic embedding[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8489-8496

[184] Nie D, Cao X H, Gao Y Z, *et al*. Estimating CT image from MRI data using 3D fully convolutional networks[C] //Proceedings of the Deep Learning and Data Labeling for Medical Applications. Heidelberg: Springer, 2016: 170-178

[185] Shin H C, Roth H R, Gao M C, *et al*. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J].

IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298

[186] Han X. MR‐based synthetic CT generation using a deep convolutional neural network method[J]. Medical Physics, 2017, 44(4): 1408-1419

[187] Bi L, Kim J, Kumar A, *et al*. Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs)[C] //Proceedings of the Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment. Heidelberg: Springer, 2017: 43-51

[188] Sun C, Myers A, Vondrick C, *et al*. VideoBERT: a joint model for video and language representation learning[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7463-7472

[189] Li L H, Yatskar M, Yin D, *et al*. Visualbert: a simple and performant baseline for vision and language[OL]. [2021-08-10]. https://arxiv.org/pdf/1908.03557

[190] Su W J, Zhu X Z, Cao Y, *et al*. VL-BERT: pre-training of generic visual-linguistic representations[OL]. [2021-08-10]. https://arxiv.org/pdf/1908.08530

[191] Lu J, Batra D, Parikh D, *et al*. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[OL]. [2021-08-10]. https://arxiv.org/pdf/1908.02265

[192] Zhen L L, Hu P, Peng X, *et al*. Deep multimodal transfer learning for cross-modal retrieval[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2): 798-810

[193] Liu X W, Li Z, Wang J, *et al*. Cross-modal zero-shot hashing[C] //Proceedings of the IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2019: 449-458