智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 以图像视频为中心的跨媒体分析与推理

黄庆明， 王树徽， 许倩倩， 李亮， 蒋树强

引用本文:
黄庆明, 王树徽, 许倩倩, 等. 以图像视频为中心的跨媒体分析与推理[J]. 智能系统学报, 2021, 16(5): 835–848.
HUANG Qingming, WANG Shuhui, XU Qianqian, et al. Image video centered cross–media analysis and reasoning[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(5): 835–848.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202105042

**您可能感兴趣的其他文章**

知识图谱的推荐系统综述
Review of recommendation systems based on knowledge graph
智能系统学报. 2019, 14(2): 207–216 https://dx.doi.org/10.11992/tis.201805001

基于用户意图理解的社交网络跨媒体搜索与挖掘
Social network cross–media searching and mining based on user intention
智能系统学报. 2017, 12(6): 761–769 https://dx.doi.org/10.11992/tis.201706075

基于文本扩展模型的网络视频聚类方法
Web video clustering method based on an extended text model
智能系统学报. 2017, 12(6): 799–805 https://dx.doi.org/10.11992/tis.201706036

基于时空域联合建模的领域知识演化脉络分析
Evolutionary path mining of domain knowledge by joint modeling in space–time domain
智能系统学报. 2017, 12(5): 735–744 https://dx.doi.org/10.11992/tis.201706023

一种基于OCC模型的文本情感挖掘方法
OCC–model–based text–emotion mining method
智能系统学报. 2017, 12(5): 645–652 https://dx.doi.org/10.11992/tis.201312032

一种多模态融合的网络视频相关性度量方法
A multi–modal fusion approach for measuring web video relatedness
智能系统学报. 2016, 11(3): 359–365 https://dx.doi.org/10.11992/tis.201603040

微信公众平台

关注微信公众号，获取更多资讯信息

## 吴文俊人工智能自然科学奖一等奖

成果名称：图像视频的多尺度表征与语义映射
获 奖 人：黄庆明、王树徽、许倩倩、李亮、蒋树强
完成单位：中国科学院大学、中国科学院计算技术研究所

### 黄庆明

教授，长期从事人工智能相关领域的研究，在媒体内容理解、图像语义检索、群体智能标注等方面开展了一系列创新性工作。主持了科技创新2030-“新一代人工智能”重大项目、国家自然科学基金重点项目和重点国际合作项目、国家973计划课题、863课题、科学院前沿科学研究重点计划等一系列国家和省部级课题的研究。在国内外权威期刊和重要国际会议上发表学术论文400余篇（近五年200余篇），其中IEEE/ACM汇刊论文和CCF-A类会议论文160余篇（近五年110余篇）。Google Scholar引用13000余次，H指数是55，相关研究成果多次获得省部级奖励，合作出版专著一部，申请和授权国内外发明专利40项。担任AVS标准工作组数字媒体内容描述专题组联合组长和IEEE 1857.6数字媒体内容描述标准的联合制定人、中国计算机学会多媒体技术专业委员会副主任、北京市图像图形学会副理事长，是IEEE会士（IEEE Fellow），IEEE电路与系统协会北京分会主席，CCF会士、理事，担任IEEE Trans. on CSVT、SCIENCE CHINA Information Sciences、自动化学报、中国图象图形学报编委，近年来10多次承担本领域主流学术会议的组织和评审工作，担任ACM Multimedia、ICCV、CVPR、ICMR、ICME、PCM、PSIVT、BigMM、ChinaMM等国内外学术会议的大会主席、程序委员会主席、领域主席、专题主席、评奖主席或程序委员会委员等。

## 团队简介

分别来自中国科学院大学和中国科学院计算技术研究所，是国内人工智能领域一支优秀、活跃的团队，通过多年的科研探索与技术积累，在理论、技术、应用等方面均取得很好的成果。团队在图像视频的多尺度表征与语义映射方面取得一系列创新成果。针对传统视觉表征存在的困难，提出视觉内容的多尺度表征方法，实现了通用视觉符号表征与层级渐进融合。针对海量网络图像视频标注匮乏及部分标签低质冲突等问题，建立数据-语义场同型化映射框架。针对视觉多义性、语义多态层次化和时空实体信息建模难等问题，提出语义结构渐进学习方法，实现了从粗到细粒度语义的渐进视觉语义理解。参与制定了IEEE 1857.6视频内容描述国际标准，获得了顶级国际会议的技术挑战赛冠军，研究成果应用于网络内容监测与服务等场景。团队在国内外权威期刊和重要国际会议上发表学术论文400余篇（近五年200余篇），其中IEEE/ACM汇刊论文和CCF-A类会议论文160余篇（近五年110余篇）。申请和授权国内外发明专利40余项（近5年20余项）。承担了科技创新2030-“新一代人工智能”重大项目、国家自然科学基金重点项目和重点国际合作项目、国家973计划课题、科学院前沿科学研究重点计划等国家和省部级课题的研究。

**DOI**: 10.11992/tis.202105042
**网络出版地址**: https://kns.cnki.net/kcms/detail/23.1538.TP.20210726.1059.002.html

# 以图像视频为中心的跨媒体分析与推理

黄庆明[1,2], 王树徽[2], 许倩倩[2], 李亮[2], 蒋树强[2]

（1. 中国科学院大学 计算机科学与技术学院, 北京 100049; 2. 中国科学院计算技术研究所 智能信息处理实验室, 北京 100190）

**摘　要:** 如何跨越从跨媒体数据到跨媒体知识所面临的“异构鸿沟”和“语义鸿沟”, 对体量巨大的跨媒体数据进行有效管理与利用, 是发展新一代人工智能亟待突破的瓶颈问题。针对以图像视频为代表的海量网络跨媒体内容, 借鉴人类感知与认知机理, 本文对跨媒体内容统一表征与符号化表征、跨媒体深度关联理解、类人跨媒体智能推理等关键技术开展研究。基于上述关键技术, 着力于解决发展新一代人工智能的知识匮乏共性难题, 开展大规模跨媒体知识图谱的构建及人机协同标注技术研究, 为跨媒体感知进阶到认知提供关键支撑, 进一步为跨媒体理解、检索、内容转换生成等跨媒体内容管理与服务热点应用领域提供了可行思路。

**关键词:** 跨媒体; 图像视频; 统一表征; 关联理解; 可解释推理; 人机协同; 知识图谱; 内容管理与服务
**中图分类号:** TP37　　**文献标志码:** A　　**文章编号:** 1673−4785(2021)05−0835−15

**中文引用格式: 黄庆明, 王树徽, 许倩倩, 等. 以图像视频为中心的跨媒体分析与推理 [J]. 智能系统学报, 2021, 16(5): 835–849.**
**英文引用格式: HUANG Qingming, WANG Shuhui, XU Qianqian, et al. Image video centered cross-media analysis and reasoning[J]. CAAI transactions on intelligent systems, 2021, 16(5): 835–849.**

# Image video centered cross-media analysis and reasoning

HUANG Qingming[1,2], WANG Shuhui[2], XU Qianqian[2], LI Liang[2], JIANG Shuqiang[2]

(1. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China; 2. Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: How to surpass the heterogeneity gap and semantic gap between the cross-media content and cross-media knowledge, and how to manage and utilize the huge amount of cross-media data effectively are urgent bottleneck problems of developing a new generation of artificial intelligence. Aiming at massive online cross-media content represented by image video and by referring to human perception and cognition mechanisms, this paper undertakes studies on such key technologies as unified representation and symbolic representation of cross-media content, deep correlative understanding of cross-media and human-like cross-media intelligent reasoning. Based on the above technologies, this paper focuses on solving the common problem of knowledge shortage in the development of a new generation of artificial intelligence and carries out a research on the construction of large-scale cross-media knowledge graph and the human-machine cooperation based labeling technology, to provide strong support for the advancement from cross-media perception to cognition and further provide feasible solutions towards cross-media content management and popular service applications, e.g., cross-media content understanding, retrieval, content transformation and generation, etc.

**Keywords**: cross-media; image video; unified representation; correlative understanding; explainable reasoning; Human-computer collaboration; knowledge graph; content management and service

人类通过多模态协同的方式对世界进行感知与认知。视觉是生物获取环境信息的一种主要方式, Hubel 和 Wiesel 通过生物学实验发现, 高级生物通过不同复杂度的组织细胞对视觉信息进行逐步提取与整合, 实现视觉场景解构与结构化感知[1]。受上述研究启发, Marr[2] 建立了完整的、可实现的视觉计算理论框架。在语言方面, Chomsky[3] 提出了研究人类语言机能的研究范式, 并为计算机模拟语言生成奠定了理论基础。心理学实验表明, 视觉与听觉之间存在复杂的相互作用关系, 即麦格克效应[4]。人类大脑的信息处理机制以图、文、声等多模态协同方式进行。基于人脑

**收稿日期:** 2021−05−27.　**网络出版日期:** 2021−07−26.
**基金项目:** 科技创新 2030-新一代人工智能重大项目 (2018AAA0102000); 国家自然科学基金项目 (62022083, 61976202, 61771457, 61732007).
**通信作者:** 王树徽. E-mail: wangshuhui@ict.ac.cn.

强大的多模态信息抽象能力，人类的认知过程体现为将多模态信息进行层级渐进的符号概念转化和符号推理。物理符号系统假说认为智能是用计算机和心理学方法进行宏观的人脑功能模拟[5]。信息加工心理学将心理过程看作是符号序列的信息加工过程[6]。心物同形论认为认知是对物理现实到人类知觉现实的复杂同型转换过程，而这一理论被大量借鉴到视觉计算领域[7]。最新的人脑结构研究表明，人类大脑当中通过各种结构的连接组成功能区域来实现从连接到认知的转换[8]。上述感知与认知理论框架是发展人工智能理论与方法研究的重要依据。

随着信息技术的不断发展，人类社会已全面进入网络互联时代。网络用户群体数量的不断增长，以及手机、摄像头等具有强大环境感知能力的终端设备的不断普及，使得对物理世界与网络世界的记录越来越翔实，并呈现跨模态、跨数据源的复杂关联特性，即不同模态、不同来源的图像、视频、文本、音频通过多源互补方式刻画同一对象与事件信息。如何对体量巨大的跨媒体数据进行有效管理与利用，是应对重大变化的信息新环境的迫切需要，也是发展新一代人工智能亟待突破的瓶颈问题。

在海量跨媒体内容当中，超过 90% 的内容通过图像、视频等视觉方式呈现，以图像视频为中心的跨媒体分析推理技术近年来引发了学术界和工业界的广泛关注和深入研究[9]。跨媒体分析推理的研究目标是在对视觉、语言等不同模态信息的语义贯通理解基础上，实现“举一反三”的类人智能推理，是促使人工智能从感知进阶到认知并走向类人智能的关键，也是信息科学、计算科学、神经科学、认知科学交叉的国际前沿科学问题。传统跨媒体处理方式是通过单一模态分析方法，如图像视频处理，自然语言处理，语音识别等，对特定模态通道的语义进行独立分析，然后进行结果融合。这一方式导致了对跨媒体内容的语义理解局限粗浅，难以从其中获取充分全面的知识，无法应对开放复杂的跨媒体内容演化和多元化的跨媒体应用场景。近年来由于直播、短视频推荐等新应用的出现和流行，数据的爆炸增长和内容的良莠不齐对网络跨媒体数据管理与内容服务造成了巨大挑战。

针对以图像视频为代表的海量网络跨媒体内容，借鉴高级生物的感知与认知机理，团队对跨媒体内容统一表征与符号化表征、跨媒体深度关联理解、类人跨媒体智能推理等关键技术开展研究；基于上述关键技术，研究团队着力于解决发展新一代人工智能的知识匮乏共性难题，开展大规模跨媒体知识图谱的构建及人机协同标注技术研究，为跨媒体感知进阶到认知建立理论支撑，进一步为多模态分类、跨媒体检索、事件发现与预测等跨媒体内容管理与服务热点应用领域提供了可行思路。

## 1　研究总体框架

跨媒体由不同来源、不同模态的信息以交织融合的方式产生与演化。跨媒体不同模态信息的异构性为跨媒体统一计算带来了“异构鸿沟”难题。另一方面，相比于传统单一媒体，跨媒体内容中蕴含更为丰富的语义信息，然而跨媒体数据到语义知识之间存在较大的“语义鸿沟”，导致对跨媒体理解的粗浅和片面。针对海量跨媒体的形式异构、内容复杂、动态演化等特点，研究组针对以图像视频为中心的跨媒体分析推理理论与方法开展了深入研究，建立了跨媒体分析推理研究的通用框架与范式，如图 1 所示。具体而言，研究组通过符号化与统一表征、深度关联理解、类人智能推理等方式构建了从数据到知识的归纳通路，通过粗粒度图谱构建、细粒度图谱构建和人机协同知识标注平台实现从知识到数据的演绎通路，最后在跨媒体归纳推理和演绎推理技术框架上，构建跨媒体分析推理引擎技术原型系统，为内容管理与服务提供技术支撑。

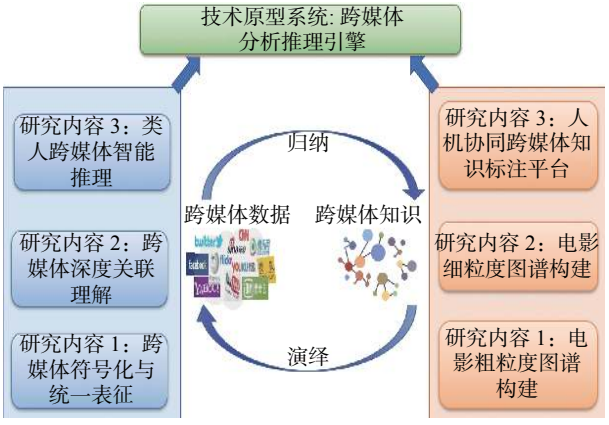



**图 1　以图像视频为中心的跨媒体分析推理技术框架**

**Fig. 1　Cross-media analysis and reasoning framework centered on images and videos**

## 2　跨媒体统一表征与符号化表征

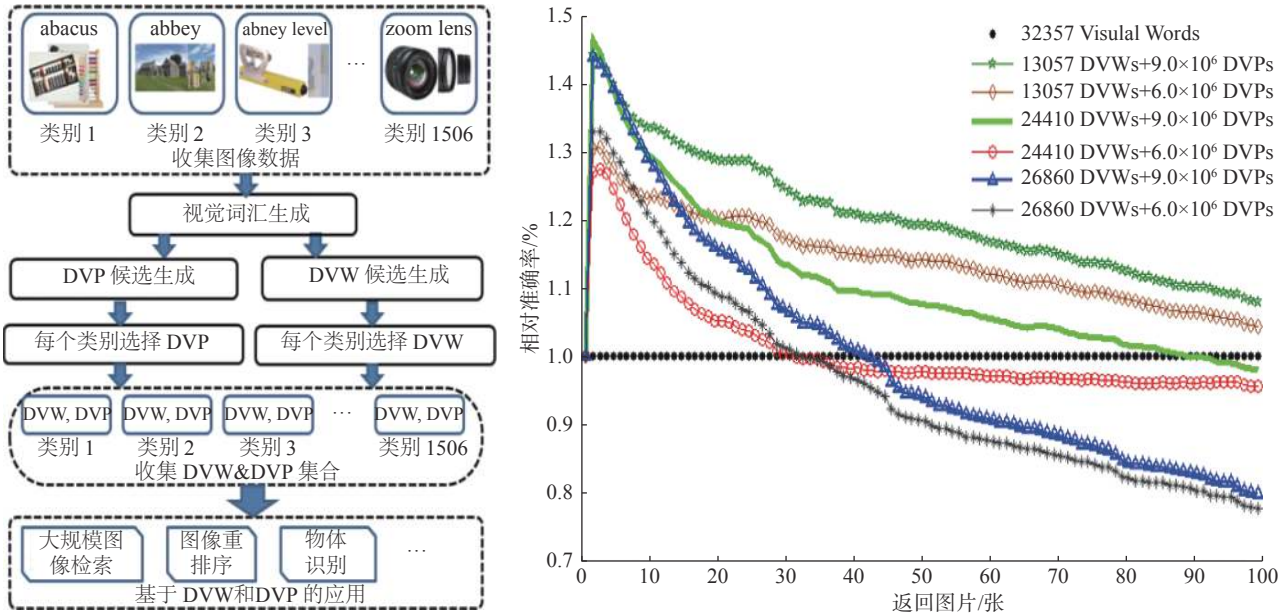
网络跨媒体内容中包含大量的视觉及图文联合表达信息，对这些信息的统一表征是实现跨媒体统一计算的基础性问题，而将跨媒体信息进行符号化转换则是支撑跨媒体推理和认知的关键。然而，尽管近年来图像分类与检测技术取得了一定进展，但对跨媒体当中的视觉信息的符号化转

换精度仍处在较低水平。进一步深入分析, 针对视觉模态与文本模态的符号化表征方式之间存在的显著差异, 也为跨媒体统一计算与符号化表征造成了本质困难。

为此, 研究组近年来开展了如下的研究工作。针对视觉内容的局部、浅层表征在描述性、显著性和判别性不足等难题, 借鉴生物视觉感知理论, 对视觉表征进行视觉空间扩展和纵向特征层级融合。引入视觉内容上下文, 模拟生物神经元信息传递规律, 建模视觉基元间相关性和信息传递关系, 建立了视觉内容的通用符号表征体系。模拟生物视觉层级信息传递过程, 提出视觉层级表征的递进式融合方法, 实现了动态复杂时空环境下的多尺度视觉目标高效聚焦与跟踪。针对图像和文本内容的异构性问题, 提出跨媒体符号化统一表示及调和统一表示方法, 实现了图文模态当中从局部到整体的内容语义对齐表示。

### 2.1　多尺度显著性视觉表征

针对视觉内容匹配与检索的需求, 提出了描述性视觉单词和视觉短语的通用提取算法框架。对视觉单词的空间近邻关系统计矩阵上的随机游走稳态结果进行挖掘, 得到视觉场景中语义显著的单词集合与频繁共现的视觉词对。如图 2 所示, 本文方法通过视觉单词及其多尺度组合刻画视觉物体和场景, 具有可比拟文本词和短语的强描述能力。




(a) 描述性视觉单词与短语的提取方法过程　　(b) 本文方法进行图像检索实验的比较结果

**图 2　描述性视觉单词和视觉短语的通用提取算法框架**

**Fig. 2　Descriptive visual words visual phrases generation framework**

所提方法可用于检索排序等任务, 实现了视觉单词表征能力的阶跃, 与传统视觉单词相比具有显著精度优势, 检索的平均精度均值 (mean average precision, MAP) 相对提高 19.5%, 重排序精度相对提高 12.4%, 处理速度快 11 倍以上[10]。

### 2.2　视觉目标与多模态符号表征

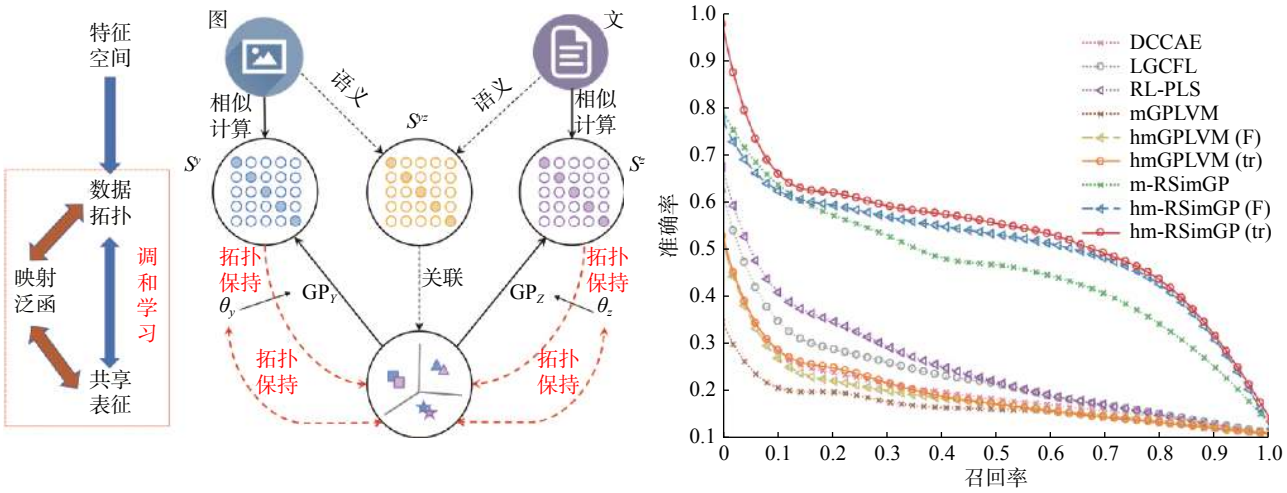
针对复杂的图文内容, 提出一种图像和文本的多粒度符号信息建模表示方法, 将图像利用物体检测技术提取到包含显著物体的图像区域并编码成视觉符号表征, 实现图像–文本的联合自注意统一表征, 并分别将图像和文本映射到隐含联合表示空间。使用 Wordpiece Token 得到文本词汇、短语、句子符号表示, 并使用自注意机制分别学习图像和文本内小块的关联, 进一步聚合小块的信息得到图像和文本的隐含空间表示。其中建模自注意机制的层包括多头自注意力子层和对每个位置的前馈网络子层。使用难例挖掘配合优化三元组损失和体现数据高阶结构特性的三角损失学习图像和文本到隐含空间的映射函数。基于该算法进行了图像文本匹配检索的实验, 在 FLICKR30K 数据集上性能超过当时最佳算法, 在 MSCOCO 数据集上性能和最优算法相当, 并且检索速度更快[11]。

### 2.3　跨媒体调和学习与统一表征

跨媒体数据对象之间存在复杂的关联关系。考虑到异构媒体数据内容和结构的复杂关联, 本项目突破传统数据拟合学习的桎梏, 提出了一种基于高斯过程隐变量模型的非线性关联学习框架, 通过跨模态数据的相似度信息来表示数据间的拓

扑结构，并通过设计合理的正则约束，使得跨模态观测空间的拓扑关系能够被有效通过隐含子空间进行保持，从而实现了拓扑保持的跨模态表示学习；此外，所提方法还能够利用跨媒体对象间的语义关系作为先验知识来指导跨模态表示的学习，实现了异构数据间的有效关联建模；在海量跨模态数据库上的多视角分类和跨模态检索等任务上的算法评测结果表明所提方法具有较好的性能表现。

如图 3 所示，进一步，通过深入挖掘跨模态对象间的内在联系，对跨模态数据间的不同关联结构构建了一种调和约束，以隐含一致表示的拓扑表示为桥梁，建立了跨模态高斯隐变量模型的参数空间，更好地实现了异构信息间的共享机制，在 4 个标准数据集上进行的大量实验结果表明了所提非线性非参数跨模态实体关联方法相比于传统线性、参数化及深层非线性的跨模态统一表征方法具有更好的模型容量，能够更有效和精确地对跨模态数据对象的深层高阶非线性关系进行刻画[12]。




(a) 跨媒体调和统一表征学习整体思路　　(b) 所提方法在进行跨模态搜索时的结果曲线

**图 3　基于高斯过程隐变量模型的跨媒体调和学习**

**Fig. 3　Harmonized multimodal learning with gaussian process latent variable models**

## 3　跨媒体深度关联理解

与传统单模态内容理解方式不同，跨媒体依赖于对不同模态内容的综合理解。同时，由于模态互补性、异构性和信息不均衡性，针对特定模态的独立语义分析容易造成对跨媒体理解的粗浅、片面等现象。为此，研究组针对全局、层次化、细粒度的跨媒体语义理解和语义保持的内容转换生成等跨媒体深度语义关联理解技术开展系统深入的研究，目标是从复杂跨媒体内容中获取全面、深入的语义信息，并进一步实现跨模态内容的演绎生成，这也是使机器具备类人跨模态信息转换的重要技术。

### 3.1　海量高维数据场的全局语义映射

针对海量网络图像视频缺乏高质量语义标注及部分标签低质冲突等难题，借鉴格式塔心物同型论，在图像视频数据上构建数据场 (特征关联拓扑图)，并在数据的不完全语义信息上构建语义场，进而建立数据拓扑结构与语义向量结构之间的数据–语义场同型化映射框架。

从散度场角度，基于数据场多层拓扑信息扩散建模思想，提出了可扩展半监督诱导式多核学习方法，融合多种互补视觉特征提高学习能力。通过多核融合的近似近邻搜索来确定有信息量的紧凑无标注训练数据子集。通过无标注域的条件期望一致性实现在无标注数据的语义标签扩散，学习过程快速有效。所提方法具有良好的理论收敛特性，相比于传统方法具有更低的算法复杂度，在处理图像分类和个性化图像重排序时具有更好的性能表现，需要的用户交互更少。所提方法[13] 是利用无标注网络数据进行半 (弱) 监督视觉学习的早期工作之一。

从旋度场角度，提出了针对海量无序标注的群体语义修正模型，从旋度场角度对标注不一致性进行建模和因子化。基于成对比较的随机图霍奇排序，构建 Erdös-Rényi 随机图和随机正规图逼近，从不完整及不平衡的数据、视频的质量分值和用户判断不一致性中得到成对比较数据的霍奇分解，实现了群体不一致标注的精确修正。在不同的群体标注数据量下，证明了两种随机图设计都具有良好的采样近似特性。在大规模直播视频质量评估任务上验证了所提方法的有效性，该方法也适用于标注质量难以控制的网络众包信息处理。该项研究为群体语义标注的组织实施方式提供了指导，为将群体智能引入到图像视频理解领域提供了理论保证和关键技术[14]。

### 3.2　层次化、细粒度语义理解

借鉴人类的层次化、概念化、实体化渐进认知过程, 建模层次化语义关联结构, 建立面向极多类数据的层次化视觉特征与层次分类联合学习框架, 实现了从粗粒度语义到细粒度语义的渐进图像视频语义理解。

1) 提出了局部到整体的视觉语义层次化表示方法, 对视觉表现和语义概念间的概率隶属关系进行建模。通过组稀疏编码, 获得更加准确的图像层面的稀疏表示, 利用混合范数正则化学习具有结构稀疏特性的判别性视觉概念隶属度分布。在不同概念层级上进行投影和距离计算, 实现了一种新的图像语义度量。所提视觉语义描述符合人类语义理解习惯, 具有天然的可解释性, 适用于包括大规模语义图像搜索, 图像标注和语义图像重排序等在内的主流视觉应用场景[15]。

2) 提出了层次化语义类别指导的视觉特征学习方法。对于在层次化类别结构当中的每个中间节点, 同时学习一个判别性字典和分类模型, 不同层次上的字典通过挖掘不同粒度的判别性视觉特性学习得到。低层细粒度类别的字典集成了其祖先节点的字典, 在低层的类别通过所构建的字典中的多尺度视觉共同描述。在主流大规模极多类数据集上的实验表明所提方法在当时取得了最高的识别精度, 在处理极多类分类问题时达到更好的精度–效率折衷。研究成果[16]被 CVPR 大会主席、IJCV 副主编等多名国际重量级学者重点引用。

### 3.3　跨媒体内容转换生成

跨媒体内容转换生成是在不同模态的深入理解基础上, 在语义保持的约束下进行不同模态的内容转换生成, 如从视觉到文本的内容转换生成(图像/视频概述[17-18]), 从多模态到视觉的内容转换生成 (图像内容生成[19])。

在从视觉到文本的内容转换方面, 视频自然语言描述任务是在对视觉内容理解的基础上, 生成对视觉内容的自然语言描述, 是一种更深层次的跨媒体内容理解任务。当前的方法通常是引入跨模态注意力机制, 动态地整合输入信息进行建模, 最终生成与源输入语义对应的语言描述或者图像内容。然而, 现有方法忽略了不同模态内容的结构信息, 从而导致了语言描述结果不精确且效率低下等难题。

为克服这一问题, 在自然语言描述方面, 利用句法信息的先验知识来指导视频描述中多模态特征的融合, 可设计一种语法指导的分层注意力模型。引入了一种分层注意力机制, 同时利用语义和句法线索来整合 2D 图像特征、3D 运动特征和句子上下文特征。该模型包含一个内容注意力模块和一个句法注意力模块, 两者分别从时序的维度和模态的维度对上述特征进行聚合。如图 4 所示, 整个内容转换过程包含 3 个部分: 特征编码器、分层注意力模块、描述生成器。在大规模公开数据集上的实验表明, 合理利用 2D 图像和 3D 运动特征有利于视觉单词的生成, 而有效利用句子上下文特征有利于非视觉单词的生成[17]。

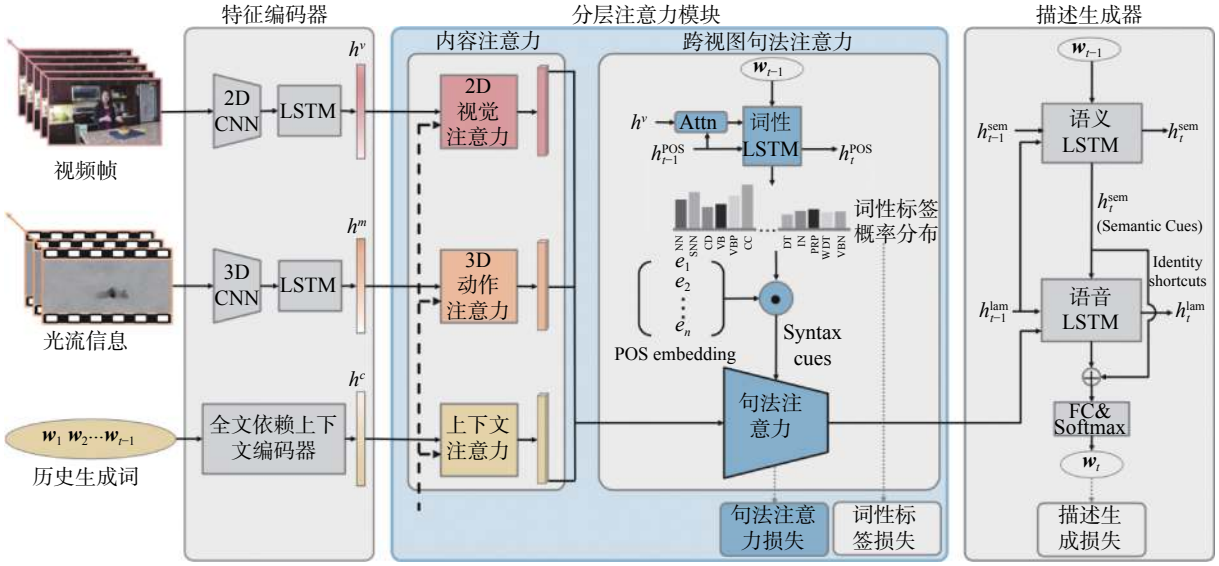



**图 4　句法指导的视频概述生成模型框架**

**Fig. 4　Syntax-guided video caption generation framework**

此外, 在视频描述任务中, 具有最好性能的处理方式为基于注意力的模型, 它们通过将显著的视觉成分和句子进行准确关联。然而, 现有的研究遵循一般化的视觉信息处理过程, 即在等间隔采样的视频帧上进行视觉表观特征提取和运动信息特征提取, 从而不可避免地遇到视觉信息表

征冗余，对内容噪声敏感和不必要的运算开销等难题。

为此，提出一个即插即用的选帧网络 PickNet，在视频概述过程对包含更多信息量的视频帧进行选择。如图 5 所示，基于标准的编码器–解码器结构，设计了一种基于强化学习的序列化网络训练过程，其中每次帧选择的奖励被设计为最大化视觉多样性和最小化句子生成与真实句子之间的差异性。得到正向奖励的帧选择候选结果将被选择，并且编码器–解码器的隐含表示将被更新用于未来的处理过程。这个过程一直持续直到整个视频序列处理完毕。相应地，一个紧凑的帧子集能够被有效选择来对视觉信息进行表征并且无性能损失地进行视频概述。实验结果表明所提方法在仅仅选择 6~8 帧的情况下就能获得与传统方法相近的视频概述结果[18]。

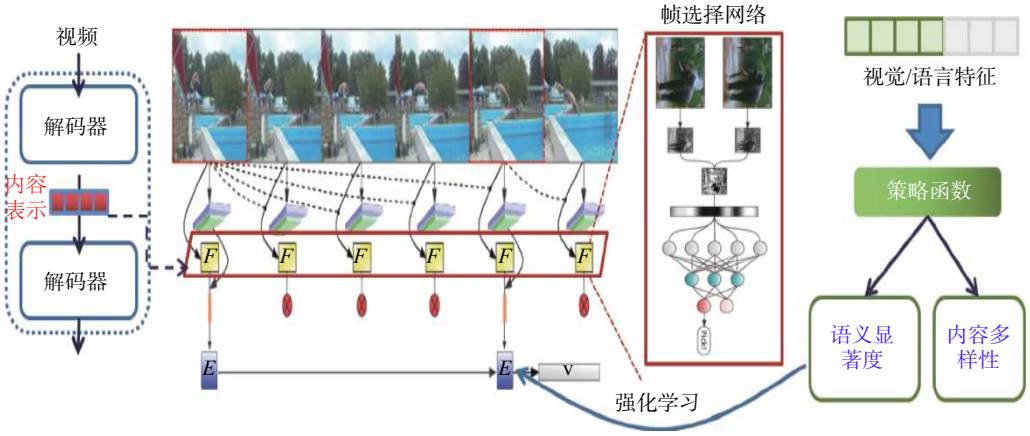



**图 5　基于帧选择的高效视频概述方法**

**Fig. 5　Efficient video captioning based on PickNet**

在从文本到视觉的内容转换生成方面，其核心难点问题在于文本信息的信息量远远小于视觉模态的信息量，且文本与视觉模态的异构性也为这一任务带来了更大的挑战。

针对上述问题，提出一种增量推理的生成对抗网络，通过推理目标图像中视觉的增量和指令中知识信息的增量之间的一致性，来保证生成结果的正确性。如图 6 所示，该模型包括 3 个模块，分别是指令编码器、图像生成器和推理判别器。对于指令编码器，分别采用单词级别和指令级别的 GRU 网络作为编码器去对当前指令和历史指令进行分析，得到知识信息的表示。对于生成器，采用多层感知机将上述的表示投影到语义增量的特征图，然后将其与原始图像的特征图进行合成。之后这个合成的表示和原始图像通过图像解码器来生成目标图像，其中原始图像作为一种具有指示作用的辅助。最后，利用推理编码器来推理历史视觉信息，保持视觉增量和当前指令的一致性。其中视觉增量从原始图像和目标图像的特征图中提取。通过采用多模态条件判别器对上述的一致性进行衡量，保证了生成图像的逻辑合理性。通过将图像中的视觉增量和用户意图的语义增量进行连接，解决了文本与图片交互生成的问题[19]。

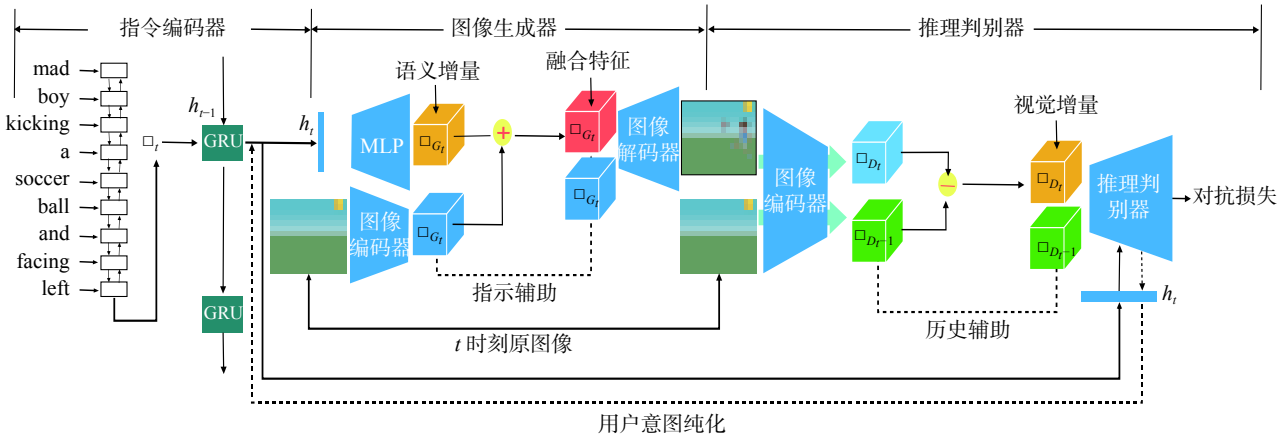



**图 6　基于增量推理的图片内容生成对抗网络**

**Fig. 6　Generative adversarial network with linguistic instruction by increment reasoning**

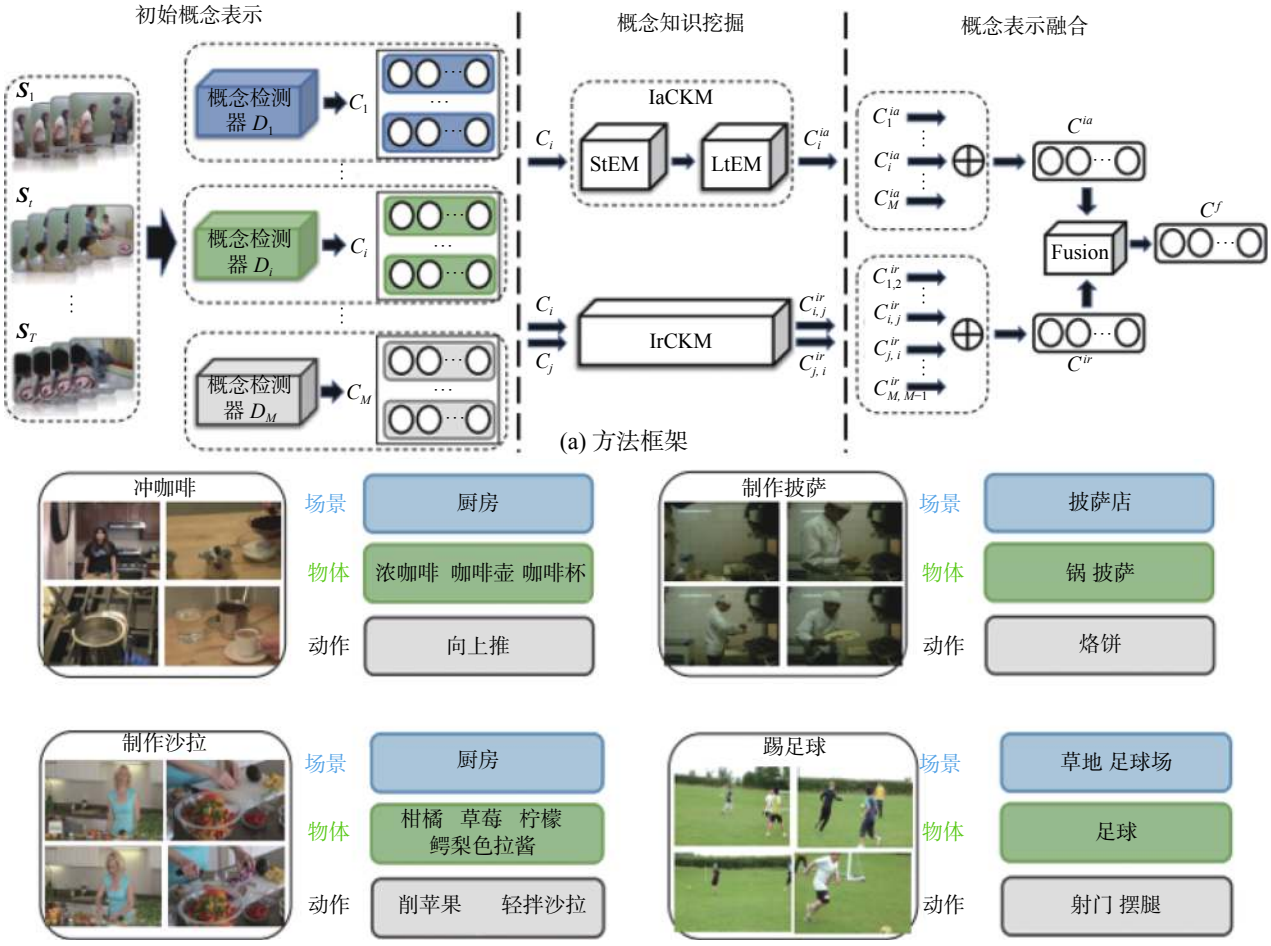
## 4　类人跨媒体智能推理

推理的本质是基于某些前提条件找到结论的过程, 是人类有别于其他生物的高级思维能力。从计算与人工智能角度而言, 实现机器推理的关键是在现有数据表征结构 (符号、向量、矩阵) 基础上, 对数据对象之间的关联性进行人类可解释的挖掘、补全与推断。然而, 实现机器的类人推理目前仍存在较大困难。首先, 机器对多模态信息的符号化转换未能达到人类水平, 从而为后续的推理任务带来阻碍。进一步来看, 机器对信息的处理方式与生物神经系统存在本质区别, 机器以单向的前馈或反馈机制为主要方式, 而人类认知系统中的信息处理通路与交互反馈更加复杂。最后, 人类所具有的举一反三甚至直觉顿悟等能力目前机器尚不具备。

基于上述现状, 研究组近年来对类人跨媒体智能推理技术开展了初步的研究。具体而言, 跨媒体机器推理的“类人”特性可体现为机器推理的可解释性、推理过程的人机可协同性以及主动交互性等。其中, 可解释性推理着力于解决现有数据驱动方法机理难以解释且难以泛化的固有缺陷, 实现表征、组件和结果的可解释性, 提高人机互信水平; 人机协同群智推理是在海量用户产生噪声知识的基础上, 通过数据学习和知识指导实现潜在实体对象关联的推断与补全, 提高跨媒体知识量和稠密度; 主动交互式推理是在人机之间充分的多模态信息交换基础上, 完成各类语义标注、内容转换生成、事件预测等任务。

### 4.1　可解释跨媒体推理

目前, 大多数视频事件分析算法都是基于端到端的深度模型, 具有黑盒属性 (black-box), 阻碍了算法的实际应用。一种可解释性视频事件分析的方法是基于概念表征进行事件分析。但是现有基于概念表征的视频事件识别方法仅利用简单的池化方法处理视频帧的概念表征以获取整个视频的概念表示, 未充分考虑概念的时序存在模式、概念间的关系以及概念与事件间的关系。基于此, 如图 7 所示。




(a) 方法框架

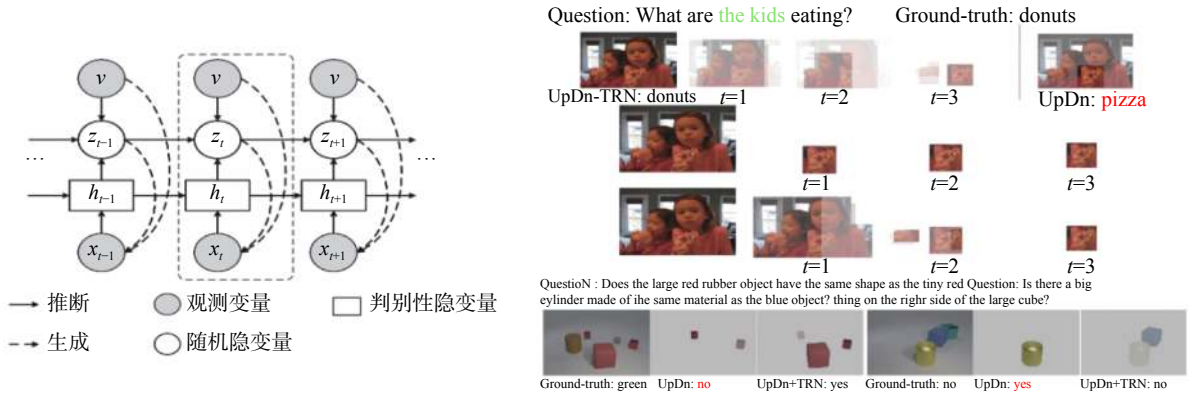
(b) 可解释视频事件识别例子

**图 7　概念挖掘网络**

**Fig. 7　Concept knowledge mining network**

本文利用场景、物体、动作概念检测器获取初始概念表征，提出概念知识挖掘网络，研究概念与事件间的依存关系，从而获取丰富且完备的视频概念表征，进行可解释性的视频事件识别。概念知识挖掘网络主要包含初始概念表征的提取，域内概念知识挖掘和域间概念知识挖掘以及概念表示融合模块。时序概念感受野动态挖掘网络的核心模块是时序动态卷积。时序动态卷积包括系数生成和结果融合两个模块，系数生成模块可以根据具有不同感受野的卷积核的输出结果生成加权系数，用于融合不同时序感受野下的概念表征，从而得到完备的视频概念表征。实验结果表明，所提算法[20]在 FCVID、ActivityNet 等主流大型事件识别数据集上均取得较好的事件识别性能，同时所提算法也可以对事件识别结果进行可解释性分析。

在跨媒体问答任务当中，针对现有隐式推理方法缺乏可解释性，而显式推理方法需要额外的标注信息的问题，从统计建模的角度出发，分析两类方法优化过程的主要差异。分析结果表明隐式推理方法缺乏足够解释性的根本原因是缺少对推理过程的直接建模。如图 8 所示，考虑到自然监督条件下缺少回答程序的标注，假设问题文本之下存在一组隐变量 $z$ 代表推理过程，并重构优化问题为优化问题、答案以及推理过程隐变量的联合分布，对推理过程进行直接建模。通过变分推断方式对优化问题进行求解，并采用基于变分自编码器的方法对分解得到的模块进行建模，模型的每个部分都是从原始的联合分布建模推导得到，保证了建模过程可解释性。将本研究推导得到的模型与现有的 3 种使用不同融合策略的隐式推理方法进行结合，在真实数据集以及合成数据集上进行实验，都取得了性能的提升，并且在推理过程可视化方面较基线更可解释[21]。




(a) 方法示意　　(b) 过程可视化

**图 8　过程可解释的跨媒体问答模型**

**Fig. 8　Interpretable visual question answering**

## 4.2　人机协同群智推理

随着互联网的迅速发展，人们可接触到的数据量日益增长。为缓解信息过载问题、改善用户体验，推荐系统得到广泛应用。然而，传统推荐方法的性能易受到数据稀疏性和冷启动问题的制约。为此，将知识图谱作为辅助信息的推荐算法得到大量关注。现有结合知识图谱的推荐算法大多使用实数向量在欧氏空间中进行建模，然而，实数向量的内积不具备内在的反对称性且表达能力有限。为此，研究组提出了基于四元数的协同知识图谱推荐网络[22]，其框架如图 9 所示，将用户–项目交互矩阵及知识图谱构建为协同知识图谱，利用四元数及其汉密尔顿乘积实现三元组旋转匹配的语义规则，并实现结合注意力机制的偏好传播与聚合方法，从而进一步提高个性化推荐的精准程度。

具体而言，利用四元数汉密尔顿乘积可建模旋转的性质，设计三元组旋转匹配的语义规则。为进一步提升推荐效果，可以采用结合注意力机制的偏好传播与聚合方法。沿协同知识图谱关系路径，利用每个实体的邻居实体信息辅助偏好的学习，使网络更适用于推荐任务。利用三元组的可信度分数度量每个实体和其邻居实体连接的强弱程度，进而求取该路径的注意力分数。基于每个实体本身的嵌入和其邻居嵌入得到每个实体的最终嵌入。

## 4.3　主动、交互式推理

对于跨模态相关的任务，现有模型往往都需要较大规模的标注数据集来训练模型，且要求数据集内的视觉内容足够丰富，文本描述足够详细，以使得模型能够很好地感知跨模态信息并将其关联理解。但是，数据的标注需要消耗很大的

人力物力以及时间，而跨模态数据又因为涉及不同模态的理解，对标注的要求高于纯视觉任务(分类、分割等)，标注跨模态数据集更是代价极大。为了解决这个问题，可将跨模态任务和主动学习相结合进行推理[23]。

为此，提出一种面向图像描述生成任务的结构化语义对抗主动学习框架，利用主动学习挑选值得标注的、更有指导意义的数据，从而在减少标注的花费的同时，又能够使模型推理学习到最有价值的数据。

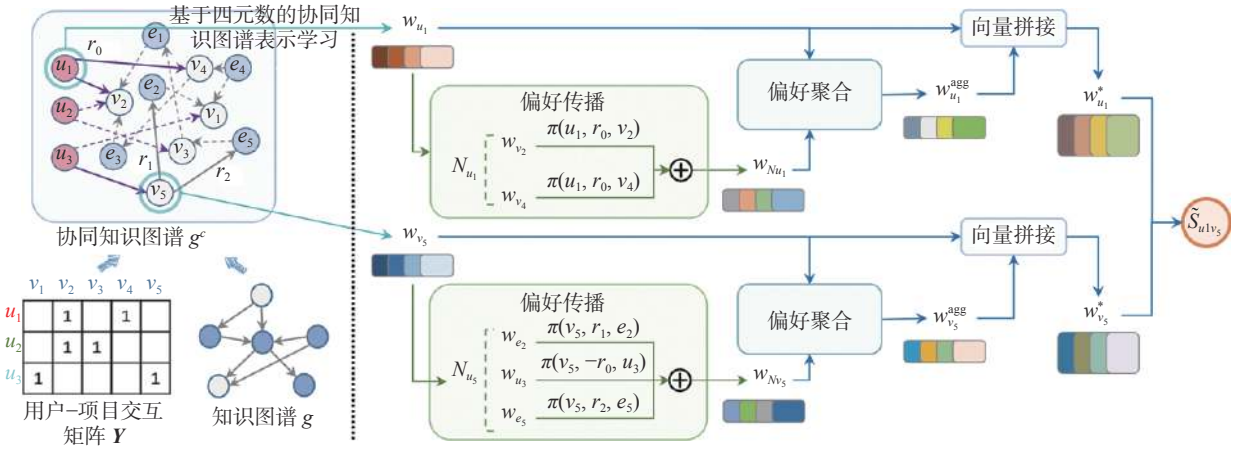



**图 9　基于四元数的知识图谱推荐方法**

**Fig. 9　Quaternion-based knowledge graph network for recommendation**

具体如图 10 所示，基于跨模态的结构化语义框架和对抗学习的主动学习模型，将视觉图像中的关键物体、物体状态和物体间联系表示为一个结构化的特征表示，进而判断样本的语义丰富度。该模型由 3 部分组成：结构化语义构建模块、多任务学习模块和标注状态判别器。结构化语义构建模块提取关键物体的区域特征，并将其编码为结构化的语义表示；之后，多任务学习模块计算了基于词级的快照损失和基于句级的重建损失，并以此更新模型；最后，状态判别器使用对抗学习机制判别样本的标注状态，并以此选取有价值的样本。作为模型关键部分的状态判别器中引入了对抗学习的机制。通过已标注样本和未标注样本在判别器内的对抗学习，使得结构化表示变得更加可分、更加充分，同时使得判别器判别有价值样本的能力更加强大。

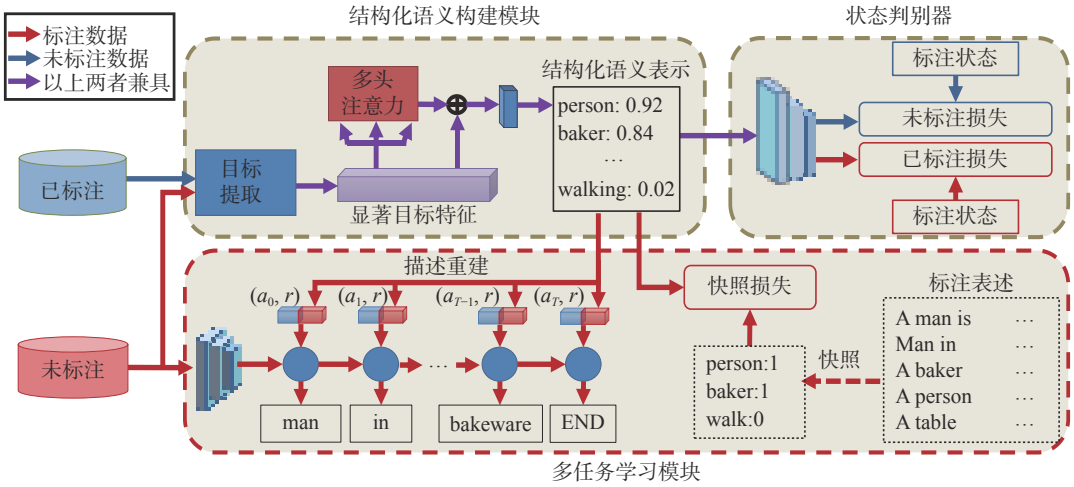



**图 10　图像概述生成的结构化语义对抗主动学习**

**Fig. 10　Adversarial active learning for image captioning**

用自然语言预测视频中潜在的未来事件是一项崭新且具有挑战性的交互推理任务，可广泛应用于安全辅助驾驶、视频监控(安防)、和人机交互等重要场合。该任务要求能够推理未来事件的不确定性和多样性，产生合理且多样化的预测和描述。对于这种跨模态交互推理，提出了隐含随机变量采样的跨模态多样性表示学习网络[24]。通过引入随机隐变量因子显式地捕获

视频中未来事件的随机性和多样性，对异质模态数据之间的复杂关联关系进行建模，以此生成合理的且多样化的自然语言语句来描述潜在的未来事件。引入隐变量因子分别对事件宏观主旨以及视觉注意力进行建模。一方面，对应多种潜在的事件主旨，生成多样性的语言描述；另一方面，在特定事件主旨下，模型使用随机性注意力机制，针对性地改变视频关注点，更精准地定位视频中的线索内容。这两个关键点使得提出的算法能够全面地且精准地对视频未来事件进行预测和描述。

如图 11 所示，引入一种宏观的隐变量因子，并使其符合特定的高斯分布，以对预测事件的宏观特性 (主旨、表达风格) 进行表征和决策。技术效果：在对视频进行预测性描述时，模型首先对此宏观的隐变量因子进行随机采样，进而根据不同采样值生成多样性的语言描述，对应多种潜在的事件；还引入一种微观的隐变量因子，构建一种随机性的注意力机制。此注意力机制模拟视觉关注点的随机性变化，根据事件宏观特征 (主旨) 针对性地挖掘可见视频中的细节线索。技术效果：在对视频进行预测性描述时，模型对此微观的隐变量因子进行采样，进而根据不同采样值改变视频内容的关注点。

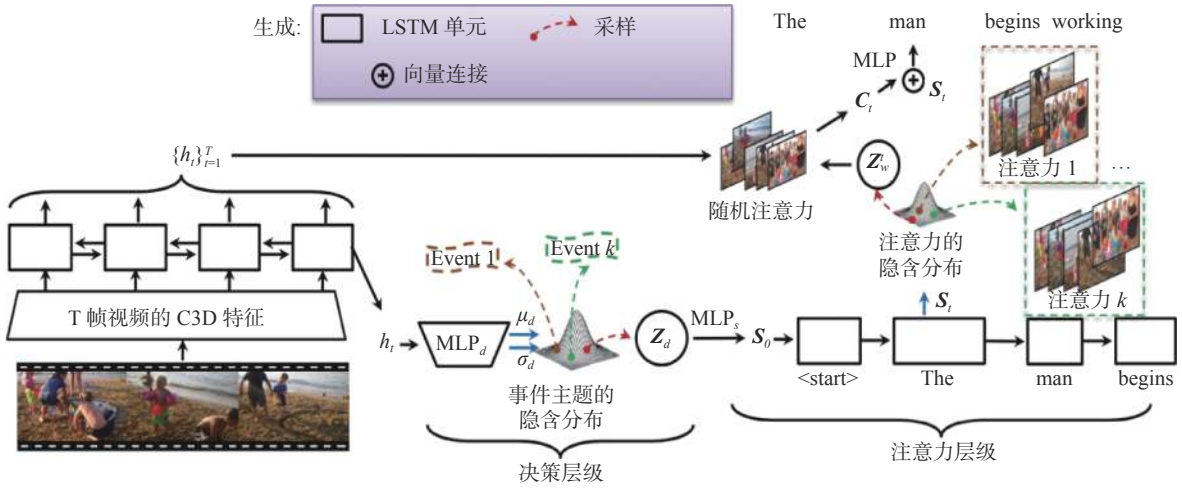



**图 11　跨模态推理网络生成过程示意**

**Fig. 11　Generation process of the proposed structured stochastic recurring network**

## 5　跨媒体知识图谱构建与演化

随着移动互联网渗透到社会生活的各方面，各大网络平台跨媒体数据呈现爆炸性增长和快速演化态势。然而，从碎片化数据难以直接提取系统完备的跨媒体知识，相关领域仍然面临跨媒体知识匮乏的困境。相比于传统的知识图谱构建任务，跨媒体知识工程面临更加严峻的技术挑战：1) 跨媒体知识图谱的构建依赖于不同模态实体的有效获取，虽然目前计算机已经能够识别各类视觉物体和文字实体，但距离通用的实体检测仍具有较高的技术难度，其主要难点在于对于新增实体无法有效识别；2) 不同模态的实体之间的关系种类与层级繁多，依赖全自动的数据关联分析技术虽然能够在短期内扩充图谱的知识条目规模，但总体而言存在知识重复、冗余、质量较低且系统智能演化缓慢的不足，而另一方面单纯依赖人工标注的方式构建的跨媒体知识图谱存在标注缓慢、效率低下等问题，难以有效适应跨媒体内容的动态演化。

为了应对上述挑战，研究组以电影知识为核心，构建了人机协同的跨媒体知识加工和演化更新基本技术框架。具体而言，所构建的跨媒体知识图谱包含两个层面的知识：1) 粗粒度知识，围绕特定的电影刻画了大量的属性知识和时空共现信息，如演员、导演、海报、评论、相关报道、影视主题曲等；1) 细粒度知识，针对某个电影内容，刻画电影镜头中的人物、表情、动作、关系、交互方式、字幕语义等知识。通过粗粒度和细粒度两个层级的知识标注，形成了以电影为主题的多层次多粒度跨媒体知识，从而为后续的跨媒体分析推理提供支撑。

### 5.1　粗粒度电影知识图谱

跨媒体知识工程旨在通过收集分布在互联网各影片平台中的跨媒体数据，从而构建节点规模亿级的电影粗粒度知识图谱数据库，并通过开放网络接口的方式实现高效的知识共享服务。

为保证知识图谱的体量与数据质量，图谱中数据主要来自国内外主流的视频网站、影片数据

库、视频评论网站，例如 IMDB、Amazon、Metacritic、AllMovie、Rotten Tomatoes、豆瓣、微博、哔哩哔哩、维基百科等。其中主要包含以下数据类型：

1) 文本：影片基本信息、演员信息、角色台词、专业影评、用户评论、新闻；

2) 图片：剧照、演员头像、演员其他照片等；

3) 视频：精彩片段、幕后花絮、视频报道等；

4) 音频：电影片头曲、主题曲、插曲等；

5) 关系：剧照中的角色关系、影片剧集关系、用户观影记录等。

构建电影粗粒度知识图谱的技术框架如图 12 所示，主要包括数据源调研、数据采集、数据存储与服务接口、数据演示等 5 部分，其中技术难点主要包括数据对齐和数据结构设计。

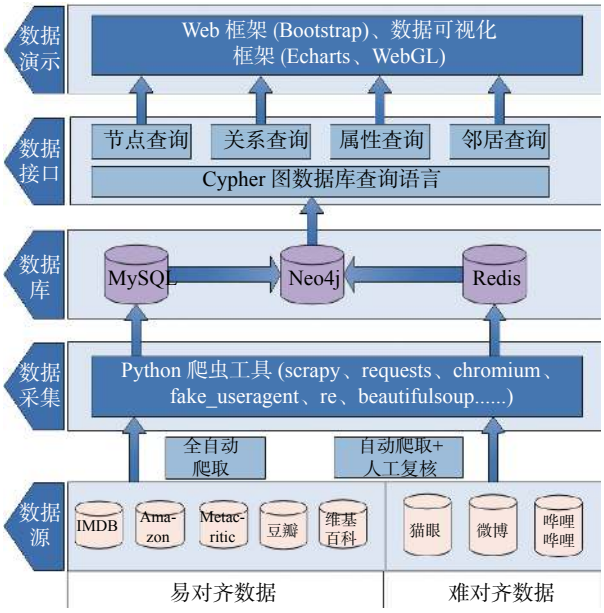



**图 12　粗粒度跨媒体知识图谱构建系统**

**Fig. 12　Coarse-grained cross-media knowledge graph construction system**

一方面，未对齐的数据不仅会产生冗余、低信息量的节点，更可能降低图谱中知识的可靠性。具体而言，同一影片在不同平台的名称并不一定相同，例如《肖申克的救赎》《月黑风高》《刺激 1995》均可以代表同一部电影作品；在一个平台中，相同的名称可能对应不同的影片，例如在 IMDB 中通过关键词“Terminator”将会同时检索到电影《终结者》及于 1991 年和 2001 年发行的两部影片。为保证数据高度对齐，同时考虑到 IMDB 的权威性与完备性，拟优先获取来自两类平台的数据：

1)IMDB 页面中包含的外部链接，例如 IMDB 电影页面包含的 Amazon 商品页面、Metacritic 电影主页等，如图 13(a) 所示。

2) 外部链接中包含 IMDB 电影页面的平台，例如豆瓣、维基百科等，如图 13(b) 所示。


(a) Metacritic 主页


(b) 豆瓣主页

**图 13　爬取的两类网站**

**Fig. 13　Two types of websites crawled**

另一方面，不合理的数据结构不仅会降低知识检索的效率，更导致大量孤立节点，降低图谱知识密度。在知识图谱中，存在节点、关系和属性 3 种数据存储方式。如图 14 所示，将具有唯一性的数据存储为节点，如影片、演员、编剧角色、公司、新闻、专业影评、用户评论、剧照、主题曲、花絮彩蛋等；将重复性很高的数据存储为属性，如性别、年龄、影片时长、发行年份、电影分类、评分等；将具有重要语义信息的数据存储为关系，如导演了、参演了、评论了、获得奖项、想看过再看等。通过设置合理的数据存储方式，一方面可减少语义信息较少的冗余节点，另一方面可减少相邻节点过少的孤立节点，从而提高图谱中有效知识的密度。

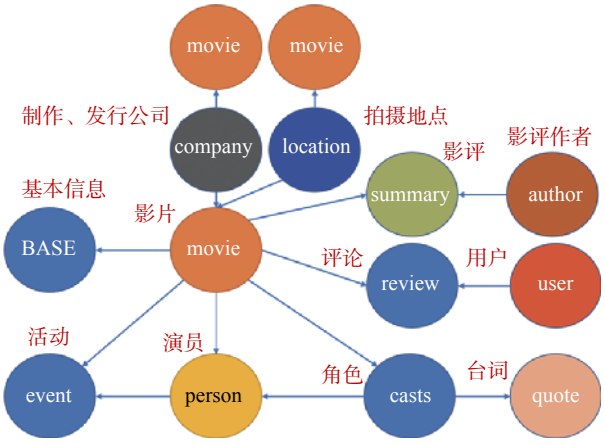



**图 14　电影粗粒度知识图谱节点关系示意**

**Fig. 14　Diagram of node relationship of coarse-grained film knowledge graph**

跨媒体知识工程的发展前景为实现跨媒体知识的自主进化和泛化。一方面，针对跨媒体数据快速更新演化的特点，如何及时准确地提取跨媒体信息，实现跨媒体知识的自主增长与进化，成为亟待解决的问题；另一方面，针对图谱中知识缺失问题，需设计高效的跨媒体知识补全算法，通过有效的知识泛化提高图谱中跨媒体知识的密度。

### 5.2　细粒度电影知识图谱与人机协同知识标注

结合跨媒体分析推理技术需求，开展构建细粒度跨媒体知识图谱，并针对电影知识图谱构建任务目标，建立了如图 15 所示的人机协同知识标注系统。由于电影包含了丰富的图、文、声多模态信息，故选择电影视频 (包括电影、纪录片等不同类型) 作为跨媒体知识图谱构建的基础数据。该跨媒体分析推理引擎中包含完备的电影视频语义概念抽取的技术，包括视频事件识别、动作识别、语音识别、表情识别、人脸识别、OCR、场景分类、物体检测、字幕解析等算法模块，对已经进行镜头分割与聚合的视频数据进行语义粗标注。在电影粗标注的基础上，搭建了群智标注平台 (如图 15 所示)，引入人类智慧对标注结果进行定期的纠错和更新，并利用修正过的视频语义概念信息对事件识别、动作识别、场景分类等语义粗标注模型进行进一步更新和优化，改善语义粗标注算法在后续标注当中的准确率。目前，已完成了 264 部中文电影的知识标注，知识图谱的规模达到节点 2 707 350，边 (事实) 4 159 220 的规模。

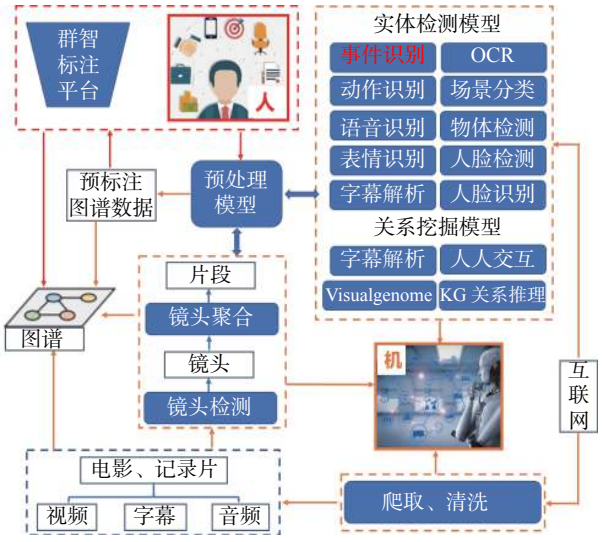



**图 15　人机协同细粒度图谱标注系统**

**Fig. 15　Fine-grained knowledge graph labeling system with human-computer collaboration**

在后续研究工作当中，将进一步对 500 部英文电影和纪录片进行知识标注，并不断完善和优化跨媒体分析推理引擎。

### 5.3　跨媒体分析推理引擎

基于上述关键技术研究成果，以亿级跨媒体数据的跨媒体统一表征索引与检索为基础，构建跨媒体分析推理引擎。该计算引擎的目标主要有 3 个层面：1) 验证跨媒体知识对跨媒体分析推理的支撑作用；2) 通过跨媒体分析推理进一步提高跨媒体知识生产效率；3) 通过跨媒体分析推理各技术子系统的集成，进一步突出可解释、可泛化分析推理的技术特色。目前该跨媒体分析推理引擎主要包含如下几个部分：

1) 跨媒体统一表征与检索，实现亿级以上跨媒体数据的统一可度量表征、高效索引与检索；

2) 跨媒体问答引擎，包括人问机答和机问人答两部分，为图谱演化、内容服务提供支持；

3) 跨媒体迁移引擎，针对内容监测与内容服务的多元化应用场景，利用无标注或者少量标注跨媒体数据进行迁移与泛化学习，满足各类开放域应用需求。

视频计算引擎，针对各类网络视频内容，实现内容识别预警，跨模态的内容搜索定位和视频事件的推理预测，为视频内容监测与服务提供技术支撑。

## 6　结束语

本文介绍了研究组近年来以图像视频为中心的跨媒体分析与推理方面的研究工作，提出了统一表征、关联理解、类人推理等方法，开展构建跨媒体知识图谱和支撑知识图谱构建的各类技术平台，初步建立了数据驱动与知识指导相结合的跨媒体分析推理技术框架。可服务于跨媒体内容管理与服务等应用场景。

从跨媒体分析推理技术的发展前景来看，跨媒体知识的自主高效演化与可解释、可泛化的类人跨媒体分析推理将继续成为未来相关领域的前沿热点研究问题，也是通向强人工智能的关键瓶颈。为此，结合领域前沿研究趋势，研究组针对跨媒体分析与推理的核心难题进行集中研究与攻关，拟从以下几个方面具体开展未来工作：

1) 现有技术已能从不同模态数据当中检测出不同类型的实体，如人物、物体、地标建筑、事件、主题等，然而，相比于实体数量，跨媒体知识图谱中的关系知识表示数量规模仍有待提高 (平均每个实体包含的关系知识三元组不超过 5 条)。为从根本上提高跨媒体连接知识的稠密度，研究组拟对多模态实体提纯、链接与多模态关系的发现、补全等前沿技术问题进行定义并构

建相应的数据集, 并在此基础之上形成相关的技术突破。

2) 针对跨媒体知识演化更新缓慢的难题, 深入开展深度人机协作的跨媒体分析推理技术, 并在跨媒体知识图谱构建当中构建相应的原型系统。拟从个体协作和群体协作两个层面开展相应研究。在人机个体协作方面, 针对跨媒体事件语义理解、内容转换与生成等复杂跨媒体任务特点, 以主动学习和问答交互式学习为技术主线, 开发人机问答博弈系统, 建立人问机答和机问人答相互博弈演化的跨媒体知识更新框架。在人机群体协作, 深入研究群智挖掘与推荐技术, 实现“标一当百”的跨媒体群智计算, 并将其应用在跨媒体知识工程当中。

3) 突破现有跨媒体知识图谱当中由于大量时空偶发贡献造成的虚假或不合理关联信息, 进一步开展跨媒体事理关系图谱的构建, 刻画深层次跨媒体事理知识。

4) 深入研究机理、过程、结果可信的鲁棒跨媒体分析推理理论方法。结合符号主义与联结主义, 研究数据去偏置的因果学习和因果辨识与组合推理等关键技术, 实现公平、可信、可泛化的跨媒体分析推理技术框架。

## 参考文献:


[1] HUBEL D H, WIESEL T N. Early exploration of the visual cortex[J]. Neuron, 1998, 20(3): 401−412.

[2] MARR D. Vision : a computational investigation into the human representation and processing of visual information[M]. Cambridge: The MIT Press, 2010.

[3] CHOMSKY N. Aspects of the theory of syntax[M]. Cambridge, MA: The MIT Press, 1965.

[4] MCGURK H, MACDONALD J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746–748.

[5] NEWELL A, SIMON H A. Computer science as empirical inquiry: symbols and search[J]. Communications of the ACM, 1976, 19(3): 113–126.

[6] PETERSON G W, SAMPSON J R JR, REARDON R C. Career development and services: a cognitive approach [M]. Thomson Brooks/Cole Publishing Co, 1991.

[7] DESOLNEUX A, MOISAN L, MOREL J M. From gestalt theory to image analysis[M]. New York: Springer, 2008.

[8] PARK H J, FRISTON K. Structural and functional brain networks: from connections to cognition[J]. Science, 2013, 342(6158): 1238411.

[9] 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.
WANG Shuhui, YAN Xu, HUANG Qingming. Overview of research on cross-media analysis and reasoning technology[J]. Computer science, 2021, 48(3): 79–86.

[10] ZHANG Shiliang, TIAN Qi, HUA Gang, et al. Descriptive visual words and visual phrases for image applications[C]//Proceedings of the 17th ACM International Conference on Multimedia. Beijing, China, 2009: 75−84.

[11] WU Yiling, WANG Shuhui, SONG Guoli, et al. Learning fragment self-attention embeddings for image-text matching[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 2088−2096.

[12] SONG Guoli, WANG Shuhui, HUANG Qingming, et al. Harmonized multimodal learning with gaussian process latent variable models[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(3): 858–872.

[13] WANG Shuhui, HUANG Qingming, JIANG Shuqiang, et al. $S^3$MKL: scalable semi-supervised multiple kernel learning for real-world image applications[J]. IEEE transactions on multimedia, 2012, 14(4): 1259–1274.

[14] XU Qianqian, HUANG Qingming, JIANG Tingting, et al. HodgeRank on random graphs for subjective video quality assessment[J]. IEEE transactions on multimedia, 2012, 14(3): 844–857.

[15] LI Liang, JIANG Shuqiang, HUANG Qingming. Learning hierarchical semantic description via mixed-norm regularization for image understanding[J]. IEEE transactions on multimedia, 2012, 14(5): 1401–1413.

[16] SHEN Li, WANG Shuhui, SUN Gang, et al. Multi-level discriminative dictionary learning towards hierarchical visual categorization[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 383−390.

[17] DENG Jincan, LI Liang, ZHANG Beichen, et al. Syntax-guided hierarchical attention network for video captioning[J]. IEEE transactions on circuits and systems for video technology, 2021( 99):1.

[18] CHEN Yangyu, WANG Shuhui, ZHANG Weigang, et al. Less is more: picking informative frames for video captioning[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 367−384.

[19] LIU Zhenhuan, DENG Jincan, LI Liang, et al. IR-GAN:

image manipulation with linguistic instruction by increment reasoning[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 322−330.

[20] QI Zhaobo, WANG Shuhui, SU Chi, et al. Towards more explainability: concept knowledge mining network for event recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 3857−3865.

[21] HAN Xinzhe, WANG Shuhui, SU Chi, et al. Interpretable visual reasoning via probabilistic formulation under natural supervision[C]//Proceedings of the 16th European Conference. Glasgow, UK, 2020: 553−570.

[22] LI Zhaopeng, XU Qianqian, JIANG Yangbangyan, et al. Quaternion-based knowledge graph network for recommendation[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 880−888.

[23] ZHANG Beichen, LI Liang, SU Li, et al. Structural semantic adversarial active learning for image captioning[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 1112−1121.

[24] YANG Shijie, LI Liang, WANG Shuhui, et al. Structured stochastic recurrent network for linguistic video prediction[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 21−29.

**作者简介：**

黄庆明，教授，博士生导师，主要研究方向为多媒体分析与计算机视觉。IEEE Fellow，享受国务院政府特殊津贴，IEEE TCSVT、自动化学报等期刊的编委，获吴文俊人工智能自然科学奖一等奖 (第一完成人)。主持科技创新 2030−“新一代人工智能”重大项目、国家自然科学基金重点项目和重点国际合作项目、国家 973 计划课题、科学院前沿科学研究重点计划等项目多项。发表学术论文 170 余篇。

王树徽，研究员，博士生导师，主要研究方向为跨媒体分析推理与图像视频理解。获 2020 年吴文俊人工智能自然科学一等奖 (第二完成人)、CCF 科学技术奖 (2012)、全国多媒体大会最佳论文奖等。发表学术论文 50 余篇。

许倩倩，副研究员，主要研究方向为数据挖掘和机器学习。获吴文俊人工智能自然科学奖一等奖 (第三完成人)、中国人工智能学会最佳青年科技成果奖、中国图象图形学学会石青云女科学家奖、吴文俊人工智能优秀青年 ACM 中国 SIGMM 新星奖、中国人工智能学会优秀博士学位论文、中科院百篇优秀博士学位论文、CCF-腾讯犀牛鸟科研金、首届 CAAI-华为 MindSpore 学术奖励基金等。发表学术论文 40 余篇。

# “中国人工智能学会-华为 MindSpore 学术奖励基金”第二期发布

# The 2nd issue of “Chinese Association for Artificial Intelligence-Huawei MindSpore Academic Award Fund” were released

《中国人工智能学会-华为 MindSpore 学术奖励基金》是由中国人工智能学会和华为技术有限公司共同发起，面向高校及科研院所的 AI 科研人员搭建学术交流平台，提供经费、算力、技术支持等服务，推动 MindSpore 在 AI 领域科研的应用，并支持基于 MindSpore 框架的国际国内高水平会议和期刊的学术论文发表，激励原创性科学研究开展，构建中国人工智能科学研究的全球影响力。入选项目将分两类进行资助：

A 类：额度 9 万，计划支持项目数量不多于 30 个；

B 类：额度 18 万，计划支持项目数量不多于 10 个。

具体内容请按《申请指南》提交申报材料。

申报开始时间：2021 年 9 月 1 日

申报截止时间：2021 年 10 月 10 日

发布评审结果时间：2021 年 10 月 30 日

项目完成时间：2022 年 10 月 30 日前

专属申报邮箱：xsjljj@caai.cn

学会联系电话：010-82686683

联系人: 邹亚茹 1312112388

详情请登录中国人工智能学会网站。