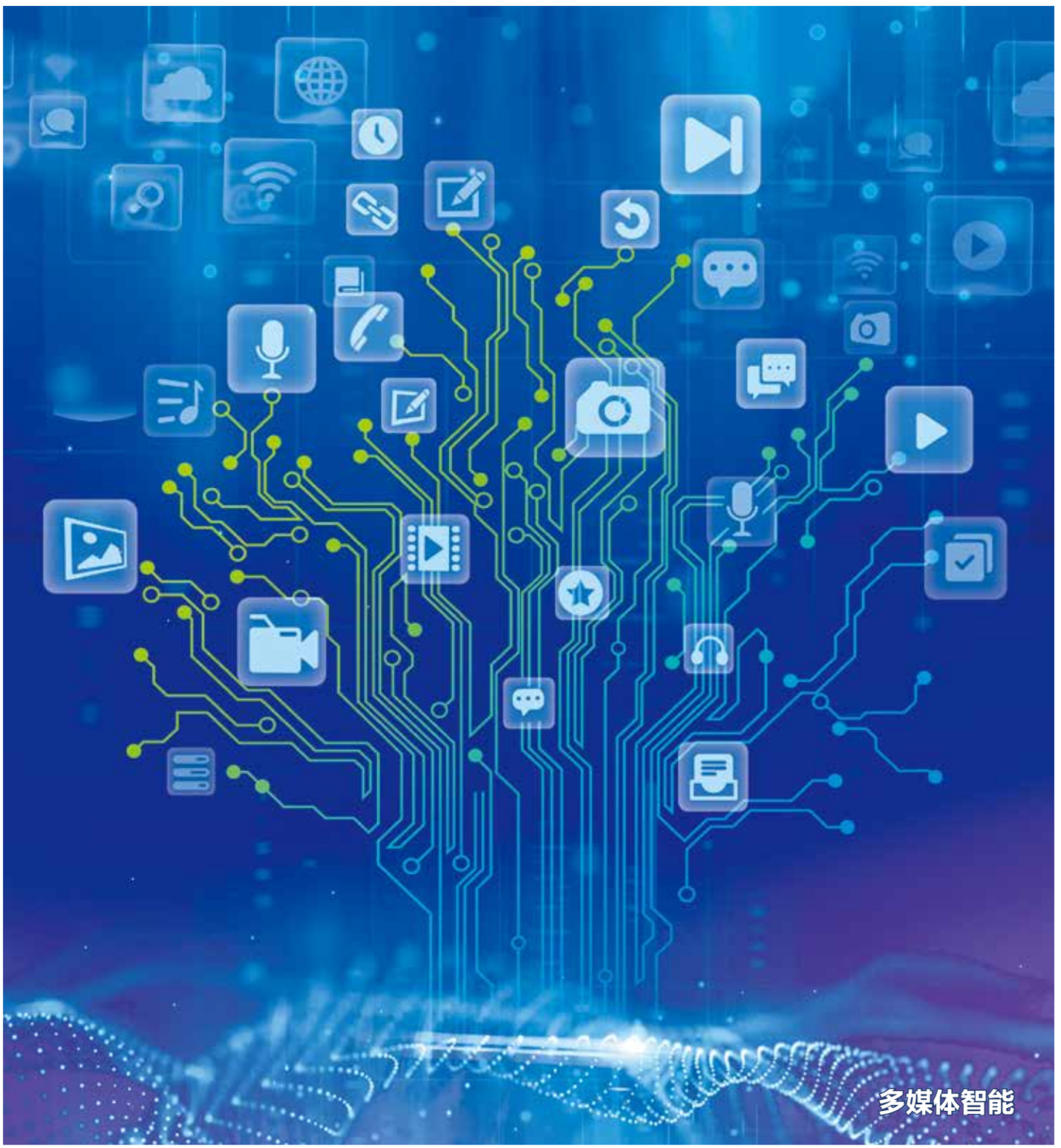
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第9期（总第317期）

2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

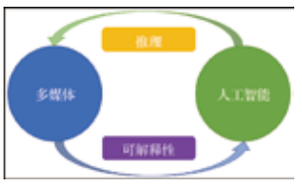
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年9月 第27卷 第9期（总第317期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

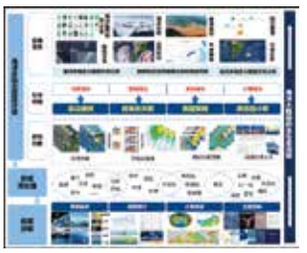
目次



多媒体智能：当多媒体遇到人工智能(第2551页)

学者观点

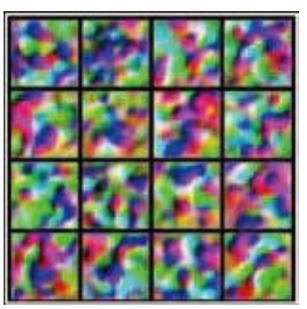
综述



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)

多媒体智能安全

目标智能检测

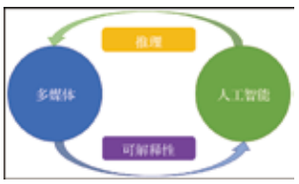


基于真实数据感知的模型功能窃取攻击(第2721页)

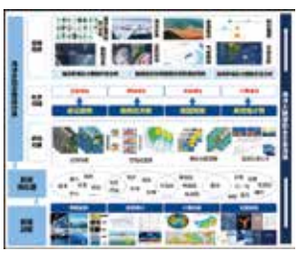
多媒体分析与理解

Volume 27, Number 9
Published September 16, 2022

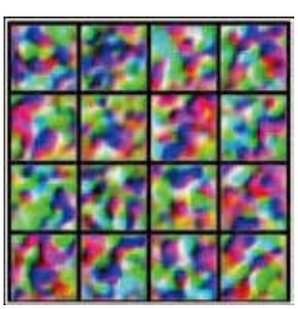
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.7　　**文献标识码**:A　　**文章编号**:1006-8961(2022)09-2574-15

论文引用格式:Yang Y, Zhuang Y T and Pan Y H. 2022. The review of visual knowledge: a new pivot for cross-media intelligence evolution. Journal of Image and Graphics,27(09):2574-2588(杨易,庄越挺,潘云鹤. 2022. 视觉知识:跨媒体智能进化的新支点. 中国图象图形学报,27(09):2574-2588)[DOI:10.11834/jig.211264]

视觉知识:跨媒体智能进化的新支点

杨易[1*],庄越挺[1],潘云鹤[1,2]

1. 浙江大学计算机科学与技术学院,杭州　310027;2. 之江实验室,杭州　310027

摘　要:回顾跨媒体智能的发展历程,分析跨媒体智能的新趋势与现实瓶颈,展望跨媒体智能的未来前景。跨媒体智能旨在融合多来源、多模态数据,并试图利用不同媒体数据间的关系进行高层次语义理解与逻辑推理。现有跨媒体算法主要遵循了单媒体表达到多媒体融合的范式,其中特征学习与逻辑推理两个过程相对割裂,无法综合多源多层次的语义信息以获得统一特征,阻碍了推理和学习过程的相互促进和修正。这类范式缺乏显式知识积累与多级结构理解的过程,同时限制了模型可信度与鲁棒性。在这样的背景下,本文转向一种新的智能表达方式——视觉知识。以视觉知识驱动的跨媒体智能具有多层次建模和知识推理的特点,并易于进行视觉操作与重建。本文介绍了视觉知识的3个基本要素,即视觉概念、视觉关系和视觉推理,并对每个要素展开详细讨论与分析。视觉知识有助于实现数据与知识驱动的统一框架,学习可归因可溯源的结构化表达,推动跨媒体知识关联与智能推理。视觉知识具有强大的知识抽象表达能力和多重知识互补能力,为跨媒体智能进化提供了新的有力支点。

关键词:跨媒体智能;视觉知识;视觉概念;视觉关系;视觉推理

The review of visual knowledge: a new pivot for cross-media intelligence evolution

Yang Yi[1*], Zhuang Yueting[1], Pan Yunhe[1,2]

1. College of Computer Science and Technology, Zhejiang University, Hangzhou　310027, China;
2. Zhejiang Laboratory, Hangzhou　310027, China

Abstract: We review the recent development of cross-media intelligence, analyze its new trends and challenges, and discuss future prospects of cross-media intelligence. Cross-media intelligence is focused on the integration of multi-source and multi-modal data. It attempts to use the relationship between different media data for high-level semantic understanding and logical reasoning. Existing cross-media algorithms mainly follow the paradigm of "single media representation" to "multi-media integration", in which the two processes of feature learning and logical reasoning are relatively disconnected. It is unlikely to synthesize multi-source and multi-level semantic information to obtain unified features, which hinders the mutual benefits of the reasoning and learning process. This paradigm is lack of the process of explicit knowledge accumulation and multi-level structure understanding. At the same time, it restricts the interpretability and robustness of the model. We interpret new representation method, i. e., visual knowledge. Visual knowledge driven cross-media intelligence has the fea-

收稿日期:2022-01-11;**修回日期**:2022-05-18;**预印本日期**:2022-05-25

*通信作者:杨易　yangyics@zju.edu.cn

基金项目:国家重点研发计划资助(2020AAA0108800);中央高校基本科研业务费专项资金资助(226-2022-00051)

Supported by: National Key R&D Program of China (2020AAA0108800); Fundamental Research Funds for the Central Universities (226-2022-00051)

tures of multi-level modeling and knowledge reasoning. Its built-in mechanisms can implement operations and reconstruction visually, which learns knowledge alignment and association. To establish a unified way of knowledge representation learning, the theory of visual knowledge has been illustrated as mentioned below: 1) we introduce three key factors of visual contexts, i. e. , concept, visual relationship, and visual reasoning. Visual knowledge has capable of knowledge representations abstraction and multiple knowledge complementing. Visual relations represent the relationship between visual concepts and provide an effective basis for more complex cross-media visual reasoning. We demonstrate visual-based spatio-temporal and causal relationships, but the visual relationship is not limited to these categories. We recommend that the pairwise visual relationships should be extended to multi-objects cascade relationships and the integrated spatio-temporal and causal representations effectively. Visual knowledge is derived of visual concepts and visual relationships, enabling more interpretive and generalized high-level cross-media visual reasoning. Visual knowledge develops a structured knowledge representation, a multi-level basis for visual reasoning, and realizes an effective demonstration for neural network decisions. Broadly, the referred visual reasoning includes a variety of visual operations, such as prediction, reconstruction, association and decomposition. 2) We discuss the applications of visual knowledge, and introduce detailed analysis on their future challenges. We select three applications of those are structured representation of visual knowledge, operation and reasoning of visual knowledge, and cross-media reconstruction and generation. Visual knowledge is predicted to resolve the ambiguity problems in relational descriptions and suppress data bias effectively. It is worth noting that these three specific applications are involved some cross-media intelligence examples of visual knowledge only. Although hand-crafted features are less capable of abstracting multimedia data than deep learning features, these descriptors tend to be more interpretable. The effective integration of hand-crafted features and deep learning features for cross-media representation modeling is a typical application of visual knowledge representation in the context of cross-media intelligence. The structured representation of visual knowledge contributes to the improvement of model interpretability. 3) We analyze the advantages of visual knowledge. It aids to achieve a unified framework driven by both data and knowledge, learn explainable structured representations, and promote cross-media knowledge association and intelligent reasoning. Thanks to the development of visual knowledge based cross-media intelligence, more emerging cross-media intelligence applications will be developed. The decision-making assistance process is more credible through the structural and multi-granularity representation of visual knowledge and the integrated optimization of multi-source and cross-domain data. The reasoning process can be reviewed and clarified, and the model generalization ability can be improved systematically. These factors provide a new powerful pivot for the evolution of cross-media intelligence. Visual knowledge can improve the generative models greatly and enhance the application of simulation technology. Future visual knowledge can be used as a prior to improve the rendering of scenes, realize interactive visual editing tools and controllable semantic understanding of scene objects. A data-driven and visual knowledge derived graphics system will be focused on the integration of the strengths of data and rules, semantic features extraction of visual data, model complexity optimization, simulation improvement, and realistic and sustainable content in new perspectives and new scenarios.

Key words: cross-media intelligence; visual knowledge; visual concepts; visual relationships; visual reasoning

0 引 言

跨媒体智能是人工智能的一个重要研究领域。人类善于综合视觉、听觉和语言文字等多种信号进行认知和推理。当人类融合多种感知途径形成对某个事物的综合理解后，这些多方面感知信号之间能够互相触发、彼此增强。在人工智能的研究中，通过多种媒体方式进行信息感知、融合、表达和推理是跨媒体智能的典型特征。跨媒体智能不仅单独处理不同来源、不同模态的数据，还对它们进行多来源、多模态融合与增强；不仅要求完成简单的识别、检测和定位，还能够进行更复杂的理解与推理等高阶智力活动。跨媒体智能展现出与人类认知和思考的高度相似性，正逐渐成为新一代人工智能研究中备受关注的一个重要方向。

本文首先调研跨媒体表达的研究现状，并分析现有相关研究的局限性。作为跨媒体智能的重要研究方向之一，跨媒体表达经历了手工设计和深度学习两个阶段。这一发展轨迹与人工智能其他领域相

似。无论是手工设计还是深度学习方法,绝大部分跨媒体表达研究以单一媒体数据下的知识表达为基础,分别获取多个模态的特征,然后将这些多模特征并行映射到模态共享的特征空间,进行特征关联与融合。这种自底而上融合的特征表达方式给跨媒体知识表达带来了很大的局限性,缺乏系统性的可解释与高阶推理能力。随后,本文分析一种跨媒体智能研究的新途径——视觉知识(Pan,2019)。视觉知识不仅关注图像或视频等视觉信号,以及基于这些信号提取或学习得到的特征,而且以视觉概念(通常由典型和范畴构成)(Pan,2019)为研究要素,联合符号化知识与逻辑推理、深度学习技术、知识图谱以及手工构造的知识(如结构化信息)等多种知识表达手段,将与视觉主体相关的音频、语言等信号进行联合建模与推理。这种性质与人类在理解和推理多媒体信号时的处理流程是相似的——以信息量最高的视觉为主导,并在其基础上关联语音、文字等感知与理解。这些声音、文字等其他形式的信息又能适时地促进、增强对环境和目标的视觉理解。这种多模态知识表达互相增强的性质,也正是视觉知识具有的多重知识表达能力(Pan,2020;Yang 等,2021)。

考虑到现有跨媒体智能算法鲁棒性弱、泛化性与可解释性不足等基础性问题,构建视觉知识驱动的新型视觉表达理论,提升视觉知识挖掘与提取的自动性和可解释性势在必行。视觉知识作为一个重要的新兴研究方向,对跨媒体智能进一步发展至关重要。视觉知识与多重知识表达的结合,有望成为跨媒体智能研究的新支点。

1 跨媒体表达现状与局限

常见的多媒体数据包括图像、视频、音频和自然语言等。跨媒体特征表达的基础是对单一媒体数据进行表达。对于单一媒体表达,手工设计特征和深度学习特征都已有大量研究,并在各自发展阶段均取得较大进展。尤其是深度学习技术的出现,促进了跨媒体研究成果在很多领域得到广泛应用。

尽管取得了长足的进步,现有跨媒体研究依然有其局限性。原因在于现有跨媒体表达主要遵循单媒体表达到多媒体融合的范式,特征学习和逻辑推理两个过程相对割裂,无法综合多源多层次的语义信息以获得统一特征,阻碍了推理和学习过程的相互促进和修正。这类范式缺乏显式知识积累与多级结构理解的过程,同时限制了模型可信度与鲁棒性。具体来讲,目前跨媒体表达的局限性主要体现在模型可信与可解释能力弱、层次建模与结构理解不足和推理认知与迁移效果欠佳等方面。

1)模型可信与可解释能力弱。现有跨媒体表达通常先独立提取不同媒体各自模态的特征,再进行跨媒体特征融合。这类方式易于模型训练,但由于缺乏统一表达,造成模型过拟合且难以归因。对于图像或视频数据,SIFT(scale-invariant feature transform)(Lowe,2004)、HoG(histogram of oriented gradients)(Dalal 和 Triggs,2005)和 IDT(improved dense trajectories)(Wang 和 Schmid,2013)等传统手工特征抽取技术利用关键点或边缘信息获取局部特征描述符。这类特征描述符具有一定的可解释性,但不具备可学习能力且拟合能力弱。随着深度学习技术的突破,主流方法大多使用深度学习技术提取单一模态特征。其中图像的特征抽取通常使用卷积神经网络(Krizhevsky 等,2012)。

语言特征通常使用词向量模型(Mikolov 等,2013)、长短时记忆模型(Hochreiter 和 Schmidhuber,1997)或 Transformer 模型(Vaswani 等,2017)。获得单一媒体的特征后,跨媒体学习的典型思路是将多种不同模态特征映射到跨媒体共享的特征空间。这一过程需要将多模态特征的学习融入到统一的学习框架中,并在模型优化过程中挖掘跨媒体数据间的内在关联。例如,图像与自然语言之间的跨媒体表达通常将图像特征和语言特征映射到同一特征空间,并使用特定的损失函数约束图像和语言在这个特征空间的相似性(Frome 等,2013;Zheng 等,2020a)。虽然深度神经网络具有强大的表征学习能力,但在一些情况下特征过拟合现象严重。深度神经网络模型较难可视化且参数庞大,可解释性不足。另外,跨媒体深度学习方法涉及多个模态的建模,往往需要利用神经网络提取每个模态的特征,整体模型参数量相比于单模态模型更多,训练过程中更易出现过拟合现象(Wang 等,2020a),导致模型预测更难系统性解释。经过跨媒体融合后的特征更为抽象,为数据归因、模型解释带来了更多困难。

2)层次建模与结构理解不足。以图像为例,常见的跨媒体表达技术为了获取更加丰富的图像特

征，会采用多尺度图像特征或对图像语义的关系建模(Li 等，2019)。除了图像和语言分别抽取特征外，一些方法(Wang 等，2019；Wu 等，2018)在抽取图像和语言特征过程中引入图像和语言信息的互相流动机制，提升图像和语言的跨媒体表达能力。这些研究取得了较大进展，但忽略了识别过程中显式建模层级化信息。现有跨媒体表达在分别完成各个模态特征提取后，继而进行信息融合，并不具备由浅及深和分层次融合多媒体知识的能力。相比之下，人类认知更倾向于一个抽象程度由浅及深的过程，在信息处理时逐渐移除琐碎细节并保留重要元素。例如，当人类识别一个特定动物种类时，首先倾向于观察它的外貌、聆听它的声音，形成直观的感知，获得颜色、尺寸和纹理等较为细节的感官信息。这些可感知的信息处于一个相对较低的抽象层次。基于这些感官信息，人类可以融入一些抽象层次更高的知识，例如生活习性、生物分类学信息等。在这个例子中，视觉、听觉信息是抽象层次较低的知识，符号化的语言、文字蕴含抽象层次较高的知识。人类这种基于多重知识表达的认知过程先依靠视觉和听觉获取低抽象的感官信息，再依靠符号化表达的语言文字获取高抽象的生活习性和分类学信息，对于人类充分利用多媒体信息，形成对环境和事物全面认知至关重要。

3)推理认知与迁移效果欠佳。无论手工设计还是深度学习方法，跨媒体表达方法大都遵循自底而上融合的范式：先分别在不同的数据模态下学习相应的单模态特征，再将这些特征映射到同一个模态共享的特征空间中进行跨媒体融合。一个代表性的思路是通过探索数据之间的关联和子空间学习获取更加准确的统一表达。Hardoon 等人(2004)提出基于线性变换的典型相关性分析方法(canonical correlation analysis，CCA)，通过成对跨媒体数据的相关性学习映射矩阵，将处于异构空间中的多媒体特征映射到同构空间，获取可以进行相似性对比的跨媒体表示。除了 CCA，还有一些研究采用图的形式进行跨媒体建模(Yang 等，2012b)。另有一些工作将跨媒体图建模与子空间方法相结合(Yang 等，2008)。尽管在利用手工特征进行跨媒体表达上进行积极尝试且取得很多进展，这类方法在某些特定领域(如跨媒体推理)的性能依然相对较弱。

当前跨媒体表达技术在一个模态共享的特征空间对所有模态进行融合，提升了综合表达能力，但鲜能利用一种模态信息对另一种模态信息进行特征增强和推理。这种范式虽然在一些应用场景能够满足特定的跨媒体信息融合和交互需求，但与人类处理跨媒体数据相比，推理能力薄弱，对媒体之间的信息增强和关联能力不强，无法有效进行跨场景迁移。相比之下，人类并不是简单地对多媒体信息进行融合，而是在融合中利用不同模态信息相互促进。例如，通过融合“汽车可以喷涂为各种颜色”的高层语义信息和“一辆红色汽车”的图像信息，一个从未见过其他汽车的儿童也能识别出黑色汽车。在这个例子中，自然语言表达的符号化知识促进了视觉信息的泛化。然而，跨媒体知识表达尚不能在不同媒体信息之间形成有效的相互增强。

上述不足限制了当前跨媒体知识表达能力的进一步提升，成为跨媒体智能发展的主要障碍，亟需更合理、更灵活和更复杂的跨模态知识表达来推动跨媒体智能的进步。

2　视觉知识理论

视觉知识(Pan，2019)是一种有望提高跨媒体知识表达能力，进一步推动人工智能(特别是跨媒体智能)发展的新框架。视觉知识对跨媒体表达具有支撑和促进作用。需要指出的是，视觉知识理论上不仅可以促进跨媒体表达的研究，也可以支撑和提升诸如智能创作、逻辑推理等更为广泛人工智能领域的研究和应用。图 1 展示了视觉知识的基本要素及其优势。

2.1　视觉知识的要素

目前，视觉知识的基本要素包括视觉概念、视觉关系和视觉推理。

2.1.1　视觉概念

视觉知识以视觉概念作为基本单元。视觉概念具备结构化和可解释的特性，保证知识建模可以外推，为跨媒体分析提供可归因的推断结果。

1)典型与范畴。视觉概念具有典型与范畴结构。典型(prototype)是某类样本中最常见的一种模式，作为视觉概念的核心表示，描述事物的典型特征。范畴为典型中各种参数的变化域，也可作为典型与若干非典型形状、色彩构成的综合场(Pan，2019；潘云鹤，1996)。针对视觉概念进行典型与范

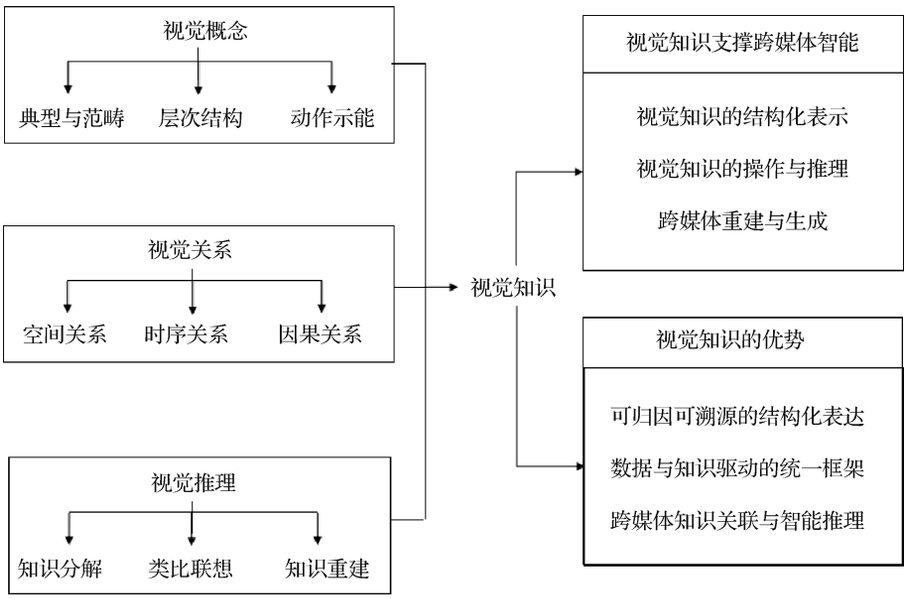


图 1　视觉知识的基本要素及其优势

Fig. 1　The main elements of visual knowledge and its advantages

畴的分解,有助于更准确的视觉概念分布估计,并有效实现典型归纳与范畴迁移。例如,Snell 等人(2017)将类内样本的特征平均理解为类别典型。在小样本条件下,该典型特征相比于样本特征更为鲁棒。Wang 等人(2021)引入一组动词典型用于描述具身动作的主要运动模式,并利用该典型辅助场景内物体特征的分解,最终选择出准确的交互物体。Zheng 等人(2019)分离了行人图像中的两种表征,即典型特征(外观特征)和结构特征(人物体态),通过交换不同行人的特征,生成行人范畴内的新图像,扩充训练样本。解耦外观特征和结构特征,有助于提升合成行人图像的鲁棒性与可靠性,实现高质量新图像合成。Zhu 和 Yang(2022)为少样本学习设计了标签独立存储器,用于缓存特定类的知识,其中每个类特征的聚合可以理解为该类视频的典型。这种典型特征对嘈杂的视频具有更强的鲁棒性。范例(exemplar)学习也可以理解为视觉概念典型/范畴建模的研究。例如 Yang 等人(2013a)通过自适应地选取范畴之外的训练数据来提升少样本条件下的复杂事件检测性能。

2)层次结构。视觉概念的层次结构(Pan,2019)研究包含多尺度样本理解、多层次类别抽象和多模态主次分析等。

多尺度理解在视觉分析领域已广泛使用。例如,Lazebnik 等人(2006)引入金字塔结构用于多尺度图像特征学习,对物体形变具有较强鲁棒性。在视频分析领域,长视频内容的层次化表达可有效减少输入信息流的长度,有效挖掘更长范围内的视频时序结构(Pan 等,2016)。Yang 等人(2013b)利用数据的层级流形结构提供更为鲁棒的多媒体语义理解。Zhu 等人(2022)提出一种跨层的注意力机制以实现相邻帧间多层级信息的探索,该跨层注意力模块决定了不同卷积层的权重。获得多尺度权重后,融合来自多尺度的上下文知识为动作识别提供了高效的特征。Zhu 等人(2022)考虑了从多个视频间获取共享的多尺度信息,这类多尺度信息具有全局一致性,降低了单个样本可能带来的数据偏差,从而获得更稳定且易识别的多尺度特征。

多层次类别抽象用于建立简单概念到复杂概念的层级关系。复杂视觉概念的多层次结构表示有助于概念分解与重组、快速概念拓展与新视觉概念理解。复杂概念往往由简单概念经过非线性组合构成。有效利用视觉概念间的层次先验,考虑多尺度多任务关系,可有效降低模型训练难度,提升视觉概念表达的丰富度。

多模态主次分析利用人类感知过程中以视觉信息为主导的特点,采用视觉信号主导,并以其他信息如声音和语言进行辅助。一般认为,人类接收的信息大部分来自视觉信号(图像或视频),承载了更丰富更细致的感官信息。然而传统方法建模各媒体信

息时仅考虑其并列关系，未考虑模态主次信息。本文认为人工智能尽管与人类智慧有很大差异，但发展跨媒体智能时，以视觉信息为主导、其他信息辅助的特点依然值得借鉴。

3）动作示能。视觉概念除了描述事物的形状、色彩和语义等，还需表达人类与物体间的交互关系。Gibson（1977）指出示能是环境或物体的可供性，即环境或物体可以提供的功能或用途。物体的示能表征了物体与人类间潜在的可交互行为。例如，杯子具有“可握”的示能，椅子具有“可坐”的示能。视觉概念包含动作示能的理解，主要涉及物体形状、语义与人类动作的关联。例如，Nagarajan 等人（2020）提出描述环境示能的拓扑图结构，有效预测未来可能发生的动作。在具身视觉问题上，Fan 等人（2022）在人类操纵物体时不仅考虑手和物体，同时引入人类意图作为参考，对视觉动态和对象位置变化进行建模，从而有效识别交互动作。Wang 等人（2020b）在动作与物体间通过共生注意力机制进行联合时空关系推理，实现更准确的具身交互理解。

具备典型与范畴结构、层次结构和动作示能的视觉概念将大幅度提升模型鲁棒性，实现人机交互、增强现实等场景下的高效应用。

2.1.2 视觉关系

在视觉概念的表达上，视觉关系表示了视觉概念间的关联情况，为更为复杂的跨媒体视觉推理提供有效基础。视觉关系包含空间关系、时序关系与因果关系，但并不仅限于此。传统视觉关系旨在捕获图像中成对物体间的各种交互。研究人员应着重将成对视觉关系拓宽至多物体级联关系，并有效统一跨媒体时空与因果表达。

1）空间关系。视觉内容中最常见的空间关系包含显式的位置关系或隐式的动作关系。常见的位置关系包括“在……之上”、“在……旁边”等。隐式动作关系描述了物体间或物体与人类间的动作相关的位置信息。如“骑”描述了物体甲在物体乙之上，并且表现出“骑”的动作（“骑在马上”或“骑在自行车上”）。Krishna 等人（2017）引入一个大规模的空间关系数据集用于视觉关系建模。视觉关系的识别往往需考虑配对物体间的相关性。例如，Chang 等人（2018）使用关系网络进行场景中人物关系建模。Zheng 等人（2020a）统一了地面视角、无人机视角和卫星视角的视角表达。这些研究主要关注图像中的静态关系。

2）时序关系。时序变化为视觉关系在时间维度上带来了多样性。时空联合关系建模带来了诸多挑战。其中涉及到单个物体在时序上的变化，以及物体间空间关系在时序上的变化，这种动态变化的集合构成了物体间复杂且细微的时序关系。动态时序关系包含人类社交、物体运动等动态关系。例如，物体逐渐靠近墙面，随后，物体碰撞墙面后，开始远离墙面。这类关系的表达需有效理解运动信息并捕捉运动情况的变化。Ji 等人（2020）引入一个大规模时空关系数据集用于时空语义关系建模。Fan 等人（2020）基于动态点云对时间与空间进行解耦，从而对 3 维空间中的运动进行建模与理解。

3）因果关系。因果关系是事件原因与结果的联系。跨媒体数据往往存在视觉偏差。数据偏差指在数据集中某些成分比其他成分出现的比例、权重更大。数据偏差不仅降低了训练模型的预测精度，有时甚至会导致公平性方面的问题。例如，人脸识别训练数据中某种人比例偏多时，会导致模型对其他人种的识别不够友好。这种数据偏差问题在跨媒体智能中尤为严重，因为跨媒体数据的主要获取渠道之一是互联网，而互联网上的数据是非规范化的，存在严重的数据分布不均衡，甚至存在局部数据重复与错误标注等问题。因果关系的建模有助于消除嵌入在跨媒体数据中的偏差，并量化事物间的因果影响。干预和反事实推理是提供无偏预测的常用工具。

2.1.3 视觉推理

视觉知识建立在视觉概念与视觉关系的基础上，可赋能更具解释性与泛化性的抽象跨媒体视觉推理。视觉知识提供了结构化知识表达，为视觉推理提供多方面的解释基础，对神经网络决策背后的推理逻辑提供有效解释。本文的视觉推理广义上包含各种视觉操作，如预测、重建、联系和分解等。

1）视觉知识分解。视觉概念包含层次与结构，具有分解性与合成性。通过简单概念的组合，人类可以构造复杂概念并创建多功能系统。另一方面，人类可以快速将复杂事物进行分解，并将陌生的事物分解为熟悉的组件。视觉知识分解旨在捕获视觉内容中显著或具有解释性的因素，将抽象知识解藕成独立、易解释的概念。研究视觉知识分解的机理有助于深刻理解数据生成过程及其潜在的因果关

系,帮助提炼重要的视觉信息,并创建更泛化的知识表达。基于自监督的视觉分解研究或是大规模自动提取视觉知识的有效途径。

2)知识类比联想。人类具有识别概念之间关系并类比推断至超越已有概念的能力。知识类比是推理的重要步骤。例如,玫瑰之于花,相当于猫之于什么?人类可以推理出答案是动物,并理解玫瑰之于花为从属关系。类比联想涉及对视觉知识的操作,但并不限于2.1.2节提到的空间、时序与因果关系。基于类比联想的推理方式通过实例组合的形式,将隐式关系包含在推理过程中。类比联想的研究将视觉知识中的关系建模推广到逻辑关系、从属关系等更为抽象的关系。

3)视觉知识重建。视觉知识重建指根据视觉知识表达重构出原始视觉内容。视觉重建过程是视觉知识表达的逆过程。视觉知识重建不仅需要重建视觉概念的形状、结构等典型信息,且需根据视觉概念的范畴进行可控的多样性内容生成。视觉知识重建不仅包含静态2维图像、3维几何生成,也包含连续动作变化的模拟。视觉知识重建亦可用于视觉知识表达质量的评估,并为可解释视觉概念提供有效工具。

2.2 视觉知识应用于跨媒体智能的研究思路

视觉知识理论旨在建立统一的知识表达方式。一些研究尽管尚未正式引入视觉知识的概念,但视觉知识的概念、优势和特点已初步运用于跨媒体智能,并取得了良好效果。本文认为,将视觉知识应用于新的跨媒体智能任务能够带来潜力,选取视觉知识的结构化表示、视觉知识的操作与推理和跨媒体重建与生成等3项任务展开样例研究,讨论视觉知识的应用并进行分析与展望。值得指出的是,这3项具体任务仅是视觉知识在跨媒体智能中的部分例子,还存在其他更多的相关任务。随着基于视觉知识的跨媒体智能的发展,将出现更多全新、更具挑战性的跨媒体智能应用和任务。

2.2.1 视觉知识的结构化表示

结构化特征具备可解释属性。虽然手工构造特征对多媒体数据进行抽象和刻画的能力相比基于深度学习的特征较弱,但这些描述子往往具有更强的可解释性。有效结合手工构造特征和深度学习特征进行跨媒体表达建模是视觉知识表达在跨媒体智能领域的一个典型应用。视觉知识的结构化表示有助于模型可信性与可解释性的提升。

1)结构关系图表示。视觉概念的表示通常可以依赖于一些符号化的表达以修饰视觉概念主体。例如,Pan(2019)指出苹果是由果核、果肉、果皮和果蒂等子概念组成的结构。这种空间结构关系可以由自然语言或关系图的形式表达,以获得对苹果视觉概念更准确的描述。这些可以符号化表达的信息表达都是建立在视觉信息主体之上的,与人类以视觉信息为主导进行信息获取的特点更加接近。

较为常见的关系图结构是场景图表示。场景图是视觉知识的一种表达形式,以带权有向图的结构形式表征图像中所有对象及对象间的交互关系,如图2所示。其中,图的节点表示对象实体,图的边表示对象间关系。节点上含有对象类别等信息,每条边含有关系的主客体指向和关系类型等信息。相比于图像分类、目标检测等视觉任务,场景图需要更高的感知层次,充分描述图中对象实体信息和交互信息。具体而言,场景图需要先获得局部特征表达,再对多个局部特征表达进行组织与整理,归纳出每个局部与其他局部、局部与全局之间的关系,形成层次化、结构化的知识表达。场景图因优良的结构化特性和丰富的语义信息量,在看图说话(Li等,2017)、视觉问答系统、图像生成(Johnson等,2018)和图像检索(Johnson等,2015)等任务中均有广泛研究。

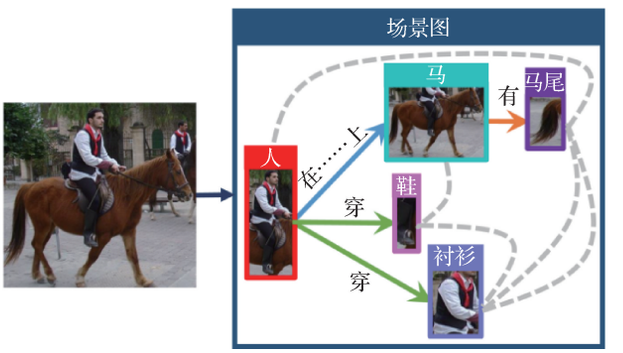


图2 视觉知识的场景图

Fig. 2 The scene graph representation of visual knowledge

为了满足场景图对局部间信息交互的需求,现有研究(Xu等,2017)充分利用场景图的连接特性,提出将图中不同节点的信息充分交互,使每一个节点都能获取、处理和传递全图其他节点信息,实现了基于图的全图语义理解。Yang等人(2018)则尝试从图的角度处理场景中信息,先选择可行候选节点,后使用图网络串连处理。Zareian等人(2020)将场

景图定义为基于图像内容的知识图谱,并将其作为图像与知识图谱之间的桥梁,实现场景与知识图谱的细粒度异构互连。与 Zareian 等人(2020)不同的是,视觉知识旨在解决传统视觉交互描述的模糊性以及歧义性问题,除了借助知识图谱中的丰富语义信息,同时利用视觉概念的动作要素,实现局部深度特征和相对关系的符号化知识表达,降低视觉场景与知识图谱关联时的模糊性与歧义性。

在目前研究中,未能解决的一个关键问题是交互描述的模糊性。自然语言描述是具有模糊性并易于产生歧义的。同样的场景,相同的交互可以存在不同的多种描述。例如,表示位置交互时,“坐在……上”、“在顶端”等在一定程度上可以互换。再例如,考虑到图像作为一种 2 维数据,从 2 维视角来看,任何图像场景内的关系均可以标注为“在……旁边”。这种描述模糊性直接反映在数据中,就会存在大量的歧义性标注。仅根据标注进行纯数据驱动学习则会导致场景图的描述产生歧义或错误。

视觉知识有望改善甚至解决这种关系描述中的歧义问题。在视觉知识中,视觉概念天然具有动作这一要素(Pan,2019),例如动物的头、肢和躯等结构包含动作及其关系等表达。动作这一要素能够更准确地刻画不同局部与局部之间的关系。例如,“坐在……上”和“在顶端”两个相近的描述,通过局部深度特征和相对位置的符号化知识表达依然很难区分,而通过动作相关的描述,能够直接分析与“坐”动作相关的状态,如腿是否弯曲、关节是否打开等,从而区分“坐在……上”和“在顶端”两种状态。

2)知识先验。视觉知识对数据偏差可进行有效抑制。例如,在人脸识别中可以通过注入“肤色无关身份确认”的抽象知识来矫正不均衡数据集上人脸识别模型偏向个别肤色的偏差。Tang 等人(2020)利用结构化表达(类似视觉知识的结构信息)减少推断偏差。Li 等人(2021)同时利用文本和视觉内容描述事件特征,进行网络谣言检测,利用视觉信息抑制文本中可能出现的不准确描述的偏差。传统的 3DMM(3D morphable models)(Blanz 和 Vetter,1999)对人脸或身体进行参数化建模,可以作为附加输入生成高质量人脸或动画。在自动驾驶场景中,Li 等人(2020)将空间位置的频率信息作为分布先验,筛选出能够促进域迁移的正样本,并减少负样本与迁移的影响。Yang 等人(2012a)在多标签分类过程中考虑标签相关性用于建模类间共享结构。Li 等人(2021)利用元学习实现属性调制的零样本学习,通过属性感知、属性增强和属性加权,减轻模型的固有偏置,实现自适应属性增强。视觉知识也旨在建模视觉属性,以提升局部知识表征的可解释性。与 Li 等人(2021)不同的是,视觉知识包含更丰富的视觉关系和结构化特性。视觉知识考虑多物体级联关系,可有效统一跨媒体时空与因果表达,实现属性知识间多样化的结构表达。

2.2.2　视觉知识的操作与推理

除了视觉知识表示方式的研究,视觉知识的操作及视觉知识推理也是研究重点之一。视觉知识的操作包括重建、关联等,具体指基于视觉概念与视觉关系的运算与推理过程。

1)知识对齐与关联。视觉知识不仅可以作为主导模态信息,更可以作为其他模态的监督信息,或是连接其他多模态的纽带和桥梁。在现实中,不同文化背景的人们对音频和文本信息的接受能力是不相同的,但不可否认的是,人类是共享一套视觉系统的。不同国家用来描述同一个物体的代词是不相同的,例如苹果和 apple。但对于中国人和英国人来说,他们看到的苹果这个物体应该是相似的。因此视觉知识可以作为一个桥梁来连接不同模态之间的信息。例如在翻译任务中,视觉信息可以作为一种弱监督的信用来监督不同语言信息的对齐。苹果这个物体和吃苹果这个动作对于不同语言背景下的人们而言,接收到的视觉知识是相似的。

视频和自然语言获取跨媒体表达的常见方式是先用卷积神经网络抽取视频特征,再使用循环神经网络获取语言文本特征,然后在两者的基础上进行联合序列融合,在高层建立视频和自然语言文本的语义关联(Zhu 等,2017)。这种模型常常需要大量的视频—文本标注数据进行训练。除此之外,由于部分视频中常常存在语言讲解,这些视频与语言之间的同步关系可以用于自监督学习。最近的一些研究,如 Zhu 和 Yang(2022)通过设计一些代理任务,如补全句子中的词、补全视频中图像帧或视频文本的相关性判断等,实现了从大量无标签视频数据中自学习视频和语言的统一特征表达。

视频与音频具有极强的内在关联性,因为视频数据中声音和图像几乎是天然同步的(Wu 等,2019),因此易于挖掘它们与额外的音频信号之间

的关联性。Wu 和 Yang(2021)研究弱监督条件下的视听视频解析,并通过音频和视觉轨道交换减少不明确标签的误导。

互联网上大量的跨模态信息以视觉知识为纽带或弱监督的信号标签来完善多模态的信息表征能力。因为有的信息是图像配合文本信息,有的是视频配合音频信息,那么对于具有相似的视觉特征的互联网信息,就可以将附着的其他模态信息进行对齐和联合,从而促进多模态的信息表征准确度。

2)多重知识融合。视觉知识自身具有多重表达的特点(Pan,2020;Yang 等,2021),能够通过组合多种知识表达方式(如手工特征表达、深度特征表达、符号化知识表达和知识图谱等)实现对多种来源、多种模态数据的多层次抽象与知识提取。这些特点使视觉知识为跨媒体表达提供了新的有效途径。

多重知识表达能力是视觉知识的内禀性质之一。为了对多种不同模态、不同来源的信息进行融合,多重知识表达需要根据这些信息的抽象层次由浅及深进行抽象与融合。从多重知识表达的观点来看,当前深度表达较多地关注相对较低的抽象层次,如纹理、形状和颜色等可感知信息(这也是目前分类、识别等感知型人工智能任务偏好深度学习方法的原因)。相比于深度表达,符号化知识表达和知识图谱则强调关注相对较高的抽象层次,如描述事件的规则、过程、不同概念或物体之间的相互关系。多重知识表达在融合不同知识时,可以考虑它们对应的抽象层次,以抽象层次由低到高的顺序进行融合。Quan 等人(2021)考虑了用于行人轨迹预测的多源信息,包括结构化深度信息、车辆速度、场景相关性和人类意图知识等,通过自适应相互作用机制挖掘了多个线索之间的内在关系。

知识图谱是一种揭示实体之间关系的语义网络,由相互连接的节点和边组成,每个节点表示一个概念或对象,边表示它们之间的关系。每两个相连节点和它们之间的边构成的 SPO(subject predicate object)三元组即构成了一条相应的知识。传统知识图谱中的对象通常为文本描述的抽象概念。多媒体知识图谱在其基础上,为对象主体添加更多的描述形式,如图像、视频和音频等。得益于互联网的迅猛发展,大量的真实数据在互联网中逐渐积累沉淀,为构建一个完备的多媒体知识图谱提供了基础。现阶段,已有 DBpedia(Auer 等,2007)、Wikidata(Vrandečić和 Krötzsch, 2014)、IMGpedia(Ferrada 等, 2017)和 MMKG(multi-modal knowledge graphs)(Liu 等, 2019)等工作通过互联网收集对应实体的多媒体信息(如图像、多语言文本描述等)来构建多媒体知识图谱。以 DBpedia 为例,其为每一个待描述实体提供一个唯一的全局标识符,并借助互联网上大量的开发者自行上传的多种模态的数据类型构建、维护这一多媒体知识图谱。借助互联网去中心化的自由发展方式,多媒体知识图谱获取并保有了远超传统知识图谱的信息量。相比传统的知识图谱,能够更加广泛地存储记录针对实体的多个媒质的多种形态的信息,提供更加丰富完善的信息描述。

多媒体知识图谱成为跨媒体知识的一种全新表现形式有两方面原因:一是对于一个对象,多媒体知识图谱可能为其提供包含图像、视频、音频和文本说明等多种模态的描述,符合视觉知识具备的多重知识表达性质。二是得益于知识图谱的内禀,它具备刻画各个对象之间的关系的能力。这种关系可能是诸如“动物→哺乳动物”这种抽象概念之间的层次关系,也可能是诸如“脸包含眼睛、鼻子、嘴”等动作结构关系。它们分别形成了构建视觉知识所需的层次结构和动作结构(Pan,2019)。

2.2.3 跨媒体重建与生成

视觉生成是用计算机图形学和计算机视觉技术生成单个或多个物体的图像、视频的技术,在数据可视化(Klawonn 等,2003;Rehm 等,2006)、计算机动画(Parent,2012)、虚拟现实(Kim,2005)和增强现实(Hainich,2006)等领域得到广泛应用。在视觉生成中,解析生成对象的部件结构,有助于获得外观、形态逼真的生成效果。而视觉知识恰好提供了这种支持。因为每一个视觉概念包含部件空间结构关系,有关动物的视觉概念则还应该有其对应常见动作的动作结构,这种视觉结构在视觉生成中能够发挥重要作用。

1)图像与动画生成。在深度学习广泛应用之前,传统计算机视觉、图形学在动画生成中就已经实质性地运用了视觉知识,例如人脸动画生成。人脸的外貌是由面颌骨骼以及覆盖之上的脸部肌肉、皮肤共同决定的。从视觉知识的观点来看,这些骨骼、肌肉和皮肤部件正是人脸这一视觉概念的层次结构。面部关节的运动、肌肉的紧张以及表情的变化

能够给同一张人脸带来丰富的外貌表现,因此它们同时也控制了人脸的动作。从这个角度来看,一些工作将人脸这一视觉概念进行层次化、动作化分解。

一些基于深度学习的 3 维重建与生成研究认为,3 维生成应该遵循光线在 3 维空间内传播的物理规律,并产生了 NeRF(neural radiance fields)(Mildenhall 等, 2020)和 GRAF(generative radiance fields)(Schwarz 等,2020)等工作,这些工作的共同点是都使用了神经网络对 3D 辐射场进行学习。而 3D 辐射场以及光线在 3D 空间内传播的设定正是对视觉知识的合理运用。

在跨媒体生成领域,Jain 等人(2022)将神经网络 3 维渲染与多模态图像文本表示相结合,仅利用自然语言描述合成各种 3 维物体。为了提高保真度和生成质量,Jain 等人(2022)引入了稀疏诱导透射率正则化、场景边界等几何先验,实现从各种自然语言字幕中生成逼真的、多视图一致的物体几何形状和颜色。通过自然语言描述,该方法允许用户通过易于创作的提示来控制生成结果的样式和形状,包括物体的材料和类别。这类方法可以为艺术家和跨媒体应用程序提供更快速的创作内容。

2)跨媒体内容生成。深度学习的快速发展,特别是生成对抗网络(generative adversarial networks, GAN)(Goodfellow 等,2014)的出现使基于深度神经网络的生成任务成为可能。深度学习在多种媒体的生成任务,如自然语言生成、图像生成和3 维模型生成等都有一套独立的学习机制,但它们在历史上是分别发展的,在不同媒体间通常很少互动。与人类相比,深度生成模型从一种模态到另一种模态之间建立新连接方面不够灵活。换句话说,深度模型一旦学会了从一种模态生成样本,使其用于生成以另一种模态为控制条件的样本通常很难并且可能需要重新训练模型。当前,深度学习方法针对的生成场景主要有文本到图像、图像到文本、文本到视频(Li 等,2018)和视频到文本(Pan 等,2016)几种形式,效果远不及人类感知的程度。因此生成模型还有很大的研究价值和提升空间。

由知识驱动的方法逐渐受到广泛关注(Gogoglou 等,2019)。由知识驱动的生成方法主要思路是将输入文本之外的各种形式知识纳入生成模型。因此,一个结构化表示的知识体系是该类方法的核心,即利用额外的知识驱动图像内容的生成,如生成物体位置、属性和类别等。这种结构化表示同样符合视觉知识对视觉概念的描述要求。另一种方法是由语言引导生成场景图作为中间表达,进而生成结构化的场景图像。Johnson 等人(2018)首次提出一个基于场景图的图像生成网络模型,使用图卷积处理输入场景图,根据物体的边界框等计算场景布局,然后将布局用级联细化网络转换成图像。这个网络整体构架基于生成对抗网络确保输出的图像真实自然。Zhu 等人(2019)根据初始图像内容选择重要的文本信息,引入动态记忆模块细化模糊图像内容,根据文本描述更准确地生成图像。

2.3 视觉知识的优势

基于视觉知识实现的跨媒体智能,具有以下 3 方面优势。

1)数据与知识驱动的统一框架。目前,前沿研究大多基于端到端的训练方式,容易受到数据偏差和优化算法的影响,鲁棒性与泛化性均有欠缺,原因在于缺乏规则与知识驱动的推理框架。视觉知识可以融合数据驱动与知识驱动的优势,实现形式化推理与表征学习的共存机制。

2)可归因可溯源的结构化表达。深度学习的黑箱性质使决策模型的可解释性较差。视觉知识为增强跨媒体智能可解释性提供了有力的手段和线索。视觉知识刻画了抽象概念之间的详尽关系,而这种刻画本身又是经过了人类抽象总结并易于理解。因此,一方面可以用深度学习对多媒体数据本身进行特征抽取;另一方面又可以依赖这些抽象关系将特征进行组织,从而融合数据驱动的浅层表观信息以及人类高度抽象的逻辑性知识,最终增强跨媒体智能的可解释性。吕露露等人(2021)利用心率和加速度等信号与运动强度的相关性,将传感器监测得到的心率和加速度等特征融入到动作表达中,提高了人体动作识别准确率,并具备较好的可解释性。成科扬等人(2021)融入了多人交互中肢体变化的规律并通过对骨架运动进行时空建模,对多人交互行为进行识别。

综上所述,结构化表征学习已应用于一些实际场景(Zareian 等, 2020; Xu 等, 2017; 吕露露 等, 2021;成科扬 等,2021),但自动化、可适配的结构化表达方法目前仍需进一步探索。传统结构化场景描述存在交互模糊与歧义的问题,造成视觉表达鲁棒性降低、精确度不足,从而影响模型可解释性。视觉

知识旨在改善场景描述的模糊性与歧义性,提升其可解释性,借助知识图谱、视觉概念和视觉关系等互补语义信息,有望实现自动化、可适配的结构化表达。

3)跨媒体知识关联与智能推理。视觉知识不仅能够融合多媒体信息,还能够利用不同媒体信息进行相互增强和推理。这种能力的基础来源于视觉知识更强的抽象表达能力和多重知识之间本身具有的互补能力。例如,将深度学习表达与符号化表达结合形成视觉知识时,深度表达能够为后者抽象细节化的表观特征,而符号表达则能够引导深度学习表达获得更好的泛化能力。视觉知识中包含的视觉概念的层次关系为知识迁移提供了一种丰富、可依赖的线索,而知识迁移是提高模型泛化能力的有效手段。具体来讲,知识迁移可以将已学习的知识迁移并应用到新的问题、任务上。而这个新的问题或任务一般与先前的任务相关,但通常又不严格相同或具有明显的域差异。例如,Roy 等人(2020)将猫科动物的外观多样性迁移到相邻的美洲狮这一视觉概念上,生成了外观多样的美洲狮,如图 3 所示。Roy 等人(2020)实现了动物间的风格迁移,但未考虑跨域差异性。视觉知识旨在利用多媒体知识图谱中的动物语义关系,为知识迁移提供可依赖、细粒度的线索。这个知识迁移可以依赖于猫与美洲狮在多媒体知识图谱中是相近节点这一事实。再如,在行人识别研究中,Miao 等人(2021)使用人体各部件之间的动作结构知识,改善了遮挡条件下行人检索的准确度。在车辆识别研究中Zheng 等人(2020b)使用

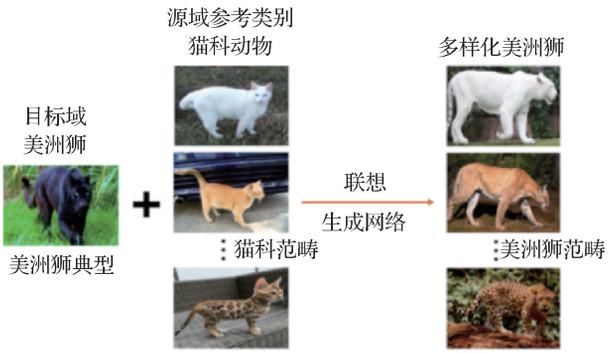


图 3　Roy 等人(2020)通过猫科动物的外观多样性,设计算法生成多样化美洲狮图像

Fig. 3　Roy et al. (2020) migrated the appearance of cats to the visual concept of cougar, resulting in a cougar with various appearances

GAN 生成对应环境下的车辆图像,通过预训练和微调改善了不同环境下车辆检索的准确度。

上述 3 项优点弥补了当前跨媒体知识表达的不足,能够为跨媒体智能提供更加合理、灵活和精细的跨模态建模。

3　视觉知识研究的展望

3.1　联合判别式与生成式学习的表达范式

判别式模型一般用于物体识别和检测,生成式模型一般用于内容生成、预测和合成。现阶段,两种模型并未有效统一整合。视觉知识旨在进一步利用生成式模型的输出,辅助判别式模型进行联合表达学习,完成判别式模型与生成式模型的高效协同训练。整合判别式与生成式模型有助于提升可解释能力,渐进地增强模型鲁棒性。

在视觉知识的联合训练框架下,未来将建立视觉知识的分解、变换、重建与合成理论。视觉知识的分解旨在获得视觉概念的组成部分。视觉知识的变换可实现视觉实体的操作与模拟,探索视觉知识重建与新知识的合成。另外,大规模跨媒体视觉知识数据集的收集和整理是未来重点工作之一,这类数据库应整合专家知识和人类先验,丰富原始数据。对于如何构建这类数据库,未来仍有广阔探索空间。

3.2　模拟仿真技术的突破

视觉知识将极大改善生成模型的效果,提升仿真技术的逼真度。未来可利用视觉知识与场景特性作为先验提升场景的表达和渲染,实现交互式的视觉编辑工具与可控的场景物体语义理解。结合数据驱动与视觉知识的图形学系统将融合数据与规则的长处,抽取视觉数据典型的语义特征,降低模型复杂度,提升仿真效率,有效产生新视角与新场景下的逼真及连续的内容。生成与仿真技术的突破将在娱乐、工业和医疗等各行业作出重要贡献。

3.3　可信跨媒体智能

视觉知识理论是提升跨媒体智能鲁棒性、泛化性和可解释性的研究基础。视觉知识理论的建立是迈向可信跨媒体智能的重要一步,将有效缓解数据算法歧视和数据偏置歧视和偏见,减轻决策偏差,提升模型的公平性。同时,视觉知识应具备稳定的进化机制,为新知识归纳、新场景理解提供终身学习能力。在司法、医疗等一些关键领域利用视觉知识结

构化、多粒度的表征解析能力和整合多源、跨领域数据的优势，使决策辅助过程可信可靠、推理过程可复查可解释，系统性地提升模型泛化能力，为可信跨媒体智能提供重要保障。

4 结　语

本文回顾了跨媒体智能已有的手工设计方法与深度学习表达方式，分析了当前跨媒体发展的瓶颈。本文以场景图、跨媒体知识图谱等研究方法为例，分析了视觉知识在跨媒体智能中的应用与优势。通过回顾与分析可以看出，具有多重知识表达能力及层次化、动作结构化的视觉知识提供了当前跨媒体智能发展亟需的重要元素。视觉知识将成为跨媒体智能进化的新支点。

参考文献(References)

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R and Ives Z. 2007. DBpedia: a nucleus for a web of open data//Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference. Busan, Korea (South): Springer: 722-735 [DOI: 10.1007/978-3-540-76298-0_52]

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. [s.l.]: ACM Press/Addison-Wesley Publishing Co.: 187-194 [DOI: 10.1145/311535.311556]

Chang X J, Huang P Y, Shen Y D, Liang X D, Yang Y and Hauptmann A G. 2018. RCAA: relational context-aware agents for person search//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 86-102 [DOI: 10.1007/978-3-030-01240-3_6]

Cheng K Y, Wu J X, Wang W S, Rong L and Zhan Y Z. 2021. Multi-person interaction action recognition based on spatio-temporal graph convolution. Journal of Image and Graphics, 26(7): 1681-1691 (成科扬, 吴金霞, 王文杉, 荣兰, 詹永照. 2021. 融合时空图卷积的多人交互行为识别. 中国图象图形学报, 26(7): 1681-1691) [DOI: 10.11834/jig.200510]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Fan H H, Yu X, Ding Y H, Yang Y and Kankanhalli M. 2020. PSTNet: point spatio-temporal convolution on point cloud sequences [EB/OL]. [2022-06-22]. https://arxiv.org/pdf/2205.13713.pdf

Fan H H, Zhuo T, Yu X, Yang Y and Kankanhalli M. 2022. Understanding atomic hand-object interaction with human intention. IEEE Transactions on Circuits and Systems for Video Technology, 32(1): 275-285 [DOI: 10.1109/TCSVT.2021.3058688]

Ferrada S, Bustos B and Hogan A. 2017. IMGpedia: a linked dataset with content-based analysis of Wikimedia images//Proceedings of the 16th International Semantic Web Conference. Vienna, Austria: Springer: 84-93 [DOI: 10.1007/978-3-319-68204-4_8]

Frome A, Corrado G S, Shlens J, Bengio S, Dean J, Ranzato M and Mikolov T. 2013. DeViSE: a deep visual-semantic embedding model//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 2121-2129

Gibson J J. 1977. The Theory of Affordances. Hillsdale: Erlbaum Associates: 67-82

Gogoglou A, Bruss C B and Hines K E. 2019. On the interpretability and evaluation of graph representation learning [EB/OL]. [2022-06-22]. https://arxiv.org/pdf/1910.03081.pdf

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Hainich R R. 2006. The End of Hardware: A Novel Approach to Augmented Reality. [s.l.]: BookSurge Publishing

Hardoon D R, Szedmak S and Shawe-Taylor J. 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 16(12): 2639-2664 [DOI: 10.1162/0899766042321814]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Jain A, Mildenhall B, Barron J T, Abbeel P and Poole B. 2022. Zero-shot text-guided object generation with dream fields [EB/OL]. [2022-06-22]. https://arxiv.org/pdf/2112.01455.pdf

Ji J W, Krishna R, Li F F and Niebles J C. 2020. Action genome: actions as compositions of spatio-temporal scene graphs//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10233-10244 [DOI: 10.1109/CVPR42600.2020.01025]

Johnson J, Gupta A and Li F F. 2018. Image generation from scene graphs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1219-1228 [DOI: 10.1109/CVPR.2018.00133]

Johnson J, Krishna R, Stark M, Li L J, Shamma D A, Bernstein M S and Li F F. 2015. Image retrieval using scene graphs//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3668-3678 [DOI: 10.1109/CVPR.

2015.7298990]

Kim G J. 2005. Designing Virtual Reality Systems. London: Springer [DOI: 10.1007/978-1-84628-230-0]

Klawonn F, Chekhtman V and Janz E. 2003. Visual inspection of fuzzy clustering results//Proceedings of 2003 Advances in Soft Computing. London, UK: Springer: 65-76 [DOI: 10.1007/978-1-4471-3744-3_7]

Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L J, Shamma D A, Bernstein M S and Li F F. 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1): 32-73 [DOI: 10.1007/s11263-016-0981-7]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 1097-1105

Lazebnik S, Schmid C and Ponce J. 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). New York, USA: IEEE: 2169-2178 [DOI: 10.1109/CVPR.2006.68]

Li G R, Kang G L, Liu W, Wei Y C and Yang Y. 2020. Content-consistent matching for domain adaptive semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 440-456 [DOI: 10.1007/978-3-030-58568-6_26]

Li K P, Zhang Y L, Li K, Li Y Y and Fu Y. 2019. Visual semantic reasoning for image-text matching//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4653-4661 [DOI: 10.1109/ICCV.2019.00475]

Li Y, Liu Z, Yao L N and Chang X J. 2021. Attribute-modulated generative meta learning for zero-shot learning. IEEE Transactions on Multimedia [DOI: 10.1109/TMM.2021.3139211]

Li Y K, Ouyang W L, Zhou B L, Wang K and Wang X G. 2017. Scene graph generation from objects, phrases and region captions//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1270-1279 [DOI: 10.1109/ICCV.2017.142]

Li Y T, Min M R, Shen D H, Carlson D and Carin L. 2018. Video generation from text//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 7065-7072

Liu Y, Li H, Garcia-Duran A, Niepert M, Onoro-Rubio D and Rosenblum D S. 2019. MMKG: multi-modal knowledge graphs//Proceedings of the 16th International Conference on Semantic Web. Portorož, Slovenia: Springer: 459-474 [DOI: 10.1007/978-3-030-21348-0_30]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Lyu L L, Huang Y, Gao J Y, Yang X S and Xu C S. 2021. Multimodal-based zero-shot human action recognition. Journal of Image and Graphics, 26(7): 1658-1667 (吕露露, 黄毅, 高君宇, 杨小汕, 徐常胜. 2021. 多模态零样本人体动作识别. 中国图象图形学报, 26(7): 1658-1667) [DOI: 10.11834/jig.200503]

Miao J X, Wu Y and Yang Y. 2021. Identifying visible parts via pose estimation for occluded person re-identification. IEEE Transactions on Neural Networks and Learning Systems: #3059515 [DOI: 10.1109/TNNLS.2021.3059515]

Mikolov T, Chen K, Corrado G and Dean J. 2013. Efficient estimation of word representations in vector space. [2022-06-22]. https://arxiv.org/pdf/1301.3781.pdf

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Nagarajan T, Li Y H, Feichtenhofer C and Grauman K. 2020. EGO-TOPO: environment affordances from egocentric video//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 160-169 [DOI: 10.1109/CVPR42600.2020.00024]

Pan P B, Xu Z W, Yang Y, Wu F and Zhuang Y T. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1029-1038 [DOI: 10.1109/CVPR.2016.117]

Pan Y H. 1996. The synthesis reasoning. Pattern Recognition and Artificial Intelligence, 9(3): 201-208 (潘云鹤. 1996. 综合推理的研究. 模式识别与人工智能, 9(3): 201-208)

Pan Y H. 2019. On visual knowledge. Frontiers of Information Technology and Electronic Engineering, 20(8): 1021-1025 [DOI: 10.1631/FITEE.1910001]

Pan Y H. 2020. Multiple knowledge representation of artificial intelligence. Engineering, 6(3): 216-217 [DOI: 10.1016/j.eng.2019.12.011]

Parent R. 2012. Computer Animation: Algorithms and Techniques. 3rd ed. San Francisco, USA: Morgan Kaufmann

Quan R J, Zhu L C, Wu Y and Yang Y. 2021. Holistic LSTM for pedestrian trajectory prediction. IEEE Transactions on Image Processing. 30: 3229-3239 [DOI: 10.1109/TIP.2021.3058599]

Rehm F, Klawonn F and Kruse R. 2006. POLARMAP-Efficient visualisation of high dimensional data//Proceedings of the 10th International Conference on Information Visualisation (IV'06). London, UK: IEEE: 731-740 [DOI: 10.1109/IV.2006.85]

Roy V, Xu Y, Wang Y X, Kitani K, Salakhutdinov R and Hebert M. 2020. Few-shot learning with intra-class knowledge transfer [EB/OL]. [2022-06-22]. https://arxiv.org/pdf/2008.09892.pdf

Schwarz K, Liao Y Y, Niemeyer M and Geiger A. 2020. GRAF: generative radiance fields for 3D-aware image synthesis//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 20154-20166

Snell J, Swersky K and Zemel R. 2017. Prototypical networks for few-shot learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4080-4090

Tang K H, Niu Y L, Huang J Q, Shi J X and Zhang H W. 2020. Unbiased scene graph generation from biased training//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3713-3722 [DOI: 10.1109/CVPR42600.2020.00377]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vrandečić D and Krötzsch M. 2014. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10): 78-85 [DOI: 10.1145/2629489]

Wang H and Schmid C. 2013. Action recognition with improved trajectories//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 3551-3558 [DOI: 10.1109/ICCV.2013.441]

Wang W Y, Tran D and Feiszli M. 2020a. What makes training multi-modal classification networks hard?//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12692-12702 [DOI: 10.1109/CVPR42600.2020.01271]

Wang X H, Zhu L C, Wang H and Yang Y. 2021. Interactive prototype learning for egocentric action recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8148-8157 [DOI: 10.1109/ICCV48922.2021.00806]

Wang X H, Zhu L C, Wu Y and Yang Y. 2020b. Symbiotic attention for egocentric action recognition with object-centric alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence: #3015894 [DOI: 10.1109/TPAMI.2020.3015894]

Wang Z H, Liu X H, Li H S, Sheng L, Yan J J, Wang X G and Shao J. 2019. CAMP: cross-modal adaptive message passing for text-image retrieval//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5763-5772 [DOI: 10.1109/ICCV.2019.00586]

Wu Y and Yang Y. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1326-1335 [DOI: 10.1109/CVPR46437.2021.00138]

Wu Y, Zhu L C, Jiang L and Yang Y. 2018. Decoupled novel object captioner//Proceedings of the 26th ACM international conference on Multimedia. Seoul, Korea (South): ACM: 1029-1037 [DOI: 10.1145/3240508.3240640]

Wu Y, Zhu L C, Yan Y and Yang Y. 2019. Dual attention matching for audio-visual event localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6291-6299 [DOI: 10.1109/ICCV.2019.00639]

Xu D F, Zhu Y K, Choy C B and Li F F. 2017. Scene graph generation by iterative message passing//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3097-3106 [DOI: 10.1109/CVPR.2017.330]

Yang J W, Lu J S, Lee S, Batra D and Parikh D. 2018. Graph R-CNN for scene graph generation//Proceedings of the 15th European Conference Computer Vision. Munich, Germany: Springer: 690-706 [DOI: 10.1007/978-3-030-01246-5_41]

Yang Y, Ma Z G, Xu Z W, Yan S C and Hauptmann A G. 2013a. How related exemplars help complex event detection in web videos?//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 2104-2111 [DOI: 10.1109/ICCV.2013.456]

Yang Y, Nie F P, Xu D, Luo J B, Zhuang Y T and Pan Y H. 2012a. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4): 723-742 [DOI: 10.1109/TPAMI.2011.170]

Yang Y, Song J K, Huang Z, Ma Z G, Sebe N and Hauptmann A G. 2013b. Multi-feature fusion via hierarchical regression for multimedia analysis. IEEE Transactions on Multimedia, 15(3): 572-581 [DOI: 10.1109/TMM.2012.2234731]

Yang Y, Wu F, Nie F P, Shen H T, Zhuang Y T and Hauptmann A G. 2012b. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. IEEE Transactions on Image Processing, 21(3): 1339-1351 [DOI: 10.1109/TIP.2011.2169269]

Yang Y, Zhuang Y T and Pan Y H. 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. Frontiers of Information Technology and Electronic Engineering, 22(12): 1551-1558 [DOI: 10.1631/FITEE.2100463]

Yang Y, Zhuang Y T, Wu F and Pan Y H. 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Transactions on Multimedia, 10(3): 437-446 [DOI: 10.1109/TMM.2008.917359]

Zareian A, Karaman S and Chang S F. 2020. Bridging knowledge graphs to generate scene graphs//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 606-623 [DOI: 10.1007/978-3-030-58592-1_36]

Zheng Z D, Ruan T, Wei Y C, Yang Y and Mei T. 2021. VehicleNet:

learning robust visual representation for vehicle re-identification. IEEE Transactions on Multimedia, 23: 2683-2693 [DOI: 10.1109/TMM.2020.3014488]

Zheng Z D, Wei Y C and Yang Y. 2020a. University-1652: a multi-view multi-source benchmark for drone-based geo-localization//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1395-1403 [DOI: 10.1145/3394171.3413896]

Zheng Z D, Yang X D, Yu Z D, Zheng L, Yang Y and Kautz J. 2019. Joint discriminative and generative learning for person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2133-2142 [DOI: 10.1109/CVPR.2019.00224]

Zheng Z D, Zheng L, Garrett M, Yang Y, Xu M L and Shen Y D. 2020b. Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(2): #51 [DOI: 10.1145/3383184]

Zhu L C, Fan H H, Luo Y W, Xu M L and Yang Y. 2022. Temporal cross-layer correlation mining for action recognition. IEEE Transactions on Multimedia, 24: 668-676 [DOI: 10.1109/TMM.2021.3057503]

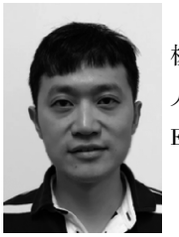
Zhu L C, Xu Z W, Yang Y and Hauptmann A G. 2017. Uncovering the temporal context for video question answering. International Journal of Computer Vision, 124(3): 409-421 [DOI: 10.1007/s11263-017-1033-7]

Zhu L C and Yang Y. 2020. ActBERT: learning global-local video-text representations//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8743-8752 [DOI: 10.1109/CVPR42600.2020.00877]

Zhu L C and Yang Y. 2022. Label independent memory for semi-supervised few-shot video classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1): 273-285 [DOI: 10.1109/TPAMI.2020.3007511]

Zhu M F, Pan P B, Chen W and Yang Y. 2019. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5795-5803 [DOI: 10.1109/CVPR.2019.00595]

作者简介



杨易,1980 年生,男,教授,主要研究方向为人工智能、跨媒体算法、计算机视觉。
E-mail: yangyics@zju.edu.cn

庄越挺,男,教授,主要研究方向为跨媒体、人工智能、计算机动画、数字图书馆。E-mail: yzhuang@zju.edu.cn
潘云鹤,男,教授,主要研究方向为计算机图形学、人工智能、CAD 和工业设计。E-mail: panyh@zju.edu.cn