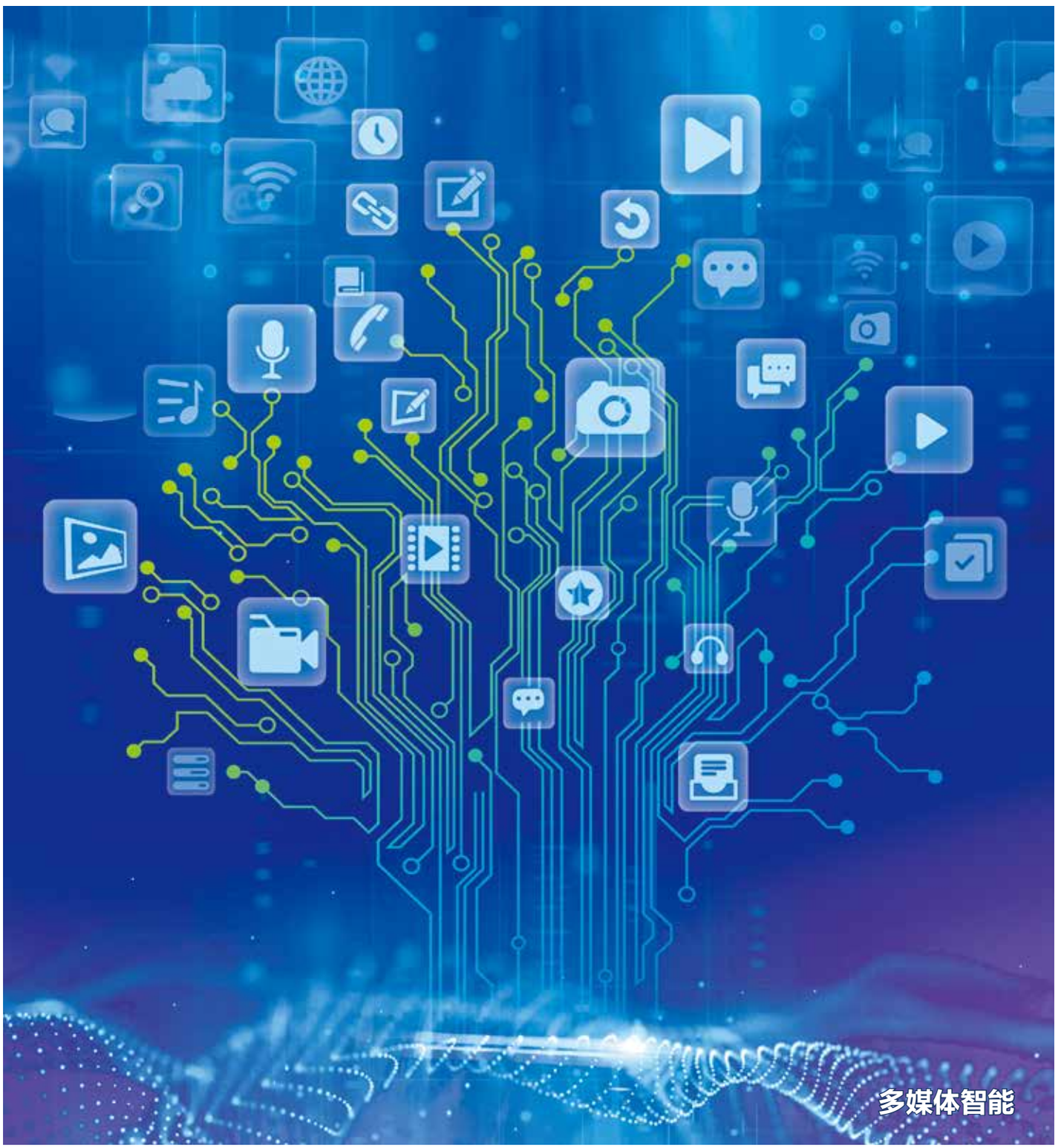
JOURNAL OF IMAGE AND GRAPHICS

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

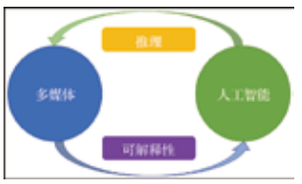
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

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

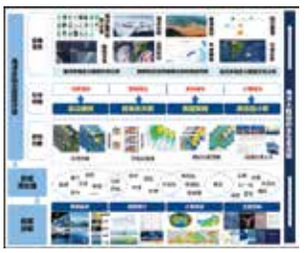
目次



多媒体智能：当多媒体遇到人工智能(第2551页)

学者观点

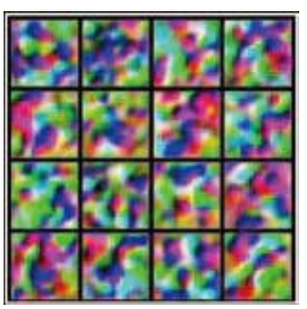
综述



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

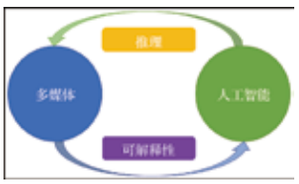


基于真实数据感知的模型功能窃取攻击(第2721页)

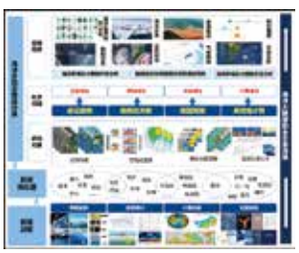
多媒体分析与理解

Volume 27, Number 9
Published September 16, 2022

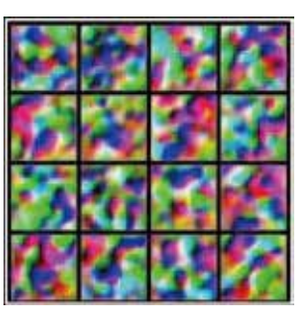
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



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)

中图法分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1006-8961(2022)09-2611-18

论文引用格式:Liao Y, Li Z M and Liu S. 2022. A review of deep learning based human-object interaction detection. Journal of Image and Graphics, 27(09):2611-2628(廖越,李智敏,刘偲. 2022. 基于深度学习的人—物交互关系检测综述. 中国图象图形学报,27(09):2611-2628)[DOI:10.11834/jig.211268]

基于深度学习的人—物交互关系检测综述

廖越[1],李智敏[2],刘偲[1*]

1. 北京航空航天大学人工智能研究院,北京　100191;2. 华中科技大学人工智能与自动化学院,武汉　430074

摘　要:　人—物交互关系检测旨在通过精细化定位图像或视频中产生特定动作行为的人,以及与其产生交互关系的物体,并识别人和物体之间的动作关系来理解和分析人体的行为。人—物交互关系检测是一个非常具有实际应用意义和前瞻性的研究方向,是高层视觉理解的关键基石。随着深度学习的发展,基于深度学习的研究方法引领了近期人—物交互关系检测研究的进步。本文一方面分析空域人—物交互关系检测任务,从数据内容场景、标注粒度两个方面总结和分析当下数据库和基准。然后从两阶段分段式方法和单阶段端到端式方法两个流派出发系统性地阐述当前检测方法的发展现状,分析两个流派方法的特性和优劣,厘清该领域方法的发展路线。其中,两阶段方法包括多流模型和图模型两种主要范式,而单阶段模型包括基于框的范式、基于关系点的范式和基于查询的范式。另一方面,对时空域人—物交互关系检测任务进行总结,分析现有时空域交互关系数据集构造与特性和现有基线算法的优劣。最后对未来的研究方向进行展望。

关键词:人—物交互关系(HOI)检测;行为理解;深度学习;目标检测;关系检测

A review of deep learning based human-object interaction detection

Liao Yue[1], Li Zhimin[2], Liu Si[1*]

1. Institute of Artificial Intelligence, Beihang University, Beijing 100191, China; 2. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: Human-object interaction (HOI) detection is essential for intelligent human behaviors analysis. Our review is focused on a fine-grain scaled image or video based human behaviors analysis through the localization of interactive human-object pairs and their recognition of interaction types. HOI detection has developed high-level visual applications like dangerous behaviors detection and human-robot interaction. Recent deep learning based methods have facilitated current HOI detection. Our critical review is carried out in terms of recent deep learning based HOI detection methods. We introduce an accelerated progress of image-level HOI detection because the growth of datasets is a key factor for the review of deep learning. First, the datasets and benchmarks of image-level HOI detection is introduced based on an annotation granularity. Therefore, the conventional image-level HOI detection datasets are assigned to three levels of instance, partial and pixel. We introduce the image collection, annotation, and statics information of every level for each dataset. Next, we analyze the conventional HOI detection methods via deep-learning-structured assignment. We summarize traditional HOI detection methods into two main folds further based on a serial architecture of two-stage fold and an end-to-end framework of one-stage

收稿日期:2021-12-31;**修回日期**:2022-06-16;**预印本日期**:2022-06-23

所有作者为共同一作;*通信作者:刘偲　liusi@buaa.edu.cn

基金项目:国家自然科学基金项目(62122010,61876177)

Supported by: National Natural Science Foundation of China(62122010, 61876177)

fold. Two-stage methods are composed of two split serial stages, where an instance detector is initial to be used for human-object detection, and a following designed interaction classifier is applied for the interaction types reasoning between the targeted human-object detection. To clarify an accurate interaction classifier, our two-stage fold methods are mostly concerned of the two stages. However, one-stage methods are melted into an end-to-end framework, where HOI triplets can be directly detected in an end-to-end manner. Additionally, one-stage methods can also be regarded as a top-down paradigm. An anchor is designed to denote interaction and first be detected in association with human and object. Specifically, we retrace the representative methods and analyze the growth paths of such two folds. Moreover, we demonstrate the pros and cons analysis of the two folds and their potentials. At the beginning, we introduce the two-stage methods sequentially. The two-stage fold into the multi-stream pipeline and graph-based pipeline is divided based on the design of the second stage. Then, the introduced one-stage methods are split into point-based, bounding box-based, and query-based contexts in terms of multiple settings of the interaction anchor. At the end, we review the progress of zero-shot HOI detection. Meanwhile, the growth analysis of video-level HOI detection is reviewed based on datasets and methods. Finally, the future directions of HOI detection are predicted as mentioned below: 1) large-scale pre-trained model-guided HOI detection: the complex HOI types are hard to be annotated for all due to multiple human-object interaction derived of various behaviors. Therefore, zero-shot HOI discovery is a challenging issue in the future. 2) Self-supervised pre-training for HOI detection: it is originated from the mechanism view, where a large-scale image-text pre-trained model hypothesis can much properly benefit for HOI understanding, and 3) efficient video HOI detection: it is hard to detect video-based HOIs efficiently for conventional multi-phases detection mechanisms. Our critical analysis reviewed deep learning based human-object interaction detection tasks systematically.

Key words: human-object interaction detection (HOI); action understanding; deep learning; object detection; interaction detection

0 引 言

人—物交互关系检测是近年来计算机视觉领域的新兴研究方向,其衍生于经典的目标检测问题,是一个针对时空域人体动作的精细化理解。任务要求检测出图像或视频中产生交互关系的人以及与其产生交互关系的物体,同时还要预测二者的动作关系并关联起来。换言之,输入图像或视频,要求算法模型在空域中输出一系列交互关系三元组〈人的检测框,物体的检测框及类别,关系类别〉,在时空域中额外输出交互关系的起止时间。该任务可以看做一种更高层次的视觉内容理解,不局限于对视觉场景中单个物体的理解,还要求理解物体之间的联系。与传统的关系检测不同,其仅关注以人为中心的动作关系,通过此来理解和分析人的行为,以及与物体之间的交互,而传统的关系检测更关注于广泛的空间位置关系,通过建模目标之间的关系来构建场景图,进行结构化理解。近些年,人—物交互检测已经形成一个独立且热门的研究方向,更有赶超传统关系检测的研究热度。此外,人—物交互关系检测有更广泛的应用场景,其研究成果广泛应用于视频监控、智能车舱、人机交互和智能机器人等多种应用场景,亦可以支撑更高层级的视觉内容理解任务,如视觉稳定、图像描述等。特别是在视频监控中的非机动车以及机动车危险驾驶行为检测领域,人—物交互关系检测成为主要解决方案。随着深度学习的发展以及大规模数据库(Gupta 和 Malik,2015;Chao 等,2018;Liao 等,2020)的提出,自 2015 年起,该方向逐渐受到了研究者的关注,特别是 2019 年至今,迎来该方向发展的小高潮。从早期的两阶段序列化模型(Gkioxari 等,2018;Gao 等,2018;Li 等,2019),到时下最为火热的单阶段模型(Liao 等,2020;Zhang 等,2021),人—物交互关系检测算法在精度和速度上都取了显著的提升,其中精度提升近 4 倍,速度方面也有超实时的算法(Liao 等,2020)涌现出来,加速了人—物交互关系检测在实际场景的应用,空域人—物交互关系发展历程如图 1 所示。深度学习算法的发展也离不开大规模数据库的支持,人—物交互关系检测数据集于早期不足 1 000 幅图像,到现在已发布多组超过 3 万幅图像的数据集(Chao 等,2018;Zhuang 等,2018;Liao 等,2020),为推动该方向的发

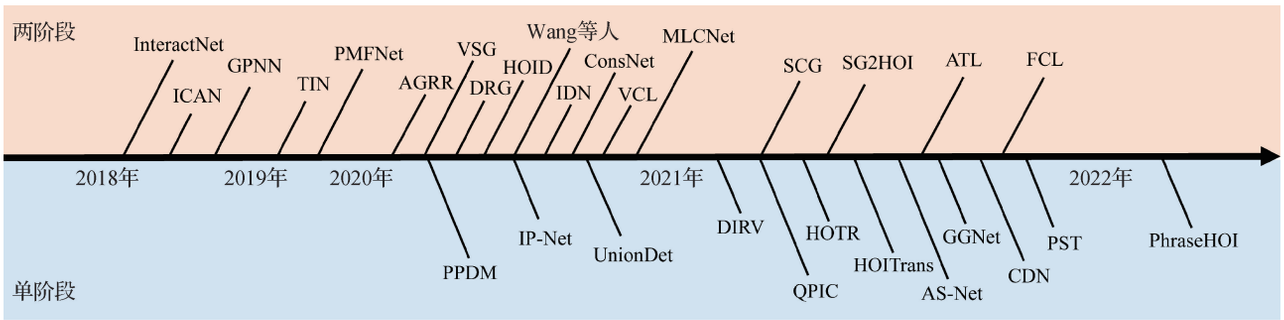


图 1　基于深度学习的空域人—物交互关系检测任务发展历程

Fig. 1　A road map of deep learning-based human-object interaction detection on spatio-domain

展做出巨大的贡献。

本文针对基于深度学习的人—物交互关系检测经典工作和最新前沿进展进行概括性的总结归纳，分别从空域和时空域人—物交互关系检测任务展开。第 1 部分，介绍空域任务。针对数据集部分，依据标注粒度将深度学习算法中常用人—物交互关系检测数据集分为 3 个层级进行阐述，分别为：实例级（instance level）、部件级（part level）和像素级（pixel level）。针对算法部分，围绕人—物交互关系检测的两个核心问题：人—物关联和交互关系预测，本文依据算法模型的结构差异将交互关系检测算法分为两阶段分段式“自底向上”（bottom-to-up）流派和单阶段端到端式“自顶向下”（top-down）流派。然后，围绕两个流派代表性的工作展开，厘清发展路线，并分析各自的优势和不足。此外，本文进一步归纳总结了空域零知识人—物交互关系检测领域的研究进展。第 2 部分，介绍时空域任务。受限于有限的时空域数据集和计算资源，该任务处于研究初期。介绍时空域行为检测数据集，分析现有识别与检测算法的优劣。接着总结目前人—物交互关系检测领域依旧面临的挑战，并以此引出人—物交互关系检测领域的未来发展趋势。

1　空域人—物交互关系检测数据库

从标注粒度的角度来归纳当下的人—物交互关系检测数据库，如图 2 所示，由粗到精可以分为 3 个层级，分别是实例级（instance level）、部件级（part level）和像素级（pixel level）。其中，实例级是对人—物交互关系检测最为基础的一种表示形式，其仅对一幅图像中的人、物体框进行实例级标注，并标注了人和物体之间的关系，但由于标注成本较低，其数据源最多，且应用最广。为了更精确地定位与物体产生交互的人体部件，部件级标注在实例级标注的基础上，进一步对人体的部件进行细粒度标注，同时标注人体部件和物体之间的交互关系。为了服务于更精细化的场景，近年来研究者进行了更为精细的像素级标注，不仅区分每个人的部件，而且对每个人体部件以及物体、场景上的像素点进行精细化类别标注。本文分别对这 3 个层级标注的数据库进行归纳和分析。

1.1　实例级人—物交互关系检测数据库

实例级标注是成本相对较低且能满足人—物交

(a) 实例级　　(b) 部件级　　(c) 像素级

图 2　注释级粒度的数据集示意图

Fig. 2　Datasets in an annotation granularity view((a) instance level;(b) part level;(c) pixel level)

互关系检测任务最基本需求的一种标注方式,也是当前主流方法采用最多的数据格式。深度学习方法中常用的实例级数据库有 4 个(如表 1 所示),按照数据库提出的时间顺序分别为:V-COCO(verbs in COCO)(Gupta 和 Malik,2015),HICO-DET(humans interacting with common objects detection)(Chao 等,2018),HCVRD(human-centered visual relationship detection)(Zhuang 等,2018)以及 HOI-A(human-object interaction application oriented)(Liao 等,2020)。其中,V-COCO 和 HICO-DET 为当前最常用的两个人—物交互关系检测基准。V-COCO 于 2015 年提出,其直接从通用目标检测数据库 COCO(common objects in context)(Lin 等,2014)选取 10 346 幅图像进行交互关系标注,其中包含训练集 5 400 幅、测试集 4 964 幅。V-COCO 一共标注了 29 种关系类型,而交互物体种类与 COCO 一致,为 80 类。其中 29 类动作关系中包含 4 种无交互对象的动作类型,该 4 种动作在人—物交互关系检测算法的研究中一般不被考虑在内。由于 V-COCO 所含图像较少,场景相对简单,随着深度学习的发展,V-COCO 难以满足通用人—物交互场景理解的需求。美国密歇根大学在 2018 年提出一个图像数量更多、交互关系更为复杂的人—物交互关系检测数据库 HICO-DET。从 Flickr 上收集了 47 776 幅有商用许可的图像,其以人为主语、包含动作类型的一系列关键词。包括训练集图像 38 118 幅,测试集图像 9 658 幅。HICO-DET 共标注了 117 类动作,80 种物体类型,其中物体类型与 COCO 一致。117 类动作和 80 种物体在 HICO-DET 累积构造了 600 种不同的人—物交互关系类型。根据在数据库中的出现频率,HICO-DET 将 600 类交互关系分为常见(non-rare)462 类和稀有(rare)138 类。由于 HICO-DET 数据库存在明显的长尾分布,如何在稀有类别上取得更好的性能也成为了人—物交互关系检测领域的一个研究热点。HCVRD 是从通用视觉关系检测数据库视觉基因(visual genome)取出与人相关的关系类型以及对应的数据组成一个人—物交互关系检测数据库,与 HICO-DET 和 V-COCO 不同,它不仅关心交互动作关系,还囊括了人和物之间的相对位置关系。以上介绍的数据库都是针对通用场景的人—物交互关系检测数据库,由于关系种类复杂,深度学习算法当前很难取得很好的性能,且大部分动作关系在现实生活中没有实际应用意义。基于此,北京航空航天大学联合商汤科技公司于 2020 年构建了一个面向实际应用场景的人—物交互关系检测数据库 HOI-A(human-object interaction for application),并基于此数据库分别在 ICCV 2019(IEEE International Conference on Computer Vision 2019)和 CVPR 2021(IEEE Conference on Computer Vision and Pattern Recognition 2021)“以人为中心的视觉理解(person in context)”研讨会上举办了两届人—物交互关系检测挑战赛。HOI-A 考虑到在实际需求中只有少数的人物和物体之间的交互关系需要被关注,如〈人,抽烟,烟〉这个行为对于智能控制公共场所吸烟具有比较重要的意义。因此,在收集数据时,着重收集了几个具有实际应用价值的人和物体之间的关系。并针对 3 种不同的场景:室内、室外和车内,分别以 RGB 和近红外(infrared radiation, IR)两种摄像头采集和网上爬取的方式收集了 38 668 幅图像。为了提高实用价值,该数据库仅标注了 11 种常见物体类别,以及 10 种常见关系类别。目前实例级人—物交互关系检测数据库虽然已经有一定的数量,但是由于标注成本高,现实生活中人—物交互场景复杂,很难覆盖现实生活中的所有需求。

表 1　人—物交互关系检测数据库概览

Table 1　The overview of human-object interaction detection datasets

数据库	数量/幅	物体种类	关系种类	标签粒度
V-COCO	10 346	80	29	实例级
HICO-DET	47 776	80	117	实例级
HCVRD	52 855	1 824	927	实例级
HOI-A	38 688	11	10	实例级
PaStaNet-HOI	110 714	80	116	部件级
PIC	17 122	141	23	像素级

1.2　部件级人—物交互关系检测数据库

在粗粒度的实例级标注监督下,网络模型容易拟合于当前数据集中的交互模式,从而导致较差的泛化能力。上海交通大学在 2020 年提出基于人体部件状态的人类行为知识引擎(human activity knowledge engine, HAKE),用来构建实例级行为标注和部件级行为标注的桥梁。HAKE 提供了约 11.8 万幅标注图像,其中包含约 28.5 万个人体实例,约 25 万个交互目标,以及 72.4 万个具有人体部件状

态的人—物交互对,大量的人—物交互样本可以显著提升交互关系描述性能。为了更好地评估人—物交互关系检测任务,该团队从 HAKE 数据集中抽取了部分标注数据,构建了 PaStaNet-HOI(Li 等,2020b)数据集。该数据集提供了约 11 万幅标注图像,其中 77 260 幅图像作为训练集,11 298 幅图像作为验证集,22 156 幅图像作为测试集。PaStaNet-HOI 摒弃了“无交互”类别,由 116 种交互关系和 80 种物体类别,组成了 520 种人—物体交互关系类别。细粒度的部件级人—物交互关系标注显著提升了模型对行为理解的能力,降低了不平衡数据的学习难度。

1.3　像素级人—物交互关系检测数据库

虽然部件级人—物交互关系标注已经提供了丰富的人体结构信息来辅助交互关系的理解,但传统粗粒度的人—物交互定位使用矩形框标定人和物体的位置,易导致目标表现形式在遮挡严重时指代不明,而且在现实生活中除了物体对象外,人和场景往往还会产生交互,而不规则的场景很难用检测框标注。因此,Liu 等人(2021)构建像素级人—物交互关系定位数据库 PIC(person in context),通过对人体部件以及物体进行像素级标注,使得遮挡时人和物体的指代更为明确。PIC 数据库在网上搜集了 17 122 幅以人为中心的图像,其中包括训练集 12 339 幅,验证集 1 916 幅和测试集 2 867 幅。PIC 数据库对每幅图像中的每个人标注了 25 个部件分割类别以及 16 个身体关键点。此外,该数据库标注了14 种人—物之间的位置关系和 9 种动作关系。PIC 数据库是目前标注最为精细的人—物交互关系检测数据库,且标注类型非常全面。但是,其任务定义对深度学习模型要求过高,对人体部件和场景进行精准分割已经是一个巨大的挑战,且实例级人—物交互关系检测已可以满足现实生活中的大部分需求。随着现实生活需求多样化,像素级人—物交互关系理解应该是未来具有潜力的一个研究方向。

2　空域两阶段分段式检测方法

两阶段人—物交互关系检测是一种实例引导的自底向上的深度学习方法,其首先检测出图像中的人、物实例,然后对人、物实例进行关联并识别对应的交互关系。具体而言,两阶段方法一般由两个串行的模块组成:目标检测模块和交互关系分类模块。给定一幅人像图像,两阶段方法首先采用一个预先在通用目标检测数据集上训练好的检测模型来检测图像中的人和物体。然后,选取具有较高置信度的检出人、物,并将选出的人、物两两组合成对亦或构造成图作为候选集,同时从检测网络中提取出对应人、物区域的深度特征。最后,采用一个特定化设计的交互关系分类网络,根据生成的候选集,输入其对应的深度特征,对人—物对逐一分类,得到其对应的交互关系类型。两阶段方法存在梯度截断特点,即目标检测器与交互关系分类网络分开训练,优化时无法端到端回传梯度信息。针对两阶段模型的研究主要集中在第 2 阶段改进,即如何更好地理解并预测出人—物之间的交互关系类型。依据第 2 阶段的不同改进思路和结构设计,可将两阶段流派划分为多流模型和图模型,如图 3 所示。本文详细介绍这两种类型的发展现状,并归纳其存在的问题和改进方向。

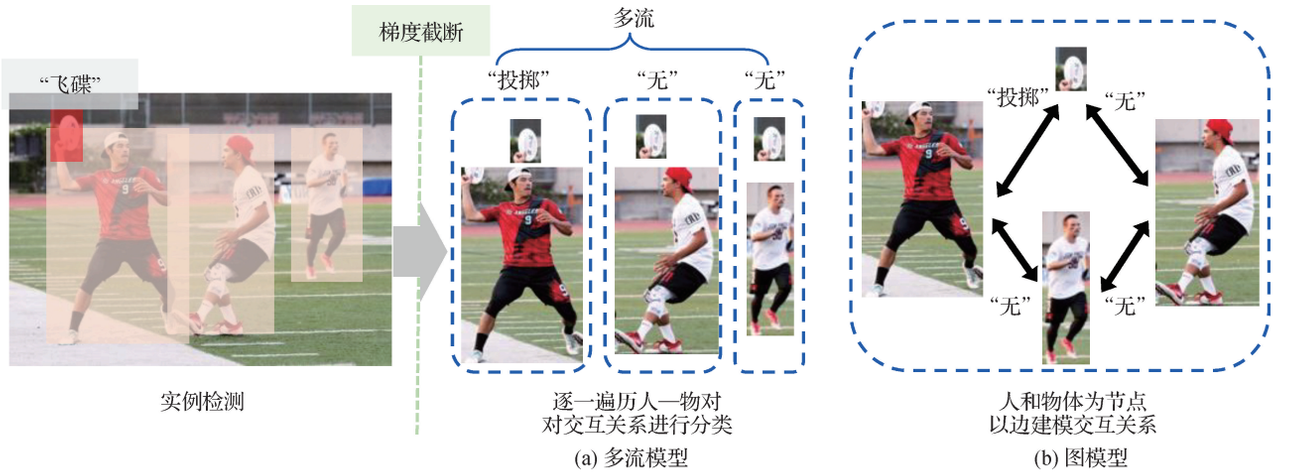


图 3　两阶段人—物交互关系检测算法框架示意图

Fig. 3　The frameworks of two-stage HOI detection methods((a) multi-stream model;(b) graph-based model)

2.1 多流两阶段模型

多流模型是深度学习在人—物交互关系检测领域最早期的探索。多流模型在第1阶段使用目标检测器获得人与物体的检测结果,并将其一一配对,构成一系列的人—物候选对,在第2阶段交互关系分类网络中,对视觉目标、交互关系分多流/多分支进行特征提取与建模,再融合分类结果,如图3(a)所示。

Gao等人(2018)认为人和物体的表观特征包含了检测交互的提示线索。于是,提出了以实例为中心的注意力网络(instance-centric attention network, ICAN),通过实例外观特征,动态学习注意力区域。这使得网络可以有选择地聚合与指定实例交互关系密切相关的特征。ICAN包含3个主要分支:人分支、物体分支和对粒度分支。人和物体分支分别以人和物体为中心,学习注意力区域,并在其引导下预测交互关系。对粒度分支通过人和物体的包围框编码空间信息,并与人的外观特征融合以预测交互关系。最终通过组合所有人—物对,融合3个分支的交互关系得分,获得人—物交互关系三元组。ICAN摒弃了手工设计注意力方法,为每一个检测实例动态提供注意力机制,关注与交互关系高度相关的图像区域,提升网络性能。

Gkioxari等人(2018)认为通过研究人的外表信息,可获得交互物体定位的重要线索。于是,提出InteractNet,利用人的外表特征来预测与当前人相关物体的具体位置。该网络在Faster R-CNN(regions with convolutional neural networks)目标检测器框架(Ren等,2017)的基础上,扩展一个以人为中心的分支。该分支作用于人的特征,判断当前人发生的交互关系,并估计每种交互关系所对应物体在图像上的位置。网络联合以人为中心的分支、目标检测分支和对粒度交互关系分支进行多任务联合训练。通过融合多流预测,有效推理交互三元组。

Li等人(2019)发现,虽然不同数据集对人—物交互关系类别有不同的定义,但是交互性知识可以跨数据集学习。于是,提出了交互性网络(interactiveness network),通过大量数据学习通用交互性知识。推理时,提出了非交互性抑制策略,有效减少非交互的人—物对分类干扰。由于交互性知识的可泛化性,交互性网络可以与任意多流人—物交互关系检测模型配合使用,以获得性能提升。

为缓解人—物交互关系复杂性带来的性能损失,Wan等人(2019)提出了姿态粒度多级特征网络(pose-aware multi-level feature network, PMFNet),利用人的姿态线索捕获关系的全局空间结构,并作为一种注意力机制来动态增强人的相关局部区域特征。这使得网络在人体姿态信息的指导下,可以同时从交互关系、视觉目标和人体部位3个层面,将不同粒度的特征融合到关系推理流程中,生成更加鲁邦和细粒度的预测结果。

Wang等人(2019b)认为捕捉人—物间微妙交互十分重要,提出了上下文注意力框架,建模人与物体间上下文感知的表征特征,自适应选择相关的以实例为中心的上下文信息,以突出显示包含人—物交互关系实例的图像区域。

Lin等人(2020)利用人与物体间的上下文兼容一致性来过滤非人物交互关系对。为了更好地区分细粒度行为间细微的差异,提出了一个行为感知的注意力机制,基于类别激活图来挖掘更利于识别人—物交互关系对的任务特征。

Sun等人(2020)认为仅基于视觉特征不足以表达复杂的人—物交互关系对,为提升理解能力,提出了一个融合空间—语义知识的多级调节网络,编码人体姿态结构和目标上下文信息在内的额外知识,通过仿射变换和注意力机制动态影响卷积神经网络特征提取流程,最终融合多模态特征识别交互关系。

人类具有较强构造感知能力以识别稀有和未见的人—物交互关系对样本,基于此,Hou等人(2021)提出了FCL(fabricated compositional learning)网络,提出一个物体制造器来生成有效的物体表达,接着合并动词特征与虚拟的物体以生成新的人—物交互关系对样本,通过扩充数据空间来解决任务中存在的长尾分布问题。

Xu等人(2022)认为基于特征裁剪的多阶段检测算法会丢失图像中的上下文信息。于是,设计了一种以人为中心的交互推理框架:在传统目标检测器输出人和物体的预测位置后,将人的位置转化成掩码与原始图像进行通道级融合,作为网络输入。在人、物体分支分别输出像素级的交互位置预测,设计索引融合策略,生成人—物交互关系三元组。该框架实现了针对指定人进行全局信息聚合和交互物体位置预测,提高检测性能。

2.2 基于图的两阶段模型

由于多流模型需要逐个处理每个人—物实例

对,一方面忽略了不同人—物对之间的关联,另一方面,逐一遍历人—物对需要线性时间复杂度。为了解决这些不足,Qi 等人(2018)首次将图网络引入到人—物交互关系检测任务中,提出了一种端到端可微的图解析网络(graph parsing neural network,GPNN)。GPNN 用图网络的节点定义人和物体,用连接节点的边编码交互关系,用邻接矩阵显式建模人—物交互关系结构。对于给定场景,GPNN 摒弃了常见的预固定图(pre-fixed),通过连接函数(link function)计算邻接矩阵并控制信息传递,在迭代更新信息的同时,自动学习出最优图结构。这使得网络可以同时适应静态图像中人—物交互关系检测和视频中的人—物交互关系识别任务。

Wang 等人(2020)认为现有基于图网络的方法将人和物体视为同类型节点,未考虑同构实体(homogeneous entities)和异构实体(heterogeneous entities)间信息传递的差异。于是,提出了语境异构图网络(contextual heterogeneous graph network),使用连接异构节点的边编码人和物体的空间关系,聚合类间信息;使用连接同构节点的边编码人和人、物体和物体间的相关性,聚合类内信息。同时,使用图注意力机制提升节点间获取信息的有效性。人—物对的空间结构信息是很重要的线索,Ulutan 等人(2020)提出了视觉空间图网络(visual-spatial-graph network,VSGNET)。该网络分为 3 个分支,视觉分支用于提取图像中人、物体和周围语境的视觉特征;空间注意力分支建模人—物对的空间关系,输出的空间注意力特征与视觉特征合并,以提高视觉特征有效性;基于图的分支建模人—物体对间的交互关系,该图网络使用节点定义人和物体,不同的是,仅将视觉不相似的人和物体节点的连接作为边,有利于生成更好的交互特征。最终,合并 3 个分支的输出,以获得交互关系分类结果。Gao 等人(2020)认为独立预测每一组人—物对缺少语境信息,容易产生歧义性,提出了双流关系图(dual relation graph,DRG)。该网络使用抽象的空间语义表达来描述每一组人—物对,同时通过双流关系图聚合场景的语境信息。空间语义表达包含人—物的空间关系以及物体类别的语义词嵌入(semantic word embedding),前者对外观和复杂场景变化较为稳定,后者可引入额外语言信息帮助稀有交互类别的训练与推理。双流关系图由以人为中心的子图和以物体为中心的子图构成,前者以每一个人的节点为中心,连接所有物体的节点,该人与不同物体间的交互检测可以相互指导并调整;后者以每一个物体的节点为中心,连接所有人的节点。双流关系图有效获取场景中具有判别性的线索,以解决仅使用局部特征带来的歧义性问题,从而提升检测器性能。

He 等人(2021)认为现有方法仅关注于视觉和语言特征,并没有利用图像中高层信息和语义关系,缺少了关键的上下文信息和详细的关系知识。于是提出 SG2HOI(scene graph to human-object interaction)网络,将场景图嵌入到上下文线索中,作为特定场景的环境上下文信息,并提出一个关系感知信息传递模块,从目标邻居中聚合场景图关系并转移至交互关系中,以提升人—物交互网络建模能力。

传统图网络建模方法从一个节点向其邻居节点传输相同或相似的信息,唯一的变量是表征连接性的可学习权重,缺少位置信息会降低网络判别能力。不于与此,Zhang 等人(2021)根据节点间的空间关系规定信息传递方向,构建了二分图来描述交互关系,以抑制检测人—物交互关系对时网络生成的虚警。

2.3 两阶段方法的特点

如表 2 所示,两阶段检测器是一种实例驱动(instance-driven)的人—物体交互检测范式。第 1 阶段,通用检测网络检测出 M 个人和 N 个物体;第 2 阶段,组合生成 $M \times N$ 组人—物对,交互关系分类器以人—物对为最小单元,进行 $M \times N$ 次前项推理,获得人—物对粒度的交互分类结果。这将导致以下问题:首先,具有交互关系的人—物对远远少于整体样本数量,不均衡的正负样本使得网络很容易过拟合,从而倾向于输出高置信度的“无交互(no-interaction)”类别标签,抑制了正样本分类结果;其次,不具有交互关系的负样本会引入额外计算消耗,时间复杂度为 $M \times N$;最后,非端到端范式阻断了两阶段中信息交互。在第 1 阶段中,通用检测网络更加关注于目标区域的边缘信息,这使得第 2 阶段中,人—物对特征缺少上下文信息,从而产生交互关系分类的歧义性。当然,两阶段范式也存在独有的优势。通过解耦目标检测和交互关系分类任务,使得每一个阶段可以聚焦于自身优化的目标,从而获得当前阶段最优结果。

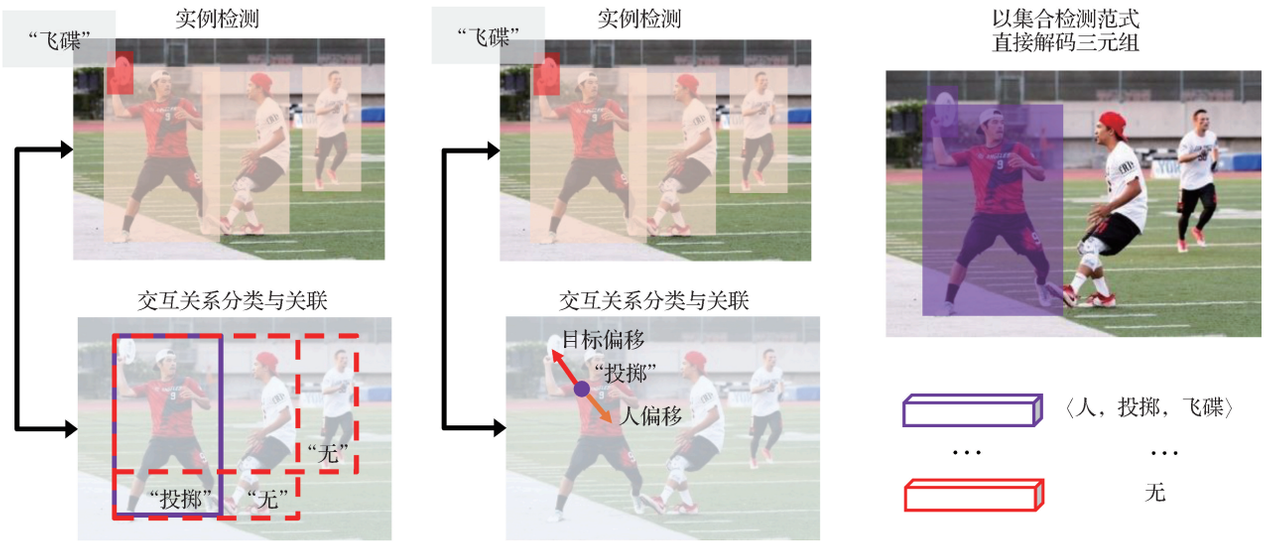
表 2　基于深度学习的人—物交互关系检测算法两种范式对比

Table 2　The comparison between two popular deep learning-based HOI detection methods

范式	流程	性能分析	效率分析
两阶段	实例驱动,先检测人、物,再预测关系	优点:可通过不同特征提升性能; 缺点:搜索空间大,受大量负样本干扰	效率普遍较低,需逐个遍历候选人—物
单阶段	关系驱动,直接预测关系,通过关系关联人、物	优点:直接定位关系,减小搜索空间; 缺点:耦合的检测和关系理解干扰性能	效率高,直接检测交互关系三元组

3　空域单阶段端到端检测方法

由于两阶段模型速度和精度受限于其分离式串行结构,受单阶段目标检测算法启发,提出了一种全新的自顶向下、关系驱动的单阶段端到端式人—物交互定位算法,并引领了近两年人—物交互关系检测的潮流。不同于两阶段方法,给定一幅图像,单阶段方法可以直接输出人—物交互三元组,不依赖于额外的目标检测(如图 4 所示)。具体而言,单阶段方法一般会预先定义一个关系交互中介,通过神经网络直接预测关系中介,然后通过关系中介关联产生交互的人—物对,并预测对应的交互关系。依据关系中介的不同设计,单阶段方法可以分为基于框、关系点和查询 3 类模型。接下来分别对 3 类模型的当前研究进展进行阐述并对其优劣加以分析。



(a) 基于框　(b) 基于点　(c) 基于查询

图 4　单阶段人—物交互关系检测算法框架示意图(依据关系中介类型不同,可分为基于框、点和查询 3 种范式)

Fig. 4　The frameworks of conventional one-stage HOI detection methods

((a) box-based; (b) point-based; (c) query-based)

3.1　基于关系点的单阶段模型

受到基于热度图的无锚(anchor-free)检测器的影响,关系点的表示方式率先开启了单阶段人—物交互关系检测的时代。Liao 等人(2020)提出将人—物交互关系三元组重定义为点的三元组,并提出并行人—物交互关系检测框架 PPDM(parallel point detection and matching)。其中对于人和物体的检测框,则采用中心点以及宽和高表示。对于人和物体之间的关系则采用两者检测框的中心点连线中点表示,作为关系中心点。因此针对人—物交互关系检测的过程可以分成两个分支,分别为:中心点检测分支和关系匹配分支。中心点检测分支对人物中心点、物体中心点以及对应检测框的宽高和关系中心点进行预测;关系匹配分支从关系中心点出发预测与物体位置和人、物位置的相对关系。当人体点、物体点与同一个关系点相匹配的时候认为得到了正

确的关系配对。在中心点检测过程中,针对关系中心点的预测隐式地完成了对于物体检测和人物检测信息的上下文信息的利用和约束,从而会强化出具有关系的物体和人物的检测结果。同时在点匹配过程中,匹配过程只在少数检测出关系中心点位置开始,从而避免了线性扫描较多〈人,物体〉配对结果,进而提高模型性能。相较于传统的两阶段模型,该模型不仅在性能上取得了大幅度提升,而且在速度上取得了近 10 倍的提升,将人—物交互关系检测推向实时。Wang 等人(2020b)提出关系点网络(interaction points network,IP-Net),也采用类似的思想来构建关系点和三元组,不过他们仍旧采用一个额外的检测器来得到检测框,仅通过关系点来关联现有的人和物体框,在性能和效率上的表现与 PPDM 仍有差距。

针对 PPDM 仅有一个点来表示关系,难以对复杂的人—物交互进行建模的问题,Zhong 等人(2021b)基于 PPDM 进一步改进,提出扫视和凝视网络(glance and gaze,GGNet)。在 GGNet 中,他们仍然采用 PPDM 的方式来定义关系点三元组。此外,为了增强关系表示,此文提出自适应地推断出一组动作感知点(ActPoints)来表示交互区域。基于此,GGNet 可以模仿人类先扫视再凝视的定位理解过程。首先,GGNet 和 PPDM 一样快速判断特征图上的每个像素是否为一个交互点,即为扫视。然后,依据扫视后的特征图,围绕每个像素点搜索一组 ActPoints,然后逐步细化这些点的位置,即为凝视。具体而言,模型先粗定位可行遍卷积核的中心,再推断可形变卷积的残差。最后,GGNet 聚合了精炼的 ActPoints 的特征,预测对应关系点的交互类别。依据这种方式,可以利用更为丰富的视觉特征来增强交互关系表示,提升关系理解,从而进一步提升了人—物交互关系检测的精准性。

3.2 基于框的单阶段模型

近年来大量工作将交互关系具象化,如基于框的单阶段模型。其基本流程为:神经网络检测图像中人与物体的同时,并行预测可能发生交互关系的边界框,通过边界框关联包含的人与目标,组成人—物交互三元组。基于框的单阶段模型摒弃了两阶段模型中时间消耗的关系分类网络,使用相对简单的关联策略,关联目标检测结果,保障整体算法精度的同时,大幅提升网络推理速度。

Kim 等人(2020)认为发生交互关系的人与物体的联合区域包含更丰富的交互信息。于是,首次提出区域级(union-level)人—物交互关系检测网络 UnionDet,并行预测联合区域,以匹配人—物关系组合。该网络由区域分支(union branch)和实例分支(instance branch)组成。区域分支预测联合区域(union box)的位置与交互关系类别,设计的前置任务子分支(pretext task)预测联合区域中包含的目标类别,可在训练中缓解网络对人的偏置;实例分支(instance branch)预测图像中出现的目标与其交互类别。最后,通过区域—实例匹配策略合并分支结果,获得人—物交互关系三元组。由于减少了两阶段算法中额外的关系网络推理流程,该网络与两阶段算法相比提高了 4～14 倍推理速度。

Fang 等人(2021)发现联合区域中引入的冗余视觉信息会干扰检测器性能,于是定义覆盖人和物体的部分包围框,且对关系分类有效的最小区域为交互区域(interaction region),并提出了一种密集交互区域选择方式,即与人、物体和关联区域的交并比满足设定阈值,密集生成锚点。基于此,提出了密集交互区域投票(dense interaction region voting)网络。该网络由交互检测器(interaction detector)和实例检测器(instance detector)并行组成。交互检测器检测交互区域(interaction region),并回归区域内存在的人和物体的包围框;实例检测器预测图像中出现的目标与其交互类别。随后,以加权定位得分(weighted localization score)作为权重,合并不同交互区域的预测结果来获得最终交互得分。该算法关注于不同尺度的密集交互区域,从而捕捉更本质的细微视觉特征。同时,相比非极大值抑制(non-maximum suppression,NMS)策略,提出的投票策略充分利用有重叠的交互区域,可显著提升预测锚点(anchor)的容错率。

将交互关系具象化的启发式策略会给网络带来歧义性,不同的人—物交互三元组可能会聚焦于相似的视觉特征。且手工设计的后处理形式复杂,检测性能对后处理中超参、阈值敏感,使得算法性能提升进入瓶颈。

3.3 基于查询的单阶段模型

自变化网络(transformer)引入计算机视觉领域,基于查询的框架在视觉任务中获得广泛应用(Carion 等,2020)。变化网络通过自注意力(self-at-

tention)和交叉注意力(cross-attention)机制获取图像的全局信息,自适应聚合上下文信息以提升人—物交互关系三元组特征的鲁棒性,缓解目标重叠、交互歧义造成的特征污染,对人—物交互关系检测任务尤为重要。2021 年以来,不断涌现出高质量的网络方案。目前研究分为两个方向,分别是重新定义人—物交互关系检测范式和充分挖掘变化网络潜力,引入多模态学习范式。本小节介绍两种方向的发展现状,并提出未来思考。

1)解码实例和交互关系的检测范式。Kim 等人(2021)将人—物交互关系检测定义为集合检测(set prediction)任务,提出了基于变化网络的 HOTR 网络。该网络由卷积神经网络、共享编码器和并行解码器组成。卷积神经网络和共享编码器提取图像特征并编码全局信息,解码器延续了单阶段算法中的并行预测结构,分为实例解码器(instance decoder)和交互解码器(interaction decoder)。前者基于目标查询(object query)检测目标;后者基于目标查询解码人—物体指针(HO pointers)和交互关系类别。其中,人—物体指针表达了由当前目标查询所负责的交互关系类别,其所属的人、物体定位区域特征,通过与实例解码器中实例特征(instance representations)计算相似度,获得匹配索引,从而关联目标位置类别和交互关系,组成人—物交互关系三元组。该算法摒弃了单阶段中基于启发式的后处理策略,用网络在线学习目标和交互关系间的关联信息,提升性能的同时,降低推理时间。

为打破已有算法中以实例为中心的局限性和交互关系定位限制,Chen 等人(2021)提出了自适应集合预测的单阶段网络 AS-Net(adaptive set-based one-stage framework)。该网络基于变换网络设计了两个并行分支:交互分支(interaction branch)和实例分支(instance branch)。交互分支预测交互向量和交互关系类别,用于匹配交互关系与目标;实例分支检测目标,并输出实例语义嵌入(instance semantic embedding),用于提升匹配性能。不同于 HOTR 在训练中学习目标和交互关系的匹配关系,AS-Net 升级 PPDM 的匹配策略,用变化网络自适应学习由人的中心指向物体中心的交互向量,通过启发式后处理策略匹配目标和交互关系。在全局特征指导下,网络聚焦于表达整体人—物交互关系对的特征,提升匹配性能。同时,提出了实例感知(instance-aware)注意力机制,在并行分支间传递信息,有助于学习有效的实例语义嵌入,并强化对交互推理有价值的目标特征。Zhang 等人(2021)详细分析了两阶段和单阶段范式的优缺点,提出了以级联方式解耦人—物对检测和关系分类的 CDN(cascade disentangling network)。该网络由视觉特征提取器、人—物对解码器(human-object pair decoder)和交互解码器(interaction decoder)组成。视觉特征提取器提取图像特征并编码上下文信息;人—物对解码器检测人与物体,并预测存在交互关系的可能性;交互解码器使用人—物对解码器的输出初始化目标查询,充分利用前者训练得到的先验知识,对其输出所表征的每一个人—物对进行关系类别分类。该模型保持了单阶段范式的计算效率和直接定位有关系的人—物对以提升精度的策略,同时,引入了两阶段范式回归和分类解耦的思想,在人—物交互关系检测任务上达到了最优性能。

2)联合实例和交互关系的检测范式。Zou 等人(2021)合并了实例检测和交互关系分支,基于变化网络直接预测人—物交互关系。该网络由卷积神经网络、编码器和解码器组成。卷积神经网络和编码器提取图像特征并编码全局信息;解码器解码包含每一个人—物交互关系特征的目标查询(object, query),通过共享权重的多层感知机检测人、物体,以及所属交互关系。最后,提出五元组损失对目标检测和关系分类同时监督。需要注意的是,该方案定义每一个目标查询只负责一组人—物交互关系,故关系分类时,需要额外设置一个背景类,并使用 Softmax 函数获得关系分类得分。联合实例和交互关系的新范式降低了两阶段冗余的计算量,摒弃了解码范式中人为设计的后处理策略,让整体框架更简洁直观。

Tamura 等人(2021)采用了与 Zou 等人(2021)相似的网络框架,使用查询粒度(query-wise)的注意力替换位置粒度(location-wise)的注意力,从而提高模型在困难场景下的建模能力,例如出现在人和物体包围框外的交互关系、远距离的交互关系和拥挤目标的交互关系等。与 Zou 等人(2021)不同的是,将发生于同一个人—物对的多种交互关系仅使用同一个目标查询建模,视为相同人—物对交互关系的多分类任务。分类时,无需添加背景类,使用 Sigmoid 函数获得发生于同一人—物对的交互关系多

分类结果。

Dong 等人(2021)提出基于组合查询的 PST 网络(part-and-sum transformers),不同于单粒度查询的自变化网络框架,PST 使用查询构造和注意力模块对联合区域和交互总和进行建模,明确地合并了求和查询,以更好地建模传统自变化网络中缺少的部分与总体间关系。

3)多模态学习范式。真实世界中,人与物体交互关系复杂,有效标注数据稀缺,数据常呈现长尾分布,影响神经网络建模性能。Li 等人(2022)提出一种结合短语学习和标签创作的多任务学习方法 PhraseHOI,来联合优化人—物交互关系检测任务和短语学习任务,利用语言先验知识,提升人—物交互关系检测性能。该方案由人—物交互关系检测分支(HOI branch)和短语分支(phrase branch)组成,前者检测交互目标以及交互关系类别,后者输出用于描述交互关系的语义向量,其真值通过人—物交互关系检测任务的标注自动转换而成,无需人为干预。同时,提出了有效的标签创作方法,通过语义邻居在标签空间内构造丰富的语义样本,以缓解标注数据稀缺和长尾分布对检测性能的损失。最后,使用蒸馏损失和三元组损失,将文本知识通过短语分支蒸馏至人—物交互关系检测分支,扩充共享的知识空间,提升检测性能。

解耦的双阶段范式易导致正负样本不均衡,计算复杂度高;单阶段范式后处理复杂,并行任务差异性高,难优化。基于查询框架,通过变化网络获得图像全局信息,将人—物交互关系检测任务视为集合检测任务,用解码器的目标查询(object-query)建模交互对的高维空间,有效解决两阶段效率低、单阶段难优化的问题。同时,变化网络是序列到序列(seq2seq)模型,对建模序列型数据具有天然优势,例如文字、语音数据等。未来可深入挖掘人—物交互关系检测任务中,变化网络建模多模态数据的潜力,扩充任务数据空间,可有效缓解现有的标注数据稀缺、数据呈现长尾分布等关键问题。

3.4 单阶段方法的特点

如表 2 所示,单阶段方法打破两阶段算法分离式的固有范式,可以端到端直接预测交互关系。相对于两阶段模型,单阶段模型的效率和精度都取得了显著的提升。效率的提升主要是源于端到端模型的设计,无需依靠额外的目标检测器,且可一次性检出所有交互关系;而精度提升归因于其自顶向下的结构,直接定位关系点,关注有关系的人—物交互对,尽可能避免负样本无关系人—物对的干扰。但目前的单阶段模型均为多任务协同学习的方式,不同任务之间共享特征。然而,目标检测和关系预测所关注的特征和优化方向可能存在较大的差异,多任务协同学习会让不同任务之间互相干扰,从而达不到最优效果。虽然 CDN 提出了解耦的思想,但仅仅是在特征上进行了一定程度的解耦,在优化上仍需显式的均衡。

4 空域零知识人—物交互关系检测

现实生活场景,由于人体动作行为多变,与其产生交互关系的物体种类繁多,进而,二者组合构成的人—物交互动作类型复杂多样。高昂的标注成本使得很难获取所有感兴趣的标注数据,且有标注数据通常呈长尾分布,影响模型学习性能。为有效缓解上述问题,零知识学习(zero-shot)逐渐进入研究人员视野。零知识学习包含了度量学习(metric learning)、属性学习(attribute recognition)和域迁移(domain transfer-based)等学习方法,广泛应用于图像分类、目标检测等基础视觉任务,并影响着人—物交互关系任务的发展。零知识人—物交互关系发现可以分为 4 种类型:1)交互类别发现,在此类型下,物体类别和动作类别各自都在现有数据集有标注,而被要求去检测没有在数据集里出现过的〈物体,动作关系〉交互组合类别;2)物体类别发现,顾名思义,需要去发现和检测没有出现在数据集里的新类型物体;3)动作类别发现,物体无新类型,但需要推理其和人产生的新类型的动作关系;4)开放人—物交互关系发现,在此设置下,会被要求去检测新的物体类别、动作类别和交互关系以应对开放场景。当前的方法主要关注于前两种类型的零知识发现。其中,Shen 等人(2018)通过独立表达交互关系和物体,在推理时组合生成未见过的人—物交互关系对;Peyre 等人(2019)使用语言嵌入来表达人—物交互关系对,利用大型语料库训练获得的语言嵌入相关性,检索未见过的相似结果;Hou 等人(2020)通过解构与重组交互对,以共享不同交互对中目标和交互关系特征,生成新的交互对样本。零知识学习的引入使得人—物交互关系检测任务更适应于真实世

界场景,更本质地聚焦于稀缺数据和长尾分布情况,有极大研究价值。

总体而言,数据增强成为现有零知识交互关系主流解决方案。然而该方案不够本质,收效甚微,难以应对开放场景。如何在开放场景下识别和定位已知或未知的人—物交互动作以实现零知识人—物交互关系检测,提升算法的泛化能力是人—物交互关系检测的未来研究核心问题之一。

5 时空域人—物交互关系检测

基于图像的人—物交互关系检测虽然在部分实际应用场景中取得了较好的落地成效,但对于人体连贯动作的理解依旧存在较大的困难。进而,基于视频的时空人—物交互关系检测(spatial-temporal human-object interaction detection)成为未来的重要研究趋势之一。时空人—物交互关系检测旨在定位一段视频中出现的所有人—物交互关系出现的时间段以及检测出在该段时间中每一帧人—物交互关系三元组。任务要求和难度都远高于图像人—物交互关系检测。当前针对时空人—物交互关系检测的研究较少(Wang 等, 2019a; Ji 等, 2021; Chiou 等, 2021),一方面是由于早期数据库的缺少,另一方面是计算资源开销大。

5.1 时空域交互关系数据集

Koppula 和 Saxena(2016)提出了时空人—物交互关系检测数据库 CAD120,仅用 120 个视频记录了 10 种由人执行的高层行为和 12 种物体属性,数据量与行为类别少,并不能满足真实世界感知任务。时空人—物交互关系检测深度学习的发展的一个重要时间节点为美国斯坦福大学李飞飞教授团队在 2020 年提出大规模时空人—物交互关系检测数据集 Action Genome(Ji 等,2020),其提供了近 1 万个视频,共计 82 h,标注了 35 种物体关系和 25 种交互关系,相较之前的时空人—物交互关系检测数据库 CAD120,视频数量提升近 20 倍。然而,Action Genome 数据集是基于 Charades 数据集(Sigurdsson 等,2016)标注产生,而 Charades 数据集是行为分类数据集,数据集中每一个时间片段仅仅包含一个人物主体关系,时间片段中人—物交互关系较为单一、简单。同时,Action Genome 数据集中的视频是人为设计拍摄的,真实性低于真实世界发生的交互关系。Chiou 等人(2021)基于 VidOR(video object relation)数据集(Shang 等,2019)构建了一个大型视频人—物交互关系检测数据集 VidHOI。其遵循视频和 HOI 任务中的通用协议,使用以关键帧为中心的策略,包含了 7 122 个视频,具有标注的关键帧达 7.3 M 帧,成为时空人—物交互关系检测任务新的基准。

5.2 时空域交互关系识别与检测算法

Wang 等人(2019a)构建了时空图解析网络用于识别视频中的人—物交互关系。其用图序列对视频进行编码,并学习时空关系在时间维度和空间图拓扑上的演变。Ji 等人(2021)提出了人—物交互关系变换模型,其先检测静态图像中存在的视觉线索,例如目标、关系或者人体姿态,接着通过变化网络解码这些视觉线索,检测存在交互关系的物体与其交互关系,其中视觉线索可有效帮助网络聚焦于发生交互关系的物体,避免背景物体对网络建模的干扰。Nagarajan 等人(2019)研究存在于具有交互关系的物体的“热点”,其定义为与人—物交互关系高度相关的特殊区域,并提出一个集成了动作识别、行为预测和特征定位的网络框架,直接从视频中学习物体属性,无需手工标注关键点与分割标签。Chiou 等人(2021)提出了一个简单且十分有效的算法框架时空 HOI 检测器(spatial-temporal HOI detection, ST-HOI),其首先揭示了常见的行为检测基线算法难以建模当前任务的原因,主要由于时空域池化过程中,人与物体的移动会造成特征不一致的问题,接着提出了 ST-HOI 网络,其使用人和目标的轨迹信息、姿态信息等时域信息弥补特征不一致引入的偏差,最后使用 3D 网络,逐关键帧粒度生成时空域人—物交互关系结果。然而 ST-HOI 算法依赖额外特征信息,例如轨迹、姿态信息,对数据集标注要求较高。Sunkesula 等人(2020)基于图卷积网络和分层循环神经网络提出了 LIGHTEN 网络。其首先在帧粒度上建模人与目标间的局部交互信息;接着,使用帧粒度的嵌入向量生成片段粒度的嵌入向量,然后逐级细化。嵌入向量通过图结构建模,人和物体作为场景中的图节点。LIGHTEN 网络仅使用视觉特征,摒弃深度图特征和人体 3D 姿态信息,在 CAD-120 数据集上达到了较优效果。然而,针对真实视频场景,冗余的背景帧占比高,帧间人—物交互关系复杂,LIGHTEN 网络难以逐帧构建稳定的图结构,同时,分层神经网络作用于长序列中存在长期依

赖问题,影响网络性能。

6　研究展望

基于深度学习的人—物交互关系检测自 2015 年发展至今,已取得显著的突破,人—物交互关系检测性能逐年提升,如表 3 和表 4 所示。在多个实际场景落地应用,成为图像级人体行为分析的最热门解决方案。然而人—物交互关系检测依旧面临着巨大的挑战。总体而言,当前的挑战可归纳于人—物交互关系检测的两个核心问题:人—物关联和动作关系预测。在人—物关联问题上,构造合适和通用的特征表示是较大的难题,在不同的现实人—物交互场景下,深度模型通过何种特征可以判断人—物对是否有交互,当下算法难以显式通过深度学习网络建模,都是依据黑盒算法端到端地学习来判断。而面向动作关系分类,主要有两个挑战:1)在现实生活场景中与人交互的物体种类繁多,且交互的种类也很复杂,从而导致很难通过标注一个特定化的数据集来解决开放场景的人—物交互关系检测,故而当下的交互检测模型都被限制在非常有限的场景下的固定几种交互类型;2)仅依据单幅图像对部分人体行为难精准理解,由于人体动作行为是一个持续且连续的过程,单单依据一幅图像进行动作关系分类不仅只局限于固定的关系类型,且还会产生一定层面的误解。

针对当前人—物交互关系检测算法面临的以上挑战,本文提出交互关系检测的未来发展趋势应该面向更广阔的应用场景和需求,得到更精准、可解释性更强的人—物交互关系检测算法模型。

表 3　基于深度学习的人—物交互关系检测算法在实例级人—物交互关系检测数据库 HICO-DET 的平均精度性能对比

Table 3　The performance comparison with mean average precision(mAP) among deep learning-based HOI detection methods in instance-level HOI detection dataset HICO-DET

/%

算法		主干	检测方式	默认设置			物体已知		
				全集	稀有	非稀有	全集	稀有	非稀有
两阶段	InteractNet	ResNet-50	多流	9.94	7.16	10.77	–	–	–
	GPNN	ResNet-152	图	13.11	9.34	14.23	–	–	–
	ICAN	ResNet-50	多流	14.84	10.45	16.15	16.26	11.33	17.73
	PMFNet	ResNet-50	多流	17.46	15.65	18.00	20.34	17.47	21.20
	VSG	ResNet-152	图	19.80	16.05	20.91	–	–	–
	DRG	ResNet-50	图	19.26	17.74	19.71	23.40	21.75	23.89
单阶段	Uniondet	ResNet-50	框	17.58	11.72	19.33	19.76	14.68	21.27
	IP-Net	Hourglass-104	点	19.56	12.79	21.58	22.05	15.77	23.92
	DIRV	EfficientDet-d3	框	21.78	16.38	23.39	25.52	20.84	26.92
	PPDM	Hourglass-104	点	21.94	13.97	24.32	24.81	17.09	27.12
	HOITrans	ResNet-50	查询	23.46	16.91	25.41	26.15	19.24	28.22
	GGNet	Hourglass-104	点	23.47	16.48	25.60	27.36	20.23	29.49
	HOTR	ResNet-50	查询	25.10	17.34	27.42	–	–	–
	AS-Net	ResNet-50	查询	28.87	24.25	30.35	31.74	27.07	33.14
	QPIC	ResNet-50	查询	29.07	21.85	31.23	31.68	24.14	33.93
	PhraseHOI	ResNet-50	查询	29.29	22.03	31.46	31.97	23.99	34.36
	CDN	ResNet-50	查询	31.78	27.55	33.05	34.53	29.73	35.96

注:“–”表示无此项内容。

表 4　基于深度学习的人—物交互关系检测算法在人—物交互关系检测数据库 V-COCO 的平均精度对比

Table 4　The performance comparison among deep learning-based HOI detection methods in HOI detection dataset V-COCO

方法		主干	细分类	精度/%
两阶段	InteractNet	ResNet-50	多流	40.0
	GPNN	ResNet-152	图	44.0
	ICAN	ResNet-50	多流	45.3
	TIN	ResNet-50	多流	47.8
	VSG	ResNet-152	图	51.8
	DRG	ResNet-50	图	51.0
	PMFNet	ResNet-50	多流	52.8
单阶段	Uniondet	ResNet-50	框	47.5
	IP-Net	Hourglass-104	点	51.0
	AS-Net	ResNet-50	查询	53.9
	RR-Net	Hourglass-104	点	54.2
	GGNet	Hourglass-104	点	54.7
	HOTR	ResNet-50	查询	55.2
	DIRV	EfficientDet-d3	框	56.1
	QPIC	ResNet-50	查询	58.8
	PhraseHOI	ResNet-50	查询	57.4
	CDN	ResNet-50	查询	63.9

6.1　大规模预训练模型指导的人—物交互关系检测

随着算力和数据规模的跨越式增长，通过海量数据训练大规模深度学习模型来探索深度学习的边际成为一个趋势（Devlin 等，2019）。受益于大模型强大的泛化能力，通过知识迁移的模式，诸多视觉任务在性能和泛化能力取得了显著的提升。但利用大模型来辅助人—物交互关系检测的研究仍旧处于相对空白的阶段。由于现实生活中人—物动作种类繁多，同时与其产生交互关系的物体复杂，进而，所产生的交互三元组类型很难通过有限的数据标注来覆盖。因此，在现实应用场景中，长尾数据分布以及新人—物交互关系类型发现成为亟需解决的问题。虽然现有方法通过数据增强以及特定损失函数设计等方法取得了些许进展，但其性能始终受限于固定的数据集标注。

因此，本文认为未来的发展趋势是充分挖掘大规模视觉语言预训练模型与人—物交互关系检测任务之间关系，指导提升后者性能。人—物交互关系三元组可以自然地构成〈主语，谓语，宾语〉文本，将人—物交互关系检测任务与语言任务相结合，可利用图像文本数据获取成本低、标注方式简单等特点，低开销提升前者任务性能。随着语言任务数据量的增加，视觉特征的泛化性逐渐增强，以提升对未知或开放场景的人—物交互关系理解能力。此外，随着近期视觉语言大规模预训练模型的快速发展，通过用数十亿数据训练的视觉语言模型打破了视觉和语言的鸿沟，因此，通过知识迁移相关算法，在不增加额外计算开销的情形下，利用视觉语言预训练模型来指导人—物交互关系预测，提升模型对人—物交互关系检测的精准性和泛化性是未来的重要研究方向。

6.2　基于无监督人—物交互关系检测预训练模型

网络预训练权重对网络性能优化有着至关重要的影响。现有分类、检测基线网络常基于 ImageNet、COCO 数据集预训练权重初始化，且在下游任务中取得了显著性能提升。在人—物交互关系检测任务中，基线网络基于 COCO 检测数据集预训练权重初始化，可有效提升人和物体的检测结果。

然而，仅关注目标检测性能是不足的，人—物对匹配与关系分类作为人—物交互关系检测任务中的重要组成部分，鲜有预训练模型构建关系知识。在前沿框架中，基线网络基于 COCO 检测数据集预训练网络初始化，使得模型更加关注于自然场景下的目标特性，不利于划分发生交互关系的目标与背景目标分界线，损害算法性能。随着硬件平台的升级，大规模多模态预训练研究逐渐成为研究热点（Li 等，2020a；Lu 等，2019）。然而，大规模预训练模型通常为图像分类、图像检测和视频理解等任务定制化设计，并不完全适配关系分类任务。

为解决上述局限性问题，通过设计专用的代理任务建模人—物交互关系，采用无监督训练策略生成大型跨模态交互关系预训练网络是未来的研究重点。预训练模型可有效提升关系建模网络的泛化性，将领域整体算法的性能提升至新的高度。

6.3　基于端到端优化的时空人—物交互关系检测

高性能高效率的视频时空人—物交互关系检测算法是未来研究重点。受到传统目标检测领域启发，未来视频时空人—物交互关系检测算法可分为

帧粒度多阶段检测算法与视频粒度端到端检测算法。

帧粒度多阶段检测算法首先通过视频行为定位算法获得推荐行为生成区域，降低无剪辑视频中对背景冗余帧的计算消耗，再利用空域人—物交互关系检测算法检测出帧粒度上发生的交互关系，最后通过管道链接等后处理策略在时序中建立关联，以获得时空人—物交互关系管道。该方案将时空检测任务解耦，可灵活引入现有图像领域人—物交互关系检测算法研究经验。然而，该方案流程复杂，基于手工设计的姿态、轨迹等信息提取时耗长，多阶段检测算法无法端到端优化，容易产生时序上过分割或欠分割问题。同时，该方案将每一帧视为独立同分布，很难区分对时序强依赖的行为，例如〈人，上，车〉与〈人，下，车〉，〈人，推开，门〉与〈人，关闭，门〉，这些行为在静态图像中具有高度相似性，但在时间序列中，需要考虑场景中对象相对于人的方向变化。

视频粒度端到端检测算法将会是未来的主流，我们预测该算法使用 3D 或更高维算子直接处理无剪辑视频序列，融合时序信息以生成高质量时空人—物交互关系管道。3D 卷积网络运算量大，受限于计算资源，如何能建模时空关系，降低网络维度，用更加优美的方式端到端预测人—物交互关系管道成为未来重要研究趋势。在时空域人—物交互关系识别过程中，目标的光流、人体姿态、人体部件的移动或相互关系蕴含着丰富的交互信息，通过目标光流可以获得时序上目标移动轨迹，人体姿态与部件可以获得细粒度交互信息。未来可以进一步与时空域人—物交互关系检测任务相融合。设计端到端网络时，面对性能与计算消耗的权衡，考虑额外补充信息可有效缓解网络规模降低对时序信息建模的影响。

在人—物交互关系检测的未来发展方向上，时空交互检测必定会添上浓墨重彩的一笔，不仅由于其更精准的动作关系表示，同时也源于其更广阔的应用场景。

7 结 语

人—物交互关系检测是智能人体行为分析的重要研究方向，也是计算机视觉领域的热门课题之一。随着深度学习的发展和大规模数据库的提出，人—物交互关系检测在基础理论、技术方法和应用落地等方面取得了显著的进步。本文先从空域人—物交互关系检测任务出发，从数据库和方法两个方面概述了前沿进展。在数据库部分，从标注粒度出发，分别对实例级、部件级和像素级 3 个层级标注的数据库进行了总结概括和对比。在方法层面，从网络结构切分，将现有方法分为两阶段序列式和单阶段端到端式两个流派，进行归纳、分析、比对。接着，介绍了时空域人—物交互关系检测任务，总结并分析了数据集与前沿算法的性能与优劣。最后，从基于深度学习的人—物交互关系检测现存的研究挑战，以及从技术难点出发，对其未来发展的趋势进行预测。希望通过本文对基于深度学习的人—物交互关系检测的总结概述分析，辅助相关方向的研究人员快速了解和掌握该方向的研究现状和难点、通点，并通过对未来方向的预测给研究人员以启发。

参考文献(References)

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chao Y W, Liu Y F, Liu X Y, Zeng H Y and Deng J. 2018. Learning to detect human-object Interactions//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, USA: IEEE: 381-389 [DOI: 10.1109/WACV.2018.00048]

Chen M F, Liao Y, Liu S, Chen Z Y, Wang F and Qian C. 2021. Reformulating HOI detection as adaptive set prediction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9000-9009 [DOI: 10.1109/CVPR46437.2021.00889]

Chiou M J, Liao C Y, Wang L W, Zimmermann R and Feng J S. 2021. ST-HOI: a spatial-temporal baseline for human-object interaction detection in videos//Proceedings of 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval. Taipei, China: ACM: 9-17 [DOI: 10.1145/3463944.3469097]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dong Q, Tu Z W, Liao H F, Zhang Y T, Mahadevan V and Soatto S. 2021. Visual relationship detection using part-and-sum transformers with composite queries//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3530-3539 [DOI: 10.1109/ICCV48922.2021.00353]

Fang H S, Xie Y C, Shao D and Lu C W. 2021. DIRV: dense interaction region voting for end-to-end human-object interaction detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, USA: 1291-1299

Gao C, Zou Y L and Huang J B. 2018. ICAN: instance-centric attention network for human-object interaction detection [EB/OL]. [2021-12-16]. https://arxiv.org/pdf/1808.10437.pdf

Gao C, Xu J R, Zou Y L and Huang J B. 2020. DRG: dual relation graph for human-object interaction detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 696-712 [DOI: 10.1007/978-3-030-58610-2_41]

Gkioxari G, Girshick R, Dollár P and He K M. 2018. Detecting and recognizing human-object interactions//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8359-8367 [DOI: 10.1109/CVPR.2018.00872]

Gupta S and Malik J. 2015. Visual semantic role labeling [EB/OL]. [2021-12-16]. https://arxiv.org/pdf/1505.04474.pdf

He T, Gao L L, Song J and Li Y F. 2021. Exploiting scene graphs for human-object interaction detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15964-15973 [DOI: 10.1109/ICCV48922.2021.01568]

Hou Z, Peng X J, Qiao Y and Tao D C. 2020. Visual compositional learning for human-object interaction detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 584-600 [DOI: 10.1007/978-3-030-58555-6_35]

Hou Z, Yu B S, Qiao Y, Pneg X J and Tao D C. 2021. Detecting human-object interaction via fabricated compositional learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14641-14650 [DOI: 10.1109/CVPR46437.2021.01441]

Ji J W, Desai R and Niebles J C. 2021. Detecting human-object relationships in videos//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8086-8096 [DOI: 10.1109/ICCV48922.2021.00800]

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]

Kim B, Choi T, Kang J and Kim H J. 2020. UnionDet: union-level detector towards real-time human-object interaction detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 498-514 [DOI: 10.1007/978-3-030-58555-6_30]

Kim B, Lee J, Kang J, Kim E S and Kim H J. 2021. HOTR: end-to-end human-object interaction detection with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 74-83 [DOI: 10.1109/CVPR46437.2021.00014]

Koppula H S and Saxena A. 2016. Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1): 14-29 [DOI: 10.1109/TPAMI.2015.2430335]

Li L J, Chen Y C, Cheng Y, Gan Z, Yu L C and Liu J J. 2020a. HERO: hierarchical encoder for video + language omni-representation pre-training//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. [s. l.]: ACM: 2046-2065 [DOI: 10.18653/v1/2020.emnlp-main.161]

Li Y L, Xu L, Liu X P, Huang X J, Xu Y, Wang S Y, Fang H S, Ma Z, Chen M Y and Lu C W. 2020b. PaStaNet: toward human activity knowledge engine//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 379-388 [DOI: 10.1109/CVPR42600.2020.00046]

Li Y L, Zhou S Y, Huang X J, Xu L, Ma Z, Fang H S, Wang Y F and Lu C W. 2019. Transferable interactiveness knowledge for human-object interaction detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3580-3589 [DOI: 10.1109/CVPR.2019.00370]

Li Z M, Zou C, Zhao Y, Li B X and Zhong S. 2022. Improving human-object interaction detection via phrase learning and label composition [EB/OL]. [2021-12-16]. https://arxiv.org/pdf/2112.07383.pdf

Liao Y, Liu S, Wang F, Chen Y J, Chen Q and Feng J S. 2020. PPDM: parallel point detection and matching for real-time human-object interaction detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 479-487 [DOI: 10.1109/CVPR42600.2020.00056]

Lin T Y, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Lin X, Zou Q and Xu X X. 2020. Action-guided attention mining and relation reasoning network for human-object interaction detection//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: [s. n.]: 1104-1110 [DOI: 10.24963/ijcai.2020/154]

Liu S, Wang Z T, Gao Y L, Ren L J, Liao Y, Ren G H, Li B and Yan S C. 2021. Human-centric relation segmentation: dataset and solution. IEEE Transactions on Pattern Analysis and Machine Intelligence:#3075846[DOI: 10.1109/TPAMI.2021.3075846]

Lu J S, Batra D, Parikh D and Lee S. 2019. ViLBERT: pretraining

task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: 13-23

Nagarajan T, Feichtenhofer C and Grauman K. 2019. Grounded human-object interaction hotspots from video//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8687-8696 [DOI: 10.1109/ICCV.2019.00878]

Peyre J, Sivic J, Laptev I and Schmd C. 2019. Detecting unseen visual relations using analogies//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1981-1990 [DOI: 10.1109/ICCV.2019.00207]

Qi S Y, Wang W G, Jia B X, Shen J B and Zhu S C. 2018. Learning human-object interactions by graph parsing neural networks//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 407-423 [DOI: 10.1007/978-3-030-01240-3_25]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/tpami.2016.2577031]

Shang X D, Di D L, Xiao J B, Cao Y, Yang X and Chua T S. 2019. Annotating objects and relations in user-generated videos//Proceedings of 2019 on International Conference on Multimedia Retrieval. Ottawa, Canada: ACM: 279-287 [DOI: 10.1145/3323873.3325056]

Shen L Y, Yeung S, Hoffman J, Mori G and Li F F. 2018. Scaling human-object interaction recognition through zero-shot learning//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA: IEEE: 1568-1576 [DOI: 10.1109/wacv.2018.00181]

Sigurdsson G A, Varol G, Wang X L, Farhadi A, Laptev I and Gupta A. 2016. Hollywood in homes: crowdsourcing data collection for activity understanding//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 510-526 [DOI: 10.1007/978-3-319-46448-0_31]

Sun X, Hu X W, Ren T W and Wu G S. 2020. Human object interaction detection via multi-level conditioned network//Proceedings of 2020 on International Conference on Multimedia Retrieval. Dublin, Ireland: ACM: 26-34 [DOI: 10.1145/3372278.3390671]

Sunkesula S P R, Dabral R and Ramakrishnan G. 2020. LIGHTEN: learning interactions with graph and hierarchical temporal networks for HOI in videos//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 691-699 [DOI: 10.1145/3394171.3413778]

Tamura M, Ohashi H and Yoshinaga T. 2021. QPIC: query-based pairwise human-object interaction detection with image-wide contextual information//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10405-10414 [DOI: 10.1109/CVPR46437.2021.01027]

Ulutan O, Iftekhar A S M and Manjunath B S. 2020. VSGNet: spatial attention network for detecting human object interactions using graph convolutions//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13614-13623 [DOI: 10.1109/CVPR42600.2020.01363]

Wan B, Zhou D, Liu Y F, Li R J and He X M. 2019. Pose-aware multi-level feature network for human object interaction detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9468-9477 [DOI: 10.1109/ICCV.2019.00956]

Wang H, Zheng W S and Ling Y B. 2020. Contextual heterogeneous graph network for human-object interaction detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 248-264 [DOI: 10.1007/978-3-030-58520-4_15]

Wang N, Zhu G M, Zhang L, Shen P Y, Li H S and Hua C. 2019a. Spatio-temporal interaction graph parsing networks for human-object interaction recognition//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 4985-4993 [DOI: 10.1145/3474085.3475636]

Wang T C, Anwer R M, Khan M H, Khan F S, Pang Y W, Shao L and Laaksonen J. 2019b. Deep contextual attention for human-object interaction detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5693-5701 [DOI: 10.1109/ICCV.2019.00579]

Wang T C, Yang T, Danelljan M, Khan F S, Zhang X Y and Sun J. 2020b. Learning human-object interaction detection using interaction points//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4115-4124 [DOI: 10.1109/CVPR42600.2020.00417]

Xu K L, Li Z M, Zhang Z J, Dong L Z, Xu W H, Yan L X, Zhong S and Zou X. 2022. Effective actor-centric human-object interaction detection. Image and Vision Computing, 121: #104422 [DOI: 10.1016/j.imavis.2022.104422]

Zhang F Z, Campbell D and Gould S. 2021. Spatially conditioned graphs for detecting human-object interactions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13299-13307 [DOI: 10.1109/ICCV48922.2021.01307]

Zhong X B, Qu X, Ding C X and Tao D C. 2021b. Glance and gaze: inferring action-aware points for one-stage human-object interaction detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13229-13238 [DOI: 10.1109/CVPR46437.2021.01303]

Zhuang B H, Wu Q, Shen C H, Reid I and van den Hengel A. 2018. HCVRD: a benchmark for large-scale human-centered visual relationship detection//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA:

AAAI: 7631-7638

Zou C, Wang B H, Hu Y, Liu J Q, Wu Q, Zhao Y, Li B X, Zhang C G, Zhang C, Wei Y C and Sun J. 2021. End-to-end human object interaction detection with HOI transformer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11820-11829 [DOI: 10.1109/CVPR46437.2021.01165]

作者简介

廖越,1994 年生,男,博士研究生,主要研究方向为人—物交互关系检测和目标检测。
E-mail: liaoyue.ai@gmail.com

刘偲,通信作者,1985 年生,女,教授,主要研究方向为以人为中心的视觉理解和多模态分析。
E-mail:liusi@buaa.edu.cn

李智敏,1997 年生,男,硕士研究生,主要研究方向为人—物交互关系检测与视频理解。
E-mail:zhiminli.cn@outlook.com