视频中的未来动作预测研究综述

张天予[1),2),3)] 闵巍庆[1),2),3)] 韩鑫阳[3)] 蒋树强[1),2),3)] 芮 勇[4)]

[1)](中国科学院智能信息处理重点实验室 北京 100190)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](中国科学院大学 北京 100049)

[4)](联想集团 北京 100085)

摘 要 预测未来是人类与生俱来的能力,也是实现人工智能的重要手段.近年来,视频中的未来动作预测逐渐成为计算机视觉领域的研究热点,具有重要的理论研究意义,并在安防监控、自动驾驶、家庭服务、工业协作以及虚拟现实等方面有着广泛的应用前景.本文对视频中的未来动作预测领域进行综述,首先明确定义了未来动作预测的研究框架.随后概述了该领域的发展历史,并重点介绍了短期动作预测和长期动作预测两种主要的问题形式.然后从模型结构、数据模态、算法策略和预测对象等不同维度对主要方法和技术进行了总结.接下来简要归纳了视频中的未来动作预测领域常用的数据集,并给出了不同方法在主流数据集上的性能对比和分析.最后本文围绕扩展现有数据集的规模和多样性、缩短模型的推理时间、从无标注或少量标注数据中学习等未来可能的研究方向进行了总结和展望.

关键词 未来动作预测;短期预测;长期预测;深度学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.01315

A Survey on Future Action Anticipation in Videos

ZHANG Tian-Yu[1),2),3)] MIN Wei-Qing[1),2),3)] HAN Xin-Yang[3)] JIANG Shu-Qiang[1),2),3)] RUI Yong[4)]

[1)](*Key Lab of Intelligent Information Processing, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*University of Chinese Academy of Sciences, Beijing* 100049)

[4)](*Lenovo Group Ltd, Lenovo Corporate Research, Beijing* 100085)

Abstract Predicting the future is an innate ability of mankind, which is also a significant step towards artificial intelligence. In recent years, future action anticipation in videos has become a research hotspot in the field of computer vision. It has important theoretical research significance, as well as wide applications including security monitoring, autonomous driving, home service, industrial assistance and virtual reality. In this paper, we provide a comprehensive overview of future action anticipation in videos. We first define the research framework of future action anticipation in videos. Then we summarize the development history in which we focus on two main paradigms, namely short-term action anticipation and long-term action anticipation. Then we introduce the prevailing methods and techniques from different dimensions such as architectures, modalities, learning strategies and predicted targets. Next, we briefly summarize the commonly used datasets in this area, and provide the performance comparison and analysis of different methods

收稿日期:2022-05-09;在线发布日期:2023-01-15.本课题得到科技创新 2030-"新一代人工智能"重大项目(2018AAA0102500)资助.**张天予**,博士研究生,主要研究方向为多媒体内容分析和理解、视频中的未来动作预测.E-mail:tianyu.zhang@vipl.ict.ac.cn.**闵巍庆**,博士,副研究员,中国计算机学会(CCF)高级会员,主要研究方向为多媒体内容分析和理解、食品计算等.**韩鑫阳**,本科生,主要研究方向为多媒体内容分析和理解、视频中的未来动作预测.**蒋树强**(通信作者),博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为图像/视频等多媒体信息的分析、理解与检索技术和多模态智能技术等.E-mail:sqjiang@ict.ac.cn.**芮 勇**,博士,教授,中国计算机学会(CCF)会士,主要研究方向为多模态检索、知识挖掘等.

on the mainstream datasets. Finally, we prospect future research directions of future action anticipation in videos, including extending the scale and diversity of existing datasets, shortening the inference time of models and learning from data with few or without action annotations.

Keywords future action anticipation; short-term anticipation; long-term anticipation; deep learning

1 引 言

作为人类智能的自然延伸和拓展，人工智能已经渗透到人类生活的方方面面. 近年来，深度学习技术的飞速发展使得机器在图像识别等感知任务层面达到甚至超越了人类水平，但机器在解决推理、规划、联想、创作等复杂的认知智能化任务时与人类水平还存在明显的差距[1]. 人类大脑有想象、模拟和评估未来可能发生事件的能力，这对于人类决策和规划现实世界中的复杂问题具有重要的意义，同时也是机器实现更高层级的认知性智能亟需获取的能力. 另一方面，随着视频获取设备的普及以及大数据技术的发展，视频数据已成为互联网时代海量信息的主要载体，如何使得机器具备根据观察到的视频片段预测出尚未发生的动作的能力显得尤为重要. 因此视频中的未来动作预测逐渐获得研究人员的关注，成为计算机视觉的研究热点，在安防监控、自动驾驶、家庭服务、工业协作以及虚拟现实等诸多领域有着广阔的研究前景. 例如，在工业协作领域，机器可以根据工人执行工序的画面引导后者进行正确的操作，并且提前发出警报，避免危险动作的发生. 由此可见，视频中的未来动作预测对于改善人类的生产生活方式具有深远的意义，是一个值得深入探索的研究方向.

视频中的未来动作预测存在着许多研究难点. 如图 1 所示，相较于计算机视觉领域广泛研究的动作识别任务[2-4]，动作预测任务有着本质区别，需要对不确定的未来做出假设. 具体而言，动作识别针对已发生的、可观察到的视频片段打上动作标签，是对既成事实的内容进行分析，而动作预测需要在理解已发生动作的基础上预测接下来可能发生的动作. 由于未来的内容是无法观察到的，有着天然的不确定性，因此可能存在多个合理的候选动作. 这些不确定因素的干扰，增加了基于视频的未来动作预测研究的难度. 此外，视频中的未来动作预测还存在着视频数据本身带来的挑战. 相较于静态的图像数据，视频数据由于增加了时序维度更具复杂性和冗余性，往往存在背景变化、相机抖动等问题，对于提取有效的信息造成了难度. 而且视频动作往往存在类内差异大、类间差异小的特点，即某些动作类别内部的差异很大，例如针对“跑步”这一类别的动作，不同的人在执行时呈现的视觉内容往往大不相同；而某些动作类别之间的差异却很小，例如“洗碗”和“洗盘子”等细粒度的动作.

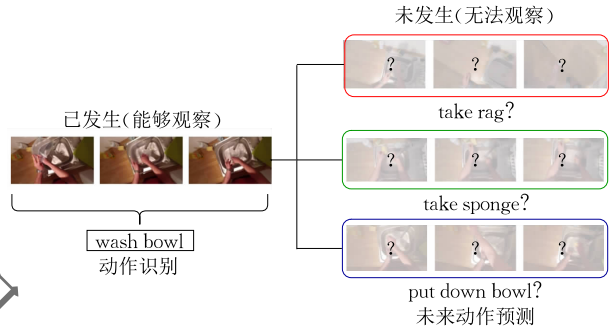


图 1 未来动作预测与动作识别对比

视频中的未来动作预测隶属于视频理解的研究范畴，紧密相关的研究课题大致可以分为早期动作识别以及视频中的其它类型预测两部分. 早期动作识别(early action recognition)[5-6]也叫早期动作预测(early action prediction)或者动作预测(action prediction/anticipation)，要求机器尽可能早地预测出正在发生的动作，即给定某个动作的一部分片段(如前 10%的部分)，判断出该动作的类别. 相较之下，未来动作预测需要预测出尚未开始的未来动作，而非简单地根据不完整的动作片段推断出完整动作的类别，因而不确定性更大. 视频中的其它类型预测包括：(1) 视频中的像素预测(video pixel prediction)[7-8]. 根据观察到的视频序列，为潜在的未来图像生成像素值，对已有的视频帧进行外插预测；(2) 视频中的轨迹预测(video trajectory prediction)[9-10]. 将观察到的视频处理成坐标序列，以坐标点而非原始像素的形式预测未来的序列；(3) 视频中的人体姿态预测(video motion/pose prediction)[11-12]. 和轨迹预测类似，专注对人体骨骼关键点运动的研究，以坐标点的形式预测人体未来的运动；(4) 视频中的物体和

区域预测(video object and region prediction)[13-14]. 常用于人物交互(human-object interaction)的场景，根据观察内容预测接下来可能操作的物体，输出物体类别和对应区域. 针对第一人称视频，部分研究工作预测未来的人眼关注(eye gaze)区域[15]或者手活动的区域[16]. 这些子课题与未来动作预测相似，都需要机器根据观察到的视频信息去预测不确定的未来信息. 但与它们相比，未来动作预测所研究的对象是语义信息更加丰富的动作类别，而不仅仅是图像中的像素或坐标点，这使得机器需要建模和推理更加复杂和抽象的语义线索，更具有挑战性.

作为新兴的研究课题，目前涉及视频中的未来动作预测的综述文献较少. Rasouli[17]总结了深度学习方法在基于视觉的预测中的应用，但他们并未对未来动作预测领域的研究进展进行详细介绍. Rodin等人[18]侧重于第一人称视角，对视频中的动作预测、轨迹预测、区域预测等方面进行了概括总结，但缺少对第三人称视角下的动作预测的总结和梳理. Zhao等人[19]围绕早期动作识别和未来动作预测两方面介绍了主流工作，但并未聚焦未来动作预测进行深入的分析和总结. 鉴于未来动作预测的研究意义以及所蕴含的应用价值，本文针对视频中的未来动作预测研究进行了系统性的综述. 在第2节中，首先结合研究任务、研究方法、应用场景、发展趋势等方面对未来动作预测的研究框架给出了明确的定义；在第3节中，简要回顾了视频中的未来动作预测的发展历史；在第4节中从模型结构、数据模态、算法策略、预测对象等不同维度对现有方法和技术进行了梳理和总结；在第5节中，归纳了未来动作预测领域常用的数据集和评价指标，并给出了一些代表性方法在主流数据集上的性能对比与分析；在第6节中，对视频中的未来动作预测研究领域可能的发展方向进行了总结和展望.

2　研究框架

作为新兴的研究方向，目前学术界对于未来动作预测的研究框架缺乏统一的定义. 在计算机视觉中，动作(action/activity)是一个宽泛的概念，涵盖了简单的手势运动(如“挥手”)、人与人或人与物的交互行为(如“握手”、“打电话”)等不同层级的人类活动[20]. 而预测(prediction/anticipation)是源自于心理学[21]的概念，指对未来做出假设的过程. 尽管很难给出正式的定义，本文将视频中的未来动作预测形式化表示为 $f: x \rightarrow y$，其中 x 表示已经发生的一段视频，y 表示在 x 结束时尚未开始的动作类别. 如图2所示，基于对已有研究工作的总结以及对未来工作的展望，本文构建了视频中的未来动作预测的研究框架：在研究任务层面，现有的研究工作经过近十年的发展，除了形成了短期动作预测、长期动作预测两种统一的问题设置之外，为未来动作生成句子描述也具有发展潜力，在第3节将展开介绍；在研究方法层面，现有的研究工作以深度学习模型为依托，从结合视觉信息和语义信息、利用外部信息进行知识蒸馏、多任务学习等方面探索方法性能的提升，在第4节将展开介绍. 此外，本文还展望了该领域未来的发展趋势，期望它们能够适应多样化的应用场景，具体内容将在第6节展开介绍.

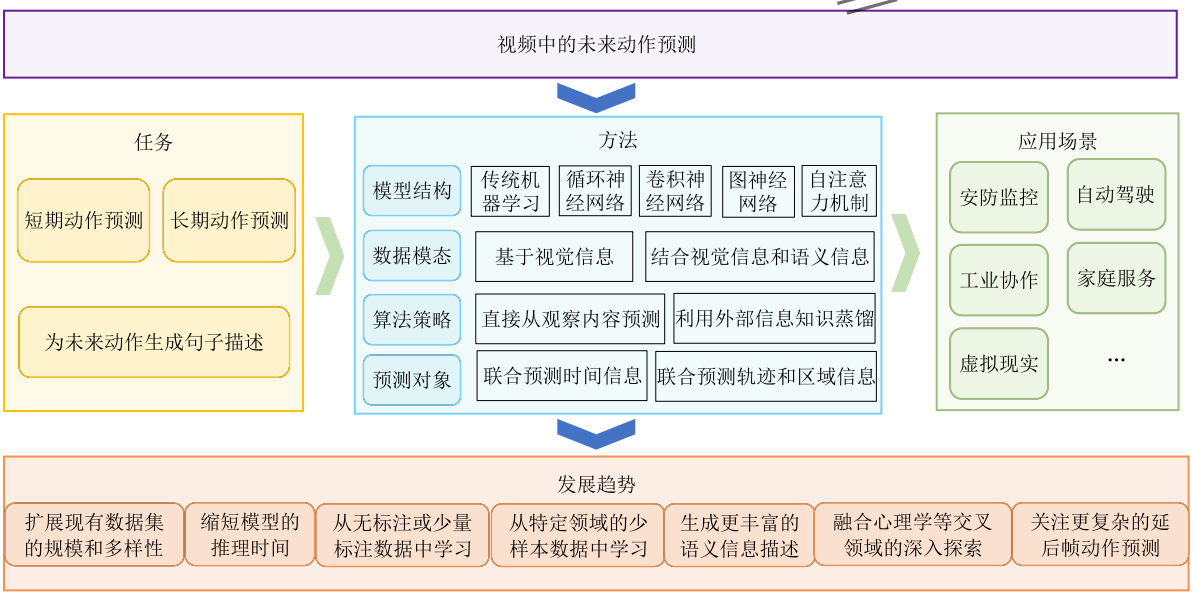


图2　视频中的未来动作预测的研究框架

3 发展历史

视频中的未来动作预测在最近十年间逐渐发展形成. 在计算机视觉的发展过程中,研究人员期望机器不再局限于对已经发生的动作进行分类,而是能够根据观察的视频内容进行时序推理,推断出尚未发生的动作. 最早的研究工作大致出现在 2012 年左右,Li 等人[22]提出将复杂的人物交互活动(例如“打电话”)分解编码成一系列动作单元(例如“拿起电话”、“接电话”、“放下电话”),并研究在给定前几个动作单元的情况下推断出后续的动作单元以及复杂活动的类别. 随后研究人员面向不同的应用场景展开探索,例如针对双人交互场景,根据一个人的动作预测另一个人的动作[23];针对视频监控场景,将未来动作预测作为子任务,与行人的轨迹、位置等信息进行联合预测[24-27];针对体育运动场景,预测球员的下一步或者球的运动轨迹和位置[28-30];针对人机交互场景,利用物体的可供性[31-32]对用户可能执行的潜在危险操作进行预测[33];针对交通场景,通过结合驾驶员的人脸画面、面向道路拍摄的视频画面、全球定位系统以及车辆速度等信息预测驾驶员的下一步动作[34-35]或预测潜在的交通事故[36-37],或者研究行人过马路的意图预测,将行人未来是否过马路视为二分类问题,并预测行人未来可能出现的位置[38]. 总而言之,研究人员针对不同类型的视频数据探索了不同形式的未来动作预测问题,但缺少公开的基准数据集和评测指标,也没有形成统一的问题形式.

2018 年,Damen 等人[39]发布了以第一人称视频拍摄的大规模无剧本数据集 EPIC-Kitchens-55,并在此基础上明确定义了短期动作预测(short-term action anticipation)问题:对于一个在 $t+\tau$ 时刻开始的动作,机器需要根据一段结束于 t 时刻的视频序列 $x=(f_1,f_2,\cdots,f_t)$,预测出该动作的类别 $y_{t+\tau}$,其中 τ 被称为预测时间(anticipation time),表示提前多长时间进行预测,在没有特殊说明的情况下通常设置为 1 s,为后续的研究工作提供了统一的形式化表示,同时弥补了先前工作存在的一些问题,例如没有形成统一的预测时间、关注的大多为“拥抱”、“到达”等粗粒度动作,缺少对细粒度动作的分析与理解. 由于短期动作预测的对象是下一个动作的类别,通常也称为下一个动作预测(next action anticipation)或简称为动作预测(action anticipation),随着深度学习的发展,逐渐涌现了许多代表性的方法和技术[40-44]. 同年,Abu 等人[45]提出了长期动作预测(long-term action anticipation)任务,也叫密集预测(dense anticipation),旨在观察一段视频的前 $p\%$ 部分 $x=(f_1,f_2,\cdots,f_p)$ 的情况下,预测出后面 $q\%$ 的内容,输出未来每帧对应的动作类别 $y=(y_{p+1},y_{p+2},\cdots,y_{p+q})$. 相较于短期动作预测,该任务预测的是序列化的多个动作,预测时间最多可以达到 5 min,因而更加复杂. 在方法和技术层面,近年来同样涌现了许多代表性的工作[42,46-47].

此外,还有一些工作[48-51]尝试将未来动作预测问题从简单地预测动作类别(如“煎鸡翅”)拓展成为未来的视频帧生成句子描述(如“将鸡翅煎至两面呈金黄色”),尽管这一类工作目前尚未形成统一的数据集和评测标准,但生成句子描述相较于动作标签能够提供更加丰富的信息,更能满足实际应用的需求,因此具有可观的发展潜力.

图 3 展示了视频中的未来动作预测研究的发展历史,可以看出 2018 年以前的工作主要围绕未来动作预测的形式进行探索,2018 年以后,逐渐形成了短期动作预测和长期动作预测两种统一的范式,随

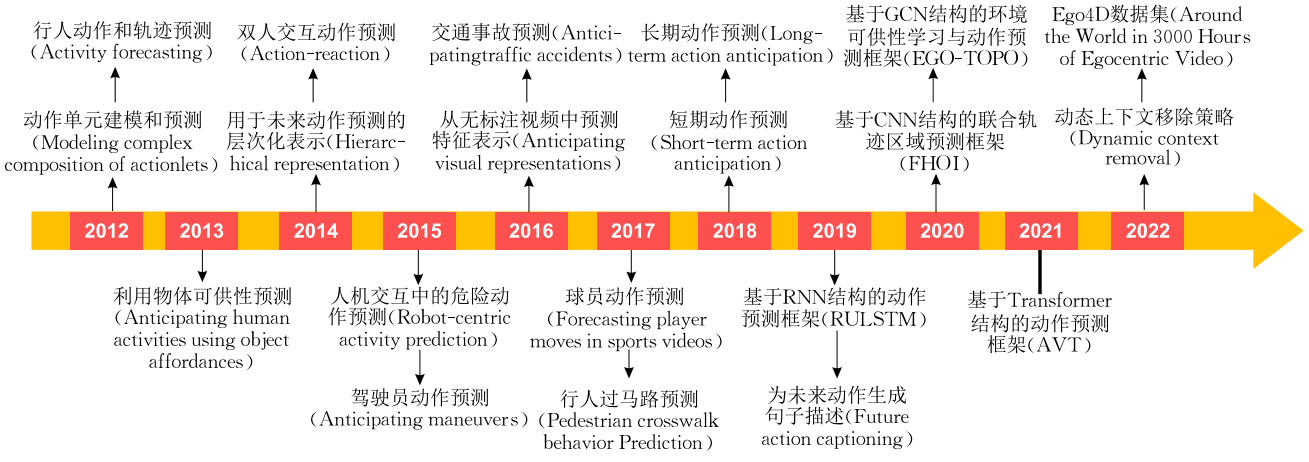


图 3 视频中的未来动作预测的发展历史

着深度学习在视频理解领域的兴起，同时涌现出了许多具有代表性的方法和技术，将在第 4 节从模型结构、数据模态、算法策略、预测对象等不同维度详细展开介绍.

4　现有主流方法

4.1　按模型结构划分

在深度学习兴起之前，研究人员普遍基于传统机器学习模型进行未来动作预测，随着深度学习技术的不断发展，以循环神经网络为代表的深度学习模型已经成为未来动作预测领域的主流方法. 本文根据核心模块所用的技术依次介绍基于传统机器学习模型的方法、基于循环神经网络的方法、基于卷积神经网络的方法、基于图卷积网络的方法、基于自注意力模型的方法.

4.1.1　基于传统机器学习模型的方法

早期的工作大多采用基于传统机器学习模型的方法. Chakraborty 等人[52]将视频中的所有动作序列视为马尔科夫随机场模型(Markov Random Field，MRF)[53]中的节点，通过执行置信度传播来计算条件分布得到下一个动作标签. Li 等人[22]提出将复杂的长时间活动分解编码为一系列有意义的动作单元，并通过概率后缀树捕获动作单元之间的不同阶马尔可夫依赖关系，对未来动作的推断可以看作是在给定前缀节点的情况下找到最有可能的后缀叶子节点. 在随后的工作[54]中，他们进一步通过将动作和物体的共现编码为复杂的符号序列实现对物体上下文信息的建模，并提出预测累积函数来描述每种活动的可预测性. Koppula 等人[31-32]基于条件随机场(Conditional Random Field，CRF)[55]建模三种类型的上下文信息，包括动作单元之间的层次结构、物体之间的时空相关性以及物体和人的运动等，并结合粒子滤波器预测未来人和物体的交互. Soran 等人[56]计算动作序列间转移概率的隐马尔可夫模型(Hidden Markov Model，HMM)[57]，学习动作之间的相关依赖关系，从而推断给定的工作流程中可能执行的下一个操作是否正确. Lan 等人[58]设计了多粒度的层次表示结构捕获不同层级的动作，并在最大化边际框架下训练支持向量机(Support Vector Machine，SVM)分类器[59]. 总体而言，基于传统机器学习模型的方法依赖人工提取特征和时序建模，往往会导致较高的时空复杂度，影响到算法的性能.

随着深度学习在视频理解领域的发展，基于深度神经网络的模型逐渐开始在未来动作预测领域得到广泛应用.

4.1.2　基于循环神经网络的方法

由于视频是序列化的数据，而循环神经网络(Recurrent Neural Network，RNN)[60]擅长处理序列数据，具有强大的序列建模能力，因此在未来动作预测领域占据了主导地位[17]，很多工作所采用的模型都是 RNN 结构，包括长短时记忆(Long-Short Term Memory，LSTM)[61]、门控循环单元(Gated Recurrent Unit，GRU)[62]等变体. 需要注意的是，基于 RNN 的方法并非只包含 RNN 结构，通常还需要先利用卷积神经网络(Convolutional Neural Network，CNN)[63]提取特征. 由于 RNN 能够处理任意长度的序列，有效地表达视频帧在时间维度的依赖关系，很多工作聚焦研究如何充分利用 RNN 结构对提取的特征进行时序建模，以更好地帮助预测未来的动作，因此本文将它们总结为基于 RNN 的方法.

早期的工作[36,64-65]使用一个 LSTM 对视频特征进行聚合，并直接通过分类器进行动作预测. Gao 等人[66]提出了基于两个 LSTM 级联的编码器-解码器结构，旨在根据过去视频序列的特征表示预测未来视频序列的特征表示，并通过线性分类器映射到未来的动作类别，此外还设计了强化学习模块鼓励模型尽早地作出正确预测. 在短期动作预测领域，Furnari 等人[40,67-68]运用了类似的思想，提出了 RULSTM 网络框架，主要由 R-LSTM(Rolling-LSTM)和 U-LSTM(Unrolling-LSTM)组成，其中 R-LSTM 用来对输入的观察视频片段进行编码，总结观察到的信息，U-LSTM 的作用对 R-LSTM 编码得到的信息进行解码，形成对于未来动作的预测. 如图 4 所示，首先利用 2D CNN(如 TSN[3])对视频片段提取特征(如捕获空间信息的彩色帧特征、捕获运动信息的光流特征)，将提取好的特征送入 R-LSTM 中进行编码，不同时刻的状态在自身网络中循环传递，使得观察内容中的信息在时间维度上聚合，U-LSTM 将 R-LSTM 最后一个单元的隐状态作为自身隐状态的输入，输出关于未来动作的特征表示，并通过线性层映射到未来的动作类别. RULSTM 的优势在于可以在未来动作开始之前的多个时刻进行预测. 图 4 展示了一个简单的例子，对于在 $t=4$ 时刻开始的未来动作，RULSTM 可以在 $t=2$ 和 $t=3$ 时刻进行预测. 在 $t=2$ 时刻，RULSTM 通过 R-LSTM 编码器对 $t=2$ 时刻之前的信息进行总结，再通过 U-LSTM 解码器得到对应 $t=4$ 时刻的预测结果，此过程中编解码器的隐状态均更新 2 次；同理，在 $t=3$ 时刻，RULSTM 通过 R-LSTM

编码器对 $t=3$ 时刻之前的信息进行总结，再通过 U-LSTM 解码器同样可以得到对应 $t=4$ 时刻的预测结果，此过程中编码器的状态更新 3 次，解码器的隐状态更新 1 次. 这充分说明 RULSTM 在实际应用中可以根据不同的输入序列长度对目标动作进行预测. 为了降低未来的不确定性，Camporese 等人[69]在 RULSTM 的基础上引入了标签平滑技术，将传统的交叉熵损失函数进行改造，鼓励模型关注与正确类别在语义空间上相近的其它类别. 考虑到观察内容和待预测的未来内容之间存在时间间隔，为了有效利用这段时间间隔中的内容，Wu 等人[43]借鉴了自监督学习的思想，在 RULSTM 的基础上增加了一条 LSTM 分支，用来预测这段时间间隔的特征表示，并将其输入到用于解码的 LSTM 中. Qi 等人[70]构造了由两个 GRU 组成的编码器-解码器结构，并同样关注观察内容和未来内容之间的时间间隔，为了避免误差在此时间段内累积，设计了自调节学习模块，一方面利用对比学习技术来突出当前帧与过去帧的不同，另一方面计算当前帧和过去数帧的相似性并重新加权拼接过去的信息，以此循环地调节序列化预测的性能. Osman 等人[71]受视频理解领域的经典模型 SlowFast 网络[72]的设计思想启发，将视频序列数据由原来的单一帧率变成由两种帧率采样组成的快帧率分支和慢帧率分支，输入到文献[40]中的两个 LSTM 网络中，使得改造后的模型能够处理不同时间尺度的数据，同时捕获快帧率对应的运动信息和慢帧率对应的空间语义信息. 为了在有限的观察片段中融入更丰富的外部知识，Liu 等人[73]在 GRU 网络的基础上加入记忆增强模块，通过记忆项的组合帮助模型重建更具判别性的未来特征表示. 在长期动作预测领域，Abu 等人[45]认为预测未来每帧的类别等价于预测未来的每个动作类别及其持续时间，他们将预测出的前一个的动作类别编码成独热(one-hot)向量及其持续时间递归地送入 GRU 网络，获得下一个待预测动作的类别以及时间长度. 在随后的工作[74]中，他们同时训练基于 GRU 的动作类别预测模型和时间长度预测模型来降低未来的不确定性. 此外，他们还引入循环一致性(cycle consistency)损失确保预测出的未来动作和过去的动作存在语义上的连贯性[47]. 为了增强 RNN 模型对于长期信息的记忆能力，Gammulle 等人[75]设计了分别负责编码视觉特征和动作标签的双流 LSTM 网络，并引入外部神经记忆单元用来捕获序列之间的长距离关系. 为了增强标注信息的丰富程度以促进模型的学习，Morais 等人[76]重新标注了现有的动作预测数据集[77]使得动作标签层级更加精细，并基于 GRU 网络设计了层级编码更新器用来建模并预测多层级的动作. 考虑到实际场景中标注获取困难，Ng 等人[78]探索更具挑战性的弱监督场景下的长期动作预测，将观察到的视频帧序列送入 GRU 编码器，通过对 GRU 解码器输出的隐状态序列应用注意力机制，递归地预测未来的动作标签. 此外，Sener 等人[49]研究针对未来视频生成句子描述，分别构造了基于 LSTM 的句子编码器和视频编码器以及句子解码器，面向烹饪视频实现对于未来菜谱的预测.

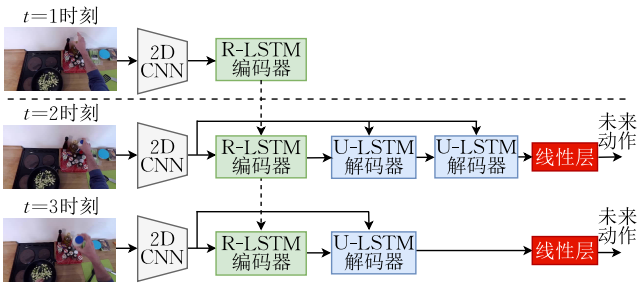


图 4 RULSTM 网络框架[40]

4.1.3 基于卷积神经网络的方法

卷积神经网络(Convolutional Neural Network，CNN)[63]是一种前馈神经网络，优势在于特征提取和表达. 一些工作采用的模型为基于 CNN 的结构. 在早期研究中，Abu 等人[45]针对长期动作预测，将观察片段中的每个动作编码成独热向量并堆叠成二维矩阵，然后通过 2D CNN 直接映射到表示未来动作的二维矩阵. 这种方法比较简单直接，与基于 RNN 的方法相比性能较低. Damen 等人[39]将动作识别中的经典模型 2S-CNN[2]和 TSN[3]直接用于短期动作预测任务，这种方法仅仅将监督信号由动作识别中的当前动作类别替换为待预测的未来动作类别，没有考虑后者与输入的视频之间存在的时间间隔，导致预测效果一般. 在短期动作预测领域，为了建立当前信息和未来信息之间的联系，Tran 等人[79]提出了一种自监督的知识蒸馏框架，在预测网络的基础上引入针对未来视频数据的识别网络，利用后者监督前者对于视频特征的学习，两者均采用基于 3D 卷积的 I3D[4]网络结构. Fernando 等人[80]采用了类似的思想，首先利用 ResNet(2+1)D[81]从观察内容中生成关于未来的预测，然后在训练过程中最大化前者和未来真实特征的相似度，使得未来的信息能够迁移到观察信息中. 如图 5 所示，Liu 等人[41]在预测未来动作的基础上扩展了预测对象，包括未来的区域和轨迹，采用多任务学习的思想提出 FHOI 框

架，首先利用 I3D[4] 或 CSN[82] 等 3D CNN 对输入视频片段提取特征，然后将特征分别输入到轨迹预测模块、交互区域预测模块得到预测出的手部的运动轨迹以及未来交互动作可能发生的区域信息，并利用这些信息帮助模型更好地预测未来动作. Zatsarynna 等人[83]将研究的重点聚焦于预测模型的运行速度，从动作检测的经典模型 TCN 网络[84]中汲取灵感，利用堆叠的时序卷积层替代循环结构，捕获输入帧之间的时序关系，改进了 RNN 不能并行计算的局限性，大大缩短了模型的运行时间. 此外，还有一些方法将卷积操作和注意力机制[85]相结合，实现在时间维度上关注更多尺度的有效信息，从而同时应用于短期和长期动作预测，Ke 等人[46]利用多尺度时序卷积来处理观察内容，并引入注意力机制更加关注时序上特定的一些视频帧，通过跨层连接高效地一次性预测多个动作. 类似地，Sener 等人[42]受非局部网络（Non-local）[86]的设计思想启发，基于 1×1 卷积和自注意力机制操作在时序上聚合当前片段和过去片段的特征，从而能够在时间维度上有效地获取多尺度的信息.

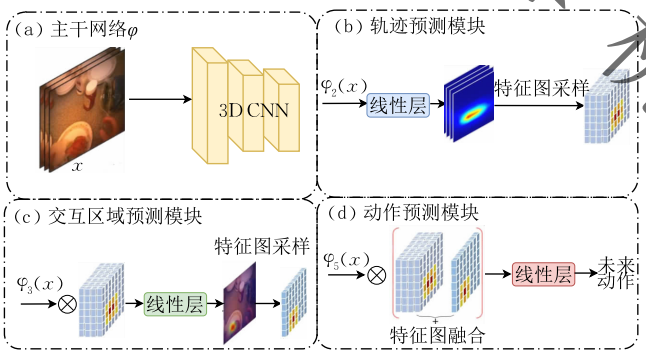


图 5　FHOI 网络框架[41]

4.1.4　基于图卷积网络的方法

近年来，图卷积网络（Graph Convolutional Network，GCN）[87]因其强大的关系建模能力在计算机视觉领域得到广泛应用. 视频中蕴含着很多依赖于图结构进行建模的复杂关系，例如视频片段与片段之间、区域与区域的关系等等，这促使一些工作设计基于 GCN 的核心模块捕获这些关系. 在短期动作领域，Huang 等人[88]将 GRU 与 GCN 相结合构造了全局关系图网络，其中 GRU 用来聚合视频片段的特征信息，GCN 用来建模视频片段与片段之间的上下文关系，能够有效地捕获过去已发生动作和未来动作之间的全局依赖关系. 为了更针对性地关注人物交互信息，Dessalene 等人[89-90]一方面利用基于 3D 卷积的 CSN 网络提取视频片段的外观特征，另一方面利用 GCN 根据手和物体可能接触的区域信息对动作之间的长期上下文关系进行建模，通过融合这些信息来同时预测接下来可能操作的物体、手和物体可能接触的区域以及未来的动作类别. 尽管取得了不错的预测效果，但和文献[41]一样都需要额外的数据标注，非常耗时耗力. 针对长期动作预测，Nagarajan 等人[91]通过从视频中划分不同的空间区域构建拓扑图，提出了 EGO-TOPO 框架，如图 6 所示，拓扑图中的每个节点都有来自不同视频的视觉上相似的区域，GCN 用来建模区域与区域之间的关系，能够有效地跨区域整合信息.

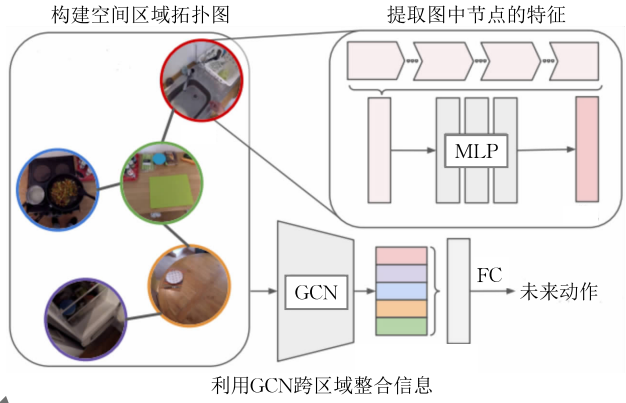


图 6　EGO-TOPO 网络框架[91]

4.1.5　基于自注意力模型的方法

Transformer[92]是一种基于自注意力机制（self-attention）的神经网络，近年来在计算机视觉领域逐渐流行[93]，一些工作也开始基于 Transformer 来构建动作预测模型. 考虑到 RNN 只能依次顺序进行计算，为了提升模型的并行计算能力，Wang 等人[94]利用 Transformer 风格的架构聚合观察的视频信息，然后利用渐进式特征生成模型生成未来的特征，最后将这些特征映射到未来的动作类别. 为了增强对于人物交互信息的利用，Roy 等人[95]设计了基于多模态特征融合的 Transformer 模型，首先提取了人物交互特征、空间全局特征和时序光流特征，将它们送入 Tranformer 编码器中，最后将编码器的输入进行拼接，用一个共享的解码器输出对未来的预测，或者采用独立的解码器分别对三种特征输出对未来的预测，再采用池化的方式得到融合后的预测分数. 如图 7 所示，Girdhar 等人[44]基于视觉 Transformer[96]提出了用于视频动作预测的 AVT 框架，主要由两部分组成：提取帧特征的编码器和聚合帧信息的解码器，其中编码器作用于每帧切分后的图像块，对应于空间注意力；解码器的作用是为每一帧输出对未来的预测，对应于时间注意力，可以利用 Transformer

同时进行特征提取和时序建模.考虑到动作预测对于视频帧序列的时序关系比较敏感,Xu 等人[97]提出了一种时间顺序感知的预训练方案,以无监督的形式促使 Transformer 模型学习视频帧的位置编码,以此提升模型对于视频帧序列的时序感知能力.总体而言,这一类方法利用 Transformer 有效地进行时序建模,并行地处理输入数据,尽管目前主要应用于短期动作预测,但在长期动作预测领域也具备可观的发展潜力.

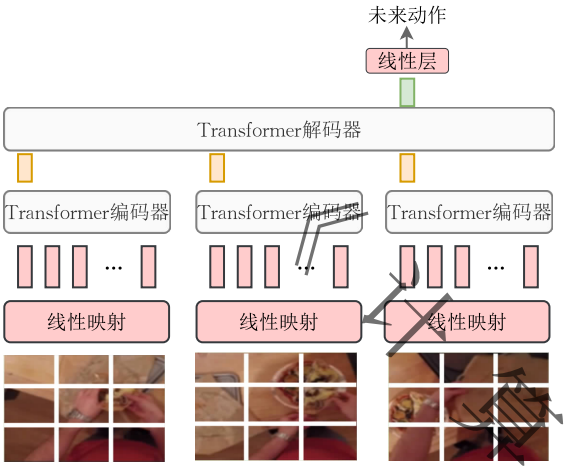


图 7 AVT 网络框架[44]

4.1.6 小结

目前研究人员普遍将未来动作预测作为有监督的分类问题处理,动作预测模型的核心在于对观察的视频片段进行特征提取和时序建模.早期的工作大多依赖手工提取特征,并采用条件随机场、贝叶斯网络、概率图模型等传统机器学习模型进行时序建模,由于手工提取特征通用性差、描述能力有限,依赖专家知识,并且传统机器学习模型的建模能力有限,因此基于传统机器学习模型的方法逐渐被基于深度神经网络模型的方法所取代.最早在图像处理领域取得成功的 2D CNN 在处理原始视频帧时无法建模时间信息,仅能捕获帧内的空间外观信息,尽管通过计算光流可以捕获相邻帧之间的运动特征[2],但光流信息对时序上下文的访问非常有限,因此 2D CNN 通常还需要和 RNN 结合,后者主要负责将若干帧的空间信息在时间维度进行聚合,并映射到未来的特征空间.与 RNN 序列建模不同,3D CNN 通过 3D 卷积核同时捕获在时间和空间维度上具有判别性的特征,可以直接作用于原始视频帧,得到相比 2D CNN 更紧凑的特征表示.GCN 则专注于捕获视频数据中的复杂结构化关系,例如视频片段与片段之间、区域与区域之间的关系,通常作用于经过 RNN 或者 3D CNN 处理过后的判别性时空特征,从而得到时间和空间维度的高层次全局关系.Transformer 则利用自注意力机制可以对给定数据的任意两个位置建立联系,因此使用起来更加灵活,既可以与 2D CNN 结构结合,对于前者已提取的特征进行时序建模,也可以直接作用于原始视频帧,先通过切分图像块在空间维度捕获特征,然后在时序维度进行特征聚合,得到具有判别性的时空特征表示.

表 1 进一步针对近年来的代表性方法总结了各自的特征提取结构、时序建模结构、特征类型以及研究重点.较早些的工作[39]仅利用 2D CNN 进行特征提取,而忽视了对于时序建模的研究.由于观察视频和待预测目标动作之间在时序层面上并不对齐,因此大部分工作将时序建模作为研究重点,使用较多的方法[40,43,71,73]是在 2D CNN 提取特征的基础上利用 RNN 进行时序建模,允许信息在时间维度上循环传递,但后一时刻的计算依赖于前一时刻信息的计算结果,容易导致误差累积,缺乏并行计算能力,逐渐被兴起的 Transformer 结构所替代[44,95-97],后者凭借自注意力机制具有更强的长距离关系捕获以及并行计算能力,并且还可以直接从原始视频帧中提取特征[44],不需要 2D CNN 的参与.由于 3D 卷积在 2D 卷积添加了对时间维度的操作,也有一些方法将特征提取和时序建模过程合二为一,利用 3D CNN 结构可以直接从原始视频帧中提取时空特征[41],通过引入注意力机制可在时间维度上获取多尺度的信息,使得模型更关注时序建模过程[46].还有一些方法会在时序建模阶段针对 RNN 或者 3D CNN 处理过后的判别性时空特征,利用 GCN 结构进一步挖掘视频中的复杂结构化关系[88,90].从特征类型来看,外观特征和运动特征是两种最基本的特征,基于 2D CNN 的特征提取器通常需要通过计算光流信息来获取运动特征,而基于 3D CNN 的提取特征提取器则通过 3D 卷积混合了时间和空间维度的信息,虽然不需要计算光流,但也容易出现优化困难以及过拟合的问题.除了外观和运动特征之外,由于通常涉及到人物交互等细粒度动作,很多工作还针对局部物体或者区域等信息进行挖掘[40-41,90,95],从而增强模型对于已观察视频片段中的视觉特征的学习和利用,具体将在下一节展开介绍.总体而言,基于深度神经网络模型的方法能够自动学习复杂抽象的特征,增强了特征提取和时序建模的能力,从而提升动作预测的性能.但相较于传统机器学习模型,深度学习模型缺乏可解析性,如何提升动作预测模型的可解释性也是未来的研究工作值得考虑的问题.

表 1　代表性方法的特征提取和时序建模比较

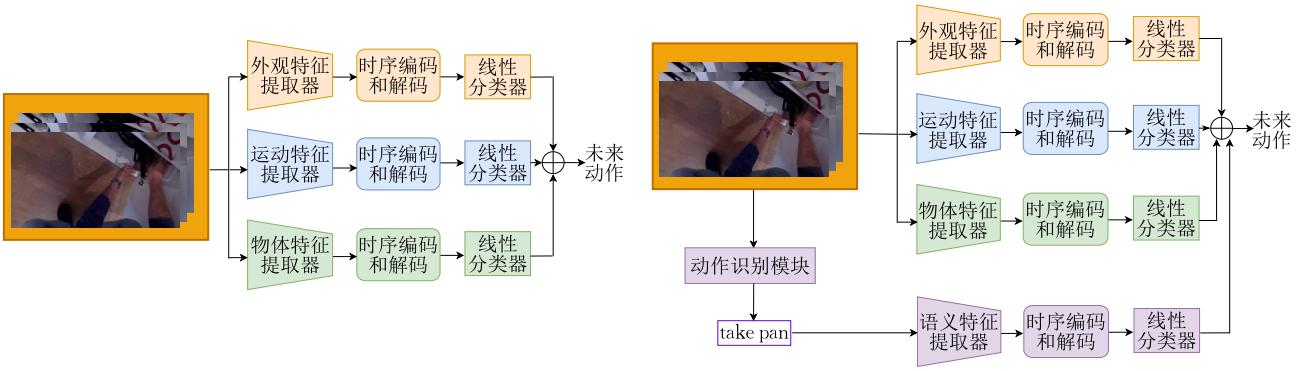
研究工作	年份	特征提取结构	时序建模结构	特征类型	研究重点
Damen 等人[39]	2018	2D CNN	—	外观，运动	特征提取
Ke 等人[46]	2019	3D CNN	3D CNN	外观，运动	时序建模
Furnari 等人[40]	2019	2D CNN	RNN	外观，运动，局部物体	时序建模
Wu 等人[43]	2020	2D CNN	RNN	外观，运动，局部物体	时序建模
Liu 等人[41]	2020	3D CNN	3D CNN	外观，运动，局部区域	特征提取
Nagarajan 等人[91]	2020	3D CNN	GCN	外观，运动，局部区域	时序建模
Osman 等人[71]	2021	2D CNN	RNN	外观，运动，局部物体	时序建模
Huang 等人[88]	2021	2D CNN	RNN，GCN	外观，运动，局部物体	时序建模
Girdhar 等人[44]	2021	Transformer	Transformer	外观，运动	时序建模
Dessalene 等人[90]	2021	3D CNN	RNN，GCN	外观，运动，局部区域	特征提取
Roy 等人[95]	2021	2D CNN	Transformer	外观，运动，局部物体	特征提取
Wang 等人[94]	2021	2D CNN	Transformer	外观，运动	时序建模
Liu 等人[73]	2022	2D CNN	RNN	外观，运动，局部物体	时序建模
Xu 等人[97]	2022	2D CNN	Transformer	外观，运动，局部物体	时序建模

4.2　按数据模态划分

视频相比静态图像数据能够提供更多的信息，包括随时间演化的复杂运动信息，同时还包含了许多丰富的物体细节信息. 除了这些具体的视觉信息之外，相对抽象的语义信息也能有效地帮助动作预测，不少工作在高级语义空间上探索动作标签推理. 本文从数据模态的维度分别介绍基于视觉信息的方法以及结合视觉信息和语义信息的方法.

4.2.1　基于视觉信息的方法

一些工作致力于从输入的视频片段中提取视觉信息，充分利用这些视觉信息实现对未来动作的预测. 早期的研究[22,31-32,52,54,56,58,98]通常基于手工提取特征，包括方向梯度直方图(Histogram of Oriented Gradient，HOG)[99]或者光流方向直方图(Histograms of Oriented Optical Flow，HOF)[100]等等. 随着深度学习的发展，研究人员普遍利用 CNN 获取视觉特征. 如图 8 所示，基于视觉的方法主要分为两部分. 第一部分是特征提取模块，在短期动作预测领域，Furnari 等人[40]提出了三种具有代表性的特征提取方式，包括(1)外观特征提取器，通过将视频帧序列输入 BN-Inception[101]网络得到空间全局信息；(2)运动特征提取器，通过对视频序列中的每两帧计算密集光流得到密集光流序列，然后将密集光流序列输入 BN-Inception[101]得到时序运动信息；(3)物体特征提取器，通过将视频帧序列输入 Faster R-CNN[102]网络得到局部物体信息. 这三种类型的视觉特征被后续工作[42-43,69-71,73,83,88,97,103-104]广泛使用. 除此之外，一些工作侧重于进一步挖掘人物的交互信息，例如 Shen 等人[105]结合人眼注释区域对视频序列进行预处理获得手部掩码信息，Liu 等人[41]和 Dessalene 等人[90]通过额外的标注数据对于手部和物体交互区域信息进行捕捉，Roy 等人[95]利用图像分割算法对视频帧提取手部特征和物体区域特征，并融合成人物交互特征表示. 由于长期动作预测关注的大多是粗粒度动作，通常不需要局部特征的参与，广泛使用的是 I3D[4]模型提取的外观特征和运动特征. 基于视觉信息的方法第二部分是时序编解码模块，作用是以第一部分提取的视觉特征作为输入，经过时序编码器得到观察内容信息的聚合表示，然后利用时序解码得到关于未来动作的特征表示，最后经过线性分类器映射到未来的动作类别空间. 由于 RNN 和 Transformer 都具有较强的时序建



(a) 基于视觉信息的方法示意图　　(b) 结合视觉和语义信息的方法示意图

图 8　基于视觉信息的方法与结合视觉和语义信息的方法比较

模能力，使用较多的时序编解码结构包括级联的LSTM[40,43,69,71,103]、GRU[70,73,88]以及 Transformer 编码器-解码器[95,97]结构. 另外，还有少部分方法[41,90]利用 3D CNN 结构将特征提取和时序编解码合二为一，由于 3D CNN 的特点在于利用 3D 卷积提取视频的时空信息，因此这些方法的重点在于视觉特征抽取，淡化了对于时序建模能力的学习. 对于不同类型的视觉特征经过不同分支进行时序编解码和线性映射之后得到的预测结果，使用较多的融合方法是模态注意力机制[40]，它能够动态地为每个分支的预测结果计算权重并进行加权求和，实现不同类型特征之间的融合.

4.2.2 结合视觉信息和语义信息的方法

一些工作在关注视频片段的视觉特征之外，还尝试挖掘动作标签的语义信息并将两者结合，获得更可靠的预测结果. 常见的研究方案有两种：(1) 从人工标注数据中提取语义信息，并通过改进传统的交叉熵损失函数来优化动作预测模型的训练过程. 例如，在短期动作预测领域，Furnari 等人[106]在损失函数中降低部分正确(例如仅名词或动词正确)的动作标签的损失，来提升 Top-k 预测值的质量. 相似地，Camporese 等人[69]采用知识蒸馏的思想，将从标注数据中提取的动作语义先验信息蒸馏到动作预测模型中，通过标签平滑技术，改造传统的交叉熵损失函数，在计算损失时赋予与正确标签语义相近的其它标签更大的权重. 此外，Roy 等人[107]将输入片段和待预测动作之间的每一个状态的语义特征视为潜在目标，设计了相应的损失函数促使模型在学习过程中不断逼近真实的动作标签；(2) 如图 8 所示，利用动作识别模型得到观察视频的动作标签信息，显式地将视觉信息和标签信息结合来进行预测. Miech 等人[108]观察到的视频片段输入 TSN 模型[3]得到当前时刻的动作标签，然后通过 word2vec 词向量模型抽取语义特征，并利用线性模型进行时序编解码得到语义信息对应的预测结果. 由于线性模型的时序能力有限，为了提升时序建模能力，Zhang 等人[103]在此基础上利用级联的 LSTM 对语义信息进行时序编解码并设计了一种从人工标注中提取语义标签信息进行预训练的策略，进一步增强了对于语义信息的挖掘. 然而过度依赖语义信息可能会导致模型倾向于记忆前后动作标签之间的语义关联，从而忽略了对于具体视觉信息的挖掘，容易产生预测偏差，Zhang 等人[104]提出反事实分析方案，旨在保留多模态信息的基础上，削弱动作标签之间的语义关联带来的副作用，使得模型更关注能够反映每个案例具体信息的视觉内容. 在长期动作预测领域，也有一些方法[75-76,109]采用类似方案(2)中视觉信息和语义信息结合的思想，这些当前时刻的语义信息更多直接来自人工标注的真实标签，不适合扩展到实际应用场景.

4.2.3 小结

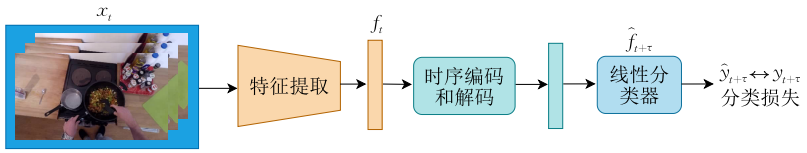
相较于静态图像数据，视频能够提供更多的数据模态，基于视觉特征的方法从观察到的视频片段中提取丰富的视觉信息，例如物体、场景、人眼关注区域、手势动作、区域、运动轨迹等多模态信息，可以通过设计多任务学习框架，利用额外的数据模态构造正则化项促进对于未来动作的预测. 此外，许多方法在视觉信息的基础上增加了对于语义信息的探索和利用，一种方式是利用从人工标注数据中提取的语义信息改造损失函数，以优化动作预测模型的训练过程，在推理时则不需要语义信息. 另一种方式是通过引入行为识别模块显式地利用当前时刻的语义标签信息预测未来的动作，与前一种方案不同，这种方案在推理时也需要利用语义信息. 实际应用中，由于语义信息需要通过动作识别模型从已观察到的视频片段中提取，这使得语义信息的可靠程度依赖于动作识别模型的性能，同时还增加了对于计算资源的依赖.

4.3 按算法策略划分

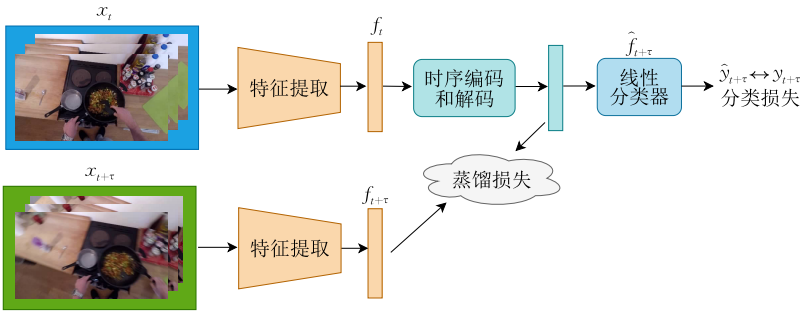
未来动作预测的实质是学习从观察内容到未来动作类别的映射，本文从算法策略的角度介绍直接从观察内容预测的方法、利用外部信息进行知识蒸馏的方法、

4.3.1 直接从观察内容预测的方法

一些工作致力于直接从观察到的视频片段建立到未来的动作标签的映射，除了用于训练的动作标签之外，未来无法观察的信息对于观察内容的贡献可以忽略不计. 图 9(a)展示了直接从观察内容预测的方法的基本流程：对于观察到 t 时刻的视频 x_t，首先经过特征提取得到 f_t，如上节所述，这里提取的特征既可以是视觉特征，也可以是语义特征；然后经过时序编码和解码得到关于 $t+\tau$ 时刻动作的特征预测值 $\hat{f}_{t+\tau}$；最后经过线性分类器预测出 $t+\tau$ 时刻的动作 $\hat{y}_{t+\tau}$，与 $t+\tau$ 时刻真实的动作标签 $y_{t+\tau}$ 计算分类损失. 除了较早采用的传统机器学习模型[22,31-32,52,54,56,58,98]，目前这一类方法主要采用的模型结构大致分为两种，一种是采用 2D CNN[39]提取特征，然后用RNN[40,45,71,74-76]或者 Transformer[44,94-95]进行时序编解码，另一种是利用 3D CNN[41,90]将特征提取和时序编解码合二为一. 总体而言，直接从观察内容预测的方法仅将未来动作预测模型直接



(a) 直接从观察内容预测的方法示意图



(b) 利用外部信息进行知识蒸馏的方法示意图

图 9　直接从观察内容预测的方法与利用外部信息进行知识蒸馏的方法比较

作用于可被观察的数据，可利用的信息相对而言较为局限.

4.3.2　利用外部信息进行知识蒸馏的方法

由于有限的观察内容难以提供全面的信息，为了提升未来动作预测的可靠性，一些研究工作受到知识蒸馏[110]的思想启发，将作用于观察内容的预测模型视为学生，尝试将外部信息迁移到有限的观察内容中. (1) 一部分工作将未被观察到的视觉信息视为教师. 如图 9(b)所示，这一类方法在训练过程中的损失函数由两部分组成，除了分类损失之外，还在特征空间计算蒸馏损失. 具体做法为：对于 $t+\tau$ 时刻的视频 $x_{t+\tau}$ 提取真实特征 $f_{t+\tau}$，通过与 t 时刻预测出的未来特征 $\hat{f}_{t+\tau}$ 计算损失，从而实现在训练过程中将未来不可观察到的知识迁移到当前的观察内容中. 在推理过程中，动作预测模型仍然仅根据 x_t 进行预测. 这一类方法的研究重点为在特征空间中构造蒸馏损失函数，在早期研究工作中，Vondrick 等人[111]针对特征向量之间的差异直接计算均方差损失. Tran 等人[79]认为特征向量丢失了空间细节信息，提出在特征图层面计算均方差损失. Huang 等人[88]针对不同时刻的特征向量构建更复杂抽象的关系图表示，在关系图层面计算 KL 距离以及均方差损失. 近来一些工作认为均方差损失的收敛性能比较差，在训练过程中难以优化，提出用相似性度量方式代替均方差损失，包括余弦相似度[73,112]以及杰卡德相似度[80]. 还有一些工作[43,70]则基于对比学习的思想设计蒸馏损失函数，将同一时刻的真实特征作为其预测值的正例，而将其它时刻的真实特征作为负例. 最近 Xu 等人[97]提出一种动态上下文移除策略，在训练过程中逐渐减少用于计算蒸馏损失的未来特征，使得模型充分学习重建未来特征的能力，从而提升时序推理能力. (2) 另一部分工作将从标注数据中提取的语义先验信息作为教师. Camporese 等人[69]在训练过程中通过标签平滑技术改造传统的交叉熵损失函数，鼓励模型学习与真实标签语义相近的其它动作标签. 为了缓解单独利用视觉信息存在的语义鸿沟现象，Zhang 等人[103]将上一时刻视频帧的语义标签作为输入，设计了基于文本信息的序列化动作预学习策略，使得模型具备预先推理出未来动作的能力. 在随后的工作[104]中，他们引入反事实分析的方案，通过从基于多模态信息获得的预测结果中扣除仅基于序列化动作学习得到的反事实预测结果，缓解动作标签之间的语义关联带来的副作用，进一步增强模型的逻辑推理能力. 总体而言，这类方法旨在充分利用可观察内容之外的信息帮助预测未来的动作，尤其可以从无标注或未剪切的视频数据学习利用未来视频的特征表示，提供更全面可靠的信息.

4.3.3　小结

未来动作预测的核心在于建立从观察内容到未来动作的映射，现有的部分方法直接从观察到的视频片段中挖掘信息，由于观察内容受到客观条件的限制，存在时间范围、语义信息等各方面的局限性，因此更多的研究工作利用外部信息进行知识蒸馏，从多个不同角度进行探索，旨在将未被观察到的视觉信息或者从标注数据中提取的语义先验信息迁移到有限的观察内容中，从而进一步提升未来动作预测的可靠性. 这一类方法目前在短期动作预测中得到广泛应用，但在长期动作预测领域只有 Nagarajan 等人[91]通过从视频中划分不同的空间区域构建拓扑图来关联数据集不同视频中的相似区域，可以看作将跨区域信息迁移到动作预测模型中实现外部信

息的知识蒸馏.因此,未来如何将这一类方法有效地运用于长期动作预测也值得探索.

4.4 按预测对象划分

研究人员在处理未来动作预测问题时,有时还会运用多任务学习的思想,除了预测未来的动作类别之外,还可能结合其它信息进行联合预测,本文从时间和空间维度将联合其它信息预测的方法划分为:联合预测时间信息的方法、联合预测轨迹和区域信息的方法.

4.4.1 联合预测时间信息的方法

在未来动作预测研究早期,大部分基于传统机器学习的方法仅关注未来的动作类别[22,31-32,52,54,56,58],随后一些工作开始同时关注下一个动作在何时发生,Mahmud 等人[98]探索了利用泊松过程这样的传统概率建模方法预测下一个动作的开始时间.在随后的工作[113]中,他们基于 LSTM 网络和全连接层等深度神经网络构建了联合预测未来动作类别和开始时间的框架,能够利用动作和物体之间的上下文关系实现对于未来多个动作及其持续时间的预测.Mehrasa 等人[114]进一步利用变分自动编码器捕获动作时间和类别标签中的不确定性,通过从先验分布中采样来生成动作序列,从而控制下一个动作类别及其发生时间的分布.Neumann 等人[115]针对未来可能发生的事件,利用启发式热图的高斯混合模型在判断未来事件是否发生的同时,预测出该事件在何时发生.由于短期动作预测问题本身设置了预测时间,即下一个动作在何时开始,并且也不关注待预测目标动作的持续时间,因此近年来联合预测时间信息的方法主要集中在长期动作预测领域,一些工作[42,45,47,74]将预测观察片段后面每帧的类别等价于预测未来的动作类别及其持续时间来进行处理,如图 10 所示,首先对观察视频通过 2D CNN 提取特征,然后通过 RNN 进行时序编解码,最后经过两个不同的线性层分别映射到动作类别空间和持续时间空间,其中未来动作的持续时间表示为占剩余时间的百分比,通过将预测未来的动作类别视为分类问题,将预测持续时间视为回归问题,实现联合时间信息的预测.总体而言,这一类工作同时关注未来动作的类别和时间信息,由于更贴近实际应用的设定,因此更有助于将模型部署到实时系统.

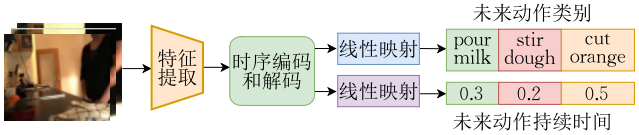


图 10 联合预测时间信息的方法示意图

4.4.2 联合预测轨迹和区域信息的方法

对于未来动作的预测本质上依赖未来的动作标签作为监督信息指导模型的学习,而动作标签是抽象的语义概念,为了获取更丰富的监督信息,很多工作在预测动作类别的同时还预测未来的轨迹或区域等信息.在早期的研究工作中,Zeng 等人[116]基于 RNN 结构研究对于未来可能发生的危险事故的预测,一方面将有无事故发生视为二分类问题,另一方面估计出事故可能发生的区域.后续一些工作围绕视频监控场景下行人的运动轨迹和动作预测展开研究.如图 11 所示,Liang 等人[26]通过 2D CNN 提取了包含场景全局信息以及人与物体在内的局部信息,并在传统动作预测模块的基础上增加了轨迹和区域预测模块,一方面通过注意力机制将 LSTM 时序编解码得到的坐标序列生成未来轨迹,另一方面将坐标位置映射至特征图,并利用目标检测算法进行边框回归生成未来区域.Chen 等人[27]进一步从空间和时间维度构建了动作与物体之间以及动作与动作之间的知识图谱,并融入到轨迹和区域预测模块中获得对于未来轨迹和动作标签的预测.由于短期动作预测所采用的大多为聚焦人物交互的第一人称视频,一些工作[41,90,95]针对第一人称视频,基于物体检测和分割算法将预测手部运动轨迹、手和物体可能接触的区域作为辅助任务,提升未来动作预测的性能.总体而言,这类工作运用多任务的思想,使得模型学习除了动作类别之外更丰富的轨迹和区域信息,比单独将动作类别作为监督信息更可靠,但同时依赖更多的人工标注信息.

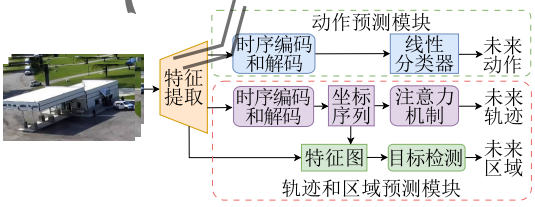


图 11 联合预测轨迹和区域信息的方法示意图

4.4.3 小结

由于未来的动作标签是抽象的语义概念,仅将其作为监督信号通常不足以使得模型充分学习.因此在视频中的未来动作预测研究领域,一些工作运用多任务学习的思想,将预测未来动作类别视为主任务,同时构造其它辅助任务,一方面联合预测出未来动作的持续时间等信息,另一方面联合预测出运动轨迹和动作发生区域等空间信息,通过包含在相关任务的监督信号中的额外信息来提升未来动作预测的可靠性.

表 2 进一步总结了各类方法的优缺点以及适用范围等信息,供读者参考.同时,表 3 从模型结构、数

表 2　各类方法的优缺点及适用范围总结

方法		优点	缺点	适用范围
模型结构	传统机器学习模型	可解释性强	手工提取特征	简单场景拍摄的小规模动作
	循环神经网络	有效进行时序建模	缺乏并行计算能力	在动作开始前的多个时刻进行预测
	卷积神经网络[41]	时空特征表达能力强	时序建模能力有限	增强视频特征表示
	图卷积网络	建模高层级的复杂关系	扩展性不足	对复杂场景拍摄的视频内容进行关系抽取
	自注意力模型[44]	并行处理输入序列	依赖大规模数据	捕获时间维度上远距离的动作信息
数据模态	基于视觉信息	充分挖掘视觉内容	缺乏语义信息进行辅助判断	缺乏明显语义性的动作
	结合视觉和语义信息	利用多模态信息	通过动作识别模型得到语义信息，增加计算开销	复杂且有明显语义性的动作
算法策略	直接从观察内容预测	实现简单	可利用的信息有限	观察内容本身较为丰富且动作类别少，不确定性小
	利用外部信息进行知识蒸馏	利用观察内容以外的信息，降低未来的不确定性	增加存储和计算开销	观察内容有限且动作类别多，不确定性大
预测对象	联合预测时间信息[40]	多任务学习，提升动作预测的可靠程度	需要额外标注，增加人工成本和计算开销	对实时性要求较高的场景
	联合预测轨迹和区域信息			关注轨迹和区域变化的场景

表 3　视频中的未来动作预测方法总结

模型结构	数据模态	算法策略	预测对象	相关论文	年份区间	应用任务
传统机器学习	基于视觉信息	直接从观察内容预测	动作类别	[22,31-32,52,54,56,58]	2012～2015	人机交互、体育运动、交通等场景中的动作预测
			动作类别＋时间	[98]	2016	
循环神经网络	基于视觉信息	直接从观察内容预测	动作类别	[36,64-66]	2016～2017	人机交互、体育运动、交通等场景中的动作预测
				[40,67-68,71,105]	2018～2021	短期动作预测
			动作类别＋时间	[113-115]	2017～2019	人机交互、体育运动、交通等场景中的动作预测
				[45,47,74,78]	2018～2020	长期动作预测
			动作类别＋轨迹区域	[116]	2017	人机交互、体育运动、交通等场景中的动作预测
				[26]	2019	行人轨迹和动作预测
		利用外部信息进行知识蒸馏	动作类别	[43,70,73]	2020～2022	短期动作预测
	结合视觉和语义信息	直接从观察内容预测	动作类别	[49]	2019	生成未来句子描述
				[107]	2022	短期动作预测
			动作类别＋时间	[75-76,109]	2019～2020	长期动作预测
		利用外部信息进行知识蒸馏	动作类别	[69,103]	2020～2021	短期动作预测
卷积神经网络	基于视觉信息	直接从观察内容预测	动作类别	[39,46,83]	2018～2021	短期动作预测
			动作类别＋时间	[45-46]	2018～2019	长期动作预测
			动作类别＋轨迹区域	[41]	2020	短期动作预测
		利用外部信息进行知识蒸馏	动作类别	[111]	2016	人机交互、体育运动、交通等场景中的动作预测
				[79-80,112]	2021	短期动作预测
	结合视觉和语义信息	直接从观察内容预测	动作类别	[42,106,108]	2019～2020	短期动作预测
			动作类别＋时间	[42]	2020	长期动作预测
		利用外部信息进行知识蒸馏	动作类别	[104]	2021	短期动作预测
图卷积网络	基于视觉信息	直接从观察内容预测	动作类别＋轨迹区域	[89-90]	2020～2021	短期动作预测
				[27]	2021	行人轨迹和动作预测
		利用外部信息进行知识蒸馏	动作类别	[91]	2020	长期动作预测
				[88]	2021	短期动作预测
			动作类别＋轨迹区域	[117]	2021	行人轨迹和动作预测
自注意力模型	基于视觉信息	直接从观察内容预测	动作类别	[44,94]	2021	短期动作预测
			动作类别＋轨迹区域	[95]	2021	短期动作预测
		利用外部信息进行知识蒸馏	动作类别	[97]	2022	短期动作预测

据模态、算法策略、预测对象等四个维度总结了本章介绍的所有研究方法，并归纳了这些论文发表的年份区间，以及所应用的任务等信息.

5 数据集与性能评估

在视频动作预测算法快速发展的同时，用于评估算法的数据集和指标也在日益完善. 表 4 总结了未来动作预测领域常用的数据集，可以看到其存在如下特点：(1) 拍摄场景以室内环境居多，其中大部分动作与细粒度的烹饪活动有关；(2) 近些年来常用的数据集以第一人称视角拍摄居多，人类的日常活动正在得到研究人员的关注；(3) 大部分数据集都是有剧本的，近年来无剧本形式采集的自然数据逐渐涌现，这些数据集所记录的动作越来越贴近真实生活，具有更大的随机性，对于机器预测而言也更具有挑战性；(4) 除了彩色图像帧之外，一些数据集还会提供深度信息、视线、手部掩膜等多模态信息；(5) 数据集规模呈现稳步上升的发展趋势，体现在视频总时长、视频数量、动作类别数、动作实例数等方面. 为了进一步方便读者使用，表 5 给出了这些数据集的下载链接.

视频中的未来动作预测主要采用的评价指标包括 Top-k 准确率、平均精准率和平均召回率(通常 k 的取值为 1 或 5). 其中 Top-k 准确率的计算方式为针对所有样本，计算真实类别在排名前 k 位的预测类别中的样本所占的百分比；Top-k 平均精准率的计算方式为首先对于某一给定类别，计算真实类别在排名前 k 位的该类别预测样本占该类别所有预测样本的比例，得到该类别的 Top-k 精准率，然后将所有类别的精准率取平均；Top-k 平均召回率的计算方式为首先对某一给定类别，计算真实类别在排名前 k 位的该类别预测样本占该类别所有样本的百分比，得到该类别的 Top-k 召回率，然后将所有类别的召回率取平均.

表 4 视频中的未来动作预测常用数据集

数据集	年份	拍摄环境	拍摄视角	拍摄方式	模态	总时长/h	视频数量	动作类别	动作实例
KTH[118]	2004	室内，室外	第三人称	有剧本	灰度图像帧	2	2391	6	2391
Hollywood2[119]	2009	室内，室外	第三人称	有剧本	彩色图像帧	7	69	12	3669
UT-Interaction[120]	2010	室外	第三人称	有剧本	彩色图像帧	—	20	6	120
TV Human Interaction[121]	2010	室内，室外	第三人称	有剧本	彩色图像帧	0.3	300	4	—
ADL[122]	2012	室内	第一人称	有剧本	彩色图像帧	—	20	32	436
GTEA Gaze[123]	2012	室内(烹饪)	第一人称	有剧本	彩色图像帧，视线，手部掩膜	9	35	40	331
MPII-Cooking[124]	2012	室内(烹饪)	第三人称	有剧本	彩色图像帧	8	44	65	5609
UCF-101[125]	2012	室内，室外	第三人称	有剧本	彩色图像帧	27	13320	5	101
50Salads[126]	2013	室内(烹饪)	第三人称	有剧本	彩色图像帧，深度	4.5	50	17	966
JHMDB[127]	2013	室内，室外	第三人称	有剧本	彩色图像帧，骨架	—	928	21	—
Breakfast[77]	2014	室内(烹饪)	第三人称	无剧本	彩色图像帧	77	1989	10	8456
THUMOS-14[128]	2014	室外	第三人称	有剧本	彩色图像帧	20	413	20	6365
GTEA Gaze+[123]	2015	室内(烹饪)	第一人称	有剧本	彩色图像帧，视线，手部掩膜	10	35	40	331
ActivityNet-200[129]	2015	室外	第三人称	无剧本	彩色图像帧	648	14950	200	23064
Charades[130]	2016	室内	第三人称	有剧本	彩色图像帧	—	9848	157	—
TV Series[131]	2016	室内，室外	第三人称	有剧本	彩色图像帧	16	27	30	6231
NTU RGB-D[132]	2016	室内	第三人称	有剧本	彩色图像帧，深度，红外	—	56000	60	114480
ActEV-VIRAT[133]	2018	室外	第三人称	无剧本	彩色图像帧	12	455	12	—
EPIC-Kitchens-55[39]	2018	室内(烹饪)	第一人称	无剧本	彩色图像帧，音频	55	432	2747	39596
Charades-ego[134]	2018	室内	第一，三人称	有剧本	彩色图像帧	34.4	2751	157	30516
EGTEA Gaze+[135]	2018	室内(烹饪)	第一人称	有剧本	彩色图像帧，视线，手部掩膜	28	86	106	10325
Something-Something-v2[136]	2018	室内，室外	第一人称	有剧本	彩色图像帧	—	220847	174	10899
YouCook2[137]	2018	室内(烹饪)	第三人称	有剧本	彩色图像帧，音频	176	2000	89	15400
Tasty Videos[49]	2018	室内(烹饪)	第三人称	有剧本	彩色图像帧	—	2511	185	21243
EPIC-Tent[138]	2019	室外	第一人称	有剧本	彩色图像帧，音频	5.4	24	12	921
BDD100K[139]	2020	室外	第一人称	有剧本	彩色图像帧，地理位置，速度计	1111	100000	—	3
EPIC-Kitchens-100[140]	2022	室内(烹饪)	第一人称	无剧本	彩色图像帧，音频	100	700	4053	90000
Ego4D[141]	2022	室内，室外	第一人称	无剧本	彩色图像帧，音频	3025	2527	110	226000

表 5　数据集下载链接

数据集	网页下载链接
KTH[118]	http://www.nada.kth.se/cvap/actions/
Hollywood2[119]	http://www.irisa.fr/vista/actions/hollywood2
UT-Interaction[120]	https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html
TVHuman Interaction[121]	http://www.robots.ox.ac.uk/nsimalonso/tv human interactions.html
ADL[122]	https://web.cs.ucdavis.edu/~hpirsiav/papers/ADLdataset
GTEA Gaze[123]	http://www.cbi.gatech.edu/fpv/
MPII-Cooking[124]	https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpii-cooking-activities-dataset/
UCF-101[125]	https://www.crcv.ucf.edu/data/UCF101.php
Breakfast[77]	http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/
50Salads[126]	https://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/
JHMDB[127]	http://jhmdb.is.tue.mpg.de/
THUMOS-14[128]	http://www.thumos.info/home.html
GTEA Gaze+[123]	http://www.cbi.gatech.edu/fpv/
ActivityNet-200[129]	http://activity-net.org/download.html
Charades[130]	http://vuchallenge.org/charades.html
TV Series[131]	https://github.com/zhenyangli/online action
NTU RGB-D[132]	http://rose1.ntu.edu.sg/Datasets/actionRecognition.asp
ActEV-VIRAT[133]	https://actev.nist.gov/trecvid19
EPIC-Kitchens-55[39]	https://epic-kitchens.github.io/2020-55
Charades-ego[134]	https://prior.allenai.org/projects/charades-ego
EGTEA Gaze+[135]	http://www.cbi.gatech.edu/fpv/
Something-Something-v2[136]	https://developer.qualcomm.com/software/ai-datasets/something-something
YouCook2[137]	http://youcook2.eecs.umich.edu/
Tasty Videos[49]	https://cvml.comp.nus.edu.sg/tasty/index.html
EPIC-Tent[138]	https://sites.google.com/view/epic-tent
BDD100K[139]	https://www.bdd100k.com
EPIC-Kitchens-100[140]	https://epic-kitchens.github.io
Ego4D[141]	https://ego4d-data.org/

接下来本文重点针对短期动作预测和长期动作预测任务分别介绍目前流行使用的数据集，并结合一些代表性方法的性能表现进行分析.

(1) 短期动作预测. EPIC-Kitchens-55[39] 是包括 55 个小时视频的第一人称视角数据集，由 32 名参与者分别在 32 个不同的厨房环境中利用头戴式相机拍摄. 由于采集过程中没有任何剧本指导拍摄者进行活动，因此该数据集记录的活动非常贴近真实生活. 该数据集包含约 39 596 个动作实例以及 2513 个动作类别，用于训练的数据来自 28 个厨房环境，包含约 28472 个动作实例，用于测试的数据分为两部分：一部分来自同样 28 个厨房环境(已知环境)，约有 8047 个动作实例，另一部分来自其它 4 个厨房环境(未知环境)，约有 2929 个动作实例. 所采用的评价指标包括 Top-1 和 Top-5 准确率、平均准确率和召回率. 不同方法在 EPIC-Kitchens-55 数据集上的性能分别如表 6 所示，从中可以看到直接套用动作识别模型的方法 2SCNN 和 ATSN[39] 取得

表 6　不同方法在 EPIC-Kitchens-55 上的性能表现

(单位：%)

方法	已知环境/未知环境			
	Top-1 准确率	Top-5 准确率	平均精准率	平均召回率
DMR[110]	1.27/0.55	7.17/4.39	0.33/0.55	0.47/0.20
ED[66]	8.08/2.65	18.19/7.57	5.69/1.35	4.33/1.38
2SCNN[39]	4.32/2.29	15.21/9.35	2.48/0.85	1.81/1.14
ATSN[39]	6.00/2.39	18.21/6.63	3.13/0.80	2.39/1.07
MCE[106]	10.76/5.57	25.28/15.71	6.05/1.99	5.11/2.39
TransR(2+1)D[108]	9.74/7.24	25.44/19.29	3.67/2.20	3.85/3.36
RULSTM[40]	14.39/8.16	33.73/21.10	7.37/3.64	7.66/4.83
IAI[103]	13.55/8.57	32.70/21.41	6.83/3.33	7.40/4.56
KDLM[69]	14.43/8.81	34.99/21.34	6.64/4.48	7.61/4.78
SRL[70]	14.24/8.88	34.61/22.06	6.45/2.84	6.34/4.33
ImagineRNN[43]	14.66/9.25	34.98/22.19	6.66/3.47	7.08/5.21
Ego-OMG[90]	6.02/11.81	34.53/23.76	4.03/4.52	5.36/5.65
FHOI[41]	15.42/9.94	34.29/23.69	6.93/4.40	7.88/5.18
MTCN[83]	15.45/8.91	34.37/21.07	6.94/4.35	8.40/4.94
MGRKD[88]	16.98/10.38	37.12/23.05	—	—
TempAgg[42]	16.64/10.04	36.06/23.42	9.64/4.92	10.05/6.26
AVT[44]	16.84/10.41	36.52/24.27	9.71/4.84	10.11/6.41
DCR[97]	17.75/10.93	38.51/24.75	9.40/5.30	10.46/6.83

的性能相对较低，这促使后续工作根据动作预测问题有针对性地设计了方法. RULSTM[40] 充分利用了 RNN 强大的序列建模能力，大幅提升了动作预测的性能. 在此基础上 KDLM[69]、SRL[70]、ImagineRNN[43] 等方法通过结合语义信息或利用未来视频帧的信息进行知识蒸馏等方式进一步提升了预测性能. FHOI[41] 通过同时预测轨迹和交互区域为预测未来动作提供了辅助信息. MGRKD[88] 利用 GCN 建模视频片段之间的上下文关系. AVT[44] 和 DCR[97] 利用近期流行的自注意力机制有效地捕捉长距离的时序依赖关系，取得了最佳的预测性能. 此外，对于不同类型的方法，从已知环境到未知环境的转变都会使得它们的性能明显降低，如何提升现有方法的泛化能力具有很大的研究空间. 表 7 进一步给出了代表性方法在 EPIC-Kitchens-55 的验证集上性能和所需计算资源的比较，验证集包含来自 28 个厨房环境的 23 493 个动作实例. 如表 7 所示，可以发现在使用相同的结构作为核心模块时，通过结合视觉和语义信息、利用外部信息进行知识蒸馏或者联合预测轨迹和区域信息等方法都可以在一定程度上提升动作预测的性能，但同时也增强了模型的复杂程度. 总体而言，随着动作预测模型越来越复杂，在预测准确率提升的同时参数量也在不可避免地增加，而参数量的增加通常会导致优化困难和出现过拟合的风险，从而对所需的计算资源提出了更高的要求. 因此，如何提升预测准确率和缓解对于计算资源的依赖之间取得平衡值得进一步思考和探索.

表 7 代表性方法在 EPIC-Kitchens-55 验证集上的性能和计算资源比较

方法	Top-1/5 准确率/%	计算资源	特点
RULSTM[40]	15.36/35.32	Nvidia Titan X	核心结构为 RNN，基于视觉信息，直接从观察内容预测
KDLM[69]	16.05/37.54	Nvidia Titan Xp	核心结构为 RNN，结合视觉和语义信息，利用外部信息进行知识蒸馏
TempAgg[42]	16.28/35.68	2*Nvidia Titan Xp	核心结构为 CNN，基于视觉信息，直接从观察内容预测
FHOI[41]	15.35/35.96	4*Nvidia Titan Xp	核心结构为 CNN，基于视觉信息，直接从观察内容预测，联合预测轨迹和区域
MGRKD[88]	17.22/37.99	4*Nvidia RTX 3090	核心结构为 GCN，基于视觉信息，利用外部信息进行知识蒸馏
AVT[44]	16.60/37.60	4*Nvidia RTX 3090	核心结构为 Transformer，基于视觉信息，直接从观察内容预测
DCR[97]	19.20/41.20	4*Nvidia RTX 3090	核心结构为 Transformer，基于视觉信息. 利用外部信息进行知识蒸馏

(2) 长期动作预测：50Salads[126] 记录了由 25 名参与者准备不同种类沙拉的 50 个视频，共包含 17 个动作类别，平均每个视频包含约 20 个动作实例，长度在 6.4 min 左右. 研究人员在该数据集上进行 5 倍交叉验证，每次留出 10 个视频用作测试. Breakfast[77] 记录了由 52 名参与者制作早餐的 1712 个视频，共包括 48 个动作类别，平均每个视频包含约 6 个动作实例，长度在 2.3 min 左右. 该数据集提供了 4 种划分，最终的性能为 4 次测试的平均值. 不同方法在 50Salads 和 Breakfast 数据集上的性能分别如表 8 和表 9 所示，所采用的评价指标均为平均召回率，通常观察某个视频前 20% 或 30% 的内容，预测其后面 10%～50% 的内容. 从中可以看到，长期动作预测非常依赖对于观察信息的时序建模，因此基于 RNN 的方法优于基于 CNN 的方法[45]，后续的研究方法在基于 RNN 的模型结构上进行改进. Time-Condition[46]、TempAgg[42]、Att-GRU[78] 等方法通过引入注意力机制，对不同时刻的视频帧信息分配不同的权重，从而更能关注时间尺度上重要的信息，这对于提升长期动作预测的性能非常有

表 8 不同方法在 50Salads 上的性能表现 (单位：%)

方法	观察前 20%/30%			
	预测后 10%	预测后 20%	预测后 30%	预测后 50%
CNN[45]	21.24/29.14	19.03/20.14	15.98/17.46	9.87/10.86
RNN[45]	30.06/21.64	25.43/20.02	18.74/19.73	13.49/19.21
Uncertainty[74]	24.86/29.10	22.37/20.50	19.88/15.28	12.82/12.31
Time-Condition[46]	32.51/35.12	27.61/27.05	21.26/22.05	15.99/15.59
Cycle[47]	34.76/34.39	28.41/23.70	21.82/18.95	15.25/15.89
TempAgg[42]	32.70/32.30	26.30/25.50	21.90/22.70	15.60/17.10
AGG[142]	39.50/39.50	33.20/31.50	25.90/26.40	21.20/19.80
Att-GRU[78]	39.32/41.73	31.39/32.73	27.01/31.44	23.88/26.39

表 9 不同方法在 Breakfast 上的性能表现 (单位：%)

方法	观察前 20%/30%			
	预测后 10%	预测后 20%	预测后 30%	预测后 50%
CNN[45]	17.90/22.44	16.35/20.12	15.37/19.69	14.54/18.76
RNN[45]	18.11/21.64	17.20/20.02	15.94/19.73	15.81/19.21
Uncertainty[74]	16.71/20.73	15.40/18.27	14.47/18.42	14.20/16.86
Time-Condition[46]	18.41/22.75	17.21/20.44	16.42/19.64	15.84/19.75
Cycle[47]	25.88/29.66	23.42/27.37	22.42/25.58	21.54/25.20
TempAgg[42]	18.80/23.00	16.90/20.00	16.50/19.90	15.40/18.60
Att-GRU[78]	23.03/26.50	22.28/25.00	22.00/24.08	20.85/23.61

帮助.另外,在给定预测时间范围的情况下,随着观察时间范围从 20%增加到 30%,各种方法的性能都有所提升;而在给定观察内容的情况下,预测时间范围从 10%增加到 50%,各种方法的性能都有所下降.如何使得长期预测方法根据有限的观察内容,在较长的时间范围内保持预测性能是该领域值得探索的核心问题.

6　发展趋势

视频中的未来动作预测经历了十年左右的发展,在定义问题形式、设计模型与算法、构建基准数据集以及完善评价指标等方面取得了一系列进展.在未来的工作中,以下几个方面值得进一步研究与探索.

6.1　扩展现有数据集的规模和多样性

现有的基准数据集大多限定在特定的场景中,例如EPIC-Kitchens-55[39]、Breakfast[47]和 50Salads[126]等数据集均拍摄于厨房环境,主要记录了与烹饪有关的活动.基于这些数据集训练的动作预测模型可能难以推广到其它环境,如户外运动或者交通驾驶场景.因此有必要扩展现有数据集的规模,使之覆盖生活中尽可能多的场景.最近 Grauman 等人[140]推出了目前最大的以第一视角拍摄的日常活动视频数据集,拍摄场景包括运动、购物、阅读、园艺、社交等,在该数据集上的未来动作预测值得探索.由于目前常用的数据集以第一视角拍摄居多,为了更好地理解人类进行的活动,未来还可以构建更多的第一视角和第三视角成对的数据集,因为第三视角更能帮助识别人类的位置和姿态等信息,更好地理解场景.此外,利用多模态数据有助于对复杂动作的理解,除了现有工作所利用的各种视觉和动作语义信息外,还可以融合音频、场景、物体属性等多模态信息进行未来动作预测,而现有的很多数据集在采集时缺少音频等重要的信息,因此将来构建数据集时还应考虑尽可能提供多种类型的模态信息.

6.2　缩短模型的推理时间

实际的应用场景不仅需要考虑未来动作预测模型的准确率,还应该将模型的推理时间(inference time)作为设计算法所考虑的重要因素之一,这样才能适应实时的流媒体(streaming)场景,对于将模型部署到实时系统也具有重要的意义,例如自动驾驶系统的反应时间越短越好.现有的大部分工作[40-47,68-71,95-97,143]忽视了模型的推理时间,在离线(offline)的设置下预测未来动作,即默认模型的推理时间为 0,这显然与实际应用的设定不符.目前已有少部分工作开始关注模型的推理时间,Zatsarynna 等人[83]基于 CNN 构建轻量级的动作预测模型,Furnari 等人[112]明确定义了流媒体场景下的短期动作预测,将推理时间作为重要的参数,并且采用知识蒸馏的方式构建高效轻量的未来动作预测模型,以缩短模型的推理时间.对于未来的研究工作而言,如何将流媒体场景与长期动作预测结合、如何构建更精简的模型以满足实时性和高效性的要求、如何取得预测准确率和效率的平衡等问题都值得深入探索.

6.3　从无标注或少量标注数据中学习

现有的动作预测模型依赖大规模标注数据集的训练,而对于大规模细粒度的动作数据集的标注通常耗时耗力,成本非常昂贵.由于自监督学习可以从数据本身获得监督信号,而视频数据包含大量的动态结构信息,一些工作[111,141]结合自监督学习的思想,尝试从当前视频帧的特征表示中预测未来视频帧的特征表示,在特征层面实现监督,不受动作标签的限制,所得到的模型能够迁移到未来动作预测等下游任务.此外,还有一些工作[78,144]结合弱监督学习的思想,尝试利用部分标签作为监督信号来进行未来动作预测.如何在缺少标注的情况下有效地学习动作预测模型具有重要的研究价值.

6.4　从特定领域的少样本数据中学习

目前用于训练动作预测模型的数据往往规模庞大[39,77,126],并且存在类别分布不均衡的问题.受到客观条件的限制,例如医疗康复领域涉及用户隐私等原因,一些特定领域所能获取的数据在样本和种类数量方面往往都存在局限性.一方面,训练样本的缺少往往会导致深度学习模型的过拟合问题,严重影响模型的泛化能力.为此有必要结合小样本学习进行研究,可以从多模态信息的角度通过动作的概率与属性推理,在数据量较少的情况下实现模型较为充分的优化,提升模型的鲁棒性.另一方面,大部分数据集的动作类别存在长尾分布的问题,一些尾部类别往往对应小概率事件,但它们在实际应用中却非常重要,例如对于危险动作的正确预测可以避免安全事故的发生.Damen 等人[140]在发布 EPIC-Kitchens-100 数据集时专门报告了针对尾部类别的预测性能,说明长尾分布问题正在受到研究人员的关注,未来值得进一步探索.

6.5　生成更丰富的语义信息描述

现有的工作一般将待预测的未来动作当作单个

标签来处理，事实上很多动作是同时发生的，例如“往锅里倒油”和“炒菜”，因此单个标签不足以概括视频的内容.这体现了动作预测的难点在于未来具有天然的不确定性，Furnari 等人[106]尝试利用多标签分类的思想改造传统的交叉熵损失函数来进行不确定性建模，但并未拓展现有的单标签评价体系.在未来，一方面可以将现有的单标签评价体系拓展成多标签评价体系，优化评测方式.另一方面，相比简单的动作类别，对于未来视频内容生成的句子描述包含了更丰富的语义信息，对于实际应用也更有帮助，例如人机交互中机器可以得到更明确清晰的指令去执行下一步的操作.目前有一些工作[48-50]在此方向上进行了初步探索，但尚未形成统一的数据集和评测标准，有待进一步完善.

6.6 融合心理学等交叉领域的深入探索

对未来动作进行预测是高层级的人类智能表现，在人类感知外部世界和作出复杂决策方面起到承上启下的作用.对于机器而言，需要对人类大脑预测潜在未来事件的过程进行模拟，而非简单地依赖从输入到输出的可观察的统计相关性，这有助于实现更高层级的认知性智能，因此结合人类心理学和生理学等交叉领域进行探索具有必要性和可行性.Zhang 等人[103]受到人脑认知模式中系统 1 和系统 2[145]的启发，构建了融入直觉和分析的短期动作预测框架，用来模拟人类无意识的直觉思考以及有意识的逻辑推理过程.另外，大脑作为复杂的智能系统，因果推理[146-147]能力是其智能的主要表现之一，近年来在人工智能和计算机视觉领域也引起了广泛关注，它致力于理解数据之间的因果关系，消除虚假关联，对于提升深度学习模型的可解释性、泛化能力以及缓解对数据的过度依赖都非常有帮助.目前将因果推理与未来动作预测研究相结合的工作相对较少[104]，未来具有很大的发展潜力.

6.7 关注更复杂的延后帧动作预测

目前未来动作预测的研究任务主要包括短期动作预测和长期动作预测，其中短期动作预测的预测时间较短，通常研究如何预测数秒之内发生的动作[39-41]，属于近邻帧动作预测的范畴.而长期动作预测[45]的时间范围则可从数秒持续至数分，如果预测时间较长，也可能属于更复杂的延后帧动作预测的范畴.从表 8 和表 9 可以看出，随着预测时间的延长，动作预测的对象由近邻帧动作变为延后帧动作，预测性能也会持续出现下降，这充分体现了延后帧动作预测的挑战性.从前文叙述可知，目前很多方法只停留在短期动作预测层面进行探索，例如 4.2.3 节提到的利用外部信息进行知识蒸馏的方法，在长期动作预测领域尚未得到广泛应用.此外，4.2.2 节提到的结合视觉信息和语义信息的方法在应用于长期动作预测时通常需要从人工标注的真实标签中获取语义信息，这显然限制其推广到实际应用场景.综上所述，在未来动作预测研究领域，针对更复杂的延后帧动作预测值得进一步关注和探索.

7 结 论

作为新兴的研究课题，视频中的未来动作预测具有重要的研究意义和应用价值.本文首先明确定义了未来动作预测的研究框架，并概述了未来动作预测的发展历史，重点介绍了短期动作预测和长期动作预测两种问题形式.随后从模型结构、数据模态、算法策略和预测对象等不同维度对主要方法和技术进行了总结梳理.然后介绍并分析了未来动作预测领域常用的数据集及性能评价.最后本文围绕扩展现有数据集的规模和多样性、缩短模型的推理时间、从无标注或少量标注数据中学习、从特定领域少样本数据中学习、生成更丰富的语义信息描述、融合心理学等交叉领域的深入探索、关注更复杂的延后帧动作预测等方面对未来动作预测的发展方向进行了总结和展望，希望能够促进该研究领域的持续发展和进步.

参 考 文 献

[1] Abdoullaev A. Trans-AI: How to build true AI or real machine intelligence and learning. Ontology of Designing, 2021, 11(4): 402-421

[2] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Proceedings of the Conference on Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 568-576

[3] Wang Li-Min, Xiong Yuan-Jun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 20-36

[4] Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6299-6308

[5] Ryoo M S. Human activity prediction: Early recognition of ongoing activities from streaming videos//Proceedings of the

IEEE International Conference on Computer Vision. London, UK, 2011: 1036-1043

[6] Cao Y, Barrett D, Barbu A, et al. Recognize human activities from partially observed videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2658-2665

[7] Oh J, Guo X, Lee H, et al. Action-conditional video prediction using deep networks in Atari games//Proceedings of the Neural Information Processing Systems. Montreal, Canada, 2015: 2863-2871

[8] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using LSTMs//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 843-852

[9] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human trajectory prediction in crowded spaces//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 961-971

[10] Gupta A, Johnson J, Fei-Fei L, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2255-2264

[11] Fragkiadaki K, Levine S, Felsen P, et al. Recurrent network models for human dynamics//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4346-4354

[12] Martinez J, Black M J, Romero J. On human motion prediction using recurrent neural networks//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2891-2900

[13] Furnari A, Battiato S, Grauman K, et al. Next-active-object prediction from egocentric videos. Journal of Visual Communication and Image Representation, 2017, 49: 401-411

[14] Nagarajan T, Feichtenhofer C, Grauman K. Grounded human-object interaction hotspots from video//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 8688-8697

[15] Zhang M, Ma K T, Lim J H, et al. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4372-4381

[16] Fan C, Lee J, Ryoo M S. Forecasting hands and objects in future frames//Proceedings of the European Conference on Computer Vision Workshops. Santiago, Chile, 2018: 1-13

[17] Rasouli A. Deep learning for vision-based prediction: A survey. arXiv preprint arXiv:2007.00095, 2020

[18] Rodin I, Furnari A, Mavroeidis D, et al. Predicting the future from first person (egocentric) vision: A survey. Computer Vision and Image Understanding, 2021, 211: 103-152

[19] Zhao H, Wildes R P. Review of video predictive understanding: Early action recognition and future action prediction. arXiv preprint arXiv:2107.05140, 2021

[20] Kong Y, Fu Y. Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230, 2018

[21] Bubic Andreja, Von Cramon, Ricarda Schubotz, et al. Prediction, cognition and the brain. Frontiers in Human Neuroscience, 2010, 25(4): 1-15

[22] Li K, Hu J, Fu Y. Modeling complex temporal composition of actionlets for activity prediction//Proceedings of the European Conference on Computer Vision. Firenze, Italy, 2012: 286-299

[23] Huang De-An, Kitani K M. Action-reaction: Forecasting the dynamics of human interaction//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 489-504

[24] Kitani K M, Ziebart B D, Bagnell J A, et al. Activity forecasting//Proceedings of the European Conference on Computer Vision. Firenze, Italy, 2012: 201-214

[25] Qi S, Huang S, Wei P, et al. Predicting human activities using stochastic grammar//Proceedings of the IEEE International Conference on Computer Vision. Zurich, Switzerland, 2017: 1164-1172

[26] Liang J, Jiang L, Niebles J C, et al. Peeking into the future: Predicting future person activities and locations in videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2019: 5725-5734

[27] Chen B, Sun X, Li D, et al. SCR-graph: Spatial-causal relationships based graph reasoning network for human action prediction//Proceedings of the International Conference on Computing and Data Science. Dortmund, Germany, 2021: 1-9

[28] Wei X, Lucey P, Vidas S, et al. Forecasting events using an augmented hidden conditional random field//Proceedings of the Asian Conference on Computer Vision. Singapore, 2014: 569-582

[29] Felsen P, Agrawal P, Malik J. What will happen next? Forecasting player moves in sports videos//Proceedings of the IEEE International Conference on Computer Vision. Zurich, Switzerland, 2017: 3342-3351

[30] Bertasius G, Shi J. Using cross-model EgoSupervision to learn cooperative basketball intention//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy, 2017: 2355-2363

[31] Koppula H, Saxena A. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation //Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 792-800

[32] Koppula H S, Saxena A. Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(1): 14-29

[33] Ryoo M S, Fuchs T J, Xia L, et al. Robot-centric activity prediction from first-person videos: What will they do to me?//Proceedings of the International Conference on Human-Robot Interaction. Portland Oregon, USA, 2015: 295-302

[34] Jain A, Koppula H S, Raghavan B, et al. Car that knows before you do: Anticipating maneuvers via learning temporal driving models//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3182-3190

[35] Jain A, Singh A, Koppula H S, et al. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture//Proceedings of the IEEE International Conference on Robotics and Automation. Stockholm, Sweden, 2016: 3118-3125

[36] Chan F-H, Chen Y-T, Xiang Y, et al. Anticipating accidents in dashcam videos//Proceedings of the Asian Conference on Computer Vision. Taipei, China, 2016: 136-153

[37] Suzuki T, Kataoka H, Aoki Y, et al. Anticipating traffic accidents with adaptive loss and large-scale incident DB//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3521-3529

[38] Rasouli A, Kotseruba I, Tsotsos J K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy, 2017: 206-213

[39] Damen D, Doughty H, Farinella G M, et al. Scaling egocentric vision: The EPIC-Kitchens dataset//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 720-736

[40] Furnari A, Farinella G M. What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 6252-6261

[41] Liu M, Tang S, Li Y, et al. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 704-721

[42] Sener F, Singhania D, Yao A. Temporal aggregate representations for long-range video understanding//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 154-171

[43] Wu Y, Zhu L, Wang X, et al. Learning to anticipate egocentric actions by imagination. IEEE Transactions on Image Processing, 2020, 30: 1143-1152

[44] Girdhar R, Grauman K. Anticipative video transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 13505-13515

[45] Abu Farha Y, Richard A, Gall J. When will you do what?—Anticipating temporal occurrences of activities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5343-5352

[46] Ke Q, Fritz M, Schiele B. Time-conditioned action anticipation in one shot//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9925-9934

[47] Abu Farha Y, Ke Q, Schiele B, et al. Long-term anticipation of activities with cycle consistency//Proceedings of the DAGM German Conference on Pattern Recognition. Basel, Switzerland, 2020: 159-173

[48] Sun C, Myers A, Vondrick C, et al. VideoBERT: A joint model for video and language representation learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 7464-7473

[49] Sener F, Yao A. Zero-shot anticipation for instructional activities//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 862-871

[50] Epstein D, Wu J, Schmid C, et al. Learning temporal dynamics from cycles in narrated video//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1480-1489

[51] Mahmud T, Billah M, Hasan M, et al. Prediction and description of near-future activities in video. Computer Vision and Image Understanding, 2021, 210: 210-219

[52] Chakraborty A, Roy-Chowdhury A K. Context-aware activity forecasting//Proceedings of the Asian Conference on Computer Vision. Singapore, 2014: 21-36

[53] Geman S, Graffigne C. Markov random field image models and their applications to computer vision//Proceedings of the International Congress of Mathematicians: Volume 1. Berkeley, USA, 1986: 1496-1517

[54] Li K, Fu Y. Prediction of human activity by discovering temporal sequence patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1644-1657

[55] Lafferty J D, McCallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the International Conference on Machine Learning. New South Wales, Australia, 2001: 282-289

[56] Soran B, Farhadi A, Shapiro L. Generating notifications for missing actions: Don't forget to turn the lights off//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4669-4677

[57] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257-286

[58] Lan T, Chen T-C, Savarese S. A hierarchical representation for future action prediction//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 689-704

[59] Schölkopf B. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: MIT Press, 2002

[60] Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1989, 1(2): 270-280

[61] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[62] Cho K, van Merrienboer B. Learning phrase representations using RNN encoder-decoder for statistical machine translation //Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[63] LeCun Y, Boser B E, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1: 541-551

[64] Wu Tz-Ying, Chien Ting-An, Chan Cheng-Sheng, et al. Anticipating daily intention using on-wrist motion triggered sensing//Proceedings of the IEEE International Conference on Computer Vision. Zurich, Switzerland, 2017: 48-56

[65] Shi Y, Fernando B, Hartley R. Action anticipation with RBF kernelized feature mapping RNN//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 301-317

[66] Gao, Jiyang, Zhenheng Yang, and Ram Nevatia. RED: Reinforced encoder-decoder networks for action anticipation // Proceedings of the British Machine Vision Conference. London, UK, 2017: 1-11

[67] Furnari A, Farinella G M. Egocentric action anticipation by disentangling encoding and inference//Proceedings of the IEEE International Conference on Image Processing. Taipei, China, 2019: 3357-3361

[68] Furnari A, Farinella G M. Rolling-unrolling LSTMs for action anticipation from first-person video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 4021-4036

[69] Camporese G, Coscia P, Furnari A, et al. Knowledge distillation for action anticipation via label smoothing//Proceedings of the International Conference on Pattern Recognition. Milan, Italy, 2021: 3312-3319

[70] Qi Z, Wang S, Su C, et al. Self-regulated learning for egocentric video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021

[71] Osman N, Camporese G, Coscia P, et al. SlowFast rolling-unrolling LSTMs for action anticipation in egocentric videos // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Montreal, Canada, 2021: 3437-3445

[72] Feichtenhofer C, Fan Haoqi, Malik J, et al. SlowFast networks for video recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 6202-6211

[73] Liu T, Lam K-M. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13904-13913

[74] Abu Farha Y, Gall J. Uncertainty-aware anticipation of activities//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea, 2019: 1-8

[75] Gammulle H, Denman S, Sridharan S, et al. Forecasting future action sequences with neural memory networks// Proceedings of the British Machine Vision Conference. Cardiff, Wales, UK, 2019

[76] Morais R, Le V, Tran T, et al. Learning to abstract and predict human actions//Proceedings of the British Machine Vision Conference. Virtual, 2020

[77] Kuehne H, Arslan A, Serre T. The language of actions: Recovering the syntax and semantics of goal-directed human activities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 780-787

[78] Ng Y B, Fernando B. Forecasting future action sequences with attention: A new approach to weakly supervised action forecasting. IEEE Transactions on Image Processing, 2020, 29: 8880-8891

[79] Tran V, Wang Y, Zhang Z, et al. Knowledge distillation for human action anticipation//Proceedings of the 2021 IEEE International Conference on Image Processing. Virtual, 2021: 2518-2522

[80] Fernando B, Herath S. Anticipating human actions by correlating past with the future with Jaccard similarity measures// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13224-13233

[81] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6450-6459

[82] Tran D, Wang H, Torresani L, et al. Video classification with channel-separated convolutional networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 5552-5561

[83] Zatsarynna O, Abu Farha Y, Gall J. Multimodal temporal convolutional network for anticipating actions in egocentric videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 2249-2258

[84] Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 156-165

[85] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2048-2057

[86] Wang X, Girshick R, Gupta A, et al. Non-local neural networks//Proceedings of the IEEV/CVF Conference on Computer Vision and Pattern Recognition. SaltLake City, USA, 2018: 7794-7803

[87] Kipf T, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico, 2016: 1715-1728

[88] Huang Y, Yang X, Xu C. Multimodal global relation knowledge distillation for egocentric action anticipation//Proceedings of the ACM International Conference on Multimedia. Chengdu, China, 2021: 245-254

[89] Dessalene E, Maynord M, Devaraj C, et al. Egocentric object manipulation graphs. arXiv preprint arXiv:2006.03201, 2020

[90] Dessalene E, Devaraj C, Maynord M, et al. Forecasting action through contact representations from first person video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021

[91] Nagarajan T, Li Y, Feichtenhofer C, et al. EGO-TOPO: Environment affordances from egocentric video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 163-172

[92] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Conference on Advances in Neural Information Processing Systems: Volume 30. Long Beach, USA, 2017: 1-11

[93] Han Kai, Wang Yun-He, Chen Han-Ting, et al. A survey on vision transformer. arXiv preprint arXiv:2012.12556, 2020

[94] Wang W, Peng X, Su Y, et al. TTPP: Temporal transformer with progressive prediction for efficient action anticipation. Neurocomputing, 2021, 438: 270-279

[95] Roy D, Fernando B. Action anticipation using pairwise human-object interactions and transformers. IEEE Transactions on Image Processing, 2021, 30: 8116-8129

[96] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021

[97] Xu X, Li Y, Lu C. Learning to anticipate future with dynamic context removal//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 12734-12744

[98] Mahmud T, Hasan M, Chakraborty A, et al. A poisson process model for activity forecasting//Proceedings of the IEEE International Conference on Image Processing. NW Washington, USA, 2016: 3339-3343

[99] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

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

[101] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 448-456

[102] Girshick R. Fast R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448

[103] Zhang T, Min W, Zhu Y, et al. An egocentric action anticipation framework via fusing intuition and analysis//Proceedings of the ACM International Conference on Multimedia. Virtual, 2020: 402-410

[104] Zhang T, Min W, Yang J, et al. What if we could not see? Counterfactual analysis for egocentric action anticipation//Proceedings of the International Joint Conference on Artificial Intelligence. Virtual, 2021: 1316-1322

[105] Shen Y, Ni B, Li Z, et al. Egocentric activity prediction via event modulated attention//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 197-212

[106] Furnari A, Battiato S, Farinella G M. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation//Proceedings of the European Conference on Computer Vision Workshops. Munich, Germany, 2018: 41-53

[107] Roy D, Fernando B. Action anticipation using latent goal learning//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 2745-2753

[108] Miech A, Laptev I, Sivic J, et al. Leveraging the present to anticipate the future in videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 2915-2922

[109] Zhao H, Wildes R P. On diverse asynchronous activity anticipation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 781-799

[110] Hinton G, Vinyals O, Dean J, et al. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015

[111] Vondrick C, Pirsiavash H, Torralba A. Anticipating visual representations from unlabeled video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 98-106

[112] Furnari A, Farinella G M. Towards streaming egocentric action anticipation. arXiv preprint arXiv:2110.05386, 2021

[113] Mahmud T, Hasan M, Roy-Chowdhury A K. Joint prediction of activity labels and starting times in untrimmed videos//Proceedings of the IEEE International Conference on Computer Vision. Zurich, Switzerland, 2017: 5773-5782

[114] Mehrasa N, Jyothi A A, Durand T, et al. A variational auto-encoder model for stochastic point processes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3165-3174

[115] Neumann L, Zisserman A, Vedaldi A. Future event prediction: If and when//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 2935-2943

[116] Zeng K-H, Chou S-H, Chan F-H, et al. Agent-centric risk assessment: Accident anticipation and risky region localization //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2222-2230

[117] Chen G, Li J, Lu J, et al. Human trajectory prediction via counterfactual analysis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 9824-9833

[118] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach//Proceedings of the International Conference on Pattern Recognition: Volume 3. Cambridge, UK, 2004: 32-36

[119] Marszalek M, Laptev I, Schmid C. Actions in context// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami Beach, USA, 2009: 2929-2936

[120] Ryoo M S, Aggarwal J K. UT-interaction dataset, ICPR contest on semantic description of human activities//Proceedings of the IEEE International Conference on Pattern Recognition Workshops. San Francisco, USA, 2010: 1036-1043

[121] Patron-Perez A, Marszalek M, Zisserman A, et al. High five: Recognising human interactions in TV shows//Proceedings of the British Machine Vision Conference. Aberystwyth, Wales, UK, 2010

[122] Pirsiavash H, Ramanan D. Detecting activities of daily living in first-person camera views//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2847-2854

[123] Fathi A, Li Y, Rehg J M. Learning to recognize daily actions using gaze//Proceedings of the European Conference on Computer Vision. Firenze, Italy, 2012: 314-327

[124] Rohrbach M, Amin S, Andriluka M, et al. A database for fine grained activity detection of cooking activities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA, 2012: 1194-1201

[125] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012

[126] Stein S, Mckenna S J. Combining embedded accelerometers with computer vision for recognizing food preparation activities //Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing. Zurich, Switzerland, 2013: 729-738

[127] Jhuang H, Gall J, Zuffi S, et al. Towards understanding action recognition//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 3192-3199

[128] Jiang Y-G, Liu J, Zamir A R, et al. The THUMOS challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding, 2017, 155: 1-23

[129] Caba Heilbron F, Escorcia V, Ghanem B, et al. ActivityNet: A large-scale video benchmark for human activity understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 961-970

[130] Sigurdsson G A, Varol G. Hollywood in homes: Crowdsourcing data collection for activity understanding// Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 510-526

[131] De Geest R, Gavves E, Ghodrati A, et al. Online action detection//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 269-284

[132] Shahroudy A, Liu J, Ng T-T, et al. NTU RGB+D: A large scale dataset for 3D human activity analysis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1010-1019

[133] Awad G, Butt A, Curtis K, et al. TRECVID 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search//Proceedings of TRECVID 2018. Gaithersburg, USA, 2018: 1-36

[134] Sigurdsson G A, Gupta A, Schmid C, et al. Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626, 2018

[135] Li Y, Liu M, Rehg J M. In the eye of beholder: Joint learning of gaze and actions in first person video//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 619-635

[136] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The "something something" video database for learning and evaluating visual common sense//Proceedings of the IEEE International Conference on Computer Vision. Zurich, Switzerland, 2017: 5842-5850

[137] Zhou L, Xu C, Corso J. Towards automatic learning of procedures from Web instructional videos//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 32. New Orleans, USA, 2018: 6665-7655

[138] Jang Y, Sullivan B, Ludwig C, et al. EPIC-Tent: An egocentric video dataset for camping tent assembly// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea, 2019: 4461-4469

[139] Yu F, Chen H, Wang X, et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2636-2645

[140] Damen D, Doughty H, Farinella G M, et al. Rescaling egocentric vision: Collection, pipeline and challenges for

EPIC-KITCHENS-100. International Journal of Computer Vision, 2022, 130(1): 33-55

[141] Grauman K, Westbury A, Byrne E, et al. Ego4D: Around the world in 3,000 hours of egocentric video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18995-19012

[142] Piergiovanni A J, Angelova A, et al. Adversarial generative grammars for human activity prediction//Proceedings of the European Conference on Computer Vision. 2020: 507-523

[143] Surís D, Liu R, Vondrick C. Learning the predictability of the future//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12607-12617

[144] Zhang H, Chen F, Yao A. Weakly-supervised dense action anticipation//Proceedings of the British Machine Vision Conference. 2021

[145] Kahneman D. Thinking, Fast and Slow. New York, USA: Macmillan, 2011

[146] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York, USA: Basic Books, 2018

[147] Pearl J, Glymour M, Jewell N P. Causal Inference in Statistics: A Primer. Hoboken, USA: John Wiley & Sons, 2016

ZHANG Tian-Yu, Ph. D. candidate. His research interests include multimedia content analysis, understanding and future action anticipation in videos.

MIN Wei-Qing, Ph. D., associate professor. His research interests include multimedia content analysis, understanding and food computing.

HAN Xin-Yang, B. E. candidate. His research interests include multimedia content analysis, understanding and future action anticipation in videos.

JIANG Shu-Qiang, Ph. D., professor. His research interests include multimedia content analysis and retrieval, image/video understanding, and multimodal intelligence.

RUI Yong, Ph. D., professor. His research interests include multimedia retrieval, and knowledge mining.

Background

The ability to predict what will happen in the future is essential for contemporary intelligent systems, which enables many practical applications, such as autonomous driving and assistive robotics. In recent years, with the rapid development of artificial intelligence technology represented by deep learning, the popularity of digital devices and the explosive expanding of video data, an increasing number of researchers have been inspired to focus on future action anticipation in videos. This paper surveys and summarizes the research and developments of this new and fast-growing research topic in artificial intelligence. During the past decade, there has formed two main problem paradigms: short-term action anticipation and long-term action anticipation. The latest research methods are based on deep neural networks, exploring the improvement of performance from many aspects such as combining visual and semantic information, using external information to distill knowledge and multi-task learning.

This work was supported by the National Key Research and Development Project of New Generation Artificial Intelligence of China, under Grant No. 2018AAA0102500. The project focuses on human-computer cooperative intelligent system and algorithms. Our group has been working on future action anticipation in videos for several years and contributes a number of papers.