深度学习在视频对象分割中的应用与展望

陈 加[1] 陈亚松[1] 李伟浩[2] 田 元[1] 刘 智[3] 何 英[4]

[1](华中师范大学教育信息技术学院 武汉 430079)

[2](海德堡大学视觉学习实验室 海德堡 69120 德国)

[3](华中师范大学教育大数据应用技术国家工程实验室 武汉 430079)

[4](清华大学深圳研究生院 广东 深圳 518055)

摘 要 视频对象分割是指在给定的一段视频序列的各帧图像中,找出属于特定前景对象的所有像素点位置区域.随着硬件平台计算能力的提升,深度学习受到了越来越多的关注,在视频对象分割领域也取得了一定的进展.本文首先介绍了视频对象分割的主要任务,并总结了该任务所面临的挑战.其次,对开放的视频对象分割常用数据集进行了简要概述,并介绍了通用的性能评估标准.接着,综述了视频对象分割的研究现状,详细地分析了当前的各种方法,并将它们划分为三大类:半监督的方法,即给出视频第一帧图像中感兴趣对象的详细人工真值标注,分割出视频剩余图像中的感兴趣对象;无监督的方法,即不给任何人工标注信息,自动识别并分割出视频中的前景对象;交互式的方法,即在分割过程中,通过人工交互式的参与,结合粗略的人工标注先验信息,进行视频对象分割.第三类方法的条件相当于前两者的折中:相对于第一类方法,它虽然需要人工的参与,但只需要少量的标注工作量;相对于第二类方法,它给视频序列中某些帧的图像适当地添加了一些人工标注信息,从而更具针对性.最后,对深度学习在视频对象分割任务中的应用,进行了总结和展望.

关键词 视频对象分割;深度学习;半监督方法;无监督方法;交互式方法

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.00609

Application and Prospect of Deep Learning in Video Object Segmentation

CHEN Jia[1] CHEN Ya-Song[1] LI Wei-Hao[2] TIAN Yuan[1] LIU Zhi[3] HE Ying[4]

[1](*Department of Education and Information Technology, Central China Normal University, Wuhan* 430079)

[2](*Visual Learning Lab, Heidelberg University, Heidelberg* 69120 *Germany*)

[3](*National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan* 430079)

[4](*Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong* 518055)

Abstract Video object segmentation refers to the technology by which the positions of all pixels belonging to the particular foreground objects in each frame of a given video sequence can be found out and labeled. This technology is one of the most important research topics in the field of computer vision. And it plays an important role in many applications of computer vision, such as 3D reconstruction, automatic driving, video editing, and so on. With the improvement of computing power, deep learning has attracted more and more attention and made significant progress in the task of video object segmentation. Firstly, this paper introduces the main task of video object segmentation and summarizes the main challenges that the task is facing. Secondly, a brief overview of the open datasets for video object segmentation task is given. Then the relevant

收稿日期:2019-06-03;在线发布日期:2020-01-17.本课题得到国家自然科学基金(61605054,61702207)、国家科技支撑计划项目(2015BAK33B02,2015BAK27B02)、华中师范大学中央高校基本科研业务费(CCNU19QD007,CCNU19TD007)资助.**陈 加**,博士,主要研究方向为视频图像分析、三维运动捕捉、VR/AR、机器人视觉.E-mail:Jacky_HIT@foxmail.com.**陈亚松**,硕士研究生,主要研究方向为视频对象分割、计算机视觉.**李伟浩**,博士研究生,主要研究方向为视频对象分割、计算机视觉.**田 元**,博士,主要研究方向为视频图像分析.**刘 智**(通信作者),博士,副教授,主要研究方向为深度学习、人工智能.E-mail:zhiliu@mail.ccnu.edu.cn.**何 英**,博士,主要研究方向为图形图像处理、机器人视觉.

benchmarks and common performance evaluation criteria are introduced. Thirdly, the research status of video object segmentation is summarized. The relevant methods are introduced and analyzed in detail. And these methods fall in one of the three following categories: the first ones are semi-supervised methods. Namely, the detailed artificial truth annotation of the interested objects in the first frame image of video sequence is given. And the interested objects in the remaining video sequence frames are segmented automatically. At present, in the video object segmentation task of a single instance, the Jaccard score of semi-supervised methods can reach more than 0.8 by taking the DAVIS16 dataset as an example. In the multi-instance video object segmentation task, for example, the DAVIS18 dataset which is widely used, the Jaccard score of semi-supervised methods has reached over 0.7. The second ones are unsupervised methods, which can identify and segment the foreground objects in video by the certain rules or models, without any manual labeling prior information. The third ones are interactive methods, based on the method of interactive rough artificial prior information. In these methods, the rough artificial prior information, such as point, bounding box, and scribble, is obtained from the interactive modules. And video object segmentation is carried out by multiple manual participations, but only a small amount of work at each time. The condition of the third kind of methods can be considered as the compromise of the former two. Compared with the first one, although it requires manual participation, it only requires a small amount of labeling work. Compared with the second one, it appropriately adds some manual labeling information to the images of some frames in the video sequence, which makes the methods more targeted for the interested objects. The best Jaccard scores of the unsupervised methods and the interactive methods can both reach 0.8 in the DAVIS16 dataset. But there are few unsupervised methods that deal with the multi instance problem of the DAVIS18 dataset. The best interactive methods can only reach 0.64 for Jaccard score in the DAVIS18 interactive dataset. Finally, the applications of deep learning in video object segmentation task are concluded, and some promising ideas are proposed from four different aspects.

Keywords video object segmentation; deep learning; semi-supervised methods; unsupervised methods; interactive methods

1 引 言

随着摄像设备和数字化存储设备的普及与广泛应用,全球范围内每天产生的视频数据总量在不断增加.视频内容的处理需求也日益增加,其中,视频对象分割的研究是计算机视觉领域十分重要的研究课题之一,在三维重建、自动驾驶、视频编辑等方面有着重要应用.早期,在视频编码国际标准 MPEG-4[1]中,采用了基于对象的编码方式,指出视频是由一系列视频对象组成的,从而引入了视频对象的概念.视频对象分割是指在给定的一段视频序列的各帧图像中,找出属于特定前景对象的所有像素点位置的技术.当前深度学习在计算机视觉的一些基础任务如图像分类[2]、目标检测[3]、语义分割[4]中都表现出了很好的效果.

视频是由图像组成的,视频对象分割与图像分割存在着紧密的联系.近年来,涌现了很多基于全监督学习和弱监督学习的图像分割方法.全监督学习图像分割方法大致可以分为以下几类:(1)基于全卷积网络(Fully Convolutional Networks,FCN)的方法.这类方法的思想最早由 Long 等人[5]提出,使用全卷积代替全连接层,可以兼容任意尺寸并能实现端到端的训练,但缺乏空间一致性,导致分割结果过于平滑不精细;(2)基于编解码的方法.这类方法改进了 FCN 方法的不足,探索了不同的解码方式,在解码阶段融入低层特征来保留细节,使得分割效果更精细,如 U-net[6]、SegNet[7]、DeconvNet[8]等.但这类方法新增了解码阶段,提高了模型复杂度;(3)基于密集连接卷积网络(Densely Connected Convolutional Networks,DenseNet)的方法.DenseNet 最早由 Huang 等人[9]提出,并被应用到图像语义分割

领域，取得了一定的效果. 典型地，FC-DenseNet[10]继承了 FCN 的思想，并结合了 DenseNet；DenseASPP[11]将 DenseNet 与空洞卷积特征金字塔池化(Atrous Spatial Pyramid Pooling，ASPP)结合. 这类方法结合 ResNet[12]的跳跃连接思想，更加密集地连接不同的卷积网络层，提高了参数和特征的利用率，减少了参数量. 然而在模型训练时，频繁的跨层连接带来了更高的显存占用率和计算量；(4) 基于多尺度特征融合的方法. 这类方法结合不同层次的语义特征和不同区域的上下文信息，提高获得全局信息的能力. 典型的方法有 PSPNet[13]、deeplab 系列[4,14-16]. PSPNet 使用金字塔池化模块获取不同尺度的特征；deeplab 系列使用空洞卷积增大感受野，获取不同尺度的上下文信息，提高精度. 然而在不同尺度下，对象的细节可能丢失，对分割结果有影响；(5) 基于注意力机制的方法. Wang 等人[17]首先提出非局部(Nonlocal)的注意力机制，挖掘每个位置像素与全局像素点之间的联系. Chen 等人[18]提出 A^2-Net，从矩阵乘法角度进行优化，降低计算量. Li 等人[19]提出了金字塔型注意力网络(Pyramid Attention Network，PAN)，挖掘局部与全局像素之间的相似关系. 此类方法根据视觉注意特点，在不增加过多参数量的情况下，选择性地筛选有效的语义特征信息，利用局部与全局像素点的联系，提高模型的效率和准确性. 由于全监督学习方法，需要具有大量精确标注的训练数据集，获取成本较高. 为了解决这一问题，很多基于弱监督学习的图像分割方法进行了相关探索，利用图像类别、边框、涂鸦等弱标签信息，或者少量标注数据，进行模型训练，降低了对精确标注数据量的需求，并取得了一定的进展. 典型的弱监督学习图像分割方法可以分为以下几类：(1) 基于不同训练策略的方法. 第一种是基于多步训练的方法. Wei 等人[20]提出了简单到复杂(STC)框架的体系结构，先利用显著性检测的结果训练初始模型，再结合弱标注信息使用简单到复杂的图像迭代训练，增强模型的泛化能力. Shen 等人[21]还使用抓取的网络数据来进行多步学习，改善了训练数据量不足的问题. 这类方法利用弱标注信息得到初始网络，并通过不同层次的图像或相关的网络图像，来逐步强化网络，提供了有效的模型训练思路. 第二种是基于编解码结构训练的方法. Hong 等人[22]提出 DecoupledNet 网络，编码阶段使用大量图像级标注来训练，解码阶段使用少量像素级标注来训练，两个阶段通过桥接层连接. 这种方法利用不同的标注信息，分别训练网络各个组件，不用迭代循环训练，便于结构调整和扩展. 第三种是基于擦除策略训练的方法. Wei 等人[23]提出了对抗擦除策略(Adversarial Erasing，AE)的方法，使用不同擦除区域的图像训练，不断改变模型的注意力区域. Hou 等人[24]提出自擦除策略的网络(Self-Erasing Network，SeeNet)，不仅减少了工作量，而且在前景和背景之间预留潜在区域，一定程度上避免了前景向背景区域的扩张. 这类方法利用有限的数据，使用擦除的方式，让模型更加全面地关注对象的各个特征，减少模型对特定区域特征的依赖. 这类方法，充分利用弱标注信息，使用不同的训练策略，取得了一定的效果，但在挖掘像素点之间的联系上，还需进一步探索；(2) 基于语义传播的方法. Ahn 等人[25]提出 AffinityNet，预测像素间的语义关联，初始生成类激活图(Class Activation Map，CAM)，通过稀疏激活随机游走来实现语义传播. Huang 等人[26]提出种子区域扩张的方法，初始化生成分割结果，再通过区域扩张实现语义传播. 这类方法使用粗分割加语义传播的形式，在语义传播或扩张的阶段，探索了像素点间的语义特征联系，但这种探索比较局部；(3) 基于不同感受野的方法. Wei 等人[27]使用不同膨胀率的空洞卷积，可以扩大感受野，得到相应的注意力图，并提出一种简单有效的抗噪融合策略. Lee 等人[28]提出基于随机推理的网络模型 FickleNet，该模型通过随机选择隐藏单元，可以产生许多不同尺寸和形状的感受野，训练出分类器，再使用 Grad-CAM 方法[29]生成位置图，将多个位置图集成得到分割结果，并将其作为样本训练分割网络. 相对于语义传播的局部特征联系，空洞卷积可以挖掘更大感受野内的语义特征联系，融合不同尺度的细节信息，提高效果；(4) 基于生成式对抗网络(Generative Adversarial Network，GAN)的方法. Souly 等人[30]将这一思想应用到弱监督的语义分割中，通过生成器，结合图像级标注生成图像，再使用生成图像、图像级标注和像素级标注信息来训练判别器. 这类方法通过生成网络和判别网络对抗的形式训练，需要的训练数据较少，但相比于一般的网络，需要更多的显存，模型较不稳定；(5) 基于显著实例的方法. Li 等人[31]利用图像边框的弱标注信息，检测实例对象，再通过迭代训练得到各个实例的分割模型. Fan 等人[32]基于显著性实例分割方法 S4Net[33]得到图像中的显著实例，每个实例包含边界框和前景掩码；再通过注意力模块，使用实例的内在属性进行类别预测；同时，使用特征提取网络获

取每个实例的语义特征，在整个数据集范围内，构建实例相似图并划分，每个子图将决定这一类实例最终的类别.这类方法利用各个对象内在的实例属性，并探索了实例间的联系，提高了分割效果.这些典型的图像分割方法，可以挖掘视频的单帧语义信息，为视频对象分割技术的发展奠定了基础.

在视频对象分割领域，虽然已经有很多传统方法[34-36]取得了一定的效果，但随着硬件的提升，深度学习的方法吸引了很多研究者的关注.如图 1 所示，根据在待分割视频中给定的人工先验信息具体程度，可以分为半监督的方法、无监督的方法和交互式的方法.另外，在图像分割中，监督学习是针对整个训练数据集的标注形式来说的，即一般意义的监督；在视频对象分割中，监督学习是针对待分割视频给出的标注形式来说的，即通过相似视频训练后的模型是否还需要使用待分割视频的标注信息进行调整，二者有一定的区别.在视频对象分割领域，半监督方法的任务定义为：给出视频第一帧图像中感兴趣对象的详细人工真值标注，自动地分割出剩余所有视频序列图像中的感兴趣对象；无监督方法的任务定义为：不给任何人工标注信息，自动识别并分割出视频中的前景对象；交互式方法的任务定义为：在分割的过程中，通过人工交互式参与，结合粗略的人工标注先验信息，进行视频对象分割.第三类方法的条件相当于前两者的折中：相对于第一类方法的条件，它减少了注释的工作量；相对于第二类方法的条件，在视频的某几帧图像中，适当地添加了粗略的人工注释信息，从而在分割感兴趣前景对象中更具针对性.在三大类条件下，根据每类方法的处理特点又可以细分为若干种不同的子方法.

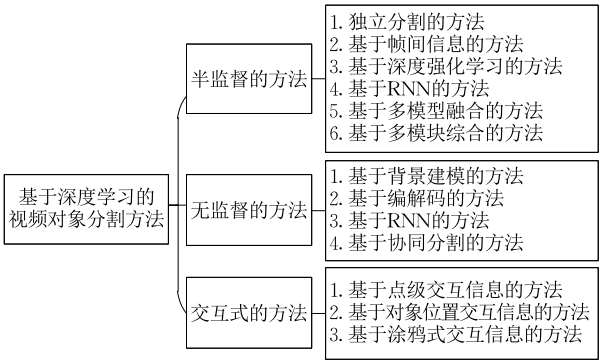


图 1 基于深度学习的视频对象分割方法分类

在视频序列中，随着时间的推移，视频不断变化，光照、视角、遮挡和图像噪声等因素为视频对象分割带来很大的挑战.虽然近几年视频对象分割领域的研究进展显著，但仍面临一些典型问题：

(1) 场景的空间复杂性

对于不同的视频序列来说，在不同的环境下，对象、背景、拍照条件等固有因素增加了场景的复杂性.在单帧图像内，存在运动模糊、相机抖动、光照不均、外观变化、对象的形变或遮挡等复杂情况，为对象空间特征的提取增加了难度，影响了对象的分割效果.

(2) 与时序信息的结合

视频序列的另一个要素就是时序信息，怎样将空间局部特征信息与时序信息相结合，是在视频序列图像之间建立信息传播机制的关键.目前的一些论文结果[37-38]体现了添加时序信息前后的差异性，然而如何基于已分割视频序列图像的信息，指导其他帧图像内的对象分割，还需要进一步的探索.对于一些视频中对象出现消失与重现的现象，如何提高模型的再识别能力，也是时序信息传播的关键.

(3) 对基础任务的依赖

很多的视频对象分割方法，将分割任务分为几个基础任务来分步处理.每个基础任务方法的性能，均在一定程度上影响着分割的结果.对每帧图像内的对象特征提取过程中，涉及到更加基础的图像分类、对象检测和静态图像分割等任务的处理过程.因此，结合视频对象分割的特点，对基础任务的解决方法进行相关的迁移与改进，是探索更好的视频对象分割方法的基础.

(4) 数据集问题

视频对象分割的最小单位是一段视频序列.对于模型训练来说，不仅需要很多的视频序列，还需要对视频的每一帧进行像素级的标注，为人工标注带来了巨大的工作量.近些年来，视频对象分割研究比较热，然而开放数据集的数量以及视频序列涵盖的场景相对有限.已存在的数据集在目标对象的个数、种类、外观变化、遮挡、运动形变等方面上仍然有所限制.真实世界中的场景复杂多样，这就需要更丰富的数据集来进行模型训练.

针对上述问题，越来越多的研究人员提出新的解决方法来优化视频对象分割的结果，同时也发布了一些新的数据集，使得数据集涵盖的场景也不断地丰富起来.本文首先介绍了目前常用的开放公共数据集，以及当前被广泛使用的分割准确率评估标准.其次，将视频对象分割方法分为三个大类：半监督的方法、无监督的方法和交互式的方法，逐个进行了详细地介绍.然后，统计并分析了各方法的实验结

果.最后,对文章进行总结,并对视频对象分割的未来研究方向进行了展望.

2　常用数据集与评估标准

2.1　常用数据集

视频对象分割方法需要一定数据集进行训练或者验证.因此,构建一定规模的数据集和评估标准,是视频对象分割的必然要求,不仅方便科研人员验证方法,也保证了视频对象分割研究的评价统一性.由于视频序列数据量很大,相应地增加了人工标注工作量.目前大部分的方法,采用的训练测试数据集主要以 DAVIS[39-40]、YouTube-Objects[41]、YouTube-VOS[42]和 SegTrack-v2[43]数据集为主,见表 1.这些数据集涵盖了各种各样的挑战,例如视频序列中的对象外观变化、被遮挡、运动模糊以及物体形变等.

表 1　视频对象分割方法的常用数据集对比

数据集名称	视频个数	标注总帧数	单帧图像对象个数
DAVIS2016[39]	50	3455	单个
DAVIS2017[40]	150	10474	多个
YouTube-Objects[41]	126	超过 20000	单个
YouTube-VOS[42]	3252	133886	多个
SegTrack-v2[43]	14	947	多个

DAVIS 数据集.该数据集被用于 DAVIS 视频对象分割挑战赛.该组织者提供了视频对象分割领域的测试基准,每年都发布带详细标注的数据集.2016 年该数据集第一次发布,只包含单个对象的视频对象分割任务,50 个视频序列,共 3455 帧详细标注图像[39].2017 年该组织者提供了 150 帧视频序列,共 10459 帧详细标注图,并提出了半监督的视频对象分割任务,在该任务中,给定待分割视频序列的首帧详细先验信息,要求输出剩余帧的多实例分割结果[40].2018 年在 DAVIS2017 年的数据集上,该组织者提出了半监督和交互式的视频对象分割任务[44].2019 年在同样的数据集上,还提出了无监督的视频对象分割任务.在该任务中,模型自动地分割出视频序列中的多个实例[45].随着比赛的开展,该数据集的数据量增加,视频对象分割任务的目标从单对象发展到多实例,序列中的对象分割难度也有所增加,如视频中出现个别实例的消失与重现、实例间遮挡更严重、运动形变程度更大等问题.在 DAVIS 数据集中,部分视频序列的单帧图像标注示例如图 2 所示.

图 2　DAVIS2017 数据集[40]示例
(第 1 列为原图,第 2 列为人工真值标注图)

YouTube-Objects 数据集.该数据集是视频对象分割领域比较常用的数据集之一,包含了 126 个视频片段、10 种类别的对象、超过 20000 帧的图像,每帧图像都进行了像素级的人工标注[41].

YouTube-VOS 数据集.2018 年,Xu 等人[42]发布了迄今为止最大数据量的数据集,一定程度上解决了领域内数据量的问题.该数据集包含了 3252 个视频片段、78 种类别对象、共 133886 帧人工像素级标注的图像.由于该数据集提出较晚,目前基于该数据集的实验测试还不是很多.

SegTrack-v2 数据集.该数据集包含了 14 个视频片段、24 种类别对象、共 947 帧图像[43].该数据集中,每一帧图像都做了像素级的标注,并覆盖了运动模糊、外观改变、复杂形变、遮挡、低速运动和对象铰接的场景.

2.2　评估标准

运用已经构建好的开放数据集,研究人员可以进行深度学习模型的训练和测试.同时需要建立一个统一的评估框架,方便评估方法的分割准确率.目前已经有一些研究者提出了比较通用的评估标准.在介绍这些评估标准之前,先说明两个通用的符号:G 是指人工标注的真值掩码,M 是指分割结果掩码.

(1)区域相似性

为了比较对象分割区域与人工标注的真值掩码的相似性,采用雅可比相似度 J 来表示算法的预测分割与真值之间的交并比.由于它不受图像的尺寸限制,就能直观反映出错误预测像素的占比,因此受到了广泛应用,也成为 DAVIS2016 挑战赛[39]的评

估标注之一.计算方式如下:

$$J=\frac{|M\cap G|}{|M\cup G|} \tag{1}$$

(2)轮廓精确度

轮廓精确度的计算公式,最早被 Galasso 等人[46]提出,后来在 DAVIS2016 挑战赛[39]上被用于轮廓的评价标准.从轮廓的角度看,视频对象分割掩码可看作是一系列封闭轮廓的集合 $c(M)$.通过计算轮廓精确度 F 测度,即对象轮廓的预测掩码 $c(M)$ 与真值 $c(G)$之间的准确率 P_c和召回率 R_c的函数,从而确定分割边界的准确率,计算方式如下:

$$F=\frac{2P_cR_c}{P_c+R_c} \tag{2}$$

这两种评估标准目前在文献中使用最多,同时它们也是 DAVIS 视频对象分割挑战赛中使用的两个指标.对于区域相似性的度量,它可以不受图像的限制,刻画分割结果与真值之间的像素重叠程度;而轮廓精确度的度量,则反映了对象边界的准确度和精细程度;二者分别从两个方面构建视频对象分割的评估指标,通常被用于算法性能评估.

3 半监督的方法

在视频对象分割任务中,基于半监督的方法主要解决的问题是,给定视频序列第一帧详细的人工标注先验信息,即对需要分割的对象进行像素级的标注,输出指定整个视频序列剩余帧的图像掩码.虽然传统的方法通过有针对性地联合多种图像特征,建立相应的模型进行视频对象分割,但分割效果有限[47].最近几年,在 DAVIS 挑战赛的推动下,基于深度学习的方法发展尤为迅速,研究者们以卷积神经网络(Convolutional Neural Network,CNN)为核心,不断提出新的模型并改进,在半监督的视频对象分割任务中取得了很好的效果.

本节将半监督的方法划分为以下六类:独立分割的方法、基于帧间信息的方法、基于深度强化学习的方法、基于 RNN 的方法、基于多模型融合的方法和基于多模块综合的方法.表 2 对这几类方法进行了比较和分析,其中,模型一般采用多步法的训练模式:首先,基于图像领域的相关数据集,训练基础特征提取模型;其次,在视频对象分割任务的训练集中,对基础模型进行再次预训练;最后,对待分割视频中,带有先验信息的第一帧图像进行增广,在增广的数据集上,对模型进行微调,使模型对当前视频的对象更具针对性.另外,有些方法可以在线微调模型:在分割的同时将已经分割的图像和对应的结果加入到训练集中,微调模型,迭代地分割整个视频序列.

表 2 半监督的视频对象分割方法对比

方法类别		代表算法	模型输入	模型训练	方法特点	优点	缺点
独立分割的方法	基于OSVOS的方法	OSVOS[38]	单帧图像	预训练,微调	将视频切分成图像帧,使用训练好的模型进行图像分割.	1.独立分割,相对来说,速度快. 2.不考虑上下文信息,一定程度上避免误差传播.	1.视频序列中的对象前后差异过大,分割效果差. 2.缺乏上下文信息,无法利用视频对象特征的时间一致性,分割效果受限.
	基于在线微调的方法	On-VOS[48]	单帧图像	预训练,在线微调	在 OSVOS 的基础上,不断地将分割结果和对应的图像加入到训练集,微调网络.		
基于帧间信息的方法	基于对象匹配的方法	FEELVOS[49]	第一帧分割图像、当前图像	预训练,微调	在一段视频序列中,虽然同一对象在各帧图像中有一定的变化,但具有相似性,通过建立模型,提取对象特征并匹配.	1.建立的模型能够挖掘图像之间的信息,指导剩余帧的分割结果. 2.在视频序列中,利用对象特征具有一定的相似性,提高分割效果. 3.相对于 OSVOS 来说,效果更好.	1.利用帧间的信息指导分割的同时,可能存在误差传播. 2.一般采用多分支的网络结构,模型复杂度增加. 3.没有再识别的功能,可能会丢失目标.
	基于光流特征的方法	MoNet[50]	相邻图像、光流图	预训练,微调	光流特征是重要的上下文信息,可以很好地描述对象的运动特征.		
	基于在线微调的方法	FAVOS[51]	当前图像,上一帧预测结果和光流图	预训练,在线微调	相对于独立分割中的在线微调方法,此类方法建立的在线微调模型,还能够结合上下文信息.		
	基于掩码传播的方法	MaskTrack[37]	上一帧和当前图像、光流图、上一帧掩码	预训练,微调	基于给定的首帧图像掩码,建立掩码传播模型,指导剩余帧的分割.		

（续　表）

方法类别	代表算法	模型输入	模型训练	方法特点	优点	缺点
基于深度强化学习的方法	RCA[52]	上一帧的分割掩码，当前图像	预训练，微调(可选)	依赖上下文信息，基于上一帧的掩码，得到对象在当前图像的初始位置边界框，再基于强化网络，指导并调整边界框的位置和大小，不断强化分割结果.	1. 将强化学习应用到视频对象分割，并取得一定效果. 2. 利用上下文中对象的位置信息.	1. 依赖于上下文中对象的位置信息，对象位置改变较大，会影响分割效果. 2. 每帧需要多步调整，耗时.
基于 RNN 的方法	文献[42]	当前图像、历史信息	预训练，微调	将 RNN 与 CNN 结合，利用 CNN 提取空间特征，RNN 提取序列特征，代表性结构：ConvLSTM、ConvGRU.	建立具有历史记忆的模型，递归处理历史信息，获取上下文信息.	1. 需记忆历史信息，占内存. 2. 递归处理过程存在误差传播.
基于多模型融合的方法	文献[53]	上一帧和当前图像、上一帧分割结果	预训练，微调	将待分割的实例对象进行分类，并对每个类别建立对应的分割模型，再将各对象实例的分割结果融合.	1. 结合上下文信息. 2. 对不同类别对象分别建立模型，具有针对性.	1. 多个模型训练时间长、消耗内存. 2. 增加对象分类过程，可能引入分类误差.
基于多模块综合的方法	PReMVOS[54]	上一帧、当前帧和下一帧的图像，对应的光流图	预训练，微调	模型包含对象检测、分割、融合上下文信息、对象再识别等模块，利用图像领域内更加基础任务的技术，综合完成视频对象分割.	1. 将视频对象分割任务分解，综合使用基础任务的先进方法. 2. 在结合上下文信息的基础上，使用再识别模块，避免丢失目标，改善分割效果.	1. 视频对象分割过程更加复杂，处理时间长. 2. 模型训练时间长.

3.1　独立分割的方法

在深度学习领域，单幅图像分割的发展更早一些，因此，很多基于单幅图像分割的方法，在一些开放数据集中取得了较好的效果[55]. 而视频序列的每一帧都是一幅图像，很直观的想法是独立地分割每一帧图像，将视频对象分割转化为单幅图像的分割问题，从而可以直接将图像分割的方法迁移过来. 在 DAVIS2016 挑战赛[39]上，OSVOS(One-Shot Video Object Segmentation)第一次被提出[38]，并取得了一定的效果，之后涌现了很多改进的方法. 这些方法，一般采用预训练、微调的训练方式，得到一个固定的模型，从而完成整个视频的对象分割. 另外还有一类方法使用在线微调的方法，在视频对象分割的同时，不断利用已分割的图像结果调整模型参数. 这类方法弥补了前者在微调模型时仅有第一帧的数据量缺乏的不足. 接下来，本节分别从这两个方面介绍相关方法.

3.1.1　基于 OSVOS 的方法

Caelles 等人[38]提出了 OSVOS 方法，该方法建立了基于 FCN[5]的双流结构，分别用于前景分割和图像的轮廓检测，通过将两个分支提取的特征融合，得到训练独立的图像语义分割结果. 该方法使用了多步法的模型训练方式：首先在 ImageNet 数据集上训练基础网络，再使用 DAVIS 训练集训练出“父网络”，最后基于第一帧注释数据微调网络，以得到最终的模型. 该模型对每一帧图像独立地处理，在视频对象分割中取得了一定的效果. Maninis 等人[56]提出了基于全卷积神经网络架构的语义 OSVOS 方法，建立了的实例分割和整体外观模型，将两个分支的分割结果融合得到最终的对象分割结果. 该方法采用的训练数据集和训练方式均与 OSVOS 相同，不考虑帧间信息而独立地分割每一帧图像，一定程度上避免了因物体遮挡或消失等因素带来的传播误差. 但后者采用了更细化的模型和处理方式，因而取得了更好的分割效果.

另外，在多个对象的视频对象分割任务中，很多方法也使用独立分割的思想，并基于 OSVOS 方法进行改进，从而取得了一定的效果. Sharir 等人[57]建立了对象检测和前景分割的双流网络，在对象检测分支中，使用基于 ResNet-101 主网络的 Faster-RCNN 结构[3]；在前景分割分支中，采用了基于 VGG16 结构[58]的全卷积神经网络；使用 IoU 阈值对候选区边界框和分割结果进行过滤，使用过滤后的结果，对前景分割分支的结果，进行再次过滤和增强，从而得到最终的分割结果. Cheng 等人[59]先训练一个基于 ResNet-101 结构[12]的通用模型用于前景和背景的分割，再基于这个通用模型对其进行微调，以学习实例分割模型；接着基于第一帧的先验信息进行数据增广，并微调网络；然后将只包含前景区域的图像，通过空间传播网络进行更精细地实例分割，最后通过连通区域感知滤波器来进一步区分实例边界. 该方法将多实例分割任务分解为前景分割

和实例分割，建立了更深层的双流分割网络，并通过后续步骤细化取得了更好的效果.

3.1.2 基于在线微调的方法

一般的基于 OSVOS 的方法，先通过预训练得到基础的分割模型，再基于第一帧的先验信息进行微调，使模型更加适应当前的视频序列. 由于这类方法仅有第一帧参与微调，训练的数据量太小而导致模型的针对性不强，分割效果不理想. 基于在线微调的方法利用模型进行视频对象分割的同时，将已经分割的图像加入到微调数据集中，不断调整模型，从而弥补了微调数据量太少的缺点，提高了视频对象分割效果. Voigtlaender 等人[48]提出了在线迭代微调的算法 On-AVOS，一定程度上提高了模型微调的数据量，克服了 OSVOS 方法不能适应对象在视频中随时间存在较大变化的缺点，在单个对象的视频对象分割任务上取得了不错的效果.

在 DAVIS2017 挑战赛[40]的针对于多个对象的视频对象分割任务上，基于 OSVOS 的在线微调方法取得了一定的效果. Newswanger 等人[60]将 OSVOS 模型作为父模型，然后在第一帧图像上进行微调，分割接下来的 10 帧，并将这 10 帧图像加入到训练集再次微调网络. 该方法使用核化相关滤波(Kernelized Correlation Filter，KCF)方法[61]得到跟踪的对象边界框，过滤掉边界框之外的粗分割信息，再结合轮廓信息得到细化分割. 然后，依次对每 10 帧分割序列进行迭代操作，从而实现在线微调网络模型. 其中，KCF 算法是基于方向梯度直方图(Histogram of Oriented Gradient，HOG)特征与核函数的方法，根据当前帧信息和上一帧信息训练出一个相关滤波器，然后与新输入的帧进行相关性计算，得到预测的跟踪结果. Voigtlaender 等人[62]对 On-AVOS 进行改进，使其适用于多个对象的视频对象分割任务. 相对于前者采用固定间隔来更新网络模型，后者基于每一帧的分割图像进行迭代调整，从而使模型调整的次数更多，训练得更充分，更适用于当前的视频对象分割.

3.2 基于帧间信息的方法

独立分割的方法将视频的每一帧图像单独处理，比较直观，虽取得了一定的分割效果，但缺乏对象在帧间的联系. 有些研究工作通过结合帧间的时序信息和对象的运动线索，指导后续视频图像的分割，以提高分割效果，该思想在后续的方法中被普遍采用. 基于帧间信息的视频对象分割方法，其一般框架如图 3 所示.

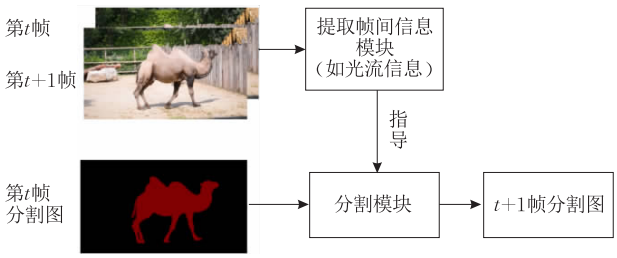


图 3 基于帧间信息的方法的一般框架

3.2.1 基于对象匹配的方法

在同一视频序列中，对象在各帧之间存在着一定的相似性. 利用深度学习的方法，提取图像的空间特征，并与具有先验信息的第一帧图像进行特征匹配，可以实现视频对象分割. Yoon 等人[63]提出了基于像素匹配的卷积神经网络方法：首先基于给定的第一帧详细先验知识，将人工标注的实例对象作为查询对象，并与当前帧的图像经过参数共享的双流网络来分别提取图像特征；然后将双流网络提取的多层图像特征连接起来并压缩，再经过全连接层进行特征融合，接着通过多层卷积结构增强对象相关性，最终得到每个像素的概率图，从而进行像素分类. 该方法借助查询匹配的思想，联合多层语义信息，使用两步法的模型训练方式，在 DAVIS16 数据集上取得了较好的效果.

除了单个对象的视频对象分割任务外，Yang 等人[64]提出了由分割网络、视觉调制器和空间调制器三个部分组成的模型，通过视觉调制器提取第一帧中标注对象的外观特征，结合空间调制器提取前一帧图像中对象位置的先验信息，对分割网络提取的当前图像特征进行增强，从而改进分割效果. Shaban 等人[65]使用改进的 OSVOS 算法和全卷积实例感知图像分割算法，对视频序列的每一帧图像生成特定的分割候选区，然后使用候选区跟踪算法实时跟踪候选区，并挑选出与第一帧标注中匹配的候选区，最后使用全连接的条件随机场(Conditional Random Field，CRF)进行更精细的分割. 该方法建立回溯机制以提高候选区跟踪的稳定性，使分割的结果更加细化. Voigtlaender 等人[49]提出了 FEELVOS 方法，能够在不依赖第一帧微调模型的情况下，简单、快速地实现视频对象分割. 该方法通过特征提取网络，提取每一帧图像的嵌入向量，通过局部和全局匹配机制，即分别与上一帧图像和具有先验信息的第一帧图像的嵌入向量进行匹配；再结合上一帧的分割结果，输出当前图像的掩码. 该方法使用更深的主网络结构，提取图像的嵌入向量，综合局部和全局信

息，从而提高了分割效果.

3.2.2　基于光流特征的方法

光流是运动图像分析的重要方法之一，是空间运动物体在观察成像平面上的像素运动的瞬时速度，是利用图像序列中像素在时间域上的变化以及相邻图像之间的相关性，来找到上一帧跟当前图像之间存在的对应关系，从而计算出相邻图像之间物体的运动信息. 在视频对象分割中，有研究者认为引入光流信息可以提高分割性能[66]. Tsai 等人[67]提出基于光流的视频对象分割算法，先进行对象检测和跟踪，在检测到对象的小区域内建立前背景的高斯混合模型(Gaussian Mixture Model，GMM)，并使用基于 VGG 网络提取的特征在线训练 SVM 模型，为像素赋予前景或背景标签；利用光流模型，将分割信息向后传播；为了弥补光流不能够精确处理对象边界的不足，独立计算分割对象的边界区域，并重新组合结果以得到对象边界清晰的光流，从而实现视频对象分割. 该方法利用 CNN 来提取特征，并使用传统方法进行分割，取得了一定的分割效果.

Jang 等人[68]构建了三叉卷积网络(Convolutional Trident Network，CTN)来进行视频对象分割：首先，基于上一帧的分割结果对前景对象进行粗定位，提取对象的 RGB 图像块，接着结合光流信息提取的前景图像块和背景图像块，使用 CTN 的编码模块进行编码，再分别通过 CTN 的三个解码模块进行解码，并基于马尔可夫随机场(Markov Random Field，MRF)的方法进行优化，得到仅包含对象的小区域分割结果再填补到原图中以完成对原图的分割. Xiao 等人[50]提出基于深度运动信息的视频对象分割网络. 该网络以相邻的三帧图像作为输入，含有两个分支结构，其中一个分支用来计算光流并提取运动信息，另一个分支用来对输入图像进行特征提取. 相邻图像的特征根据光流分支计算的结果进行对齐；并结合当前帧的图像特征得到新的特征图. 新的特征图通过两个独立的分支，分别融合运动信息，再合并得到最终的分割结果. 不同于 CTN 只利用前向计算的光流特征，该方法同时利用相邻图像计算前向和后向的光流图，双向指导视频对象分割，获得了更好的效果. 虽然上述方法使用光流模型来跟踪相邻图像中像素点的运动，在提取时序信息上取得了一定的效果，但是遇到物体位移大、遮挡、漂移等情况时效果欠佳.

3.2.3　基于在线微调的方法

类似于独立分割方法中的在线微调思想，一些方法通过建立基于帧间信息的分割模型，在分割的同时不断地迭代微调模型，不仅能够利用帧间的上下文信息改善分割效果，还可以使模型更加适合当前的视频序列. Petrosyan 等人[69]提出了边缘跟踪视频对象分割算法，分别运用 RCF 网络[70]、在线自适应视频对象分割和光流方法，进行视频对象的边缘检测，接着将三种边缘融合，并细化得到对象边缘对原图进行分割，再送到随机森林分类器对每个对象实例进行分类预测. 该方法将在线自适应分割方法与传统方法结合，先边缘检测再分割，分割结果在很大程度上受遮挡、光照不均、运动模糊等因素的影响.

Lin 等人[51]提出了光流自适应的视频对象分割方法. 先使用在线自适应的视频对象分割方法[62]对当前图像进行初步分割预测，接着基于相邻帧的光流信息计算光流预测结果，二者结合得到自适应的分割掩码，再通过在线迭代训练的自适应网络进行分割. 该方法基于第一帧和当前图像进行在线微调，虽然让模型在一定程度上适应了当前视频的变化，但微调的数据量还是较少. Guo 等人[71]提出了在线数据增广的方法，增加了模型在线微调的数据量，提高了模型对当前视频中对象和背景变化的适应性和分割效果.

3.2.4　基于掩码传播的方法

在同一视频序列中，对象在相邻的图像之间或很短的时刻内一般变化不大，同时深度学习在对象跟踪领域已取得较好的效果[72-74]，因此，很多方法采用像素级跟踪的思想学习掩码传播的模型，利用相邻图像的分割结果指导下一帧图像的分割，依次遍历整个视频序列. 最早，Perazzi 等人[37]提出掩码跟踪(Mask Track)方法，先在开放的语义分割数据集上离线训练基于 Deeplab-v2 的分割模型[4]，该模型接受当前图像和上一帧的分割结果掩码作为输入，然后基于视频的第一帧图像和标注信息，通过图像的翻转、镜像、裁剪、缩放等方式进行数据增广，并对训练的模型进行微调. 该方法建立了掩码向前传播的模型，开启了新的视频对象分割解决方法. Jampani 等人[75]提出双向的视频传播网络(VPN)提升了分割效果. 该网络结合了时序的双流网络和视频自适应滤波两个组件，并添加了空间网络结构来细化特征和提高网络的灵活性，从而融合了视频的时空信息. Hu 等人[76]提出基于编解码结构的级联细化网

络(CRN),使用 ResNet-101 作为编码阶段的主网络结构提取图像特征,在解码阶段接受光流图作为指导信息以改善分割结果.该方法使用了更深的网络结构,结合光流信息,在单个对象的视频对象分割任务上,取得了比前两个方法更好的结果.

类似于 Mask Track 方法,Suny 等人[77]基于更先进的 Deeplab-v3+语义分割网络[14],结合光流图,建立了视频对象分割模型,可实现多个对象的分割任务,取得了一定的效果.

很多半监督的方法为了使模型聚焦于当前视频的对象和背景,都基于第一帧图像的标注来微调网络;然而由于数据量十分有限,一些方法通过翻转、放缩、裁剪等方式进行数据增广并取得了一定的分割效果.后来,Khoreva 等人[78]提出了新的数据增广方式"lucid dreaming"训练掩码传播的模型,在 DAVIS2017 和 DAVIS2018 挑战赛的多实例视频对象分割任务上取得了较好的效果[79]."lucid dreaming"使用较少的数据量进行数据增广来训练模型,得到后续研究者的认可并被广泛采用.

3.3 基于深度强化学习的方法

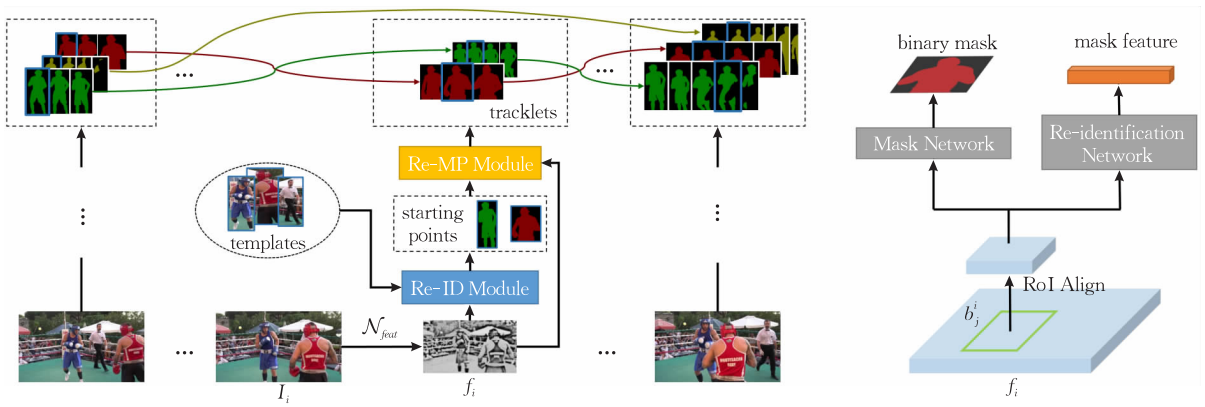
深度强化学习擅长于完成在控制与计算机视觉领域中存在的任务决策问题.由于视频图像中的对象位置信息有很强的上下文关联,Han 等人[52]将视频对象分割视为马尔可夫决策过程(Markov Decision Process,MDP),建立了基于深度强化学习(Deep Reinforcement Learning,DRL)的视频对象分割框架.先基于上一帧的掩码,获取对象在当前图像中调整的初始位置边界框,再通过强化学习网络,指导并调整边界框的位置和大小,检测对象位置,不断强化分割结果.该方法为深度强化学习应用到视频对象分割上做了一定的探索,取得了一定的分割效果,但该方法过于依赖上下文中对象的位置信息,不能很好地处理位置变化过大的情形.

3.4 基于 RNN 的方法

为了充分利用上下文信息,另一种思路是结合 RNN 方法,利用其递归处理历史信息和历史记忆建模的特点.Hu 等人[80]提出了一种递归神经网络 MaskRNN.首先,该方法基于上一帧的分割结果,预测出当前图像中的对象候选区信息,再结合相邻图像的光流信息以及当前图像,整合成网络的输入信息.其次,输入信息将经过两个深度网络结构,其中一个网络是分割网络,用来预测二值分割结果;另一个是定位网络,用来调整对象候选区的位置.最后,通过整合每个实例的输出结果得到最终的分割结果.由于当前的分割结果被用于指导下一帧图像的分割,从而实现递归的视频对象分割过程.该方法将 RNN 和 CNN 结合,更好地利用了视频的时间信息和位置先验信息,为半监督的视频对象分割任务提供了新的解决思路.

Xu 等人[42]结合循环神经网络的序列学习能力,提出了端到端的卷积长短期记忆(Convolutional Long Short-Term Memory, ConvLSTM)方法,挖掘视频序列图像之间的长期依赖关系和空间特征,从而完成视频对象分割的任务.Lattari 等人[81]也提出了将 VGG-16 结构[58]与 ConvLSTM 结构相结合的模型,而且在前者的基础上添加了视觉和空间调制器,分别提取基于第一帧先验信息的对象特征,以及基于上一帧的位置先验信息.该方法在单个对象的视频对象分割任务上取得更好的效果,另外,该方法也可以应用到多个对象的分割任务.

Li 等人[82]提出了基于注意力感知的 RNN 和再识别机制的方法,如图 4 所示,通过 RNN 获取上



(a) 视频对象分割的整体流程 (b) 再识别模块

图 4 带有再识别模块的视频对象分割流程图[82]

下文信息以提高分割的效果；利用再识别机制解决对象的大尺度变化问题，以避免对象丢失的情况.

3.5　基于多模型融合的方法

视频对象分割在实际应用场景中，往往存在不同种类的对象出现频率不一样的情形，而且不同类别对象的主要特征属性差别较大. 基于这些情况通过有区别地建立模型，可以针对性地分割对象，最后将各类分割模型的结果融合，可提高分割效果.

Le 等人[83]提出了用于视频对象分割的实例再识别流，先使用 Faster R-CNN 对人的实例进行定位，使用 DeepFlow 和可形变组件模型对非人的实例进行定位；在对象分割方面，采用多 SVM 分类器，结合已分割视频图像中的对象显著性、卷积特征、位置和颜色等特征，在各帧图像中对候选框内的对象进行分割，再通过拓扑排序筛选并融合分割结果. 之后，该团队的 Tran 等人[53]同样基于视频场景中人和非人的实例特征和数量区别较大的特点，有区别地建立分割模型，对于人的实例使用 Mask-RCNN 方法[84]进行更细致的分段分割，再将各分段的分割结果结合起来；对于非人的实例，使用该方法直接进行实例分割. 后者采用了实例分割领域更先进的实例分割方法，而且对人的实例对象进行更细致的分割，因而取得了更好的效果.

多模型融合的方法虽然取得了一定的分割效果，但对具体的视频特点有较高的要求，当视频中对象的类别只属于某一类时，其结果将与一般的分割方法没太大差别，而且存储多个模型，内存和训练时间都会很大.

3.6　基于多模块综合的方法

视频对象分割可看作是多个模块的综合问题，一些方法将视频对象分割任务分为检测、跟踪、分割和再识别等步骤，通过问题分解的思想，结合多个领域的先进方法，从而解决视频对象分割问题. Cheng 等人[85]提出了快速的视频对象分割方法：首先基于第一帧先验信息生成分段进行跟踪；再使用改进的 SegNet 网络[86]对当前的各分段跟踪结果进行分割；最后结合人工先验信息，融合得到分割结果. 该方法将待分割的完整对象生成各个分段进行跟踪和分割，显著地减小了待分割的区域，从而提高了速度，但该方法由于多环节的误差累积，分割效果有待提高.

Xu 等人[87]提出了通用的视频对象分割方法，可用来处理未知类型的对象. 该方法包含遮挡检测、对象检测、图像分割、以及结合上下文信息的 MRF 模型等步骤，一定程度上改善了遮挡的问题，但容易出现对象丢失的情况. Li 等人[88]在视频对象分割领域首次提出了自适应对象再识别模块 ReID，来防止对象丢失的情况，并结合掩码传播模块和光流图得到分割结果. Luiten 等人[54]提出了 PReMVOS 方法，先使用 Mask-RCNN 方法为视频的每一帧生成分割掩码及候选区，再使用基于 Deeplab-v3＋的细化网络为每个候选区生成精确的像素掩码，使用 ReID 模块计算候选区与第一帧对象掩码的相似度，并结合光流对第一帧的掩码进行传播，综合候选区的得分得到分割结果. 该方法使用多个不同功能的模块，综合各个模块的输出，避免了对象丢失的情况，并且分割边界更精细，在 DAVIS2018 挑战赛[44]的半监督视频对象分割任务中获得了第一名.

4　无监督的方法

虽然半监督方法在视频对象分割领域获得了较好的效果，但在分割的过程中需要结合待分割视频第一帧的详细人工先验信息，标注过程费时、费力. 无监督方法通过建立全自动的视频对象分割模型，可以改善这一问题. 其目标是，在不借助任何人工先验信息的情况下，算法自动发现视频中的前景对象，并在视频序列每一帧图像中分割出来. 本节将无监督方法主要划分为：基于背景建模的方法、基于帧间信息的方法和基于协同分割的方法. 表 3 对这几类方法进行了分析和比较. 在无监督方法中，由于没有先验信息的参与，大多数方法的模型在训练的过程中不再需要进行微调，缺乏对待分割视频的适应性，因而相对于半监督方法效果会有所下降.

表 3　无监督的视频对象分割方法对比

方法类别	代表算法	模型输入	方法特点	优点	缺点
基于背景建模的方法	文献[89]	生成的背景图像、当前图像	基于视频背景变化的特点，对背景进行建模，建立基于 CNN 的背景减法模型，进行分割.	1. 对视频背景进行分析建模，取得一定的效果. 2. 可以全自动地实时处理视频图像. 3. 模型训练无需基于第一帧信息进行微调.	1. 模型在处理背景变化较大、无规律运动、噪音较大等情况时，效果不好. 2. 只能处理单个对象的视频对象分割任务.

（续 表）

方法类别	代表算法	模型输入	方法特点	优点	缺点
基于编解码结构的方法	文献[90]	光流图、相邻图像	建立编解码结构的网络模型，在编码阶段，获取上下文的运动信息和对象的外观模型等，在解码阶段，基于编码特征，融合并恢复图像分辨率，得到分割结果.	1. 图像分割领域的先进方法经常采用编解码结构的模型，通过一定的迁移和改进，应用到视频对象分割领域，取得一定的效果. 2. 能够很好的获取上下文信息和对象的外观特征. 3. 模型训练无需基于先验信息微调.	1. 比较依赖于基于光流计算的上下文信息，会存在误差传播和累积的情况. 2. 视频前后差别较大、对象消失和重现等情况，会影响分割效果.
基于RNN的方法	文献[91]	相邻图像	建立相应的模型，提取图像的特征，通过 RNN 单元，结合历史信息.	1. 递归处理历史信息，充分利用上下文信息. 2. 模型训练无需基于先验信息微调.	1. 没有第一帧的先验信息，模型不能很好地聚焦于当前视频中对象. 2. 存在误差累积传播的问题.
基于协同分割的方法	文献[92]	具有公共对象的图像对	利用图像对中公共对象的特征相似性，同时分割两幅或多幅图像.	1. 利用公共对象的特征相似性，可同时分割出一组图像的公共对象. 2. 模型无需基于先验信息微调.	由于没有先验信息和上下文信息的参与，出现对象背景相似度太高、公共对象消失与重现等情况，都会影响分割效果.

4.1 基于背景建模的方法

当视频场景中背景的变化不是很大或者背景按照一定的规律变化时，背景具有可建模性，此时对背景建模，然后将视频的各帧图像与背景比较，就可以实现对逐个像素的前背景分类. Babaee 等人[89]先对视频序列进行粗分割生成粗略的背景模型，再结合运动检测进行调整得到细化的背景模型；对每帧图像和相应的背景，提取每个位置的图像块，使用训练好的卷积神经网络对像素点进行前景和背景的预测，并使用空间中值滤波和阈值化操作，完成视频对象的分割. 该方法虽然取得一定的效果，但仍存在明显的缺陷，如在视频的背景变化较大、光照不均、相机运动等情况下，该方法的分割效果会受到影响.

4.2 基于编解码结构的方法

很多无监督的方法通过建立编解码结构类型的网络，尽可能地学习相邻图像的运动模型和待分割对象的外观模型，实现了模型的全自动分割.

Cheng 等人[93]提出了一种端到端的网络 SegFlow，包含了使用 FlowNet 模型[94]的光流分支和全卷积神经网络分支，分别用于计算光流和图像特征提取，通过将二者结合以进行分割. Tokmakov 等人[95]介绍了 MP-Net 结构，以光流特征图作为输入，建立编解码网络来学习相邻图像的运动模式，实现了对视频对象的分割. Jain 等人[90]提出了端到端的学习框架，设计了基于 ResNet-101 模型的双流全卷积神经网络，接着分别学习图像的外观模型和基于光流的运动模型，将二者的输出融合，再经过解码得到最终的分割结果. 另外，该团队使用相同的双流编码结构，还探索了新的特征融合方式，改善了分割效果[96]. 相比之下，后面的三种方法通过预先计算光流图，进一步通过分支网络结构提取光流信息，相当于对光流信息使用更深层次网络结构进行提取，比直接使用相邻图像提取光流信息的方法效果更好；在此基础上，最后两种方法还学习了对象外观模型，联合运动和外观信息以提高分割效果.

这些基于编解码结构的方法，虽然在单个对象的无监督任务中取得了一定的效果，但其模型依赖光流信息挖掘对象的运动特征，因而遮挡、光照不均以及光流计算的误差都会使分割结果受到影响.

4.3 基于 RNN 的方法

类似于 3.3 节中介绍的方法思想，一些方法结合 RNN 方法可有效利用上下文信息，建立了无监督的视频对象分割模型. Tokmakov 等人[97]建立了具有视觉记忆的视频对象分割网络，该网络包含三个部分：基于 Deeplab-v1 的外观模型网络[16]、基于 CNN 的运动模型网络、基于卷积门控循环单元（Convolutional Gated Recurrent Unit，ConvGRU）的记忆模块. 前两者分别提取对象的外观和运动特征，之后经过记忆模块进行处理，并结合历史信息得到分割结果. 该方法无需第一帧先验信息，并通过提取对象的空间特征，并结合上下文信息，取得了一定的效果.

Xie 等人[98]将视频对象分割视为聚类问题，提出了一种新的像素轨迹递归神经网络，该网络可以学习前景像素轨迹的特征嵌入，然后利用所学习的特征嵌入，对像素轨迹进行聚类，建立了视频相邻图像间的前景对象掩码之间的对应关系.

Wang 等人[91]提出基于视觉注意机制的无监督方法，该方法先使用基于 ResNet-101 的网络结构提

取对象的卷积特征，再使用 ConvLSTM 建立的动态注意力模型来获取对象的位置信息，从而指导基于 FCN 的对象分割模块. 该方法使用了更深的 CNN 结构充分提取对象的特征，经过 RNN 单元融合历史信息，再使用 FCN 进行细化分割. 相对于前面两种方法，该方法融合了历史信息，并使用注意力机制进行特征增强从而提高了效果.

除了上述基于单个对象的无监督视频对象分割方法，Ventura 等人[99]首次针对多个对象的无监督任务进行了尝试，提出了 RVOS 方法. 该方法基于 ResNet-101 的编码结构提取 CNN 特征，在解码阶段的不同分辨率下，分别使用 ConvLSTM 结构，并结合历史信息得到最终的分割结果.

这类基于 RNN 的方法，充分利用视频的上下文信息，结合 CNN 的分割模型提取图像特征，不仅在单个对象的视频对象分割任务上取得了一定的效果，还对多个对象的视频对象分割任务进行了探索.

4.4　基于协同分割的方法

协同分割任务是指同时分割出含有公共语义对象的多幅图像. 一些方法也支持同时分割一组图像，为视频对象分割提供了新的途径. 我们建立了具有关联单元的孪生编解码网络结构[100]，以包含公共待分割对象的图像对作为输入，挖掘图像对之间的联系，当将视频序列拆分成一组图像时，该方法也提供了同时分割图像组的解决方案. Chen 等人[101]提出了基于语义感知注意力的深度对象协同分割方法. 该方法同样采用了基于 VGG16 的孪生编解码网络结构，使用注意力机制替换了关联单元来增强公共对象的特征，从而同时分割图像对. 相对于前者关联单元的计算，该注意力机制减少了计算量，可同时学习并增强多幅图像的公共对象特征，因而也适合于图像组的分割.

Lu 等人[92]提出了基于协同注意力机制的孪生网络结构，通过学习视频各帧图像之间的关联特征，来完成视频对象分割. 该网络是基于 DeepLabv3[15]的主结构，先提取两幅输入图像的特征，使用协同注意力模块增强相干特征，再经过分割模块，得到最终的分割结果.

上述三种方法实现了语义级协同分割任务，并取得了一定的成果. Hsu 等人[102]进行了更细化的探索，提出了实例级协同分割方法. 该方法将实例协同分割任务细分为两个子任务：协同峰值查找和实例掩码分割. 在前一个子任务中，该方法使用基于 FCN 的网络模型，并提出了图像协同峰值查找方法和相应的损失函数，来检测图像之间的相关区域. 在后一个子任务中，该方法使用无监督分割方法 MCG[103]来生成一系列候选实例区域，再综合前一子任务的检测得分，筛选得到协同分割结果.

这类基于协同分割的方法通过挖掘图像之间呈现的公共特征，为视频对象分割提供了新的解决思路，但图像中出现的公共对象丢失、遮挡等情况，会对分割结果带来一定的影响.

5　交互式的方法

基于交互式粗略人工先验信息的方法，在分割的过程中通过人工交互式参与，结合粗略的人工标注先验信息进行视频对象分割. 相对于半监督方法，该方法仅需要交互式添加粗略的标注信息，因而减少了标注工作量；相对于无监督方法，该方法可以利用交互式粗略先验信息指导视频对象的分割，能够让模型聚焦于视频的对象. 传统的交互式方法已经取得了一定的效果，如我们的早期研究工作[104]. 与之相比，由于标注数据量的缺乏，基于深度学习的交互式视频对象分割方法研究较晚. 近年来出现的视频对象交互分割方法中，粗略的数据标注主要有点、对象位置和涂鸦三种形式，本节将从这三个方面对交互式方法进行分类.

5.1　基于点级交互信息的方法

点是最简单的先验信息形式，在实际交互过程中可以很方便地给定，节省了交互时间. 通过点先验信息，可以指导算法分割出感兴趣对象.

Chen 等人[105]提出基于点的快速交互式分割方法，首先建立了基于 Deeplab-v2 的网络，提取图像的像素特征，然后基于用户点的输入，使用最近邻分类器对像素进行分类以得到分割结果. 该方法可以接受多种形式交互信息的输入，可以推广到处理半监督任务，其效果超越了大多数半监督方法，并在速度上有一定的提升.

5.2　基于对象位置交互信息的方法

对象位置通常以对象边界框的形式给出，标注成本较低，因而可节省人力物力. 通过对象位置的先验信息，指导模型聚焦于特定对象的区域，可避免视频背景中相似对象的干扰.

Ci 等人[106]建立端到端的可训练网络，基于给定的对象边界框信息，通过跟踪得到当前图像可能存在的对象边界框，然后将对象裁剪出来，通过基于 ResNet-101 和 ASPP 结构[15]的网络进行分割. 该方

法避免了视频背景的干扰,同时其提出的分割网络一定程度上改善了对象外观变化的问题.目前,在单个对象的视频对象分割任务中,该方法的效果接近最先进的半监督方法.

Wang 等人[107]提出了可同时实现视频目标跟踪和视频目标分割这两个任务的 SiamMask 模型.该网络是基于 ResNet-50 的孪生网络,接受两个输入:第一帧框选出的对象和待分割的当前图像,输出对象跟踪的边界框、得分和分割掩码;其中,使用基于通道的交叉关联操作,结合对象位置的先验信息,基于框选出的对象特征分割当前图像中的对象.该方法使用简单的网络结构,不需要微调,取得了不错的分割效果,而且在速度上有所提升.

5.3 基于涂鸦形式交互信息的方法

通过涂鸦的形式粗略地标识出感兴趣的对象,得到更多、更细化的人工先验信息,实现对视频中对象分割的交互式调整.

在 DAVIS2018 挑战赛[44]中,新增了交互式视频对象分割的任务.Najafi 等人[108]介绍了基于相似度学习的视频对象分割方法,该方法基于 ResNet-101 的 Deeplab-v1 模型,学习两个图像之间对应像素的相似性度量和相邻图像密集标签的变换,将对象标签从参考图像传播到视频的后续图像,该方法获得比赛的第二名.Oh 等人[109]提出了交互式视频对象分割的深度学习方法,基于 ResNet-50 模型,建立了交互网络和传播网络:在交互网络中,接收用户的输入信息和基于上一轮先验信息的图像分割结果,分割当前交互图像并提取图像特征;在传播网络中,结合上一轮先验信息的图像分割结果、上一帧掩码和交互网络提取的特征,指导视频各帧图像的分割.两个网络共同训练以相互适应,减少了交互分割过程之间的不稳定行为.该方法在比赛中获得了第一名的好成绩.基于涂鸦式先验信息的方法,以其简单的标注、节省人力的优点,取得了一定效果,但涂鸦式交互方法受到的关注较晚,分割性能与最先进的半监督方法还存在一定差距.

6 实验结果比较与分析

为了便于比较算法的效果,本节按照图 1 中的分类对半监督、无监督和交互式方法的基础网络、数据集和测试结果,进行统计和分析.其中,测试结果主要基于 2.2 节中介绍的两个评价标准.

6.1 实验结果统计

6.1.1 半监督方法的实验结果

表 4 对深度学习应用到各数据集的半监督任务中的实验结果进行统计,主要统计了基础网络、数据集及对应的分割结果.其中,需要注意的是,DAVIS2016、YouTube-Objects 和 SegTrack-v2 数据集相对统一,采用固定的验证集测试;而 DAVIS2017 由于比赛的关系,一般会区分为 DAVIS2017(val)、DAVIS2017(T-C)和 DAVIS2017(T-D),论文中未明确说明时,使用 DAVIS2017 表示.表中的基础网络主要包含:浅层 CNN、VGG[58]、FCN[5]、FC-DenseNet[10]、PSPNet[13]、ResNet[12]、Mask-RCNN[84]、Deeplab-v1[16]、Deeplab-v2[4]、Deeplab-v3[15]、Deeplab-v3+[14],以及使用到的循环神经网络结构主要包含:RNN[110]、GRU[111]和 LSTM[112].另外,本文统计了各方法的硬件环境和速度.硬件环境主要指使用的 CPU/GPU 个数和型号,其中第一个数字表示个数,当个数为 1 时省略不写;当未指明硬件环境时,填写为"—";当使用 GPU 但未指定具体型号时,填写为"GPU".速度是指处理视频中每帧图像的平均时间,单位是秒/帧,其中视频图像分辨率为 480p.

表 4 半监督的视频对象分割方法实验结果(其中"—"表示原文中未展示相关结果).

分类		文献	基础网络	硬件环境	速度/(秒/帧)	数据集	评估标准/%	
							J	F
独立分割的方法	基于 OSVOS 的方法	[38]	FCN	GPU	0.102	DAVIS2016	79.8	80.6
						YouTube-Objects	78.3	—
		[45]	FCN	NVIDIA Titan X GPU	4.5	DAVIS2016	85.6	87.5
						YouTube-Objects	83.2	—
		[57]	VGG16	—	10	DAVIS2016	80.1	—
						DAVIS2017(T-C)	49.7	—
		[59]	VGG16	NVIDIA Titan X GPU	10	DAVIS2017(T-C)	53.6	60.2
	基于在线微调的方法	[48]	FCN	NVIDIA Titan X GPU	9	DAVIS2016	85.7	—
						YouTube-Objects	77.4	—
		[60]	FCN	—	—	DAVIS2016	80.4	80.9
						DAVIS2017(T-C)	49.0	52.8
		[62]	ResNet-38	—	—	DAVIS2017(T-C)	54.8	60.5

（续　表）

分类		文献	基础网络	硬件环境	速度/(秒/帧)	数据集	评估标准/%	
							J	F
基于帧间信息的方法	基于对象匹配的方法	[63]	4 层 CNN	NVIDIA Titan X GPU	0.15	DAVIS2016	70.0	62.0
						SegTrack-v2	73.0	—
		[64]	VGG16	NVIDIA Quadro M6000 GPU	1	DAVIS2016	74.0	—
						YouTube-Objects	69.0	—
						DAVIS2017(val)	52.5	57.1
		[65]	FCN	—	—	DAVIS2017(T-C)	59.8	63.2
		[49]	Deeplab-v3+	16 Tesla P100 GPUs	0.45	DAVIS2016	81.1	82.2
						YouTube-Objects	82.1	—
						DAVIS2017(T-D)	55.2	60.5
	基于光流特征的方法	[66]	FCN	—	—	DAVIS2017(T-C)	50.7	57.1
		[67]	VGG16	Intel i7 CPU	—	SegTrack-v2	74.5	—
						YouTube-Objects	77.6	—
		[68]	VGG16	NVIDIA Titan X GPU	1.33	DAVIS2016	75.5	71.4
		[50]	Deeplab-v2	NVIDIA Titan X GPU	14.1	DAVIS16	82.0	85.5
						YouTube-Objects	81.7	—
						SegTrack-v2	72.4	—
	基于在线微调的方法	[69]	FCN	NVIDIA Titan X GPU	—	DAVIS2017(T-C)	56.7	61.1
		[51]	FCN	—	—	DAVIS2017(T-C)	58.4	62.9
		[71]	ResNet-101, VGG	—	—	DAVIS2017(T-C)	67.5	71.5
	基于掩码传播的方法	[37]	Deeplab-v2	—	—	DAVIS2016	80.3	—
						YouTube-Objects	71.7	—
						SegTrack-v2	67.4	—
		[77]	Deeplab-v3+	—	—	DAVIS2017(T-C)	57.7	62.4
		[76]	ResNet101	NVIDIA Titan X GPU	0.73	DAVIS2016	84.4	85.7
						YouTube-Objects	76.6	—
		[75]	Deeplab-v1	NVIDIA Titan X GPU	0.63	DAVIS2016	75.0	72.4
		[78]	Deeplab-v2	—	—	DAVIS2016	84.8	—
						YouTube-Objects	76.2	—
						SegTrack-v2	77.6	—
						DAVIS2017(T-C)	65.1	70.6
		[79]	Deeplab-v2	—	—	DAVIS2017(T-C)	65.1	70.6
基于深度强化学习的方法		[52]	FC-DenseNet	—	—	DAVIS2016	84.1	84.6
						YouTube-Objects	78.1	—
基于 RNN 的方法		[80]	VGG16,RNN	GPU	—	DAVIS2016	80.4	82.3
						DAVIS2017	60.5	—
						SegTrack v2	72.1	—
		[42]	VGG16, ConvLSTM	—	9	YouTube-VOS	66.9	74.1
						DAVIS2016	79.1	—
		[81]	VGG16, ConvLSTM	—	—	DAVIS2016	79.4	76.8
						DAVIS2017(T-C)	62.7	68.7
		[82]	ResNet-101	—	—	DAVIS2016	86.2	—
						SegTrack-v2	78.7	—
						YouTube-Objects	79.6	—
						DAVIS2017(T-D)	65.8	70.5
基于多模型综合的方法		[83]	PSPNet	—	—	DAVIS2017(T-C)	61.5	66.2
		[53]	Mask-RCNN	—	—	DAVIS2017(T-C)	64.1	68.6
基于多模块综合的方法		[85]	ResNet-101	NVIDIA Titan X GPU	0.6	DAVIS2016	82.4	79.5
						DAVIS2017(val)	54.6	61.8
		[87]	FCN	—	—	DAVIS2017(T-C)	66.9	72.5
		[88]	ResNet-101	—	—	DAVIS2017(T-C)	67.9	71.9
		[54]	Deeplab-v3+	NVIDIA GTX 1080 Ti GPU	—	DAVIS2016	84.9	88.6
						DAVIS2017(T-C)	71.0	78.4
						DAVIS2017(T-D)	67.7	76.1

单个对象的分割任务主要选用 DAVIS2016 数据集,多个对象的分割任务主要选用 DAVIS2017.很多文献也会选用 YouTube-Objects 和 SegTrack-v2 数据集进行测试,一般只统计区域相似性的实验结果.DAVIS2016、YouTube-Objects 和 SegTrack-v2 数据集提出较早,场景难度相对简单,各种方法的实验结果达到了一定的水平.在 DAVIS2017 数据集上,多个对象的分割任务由于难度较大,实验结果还有待提高.

6.1.2 无监督方法的实验结果

表 5 统计了无监督的视频对象分割方法的结果,其中数据集的说明同 6.1.1 节.可以看到,由于缺乏先验信息,给视频对象分割任务增添了很大难度,目前大多数的方法,主要基于单个对象的视频对象分割.相对于半监督任务来说,此类任务关注度有所降低,但在 DAVIS2016 数据集上,不使用第一帧详细先验信息的最先进方法,已经达到和半监督任务相差不到 10%的效果.而且无监督方法中,使用了 ResNet-101 或结合 RNN 的方法,其结果要明显高于其他的方法.此外,也有方法在多个对象的视频对象分割任务上进行了尝试[99].部分方法还采用了其他数据集,如 CDnet2014[113] 和基于 PASCAL VOC 自制的协同分割数据集[100].

表 5 无监督的视频对象分割方法实验结果(其中"—"表示原文中未展示相关结果)

分类	文献	基础网络	硬件环境	速度/(秒/帧)	数据集	评估标准/%	
						J	F
基于背景建模的方法	[89]	5 层 CNN	—	—	CDnet2014	—	75.12
基于编解码结构的方法	[93]	ResNet-101	NVIDIA Titan X GPU	7.9	DAVIS2016	67.4	66.7
	[95]	5 层 ResNet	GPU	—	DAVIS2016	69.7	66.3
	[90]	ResNet-101	—	—	DAVIS2016	71.51	—
					YouTube-Objects	68.57	—
					Segtrack-v2	61.40	—
	[96]	ResNet-101	GPU	—	DAVIS2016	72.82	—
					YouTube-Objects	68.95	—
					Segtrack-v2	64.44	—
基于 RNN 的方法	[97]	Deeplab-v1, ConvGRU	GPU	—	DAVIS2016	75.9	72.1
					SegTrack-v2	57.3	
	[98]	CNN,RNN	NVIDIA Titan X GPU	0.067	DAVIS2016	74.2	73.9
	[91]	ResNet101, ConvLSTM, FCN	—	—	DAVIS2016	79.7	77.4
					YouTube-Objects	69.7	
	[99]	ResNet101, ConvLSTM	Tesla P100 GPU	0.044	YouTube-VOS	44.7	45.0
					DAVIS2017(val)	23.0	29.9
基于协同分割的方法	[100]	VGG16	GPU	—	PASCAL VOC	64.5	—
	[101]	VGG16	—	—	PASCAL VOC	59.76	—
	[92]	Deeplab-v3	NVIDIA Titan X GPU	—	DAVIS2016	80.5	79.4
					YouTube-Objects	70.5	—
	[102]	FCN	—	—	PASCAL VOC	45.6	—

6.1.3 交互式方法的实验结果

表 6 统计了交互式方法的实验结果,DAVIS2018(interactive)表示在 DAVIS2018 挑战赛中,基于 DAVIS2017 的数据集制作了用于交互式对象分割任务的涂鸦式数据集,其中数据集的说明同 6.1.1 节.在单个对象的视频分割任务中,交互式的

表 6 交互式的视频对象分割方法实验结果(其中"—"表示原文中未展示相关结果)

分类	文献	基础网络	硬件环境	速度/(秒/帧)	数据集	评估标准/%	
						J	F
基于点级先验信息的方法	[105]	Deeplab-v2	—	—	DAVIS2016	77.4	79.3
基于对象位置先验信息的方法	[106]	ResNet-101	GPU	—	DAVIS2016	80.9	80.8
	[107]	ResNet-50	NVIDIA RTX 2080 GPU	0.017	DAVIS2016	71.7	67.8
					DAVIS2017(val)	54.3	58.5
					YouTube-VOS	60.2	58.2
基于涂鸦形式先验信息的方法	[108]	ResNet-101	GPU	—	DAVIS2018(interactive)	54.9	—
	[109]	ResNet-50	—	—	DAVIS2018(interactive)	64.1	—

方法可以利用粗略的先验信息，因此在 DAVIS2016 数据集上，其最先进方法的结果略好于无监督方法的最好结果. 多个对象的交互式分割任务方面，由于数据集缺乏，相关的比赛开展较晚，目前还未得到足够的重视.

6.2　结果分析

通过上述实验结果比较，针对于本文第 1 节中所提到的视频对象分割挑战，可以发现基于深度学习的解决方法主要呈现以下几个特点：

(1) 从基础网络角度来看，从最初的 FCN 和浅层 CNN，到目前普遍采用的 Deeplab-v3＋，实验结果一定程度上得益于更深的网络结构和计算机基础领域的先进方法. 对于挑战一来说，场景的空间复杂性具体表现为单幅图像背景复杂、光照不均、运动模糊、对象变化较大等情况，影响了提取对象特征，使用更深的网络结构可以更充分地提取图像特征，一定程度上使这一问题得到了改善.

(2) 对于挑战二，从实验结果上看利用图像间时序信息的方法比独立分割的方法要好，可见结合时序信息是必要的. 从方法的模型结构上看，RNN 与 CNN 结合逐渐成为新的趋势，能够较好地利用上下文信息，而且取得了不错的效果. 另外，光流特征用来获取相邻帧图像间的运动信息，也是各种先进方法中重要的支撑.

(3) 从评价标准的结果来看，为了追求更好的效果，模型越来越复杂，如越来越多的方法综合分类、检测、跟踪、分割等计算机视觉的基础模块. 通过将视频对象分割任务分解为多个子问题，并结合各子问题的特点，借鉴先进的方法，层层递进地解决综合问题，有助于改善挑战三的问题.

(4) 先验信息的详细程度对实验效果有一定影响. 无监督的方法、交互式的方法和半监督的方法，在模型分割的过程中，能够利用的先验信息详细程度依次增加，因而在单个对象的视频分割任务数据集 DAVIS2016 上，各类中最先进方法的效果依次提升.

(5) 目前的数据集已经提供了一定数量的像素级标注图像，而且最新提出的 YouTube-VOS 数据集提供了比较可观的数据量. 另外在实际训练过程中，为了解决数据量不足的问题，除了采用传统方式进行数据增广外，Khoreva 等人[78] 提出新的数据增广方式“lucid dreaming”被很多方法广为采用. 因此挑战四中数据集缺乏的问题得到一定的改善.

7　总　结

视频对象分割在三维重建、自动驾驶、视频剪辑等领域有着重要应用，但由于可供训练的数据集标注工作量较大，导致很长一段时间内发展缓慢. 在应用场景中，又面临着遮挡、形变、消失和重现等问题. 从给定人工先验信息具体程度的角度，本文将视频对象分割分为三大类：基于第一帧详细人工先验信息的方法、无先验信息的方法和基于交互式粗略人工先验信息的方法，并分别综述了其主流算法. 虽然这三大类方法在文中分开讨论，但基于视频序列的时空特性，在处理单帧图像的局部信息以及结合时序信息的细节上有很多相通的地方，因此，在各个子类方法上又有所交叉.

本文主要介绍视频对象分割的发展现状，对相关的研究方法进行了分析和对比，并总结了各类方法的优缺点. 进一步，我们详细地统计了各类方法的实验效果，对比分析了深度学习在视频对象分割领域的各项任务中的应用特点. 目前该领域还存在如下具有挑战性的研究方向：

(1) 无监督的视频对象分割任务

目前，基于第一帧详细先验信息的半监督分割任务涌现了不少的研究方法，而且取得了一定的效果，但在实际应用场景中，一般会缺少第一帧的详细先验信息. 针对这一情况，无监督的解决方案具有一定的潜力，基于统计的实验结果来看，无监督方法还有一定的提升空间. 其中，在单个对象的视频对象分割任务中已经取得一定的效果，但在多个对象的视频对象分割任务中，文献[99]进行了初步的尝试.

(2) 交互式的视频对象分割任务

在很多实际应用场景中，给定基于详细的第一帧先验信息比较困难，但借助于用户输入粗略先验信息的交互平台，实现视频对象分割的解决方案是可行的，具有较强的应用前景. 目前的交互式方法还不是很多，特别是在多个对象的交互式视频对象分割任务上还处于探索阶段，因而在这个方向上还存在很大的提升空间.

(3) 网络模型细节的改善

首先，随着对各个基础任务的研究逐步深入，视频对象分割任务在单帧图像分割问题上已经有一定的改善. 如图像的边缘检测[114]、图像分割[115-116]、对象检测[117-118]等任务的最新方法，都已经取得了较好的效果. 在未来的研究中，结合这些基础任务的先

进方法,并通过迁移学习可提高对复杂场景的鲁棒性.其次,根据 6.2 节的分析,在结合时序信息方向上,如何与 RNN 更好地结合,还需要进一步探索,而且当前先进的光流方法如文献[119],其性能还存在提升空间.

(4)提高模型的实时性

根据 6.2 节的实现结果分析,为了追求更好的效果,各类先进方法的模型越来越复杂(往往综合多个模块而成),实时性较差,而视频对象分割的实际应用场景对模型的处理速度有着较高的要求.在追求视频对象分割准确率的同时,提升方法的速度也是一个值得关注的研究点.事实上,部分较新的方法如文献[49,99,107]已经注意到模型处理速度的问题,但是可以发现其性能与相应领域最先进的方法相差较大.由于当前文献大都未进行算法复杂度分析,仅有部分文献给出了平均每帧的处理时间,据此,本文统计了相应方法的处理速度,相信在未来各个论文会越来越注重算法复杂度的分析与对比.

总体而言,本文首先描述了视频对象分割任务并归纳了现有研究中存在的普遍性问题.其次介绍了目前常用的具有代表性的训练数据集和研究方法的通用评估标准.然后基于视频对象分割任务的条件,将视频对象分割方法分为三大类:半监督、无监督和交互式的方法,并分别详细综述了其主流方法,接着对各类方法的实验结果进行统计分析.最后进行总结,并对未来研究方向进行了分析与展望.虽然深度学习在视频对象分割任务中,被越来越多的研究者关注,目前在各类任务上皆取得了明显的进展,但还有较大的提升空间,未来该领域相关的研究仍将是一个热门的方向.

致 谢 本文第一作者感谢在国外访学期间与 Adrian Hilton 教授间的学术交流与讨论.同时感谢评审专家的审稿意见和提出的建设性修改建议!

参 考 文 献

[1] Sikora T. The MPEG-4 video standard verification model. IEEE Transactions on Circuits and Systems for Video Technology, 1997, 7(1): 19-31

[2] Huang Kai-Qi, Ren Wei-Qiang, Tan Tie-Niu. A review on image object classification and detection. Chinese Journal of Computers, 2014, 37(6): 1225-1240(in Chinese)
(黄凯奇,任伟强,谭铁牛. 图像物体分类与检测算法综述. 计算机学报, 2014, 37(6): 1225-1240)

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[4] Chen L-C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on CVPR. Boston, USA, 2015: 3431-3440

[6] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on MICCAI. Munich, Germany, 2015: 234-241

[7] Badrinarayanan V, Handa A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:150507293, 2015

[8] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation//Proceedings of the IEEE ICCV. Santiago, Chile, 2015: 1520-1528

[9] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks//Proceedings of the IEEE Conference on CVPR. Honolulu, USA, 2017: 4700-4708

[10] Jegou S, Drozdzal M, Vazquez D, et al. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation//Proceedings of the IEEE Conference on CVPR Workshops. Honolulu, USA, 2017: 11-19

[11] Yang M, Yu K, Zhang C, et al. Denseaspp for semantic segmentation in street scenes//Proceedings of the IEEE Conference on CVPR. Salt Lake City, USA, 2018: 3684-3692

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[13] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, USA, 2017: 2881-2890

[14] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the European Conference on Computer Vision(ECCV). Munich, Germany, 2018: 273-284

[15] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:170605587. 2017

[16] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs//Proceedings of the International Conference on Learning Representations(ICLR). San Juan, Puerto Rico, 2016: 175-188

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

[18] Chen Y, Kalantidis Y, Li J, et al. A2-Nets: Double attention networks//Proceedings of the Neural Information Processing Systems (NIPS). Montréal, Canada, 2018: 1571-1581

[19] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation//Proceedings of the British Machine Vision Conference. Newcastle, UK, 2018: 244-251

[20] Wei Y, Liang X, Chen Y, et al. STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2314-2320

[21] Shen T, Lin G, Shen C, et al. Bootstrapping the performance of webly supervised semantic segmentation//Proceedings of the IEEE Conference on CVPR. Salt Lake City, USA, 2018: 1363-1371

[22] Hong S, Noh H, Han B. Decoupled deep neural network for semi-supervised semantic segmentation//Proceedings of the NIPS. Montréal, Canada, 2015: 1495-1503

[23] Wei Y, Feng J, Liang X, et al. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach//Proceedings of the IEEE Conference on CVPR. Honolulu, USA, 2017: 1568-1576

[24] Hou Q, Jiang P-T, Wei Y, et al. Self-erasing network for integral object attention//Proceedings of the Neural Information Processing Systems (NIPS). Montréal, Canada, 2018: 549-559

[25] Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation//Proceedings of the IEEE Conference on CVPR. Salt Lake City, USA, 2018: 4981-4990

[26] Huang Z, Wang X, Wang J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing//Proceedings of the IEEE Conference on CVPR. Salt Lake City, USA, 2018: 7014-7023

[27] Wei Y, Xiao H, Shi H, et al. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation//Proceedings of the IEEE Conference on CVPR. Salt Lake City, USA, 2018: 7268-7277

[28] Lee J, Kim E, Lee S, et al. FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5267-5276

[29] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 618-626

[30] Souly N, Spampinato C, Shah M. Semi supervised semantic segmentation using generative adversarial network//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 5688-5696

[31] Li Q, Arnab A, Torr P H S. Weakly- and semi-supervised panoptic segmentation//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 106-124

[32] Fan R, Hou Q, Cheng M-M, et al. Associating inter-image salient instances for weakly supervised semantic segmentation//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 367-383

[33] Fan R, Cheng M-M, Hou Q, et al. S4Net: Single stage salient-instance segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach California, USA, 2019: 6103-6112

[34] Li Zong-Min, Gong Xu-Chao, Liu Yu-Jie. Video object segmentation research based on features joint modeling. Chinese Journal of Computers, 2013, 36(11): 2356-2363(in Chinese)
(李宗民，公绪超，刘玉杰. 多特征联合建模的视频对象分割技术研究. 计算机学报，2013，36(11)：2356-2363)

[35] Sun Tao, Chen Kang-Rui. Video segmentation algorithm based on join weight of superpixels. Computers Science, 2016, 43(2): 302-306(in Chinese)
(孙焘，陈康睿. 基于超像素联接权模型的视频分割算法. 计算机科学，2016，43(2)：302-306)

[36] Chen H, Qian K, Wang B. Temporal coherent video segmentation with support vector machine and graph cut. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(8): 1389-1395(in Chinese)
(陈华榕，钱康来，王斌. 结合支持向量机和图割的视频分割. 计算机辅助设计与图形学学报，2017，29(8)：1389-1395)

[37] Perazzi F, Khoreva A, Benenson R, et al. Learning video object segmentation from static images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 3491-3500

[38] Caelles S, Maninis K K, Pont-Tuset J, et al. One-shot video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5320-5329

[39] Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 724-732

[40] Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 DAVIS challenge on video object segmentation. arXiv preprint arXiv: 170400675v3, 2017

[41] Jain S D, Grauman K. Supervoxel-consistent foreground propagation in video//Proceedings of the European Conference on Computer Vision (ECCV). Zurich, Switzerland, 2014: 656-671

[42] Xu N, Yang L, Fan Y, et al. YouTube-VOS: Sequence-to-sequence video object segmentation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 585-601

[43] Li F, Kim T, Humayun A, et al. Video segmentation by tracking many figure-ground segments//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Sydney, Australia, 2013: 2192-2199

[44] Pont-Tuset J, Caelles S, Perazzi F, et al. The 2018 DAVIS challenge on video object segmentation. arXiv preprint arXiv: 180300557, 2018

[45] Caelles S, Tuset J P, Perazzi F, et al. The 2019 DAVIS challenge on VOS: Unsupervised multi-object segmentation. arXiv preprint arXiv:190500737v1, 2019

[46] Galasso F, Nagaraja N S, Cardenas T J, et al. A unified video segmentation benchmark: Annotation, metrics and analysis//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Sydney, Australia, 2013: 3527-3534

[47] Wu X, Chen J, Wang M Y, et al. Shape prior based foreground segmentation with local rotation and structural changes//Proceedings of the 9th IEEE International Conference on Control and Automation (IEEE ICCA). Santiago, Chile, 2011: 1305-1310

[48] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation//Proceedings of the British Machine Vision Conference. London, UK, 2017: 383-395

[49] Voigtlaender P, Chai Y, Schroff F, et al. FEELVOS: Fast end-to-end embedding learning for video object segmentation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 9481-9490

[50] Xiao H, Feng J, Lin G, et al. MoNet: Deep motion exploitation for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 1140-1148

[51] Lin A, Chou Y, Martinez T. Flow adaptive video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 32-35

[52] Han J, Yang L, Zhang D, et al. Reinforcement cutting-agent learning for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 9080-9089

[53] Tran M T, That V T, Le T N, et al. Context-based instance segmentation in video sequences//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 28-31

[54] Luiten J, Voigtlaender P, Leibe B. PReMVOS: Proposal-generation, refinement and merging for the DAVIS challenge on video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 8-11

[55] Tian Xuan, Wang Liang, Ding Qi. Review of image semantic segmentation based on deep learning. Journal of Software, 2019, 30(2): 440-468(in Chinese)
(田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述. 软件学报, 2019, 30(2): 440-468)

[56] Maninis K K, Caelles S, Chen Y, et al. Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1515-1530

[57] Sharir G, Smolyansky E, Friedman I. Video object segmentation using tracked object proposals//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 46-51

[58] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations (ICLR). San Diego, USA, 2015: 7-15

[59] Cheng J, Liu S, Tsai Y H, et al. Learning to segment instances in videos with spatial propagation network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 31-36

[60] Newswanger A, Xu C. One-shot video object segmentation with iterative online fine-tuning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 41-45

[61] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596

[62] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for the 2017 DAVIS challenge on video object segmentation//Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 25-30

[63] Yoon J S, Rameau F, Kim J, et al. Pixel-level matching for video object segmentation using convolutional neural networks //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2186-2195

[64] Yang L, Wang Y, Xiong X, et al. Efficient video object segmentation via network modulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 6499-6507

[65] Shaban A, Firl A, Humayun A, et al. Multiple-instance video segmentation with sequence-specific object proposals//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 19-24

[66] Zhao H. Some promising ideas about multi-instance video segmentation//Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 37-40

[67] Tsai Y H, Yang M H, Black M J. Video segmentation via object flow//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3899-3908

[68] Jang W D, Kim C S. Online video object segmentation via convolutional trident network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 7474-7483

[69] Petrosyan V, Örnsberg O, Proutiere A. Video object segmentation via tracking edges and classifying segments//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 40-43

[70] Liu Y, Cheng M-M, Hu X, et al. Richer convolutional features for edge detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3000-3009

[71] Guo P, Zhang L, Zhang H, et al. Adaptive video object segmentation with online data generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 20-23

[72] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking//Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands, 2016: 472-488

[73] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking//Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands, 2016: 850-865

[74] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas Nevada, USA, 2016: 4293-4302

[75] Jampani V, Gadde R, Gehler P V. Video propagation networks //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 3154-3164

[76] Hu P, Wang G, Kong X, et al. Motion-guided cascaded refinement network for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 1400-1409

[77] Suny J, Yuy D, Li Y, et al. Mask propagation network for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 36-39

[78] Khoreva A, Benenson R, Ilg E, et al. Lucid data dreaming for object tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 375-387

[79] Khoreva A, Benenson R, Ilg E, et al. Lucid data dreaming for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 24-27

[80] Hu Y T, Huang J B, Schwing A G. MaskRNN: Instance level video object segmentation//Proceedings of the Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 10-21

[81] Lattari F, Ciccone M, Matteucci M, et al. ReConvNet: Video object segmentation with spatio-temporal features modulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 44-47

[82] Li X, Loy C C. Video object segmentation with joint re-identification and attention-aware mask propagation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 12-15

[83] Le T N, Nguyen K T, Nguyen-Phan M H, et al. Instance re-identification flow for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 13-18

[84] He K, Gkioxari G, Dollar P, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2980-2988

[85] Cheng J, Tsai Y H, Hung W C, et al. Fast and accurate online video object segmentation via tracking parts//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 7415-7424

[86] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495

[87] Xu S, Bao L, Zhou P. Class-agnostic video object segmentation without semantic re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 16-19

[88] Li X, Qi Y, Wang Z, et al. Video object segmentation with re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 1-6

[89] Babaee M, Dinh D T, Rigoll G. A deep convolutional neural network for background subtraction. Pattern Recognition, 2018, 76(4): 635-649

[90] Jain S D, Xiong B, Grauman K. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos//Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2126

[91] Wang W, Song H, Zhao S, et al. Learning unsupervised video object segmentation through visual attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach California, USA, 2019: 667-680

[92] Lu X, Wang W, Ma C, et al. See more, know more: Unsupervised video object segmentation with co-attention Siamese networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3623-3632

[93] Cheng J, Tsai Y H, Wang S, et al. SegFlow: Joint learning for video object segmentation and optical flow//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 686-695

[94] Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: Learning optical flow with convolutional networks//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 2758-2766

[95] Tokmakov P, Alahari K, Schmid C. Learning motion patterns in videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 531-539

[96] Xiong B, Jain S D, Grauman K. Pixel objectness learning to segment generic objects automatically in images and videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2677-2692

[97] Tokmakov P, Alahari K, Schmid C. Learning video object segmentation with visual memory//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 4481-4489

[98] Xie C, Xiang Y, Harchaoui Z, et al. Object discovery in videos as foreground motion clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 9994-10003

[99] Ventura C, Bellver M, Girbau A, et al. RVOS: End-to-end recurrent network for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): Long Beach, USA, 2019: 5277-5286

[100] Li W, Jafari O H, Rother C. Deep object co-segmentation//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 153-167

[101] Chen H, Huang Y, Nakayama H. Semantic aware attention based deep object co-segmentation//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 435-450

[102] Hsu K, Lin Y, Chuang Y. DeepCO3: Deep instance co-segmentation by co-peak search and co-saliency detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8846-8855

[103] Pont-Tuset J, Arbeláez P, Barron J T, et al. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 128-140

[104] Tian Yuan, Wang Cheng, Guan Tao. Interactive image segmentation method based on fuzzy c-means and graph cuts. Journal of Engineering Graphics, 2010, 31(2): 123-127(in Chinese)
(田元, 王乘, 管涛. 基于 FCM 和图割的交互式图像分割方法. 工程图学学报, 2010, 31(2): 123-127)

[105] Chen Y, Pont-Tuset J, Montes A, et al. Blazingly fast video object segmentation with pixel-wise metric learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 1189-1198

[106] Ci H, Wang C, Wang Y. Video object segmentation by learning location-sensitive embeddings//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 524-539

[107] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach California, USA, 2019: 1328-1338

[108] Najafi M, Kulharia V, Ajanthan T, et al. Similarity learning for dense label transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 14-27

[109] Oh S W, Lee J Y, Xu N, et al. Fast user-guided video object segmentation by deep networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 1-13

[110] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1529-1537

[111] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[112] SHI X, Chen Z, Wang H. Convolutional LSTM network: A machine learning approach for precipitation nowcasting//Proceedings of the Neural Information Processing Systems (NIPS). Montréal, Canada, 2015: 356-368

[113] Wang Y, Jodoin P-M, Porikli F, et al. CDnet 2014: An expanded change detection benchmark dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus Ohio, USA, 2014: 387-394

[114] Xie S, Tu Z. Holistically-nested edge detection//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1395-1403

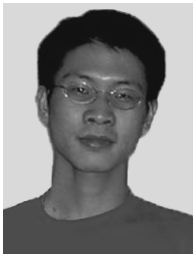
[115] Hu R, Dollar P, He K, et al. Learning to segment every thing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4233-4241

[116] Chen L-C, Hermans A, Papandreou G, et al. MaskLab instance segmentation by refining object detection with semantic and direction features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 4013-4022

[117] Li Z, Peng C, Yu G, et al. DetNet: A backbone network for object detection//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 334-350

[118] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv:180402767v1. 2018

[119] Ilg E, Mayer N, Saikia T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 1647-1655



CHEN Jia, Ph. D. His research interests include video processing, 3D motion capture, VR/AR and robot vision.

CHEN Ya-Song, M. S. candidate. His research interests include video object segmentation and computer vision.

LI Wei-Hao, Ph. D. candidate. His research interests include video object segmentation and computer vision.

TIAN Yuan, Ph. D. His current research interests focus on video processing and computer vision.

LIU Zhi, Ph. D., associate professor. His research interests include deep learning and artificial intelligence.

HE Ying, Ph. D. His research interests include graphic and image processing, robot vision.

Background

Video object segmentation is an important research topic in the fields of video image processing, analysis, computer vision, and so forth. As a key fundamental procedure, it can be widely utilized to quite a few applications, such as 3D reconstruction, autonomous driving, video editing, and so on. In the past few years, the performances of video object segmentation methods have been improved through the applications of deep learning. Many valuable strategies and methods are worth learning and studying deeply. That is why this paper reviews the research work and progress in this field in the past few years. Meanwhile, there are still some challenging problems that deserve great attentions, such as various complex scenes, better combination with temporal information, better use of algorithms of basic tasks and the generalization ability of model. To these problems, we make analyses and summaries for the existing methods and discuss the promising research directions in the future.

This review is a critical part of the NSFC project "Fusion of Video and Depth Data for Markerless 3D Human Body Motion Tracking (No. 61605054)" and "Behavior-Sentiment-Topic Joint Modeling in Multi-Scene Network Learning for Learners Interests Mining (No. 61702207)" the National Science and Technology Support Plan project "Research on Key Technologies of Science and Technology Museum Construction and Exhibition (No. 2015BAK33B02)" and Youth Academic Innovation Team project of Central China Normal University (No. CCNU19TD007). Video object segmentation is a key fundamental research task of our 3D human body motion tracking research project. At the meantime, this research topic is extremely vital to other projects. Without accurate video object segmentation, we cannot reconstruct the 3D human body Visual Hull model and then the accurate 3D human body motion tracking will not be possible, while the accurate 3D body-sense interactive technology will not be possible either, which is one of the key technologies for the exhibition project in science and technology museum.

Video object segmentation is a main research topic of our research group in Central China Normal University, Tsinghua University and the Heidelberg University. In the past few years, we have worked out some effective object segmentation methods, and have published a few of papers:

[47] Shape prior based foreground segmentation with local rotation and structural changes//Proceedings of the IEEE International Conference on Control and Automation, 2011

[100] Deep object co-segmentation//Proceedings of the Asian Conference on Computer Vision, 2018

[104] Interactive image segmentation method based on fuzzy c-means and graph cuts. Journal of Engineering Graphics, 2010.