软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.006821]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

联合吸收马尔可夫链和骨架映射的视频分割*

梁 云[1], 张宇晴[2], 郑晋图[1], 张 勇[2]

[1](华南农业大学 数学与信息学院, 广东 广州 510642)

[2](北京工业大学 信息学部, 北京 100124)

通信作者: 张勇, E-mail: zhangyong2010@bjut.edu.cn

摘 要: 因严重遮挡和剧烈形变等挑战长期共存, 精准鲁棒的视频分割已成为计算机视觉的热点之一. 构建联合吸收马尔可夫链和骨架映射的视频分割方法, 经由"预分割—后优化—再提升"逐步递进地生成精准目标轮廓. 预分割阶段, 基于孪生网络和区域生成网络获取目标感兴趣区域, 建立这些区域内超像素的吸收马尔可夫链, 计算出超像素的前景/背景标签. 吸收马尔可夫链可灵活有效的感知和传播目标特征, 能从复杂场景初步预分割出目标物体. 后优化阶段, 设计短期时空线索模型和长期时空线索模型, 以获取目标的短期变化规律和长期稳定特征, 进而优化超像素标签, 降低相似物体和噪声带来的误差. 再提升阶段, 为减少优化结果的边缘毛刺和不连贯, 基于超像素标签和位置, 提出前景骨架和背景骨架的自动生成算法, 并构建基于编解码的骨架映射网络, 以学习出像素级目标轮廓, 最终得到精准视频分割结果. 标准数据集的大量实验表明: 所提方法优于现有主流视频分割方法, 能够产生具有更高区域相似度和轮廓精准度的分割结果.

关键词: 视频分割; 吸收马尔可夫链; 长期/短期时空线索; 骨架映射网络

中图法分类号: TP391

中文引用格式: 梁云, 张宇晴, 郑晋图, 张勇. 联合吸收马尔可夫链和骨架映射的视频分割. 软件学报. http://www.jos.org.cn/1000-9825/6821.htm

英文引用格式: Liang Y, Zhang YQ, Zheng JT, Zhang Y. Video Segmentation with Absorbing Markov Chains and Skeleton Mapping. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/6821.htm

Video Segmentation with Absorbing Markov Chains and Skeleton Mapping

LIANG Yun[1], ZHANG Yu-Qing[2], ZHENG Jin-Tu[1], ZHANG Yong[2]

[1](College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China)

[2](Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: As challenges such as serious occlusions and deformations coexist, video segmentation with accurate robustness has become one of the hot topics in computer vision. This study proposes a video segmentation method with absorbing Markov chains and skeleton mapping, which progressively produces accurate object contours through the process of pre-segmentation—optimization—improvement. In the phase of pre-segmentation, based on the twin network and the region proposal network, the study obtains regions of interest for objects, constructs the absorbing Markov chains of superpixels in these regions, and calculates the labels of foreground/background of the superpixels. The absorbing Markov chains can perceive and propagate the object features flexibly and effectively and preliminarily pre-segment the target object from the complex scene. In the phase of optimization, the study designs the short-term and long-term spatial-temporal cue models to obtain the short-term variation and the long-term feature of the object, so as to optimize superpixel labels and reduce errors caused by similar objects and noise. In the phase of improvement, to reduce the artifacts and discontinuities of optimization results, this study proposes an automatic generation algorithm for foreground/background skeleton based on superpixel labels and positions

* 基金项目: 国家自然科学基金 (62072015, 61772209); 广东省科技计划 (2019A050510034); 广州市智慧农业重点实验室 (201902010081); 广州市重点研发计划 (202206010091)

收稿时间: 2022-04-02; 修改时间: 2022-07-26, 2022-09-09; 采用时间: 2022-10-21

and constructs a skeleton mapping network based on encoding and decoding, so as to learn the pixel-level object contour and finally obtain accurate video segmentation results. Many experiments on standard datasets show that the proposed method is superior to the existing mainstream video segmentation methods and can produce segmentation results with higher region similarity and contour accuracy.
Key words: video segmentation; absorbing Markov chains; long-term/short-term spatial-temporal cue; skeleton mapping network

视频分割旨在将视频序列中的运动目标与背景分离, 为识别与分析特定目标提供基础支撑. 它是视频语义解析、视觉导航等的关键技术, 在人工智能、视觉感知等领域广泛应用, 已成为计算机视觉的研究热点. 根据是否需要预先给出目标线索, 视频分割分为无监督方法和半监督方法. 本文主要针对半监督方法展开研究. 半监督方法在第 1 帧给出目标掩码, 然后通过匹配或传播等方式, 预测后续帧的目标轮廓.

传统的半监督视频分割方法大多基于光流估计[1−3]、马尔可夫随机场[4−6]等, 获取相邻帧像素间的时空一致性, 以实现视频分割. 当目标或背景剧烈变化时, 它们缺乏对场景的高级语义理解, 易导致分割错误. 后来, 人们用跟踪辅助视频分割[7−9], 通过跟踪目标[10]定位分割区域, 可降低噪声干扰, 提高分割鲁棒性. 但上述方法主要利用底层像素级或高层物体级线索, 缺乏对目标局部特征和轮廓边界的感知, 在多挑战共存时分割效果不理想. 中层视觉线索如超像素和超轨, 可准确描述目标局部细节和边缘特征, 已被应用于视频分割[11,12]、多目标跟踪[13,14]、语义分割[15]等领域, 在速度和精度方面带来提升. 因此, 本文以超像素为线索, 在帧间构建吸收马尔可夫链, 并引入跟踪辅助, 实现运动目标的预分割和优化处理, 为精准分割提供依据.

近来, 基于学习的半监督方法备受关注, 例如基于记忆网络的方法[16], 基于状态感知跟踪器的方法[7], 但因目标和背景复杂多变仍存在大量分割误差. 其中, 交互式视频分割方法[17,18]可根据用户期望, 在分割错误处通过交互提供优化线索, 能够有效降低误差, 分割精准度高. 这表明用户引导可有效校正分割中的错误. 但若要求用户按顺序逐帧勾画视频, 会严重影响分割效率, 难以处理批量视频分割和长视频分割. 因此, 本文基于不断变化的目标骨架建立引导, 在目标预分割和优化基础上, 实现视频分割精准度的再提升.

本文提出了联合吸收马尔可夫链和骨架映射的视频分割方法, 该方法包括“预分割—后优化—再提升”这 3 步. 首先, 基于跟踪确定目标感兴趣区域, 并对其超像素建立吸收马尔可夫链, 得到超像素的初始前景标签和背景标签, 实现预分割处理; 然后, 提出基于长期和短期时空线索的超像素标签优化算法, 完成后优化操作; 最后, 设计骨架映射网络, 通过骨架引导得到分割结果, 实现精准度的再提升. 主要贡献如下.

- 设计基于吸收马尔可夫链的目标“预分割”方法. 首先, 基于孪生网络和区域生成网络, 获取相邻两帧的目标位移和形变, 确定每帧的目标感兴趣区域; 然后将该区域分割成若干超像素, 并根据时空关系和随机游走模型建立吸收马尔可夫链, 得到超像素的初始前景标签和背景标签, 实现视频的目标预分割.
- 提出基于长期和短期时空线索的分割“后优化”算法. 首先设计短期时空线索模型, 并据此寻找目标短期变化规律和超像素空间分布规律, 校正误分割的孤立超像素; 然后, 构建长期时空线索模型, 并据此提出目标表观模型更新策略, 更新超像素标签优化分割效果, 降低相似物体的干扰.
- 构建基于骨架映射网络的视频分割“再提升”模型. 基于超像素标签和超像素的邻接关系, 设计前景骨架和背景骨架的自动生成模型; 并设计基于编解码的骨架映射网络, 将每帧目标的骨架作为引导注释信息, 和图像同时作为网络输入以指导目标轮廓学习, 进而得到精准的视频分割结果.

1 视频分割相关工作

当前主流的视频分割方法按照所需注释信息的不同, 主要可分为 3 大类, 即: 无监督的分割方法、半监督的分割方法、交互式的分割方法.

1.1 无监督的视频分割方法

该类方法不提供任何注释信息, 仅通过分析视频内容自动分离目标和背景, 并获取目标物体轮廓. 例如, Wang 等人[19]在整体时空域动态预测视觉注意力, 根据注意力强弱计算目标外观轮廓. Li 等人[20]构建实例嵌入网络, 计算每个像素点的嵌入向量, 通过找出属于目标的所有向量实现视频分割. Hu 等人[21]通过光流法和边界相似性计算

运动显著性, 进而识别和分割目标物体. Li 等人[22]基于双边网络判断非运动区域和稳定背景, 从而获取目标. 无监督视频分割的局限性是用户不能选择感兴趣目标, 当识别出目标的物体不符合用户期望时, 分割结果无任何应用价值.

1.2 半监督的视频分割方法

该类方法首先给出目标物体的掩码, 根据掩码标识可精准获取目标的颜色、形状、纹理等特征, 进而建立目标模型以实现视频分割. 按理论模型不同, 该类方法大致分为 3 种: ① 基于神经网络的分割方法; ② 基于目标整体运动信息的分割方法; ③ 基于中层视觉线索的分割方法.

(1) 基于神经网络的分割方法. 该类方法通过构造不同的网络结构和能量损失函数训练网络, 将视频像素分类为目标和背景. 例如, Caelles 等人[23]提出的 OSVOS, 首先运用多个视频分割数据集离线训练网络, 学习通用目标语义和粗略的前景分割, 然后用关键帧的准确掩码微调网络, 获取分割结果. Maninis 等人[24]在 OSVOS 的基础上增加了提取目标轮廓的分支, 提高了对目标细节的辨别能力, 可得到更鲁棒结果. Voigtlaender 等人[25]提出了 FEELVOS 网络, 使用单一卷积神经网络学习将像素嵌入向量的方法, 并基于局部匹配和全局匹配, 实现端到端的快速目标分割. Oh 等人[26]通过记忆网络匹配当前帧和已分割帧的像素, 获取当前帧的像素标签. 该类方法对网络结构和训练样本依赖性强, 将图像语义融合成深度特征, 对目标运动规律的考虑不足. 在多种挑战共存的复杂场景下, 该类方法易对网络提供错误引导, 导致分割欠佳.

(2) 基于整体运动信息的方法. 该类方法将目标跟踪嵌入分割网络, 利用跟踪引导分割网络感知目标大致区域. 例如, Zhou 等人[27]将检测、跟踪、分割融于同一网络, 在检测阶段选出多个目标候选框, 并对每个候选框生成粗略掩码; 在跟踪阶段嵌入高性能的 MDNet[28], 定位目标框; 在分割阶段将目标框和其粗略掩码输入网络, 学习出分割结果. Wang 等人[9]在 Siamese 跟踪网络[29]基础上, 增加多个细化模块以合并低分辨率和高分辨率特征实现目标像素级分割. Wen 等人[30]将部件级跟踪目标和视频分割集成到统一能量优化框架, 跟踪为分割提供连续运动信息, 分割为跟踪提供精确局部外观特征, 两者相互促进, 同时实现目标跟踪和分割. 这些方法通过跟踪为分割提供目标先验, 有利于网络应对剧烈变形、快速运动等挑战. 借鉴于此, 本文基于孪生网络和区域生成网络跟踪目标物体, 为构建基于吸收马尔可夫链的分割提供有效目标先验.

(3) 基于中层视觉线索的分割方法. 该类方法根据超像素、超轨等线索描述运动目标和构建视频分割方法. 例如, Wang 等人[31]基于光流计算目标像素轨迹, 并将轨迹聚类为若干超轨, 再根据高斯混合模型计算超轨中像素的目标概率. Yeo 等人[11]则根据相邻两帧超像素建立吸收马尔可夫链, 利用吸收时间确定超像素的前景和背景标签. Wang 等人[32]利用光流计算超像素的相似性, 并据此构建随机森林训练和学习超像素的分割标签. 这类方法可充分利用局部语义和传播规律, 通常不需要大量训练数据. 本文充分利用超像素和吸收马尔科夫链的优势, 构建目标的预分割网络, 为进一步精准分割目标提供依据.

1.3 交互式的视频分割方法

交互式视频分割方法, 允许用户手动标注和优化分割对象的目标边缘. 例如, Heo 等人[33]首先基于编解码结构, 根据用户涂鸦生成分割结果; 然后提出全局和局部传输模块, 将结果双向传输到其他帧. Oh 等人[17]通过交互和传播两个子网络实现视频分割, 交互子网络根据用户对关键帧的注释生成关键帧分割结果, 传播子网络根据该结果分割后续帧. Miao 等人[18]提出支持多轮交互的视频分割方法, 用户可不限帧数和次数的交互注释, 以优化分割结果. 该类方法由于可靠的用户注释更鲁棒, 但需要人工干预, 不适宜批量处理长视频. 但它们证明了注释可对分割网络提供有效引导. 本文根据预分割结果生成超像素标签, 构建目标和背景的自动骨架提取方法, 将该骨架视为注释输入骨架映射网络训练, 实现目标分割精准度的再提升.

2 本文方法

本文提出了一种联合吸收马尔可夫链和骨架映射的视频分割方法 (如图 1), 该方法主要由 4 步实现. 首先, 基于孪生网络和区域生成网络逐帧跟踪目标, 将跟踪结果向外扩展形成感兴趣区域, 并计算出该感兴趣区域的超像

素. 然后, 基于第 1 帧与当前帧的超像素, 以及上一帧与当前帧的超像素, 分别建立两条吸收马尔可夫链, 并根据吸收时间得到超像素标签, 实现视频的预分割. 接下来, 基于长期和短期时空线索的实现超像素标签优化, 解决相似物体干扰和分割不连续问题. 最后, 根据优化后分割结果和超像素线索, 自动生成前景骨架和背景骨架, 并将其作为注释信息输入骨架映射网络提升目标轮廓, 最终得到精准分割结果. 本文方法通过吸收马尔可夫链实现目标预分割, 通过长期和短期时空线索优化预分割结果, 通过骨架映射网络进一步提升分割结果, 最终形成了基于“预分割—后优化—再提升”的鲁棒视频分割新方法.

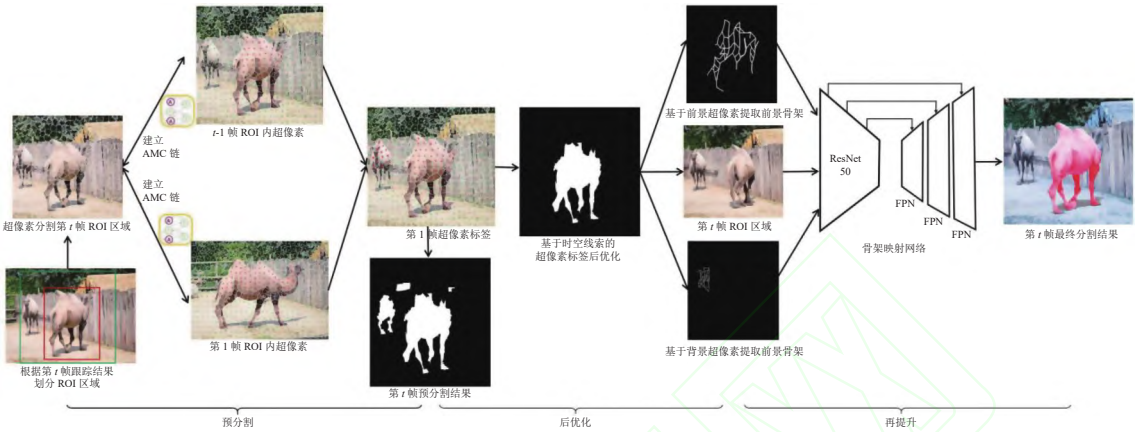


图 1 本文算法流程图

2.1 基于跟踪的目标感兴趣区域提取与超像素分割

获取每帧中目标的感兴趣区域及其超像素, 是本文视频分割方法的第 1 步. 本节介绍了如何基于目标跟踪, 实现目标感兴趣区域的提取与超像素分割. 首先, 使用文献 [34] 提出的基于孪生网络和区域生成网络的目标跟踪方法跟踪每帧的待分割目标; 然后, 扩展跟踪结果获取感兴趣区域 (region of interest, ROI); 最后, 采用文献 [35] 的 SLIC 超像素分割算法计算感兴趣区域的超像素. 本文采用的跟踪算法可快速准确的跟踪物体, 在遮挡和剧烈形变等挑战中仍有效. 超像素是由一系列位置相邻且颜色、亮度等特征相似的像素聚合成的不规则局部区域, 属于中层视觉线索. 与底层像素级线索相比, 它蕴含目标的局部细节和边缘特征. SLIC 算法在运行速度、像素紧凑度、轮廓保持方面都处于较佳水平.

本文将跟踪结果向外扩展 1.5 倍得到 ROI. ROI 包涵了目标及其紧邻背景, 在跟踪出现轻微错误时, 仍可将未被跟踪框覆盖的局部区域纳入待分割区域. 基于 SLIC 分割 ROI 时, 需指定第 1 帧超像素的分割个数, 在后续帧中, 为保证超像素大小均匀, 增加基于吸收马尔可夫链超像素标签预测的准确性, 本文根据 ROI 面积变化, 动态调整每帧的超像素个数, 调整策略如下:

$$k_n = k_{n-1} \times \frac{W_n \times H_n}{W_{n-1} \times H_{n-1}} \tag{1}$$

其中, k_n 代表第 n 帧的超像素个数, W_n 和 H_n 分别代表第 n 帧 ROI 的宽度和高度, $n-1$ 时情况类似.

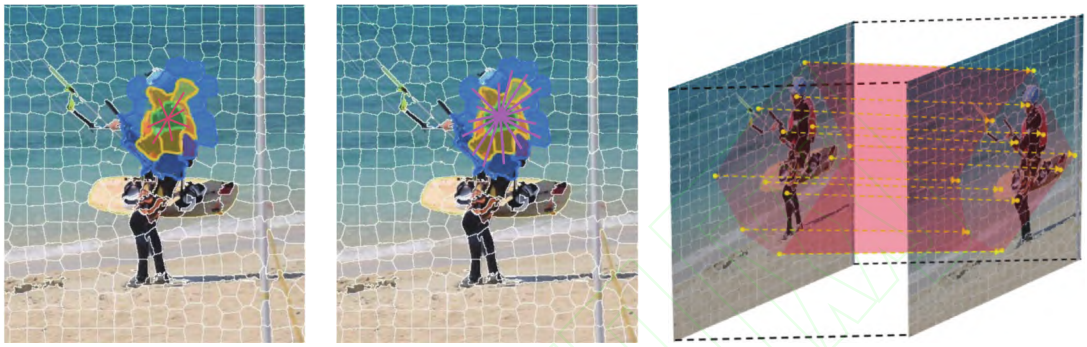
2.2 基于吸收马尔可夫链的视频预分割

得到每帧 ROI 的超像素后, 本节根据当前帧、第 1 帧、上一帧的超像素建立两条吸收马尔可夫链 (absorbing Markov chains, AMC), 进而得到当前帧超像素的吸收时间, 并据此计算超像素标签, 实现视频序列的预分割. AMC 是至少包含一种吸收状态的特殊马尔可夫链, 一条 AMC 上所有瞬态顶点, 都可以通过随机游走模型到达吸收态. AMC 能够灵活地分析和传播图像特征, 在图像匹配、分割等领域表现突出.

本文将每条 AMC 链记做 G, 令 $G=(V,E)$. V 为 G 的顶点, ROI 内每个超像素均视为 G 的一个顶点. 在 AMC 中, 所有顶点可被标记为吸收态顶点 (V^A) 或瞬态顶点 (V^N). 本文共构造了两条 AMC, 第 1 条 AMC 在第 1 帧和

当前帧间建立, 它将第 1 帧背景超像素标记为 V^A, 第 1 帧前景超像素和当前帧超像素标记为 V^N; 第 2 条 AMC 在上一帧和当前帧间建立, 它将上一帧的背景超像素同时标记为 V^A 和 V^N, 上一帧的前景超像素和当前帧超像素标记为 V^N. 上一帧背景超像素被同时标记为两种状态的顶点, 这一过程使 AMC 在双向的随机游走中, 反向校正上一帧中被误标记为背景的前景超像素.

每条 AMC 中, E 表示 G 内顶点形成的边, 本文将其分为帧内边和帧间边两类. 所有同一帧地一跳和两跳内相邻的超像素, 以帧内边相连. 假设图 2(a) 中绿色超像素为目标超像素. 一跳相邻的超像素被定位为: 目标超像素的所有紧邻超像素, 如图 2(a) 中的黄色超像素, 红色直线为一跳内相邻的帧内边. 两跳相邻的超像素被定位为: 目标超像素邻居的相邻超像素, 如图 2(b) 中蓝色超像素, 粉色直线为两跳内相邻的帧内边. 帧间边表示不同帧中超像素间的联系, 本文通过 EPPM 光流法[36]寻找两帧间超像素的对应关系, 并用帧间边连接, 它可快速匹配位置相近且颜色纹理特征相似的超像素, 如图 2(c) 中黄色虚线.



(a) 一跳内帧内边 (b) 两跳内帧内边 (c) 帧间边

图 2 AMC 中的帧内边和帧间边示意图

G 中所有边均被赋予权重, 预分割阶段根据每个顶点到达吸收态的时间计算超像素标签. 因此, G 中每个顶点的吸收时间必须有限, 即要保证所有瞬态顶点都可到达吸收态, 所以 G 所有与吸收态顶点相连的边都是单向的, 方向由瞬态顶点指向吸收态, G 中其余所有边都双向且权重相同. 边的权重代表边的两个超像素之间的相似性. 本文根据支持向量机 (support vector machine, SVM) 计算超像素特征距离, 学习超像素间的相似度, 得到边权重. 为避免目标轮廓漂移, 将提供准确的 SVM 训练数据, 本文把上一帧和第 1 帧的前景超像素作为前景数据集, 将它们的背景超像素作为背景数据集. 基于 SVM 计算边权重的过程如下.

将训练集记作 $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, x 为超像素样本的特征向量, n 为训练样本数量, y 为样本标签, 前景超像素标签为 1, 背景超像素标签为−1. SVM 使用高斯核函数作为核函数, 其定义如下:

$$k(x_i,x_j)=\left\langle \delta(x_i),\delta(x_j)\right\rangle=\exp\left(-\frac{\|x_i-x_j\|^2}{\sigma^2}\right) \tag{2}$$

其中, k 为高斯核函数, δ 为非线性特征映射函数, 将样本特征信息映射到更高维度, σ 为常数. 完成训练后, 对于每个输入样本 x_i, 都会得出其回归分数, 本文将 x_i 对应的回归分数记为 r_i, 其计算如下:

$$r_i=\langle w,\delta(x_i)\rangle \tag{3}$$

其中, w 表示支持向量机训练出的权重向量, 其维度与 $\delta(x_i)$ 维度相同. 在 AMC 中, 边权重由相连超像素的回归分数得出, 超像素 i 与超像素 j 间的边权重记作 w_{ij}, 计算如下:

$$w_{ij}=\exp\left(-\frac{|r_i-r_j|}{\varphi}\right) \tag{4}$$

其中, φ 为常量. 为得到当前帧每个超像素的吸收时间, 本文基于超像素间边的权重, 构造状态转移矩阵 P, 该矩阵由 Q 和 R 两个子矩阵构成, 其作用为增大前景超像素与背景超像素的区分度.

$$P=\begin{cases} Q_{ij}=\dfrac{\pi_t w_{ij}}{\sum_{f=1}^{N}\pi_{if}w_{if}} \\ R_{ij}=\dfrac{\pi_a w_{ik}}{\sum_{f=1}^{N}\pi_{if}w_{if}} \end{cases} \quad \text{s.t.}\begin{cases} v_i,v_j\in V^N \\ v_k\in V^A \end{cases}, \quad \pi_{if}=\begin{cases} \pi_t, & v_f\in V^N \\ \pi_a, & v_f\in V^A \end{cases} \tag{5}$$

其中, V^N 为 AMC 中的瞬态顶点, V^A 为吸收态顶点. π_t、π_a 为常数, N 为 v_i 的度. v_f 是与 v_i 有边相连的顶点, 为了保证从 V^A 开始的随机游走可以被快速吸收, 提高当前帧超像素吸收时间的区分度和标签预测的准确度, 需保证 $\pi_t<\pi_a$ 恒成立, 在本文中 π_t 值为 1, π_a 值为 5. 根据 P 和公式 (6) 计算超像素的吸收时间 F.

$$F=(I-P)^{-1} \tag{6}$$

其中, I 为单位矩阵, 其行数和列数与每条 AMC 中超像素个数相同, AMC 中第 i 个超像素的吸收时间, 即为矩阵 F 中第 i 行所有元素的和.

在当前帧和上一帧建立 AMC 后, 可得到当前帧所有超像素的吸收时间 H_1; 同理, 在当前帧和第 1 帧建立 AMC 后, 也得到当前帧所有超像素的吸收时间 H_2, H_1 和 H_2 均为 n 行 1 列的矩阵, n 为当前帧 ROI 内超像素数. 当前帧超像素的吸收时间 H 由 H_1 和 H_2 所得, 具体如下:

$$H=\begin{cases} H_1+\alpha_1 H_2, & \text{if } \left(Area_1<K_1\times Area_2 \cap Area_1>\dfrac{1}{K_1}\times Area_2\right) \\ H_1+\alpha_2 H_2, & \text{else if } \left(Area_1<K_2\times Area_2 \cap Area_1>\dfrac{1}{K_2}\times Area_2\right) \\ H_1+\alpha_3 H_2, & \text{else} \end{cases} \tag{7}$$

其中, α_1, α_2, α_3 表示 H_2 对当前帧超像素吸收时间的影响. $Area_1$ 表示当前帧 ROI 的面积, $Area_2$ 表示第 1 帧 ROI 的面积, 将 $Area$ 作为目标形变的定量评估依据, 本文实验中 $\alpha_1=1$、$\alpha_2=0.8$、$\alpha_3=0.4$、$K_1=2$、$K_2=5$. 得到当前帧所有超像素的吸收时间矩阵 H 后, 计算当前帧超像素的平均吸收时间 $\bar{H}$. 当前帧超像素中, 吸收时间小于 $\bar{H}$ 的标记为背景, 大于 $\bar{H}$ 的标记为前景.

2.3 基于长期和短期时空线索优化视频分割

根据第 2.2 节的吸收马尔可夫链, 可得到当前帧超像素的标签, 实现视频预分割. 该结果仅基于 AMC 的吸收原理和超像素分割结果, 缺乏目标整体性和运动一致性约束, 会产生标签计算错误. 为此, 本文提出了基于长期和短期时空线索的超像素标签优化算法. 其中, 基于短期时空线索, 可识别和纠正误分割为目标的局部背景; 基于长期时空线索, 可更新目标外观模型, 解决相似物体干扰问题.

2.3.1 基于短期时空线索的超像素标签优化

视频预分割时, 当背景与目标的颜色、纹理等特征非常相近时, 常导致背景超像素被错误地识别为前景. 但在视频相邻帧间, 目标的位置和形变一般变化不大. 据此, 本文引入短期时空线索, 优化误分割的超像素.

本文所提出的短期时空线索, 基于第 $k-t$ 帧到 $k-1$ 帧的分割结果实现, 用于校正第 k 帧分割, 具体如下:

$$T_k=R'_k-\sum_{i=1}^{t}R_{k-i} \tag{8}$$

其中, 矩阵 R_{k-i}, R'_k, T_k 的维度均为 $m\times n$, m 和 n 代表视频帧的长和宽, R'_k 和 R_{k-i} 均为 0-1 矩阵, 当像素标签为目标, 对应矩阵元素值为 1, 反之为 0. R'_k 为第 k 帧经第 3.2 节超像素标签预分类后的分割结果, R_{k-i} 为第 $k-i$ 帧经过预分割和后优化后的分割结果. 由公式 (8) 可知, 矩阵 T_k 中值为 1 的像素在基于马尔可夫链的预分割后被标记为前景, 而该像素在第 $k-t$ 帧到 $k-1$ 帧均为背景. 本文将矩阵 T_k 中所有值为 1 的像素归入集合 s, 并把 s 对应的超像素用集合 S 表示. 当 S 内某个超像素的一跳相邻超像素标签均为背景时, 表示它是孤立的且与目标其他部分不连续, 因此本文将其标签纠正为背景, 以实现超像素标签优化.

图 3 为引入短期时空线索的示意图. 在图 3(a) 中背景的假山区域出现一小块与熊颜色、纹理均相似的超像素时, 会将该背景的一部分超像素误分割为前景, 分割结果如图 3(c) 所示. 这块误分类的超像素对应的像素点在之

前 t 帧中标签均为前景, 且其一跳相邻的超像素标签均为背景. 所以根据短期时空线索, 将该超像素的前景标签更正为背景, 更正结果如图 3(d) 所示.

(a) 图像原图 (b) 真实轮廓 (c) 短期时空线索优化前 (d) 短期时空线索优化后

图 3 短期时空线索优化预分割结果

2.3.2 基于长期时空线索的超像素标签优化

当视频中出现与目标外观相近且相互靠近的物体时, 该干扰物常被误识别为目标 (如图 4(b) 中蓝色掩码的丹顶鹤). 仅依靠吸收马尔可夫链和短期时空线索, 在相似物体干扰时会产生分割误差. 相对而言, 目标的整体外观变化缓慢, 且具有相对稳定的运动规律. 由于本文解决的是单目标分割问题, 所以, 当预分割结果得到多个目标区域时, 相似物体带来分割错误. 为此, 本文提出基于长期时空线索的目标表观模型更新策略, 比较每个区域与目标表观模型的相似度, 将相似度最高记为前景, 其余为背景, 进一步优化超像素标签.

(a) 基于预分类超像素前景标签建立图 (b) 寻找图的连通区域 (c) 基于长期时空线索去除相似物体干扰

图 4 基于长期时空线索优化超像素前景标签

图 4 描述了基于长期时空线索的超像素前景标签优化过程. 首先, 根据分割结果的超像素构建图模型, 基于图的连通性, 计算已得结果的非连通区域数. 构建图时, 将所有前景超像素为图的顶点, 两跳内相邻的前景超像素为图的边. 图 4(a) 中红色区域为优化前的前景超像素, 共分成①②③这 3 个子区域. 子区域中黄实线、紫实线、蓝实线分别连接与橙色星号超像素、红色星号超像素、绿色星号超像素两跳内相邻的超像素. 由于区域①的橙色星号超像素和区域②的超像素在两跳内相邻, 所以区域①②合并成同一个连通区域. 而区域①③之间无两跳内相邻的前景超像素, 无法连通在一起. 因此基于图的连通性, 得到当前帧的两个候选目标区域 (图 4(b) 的红色区域和蓝色区域).

为识别准确分割目标, 本文基于已知分割结果的表观特征, 构造目标表观模型 Q_k, 并对当前帧的所有候选区域构造模型表达 T_i, $i \in n$, n 为当前帧候选区域数. 通过比较 Q_k 与 T_i 的相似性, 选取与 T_i 相似度最高者作为目标, 其余均为背景, 实现基于长期时空线索的分割优化. 本文基于归一化颜色直方图, 构造 Q_k 如下:

$$Q_k = R_{\text{learn}} \times \sum_{i=1}^{n} T_i + (1 - R_{\text{learn}}) \times Q_{k-1} \tag{9}$$

其中, $\sum_{i=1}^{n} T_i$ 为第 k 帧所有候选区域模型表达累加, Q_{k-1} 为基于第 1 帧到第 $k-1$ 帧的模型表达构造的表观模型, 即目标在第 $k-1$ 帧的长期表观模型, R_{learn} 为常数, 代表学习率, 本文中 R_{learn} 为 0.1. 比较图 4(c) 中红色区域和蓝色区域与外观模型的相似度, 发现红色区域最相似, 则其标签置为前景, 其余超像素均置为背景.

2.4 基于骨架映射网络提升视频分割

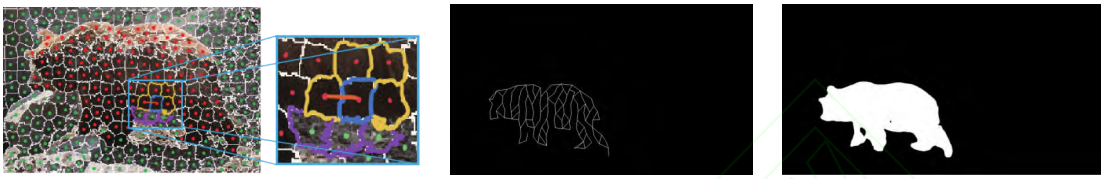
基于第 2.2 节和第 2.3 节, 可得到每帧超像素的标签, 组合标签为前景的超像素, 得到初步分割结果. 在多种挑战下, 初步分割能够粗略得到目标轮廓. 但超像素的不规则边缘会产生毛刺, 导致轮廓精确度降低; 同时少量存在的超像素标签错误, 会引入背景或丢失目标, 进而产生大量分割误差. 因此, 本文并不直接以初步分割结果作为最

终结果, 而是根据初步分割结果中的超像素标签线索, 设计骨架自动生成算法和骨架映射网络, 以骨架作为注释引导映射网络学习出精准的分割结果.

2.4.1 前景骨架和背景骨架的自动生成

在交互式视频分割中, 用户在关键帧感知并勾画目标的位置和形态. 以这些勾画作为引导, 可较精准地分割出关键帧的目标轮廓, 并在后续帧通过轮廓传播实现目标分割. 虽然有交互引导的分割效果较好, 但过多交互会导致人力和时间成本太高. 因此, 本文基于超像素标签自动生成前景骨架和背景骨架, 并将骨架代替交互式分割中的用户注释, 进而构建骨架传播网络, 实现视频初步分割的再优化.

本文基于前景超像素间的邻接关系提取目标前景骨架, 具体为: 对每个前景超像素一跳内的邻接超像素, 根据它们之间的空间位置关系, 按照上、下、左、右的顺序, 依次连接其邻接前景超像素, 每个超像素只发生一次连接, 最后形成连通骨架 (如图 5(b)). 前景骨架描述了目标的姿态、尺寸、位置等重要形态学线索.



(a) 前景骨架的构造过程 (b) 前景骨架构造图 (c) 基于前景估计的分割结果

图 5 构造前景骨架提升分割结果

图 5(a) 中前景超像素中心点为红色, 背景超像素中心点为绿色. 其中蓝色边缘超像素共有 7 个邻接超像素: 4 个标签为前景 (黄色边缘) 和 3 个标签为背景 (紫色边缘). 根据前景骨架的构造规律, 蓝色超像素连接位于其最上端的邻接目标超像素, 如图中橙色短直线所示. 根据该规律得到的整个目标的前景骨架如图 5(b). 将该骨架输入骨架映射网络后的分割结果见图 5(c).

基于第 2.2 节的马尔可夫吸收链和第 2.3 节的时空线索优化分割后, 本文能够大致识别属于目标的前景超像素, 如图 6 第 2 行红色点标记的超像素. 这些超像素通常可以覆盖目标的关键语义信息和局部细节, 而图 6 第 3 行所展示的基于分割结果自动生成的骨架图中包含这些有效信息 (骆驼的躯干和四肢、狗的基本形态、人的姿态特点).

背景骨架提取的目的在于进一步弱化相似物体和复杂背景的干扰. 本文在第 2.3.2 节中基于长期时空线索和图的连通性, 选取与目标表观模型最相似的连通区域保持其前景标签不变, 其他区域标记为前景的超像素, 其标签将被更改为背景. 本文标记所有被更改标签的超像素, 按照提取前景骨架相同的策略, 基于这些更改了标签的标记后超像素提取其背景骨架. 若某帧在基于长期时空线索优化标签前只有一个连通区域, 则该帧无背景骨架. 为区分前景骨架和背景骨架, 本文将前景骨架的灰度值设置为 1, 将背景骨架灰度值设置为 50.

2.4.2 基于骨架映射网络的视频分割

生成目标前景骨架和背景骨架后, 本文构建了骨架映射网络, 将当前帧、前景骨架、背景骨架作为网络输入, 通过训练学习, 输出当前帧的分割结果. 与第 2.2 节基于马尔可夫吸收链的预分割, 以及第 2.3 节基于时空线索的分割优化相比, 本节的骨架映射网络在前两者基础上, 实现了分割精准度的再提升.

本文的骨架映射网络采用典型的编码器-解码器 (encoder-decoder) 架构, 如图 7. 首先基于 ResNet50 设计编码器, ResNet50 是提取图像深度特征的常用结构, 其中跳跃式连接可有效避免网络退化和过拟合现象. 为处理骨架映射网络的多通道输入, 本文在编码器的第 1 个卷积层前加入一个辅助滤波器, 用于 ROI 对齐, 使网络关注到目标特征明显的感兴趣区域. 为将编码器得到的特征更有效的传入解码器, 本文去除了 ResNet50 最后的全局池化层和全连接层, 并基于跳跃连接将编码器内部获取中间特征传入解码器, 使解码器能充分利用不同深度和尺度的特征. 本文基于特征金字塔网络 (feature pyramid networks, FPN) 构造解码器, 特征金字塔网络可从多尺度, 多分辨率的融合特征中预测分割结果, 能够实现分割精度的最大化.

图 6　基于视频初步分割的前景骨架图

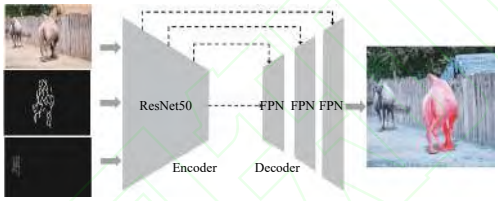


图 7　骨架映射网络结构图

图 7 展示了本文的骨架映射网络, 网络输入由当前帧、前景骨架、背景骨架组成. 3 个输入沿着通道维度串联, 形成 5 维输入张量. 其中, 待分割图像的 RGB 信息占 3 个维度; 前景骨架和背景骨架均为灰度图, 各占 1 个维度, 灰度值分别为 0 和 50. 网络的输出为与待分割图像大小相等的灰度图, 其值分布在 0–1 之间, 代表每个像素属于前景的概率. 概率大于 0.5 的像素标签为前景, 反之为背景. 训练网络模型前, 使用 ImageNet[37]预训练的权重进行初始化. 然后用 DAVIS[38], Youtube-2018[39]数据集训练网络. 为简化训练数据, 固定图像短轴为 480 像素, 并锁定纵横比. 使用交叉熵函数计算损失, 并将学习率固定为 1E–5.

3 实　验

3.1 算法实验配置

● 实验环境. 本文实验预分割、后优化两个阶段的运行环境为 Windows 10 64 位 PC 机, 16 GB 内存, Intel(R) Core(TM) i5-10300H CPU @ 2.50 GHz, 独立显卡为 NVIDIA GeForce GTX 1650, 算法开发平台为 Matlab 2015a. 再提升阶段的训练环境为 Ubuntu 18.04, CUDA 10.0, 独立显卡为 NVIDIA Tesla P100 16 GB, 算法开发平台为 Python 3.6.2, PyTorch 1.6.0.

● 实验数据集. 本文采用 Youtube-2018 数据集[38]、Davis 2016 数据集[39]和 Segtrack-v2 数据集[40]进行测试. Youtube-2018 数据集是目前数据量最多的视频分割数据集, 我们使用 Youtube-2018 验证集进行测试, 该验证集共包含 474 个视频序列, 覆盖 91 种不同类别的对象. Davis 2016 共涵盖 50 个不同的单目标视频序列, 共计 3 455 个

标注帧, 视频帧率为 24 fps, 分辨率有 480p 和 1 080p 两种可选, 本文选用 480p. Segtrack-v2 共涵盖 14 个视频共 24 个目标, 共计 976 帧, 当同一视频出现多个目标时逐个处理. 上述 3 个数据集涵盖遮挡、光照变化、画面模糊、快速移动等复杂场景. 在上述 3 个数据集中, 本文与近年主流的多种视频分割方法进行了定量和定性分析, 验证了本文方法的有效性. 同时, 开展了消融实验, 验证本文方法的鲁棒性和可靠性.

3.2 定性分析评估

本节将本文所述分割方法与当前 3 种视频分割方法 (AMCT[14], FCVOS[41], FTMU[42]) 在 Davis 2016 数据集进行定性比较. 图 8 展示了 Davis 2016 数据集内 5 个有挑战的视频序列 (breakdance、bus、bmx-trees、dance-jump、horsejump-low) 的定性比较结果, 由图 8 可知本文算法基本取得了较好的分割效果.

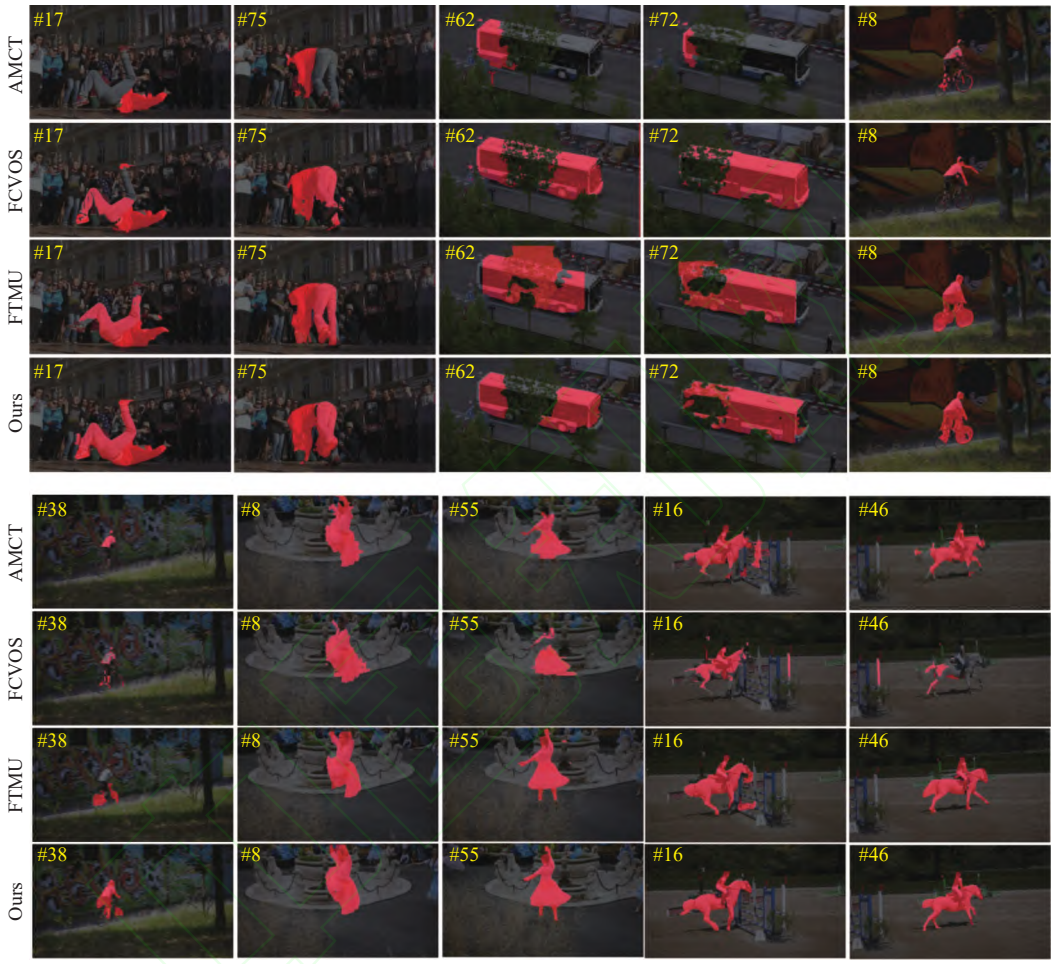


图 8 Davis 2016 数据集的定性比较结果

图 8 上半部分的第 1 列和第 2 列为“breakdance”数据集第 17 帧和第 75 帧的分割结果对比, “breakdance”中背景内包含多个与目标外观相似的人, 且目标的运动幅度在帧间变化较大. AMCT 方法基于超像素的颜色特征和位置信息实现分割, 在人腿部形变剧烈时分割失败. FCVOS 方法基于最小图割法实现像素的分类, 未考虑目标整体信息, 导致分割结果不连续, 误差较大. FTMU 方法基于强化学习实现帧间目标匹配, 其分割结果与本文相似, 精准度较高. 图 8 的第 3 列和第 4 列为“bus”数据集第 62 帧和 72 帧的分割结果对比. 该数据集中随着公交车的驾驶, 树叶会不断遮挡公交车的不同部位, 形成分割挑战. AMCT 方法仅基于上一帧的结果分割当前帧, 会在树叶遮挡时造成错误累计, 最终导致分割失败. FCVOS 方法与 FTMU 方法均分割出了公交车的大致轮廓, FCVOS 方法和

本文的分割结果相当, FTMU 方法由于在强化学习中逐渐将树叶背景学习为目标, 结果较差. 图 8 上半部分第 5 列和下半部分第 1 列为“bmx-trees”数据集第 8 帧和第 38 帧的分割结果对比. 该数据集中背景剧烈变化, 且骑车的人常被树枝遮挡, 分割挑战较大. AMCT 方法和 FCVOS 方法分别基于吸收马尔可夫链和图割法实现分割, 由于自行车部分的细节特征较多, 导致仅准确分割出骑车人的躯干部分. FTMU 方法在视频的前几帧准确学习了人和自行车的整体特征, 后续帧中随着干扰增多, 逐渐将人当成了背景. 对比可知, 本文的分割效果较稳定. 图 8 下半部分的 2、3 列和 4、5 列分别为“dance-jump”和“horsejump-low”的分割结果对比, 4 种方法均分割出了目标的大致轮廓, 但本文所提出的基于“预分割—后优化—再提升”分割方法, 在目标边缘和细节区域的分割中, 取得了和其他 3 种分割方法相比更鲁棒的分割效果.

本文与当前 2 种分割方法 (FRTM[43], STM[44]) 在 Youtube-2018 验证集进行定性比较. 图 9 展示了在 Youtube-2018 验证集中 4 个有挑战的视频序列的定性比较结果. FRTM 由目标建模和目标分割两部分组成, 该方法使用第 1 帧的真实分割结果构建目标外观模型, 使用上一帧的分割结果获得当前帧的精细分割结果. STM 方法首次将记忆网络用于视频分割, 并获得了较佳的分割效果. 该方法首先基于记忆网络读取并存储之前帧的目标特征, 然后基于非局部的注意力机制匹配当前帧的目标. 图 9 上半部分第 1、2 列为在镜头剧烈晃动挑战下的对盥洗室多个目标的分割结果. 由于镜子 (绿色掩码目标) 在第 1 帧未完整出现, 且在帧间的变化较为明显, 导致 FRTM 和 STM 算法均未能将其完整分割, 我们的方法若仅依赖基于吸收马尔可夫链的超像素分割结果, 也不能完整地分割出镜子的准确轮廓. 但基于吸收马尔科夫链的分割结果可以转化为镜子的骨架, 该信息能引导骨架映射网络分割出镜子的完整轮廓. 图 9 上半部分第 3、4 列为相似物体挑战下对两个滑板和滑板上两人, 共 4 个目标的分割结果. 仅有我们的方法能在双人遮挡现象严重时准确地区分每个目标 (第 3 列), 且即使双人遮挡现象不严重, FRTM 算法也不能很好地区分两人 (第 4 列). 但我们的方法和上述两种方法均不能非常精准地区分外观几乎完全相同的两个滑板.

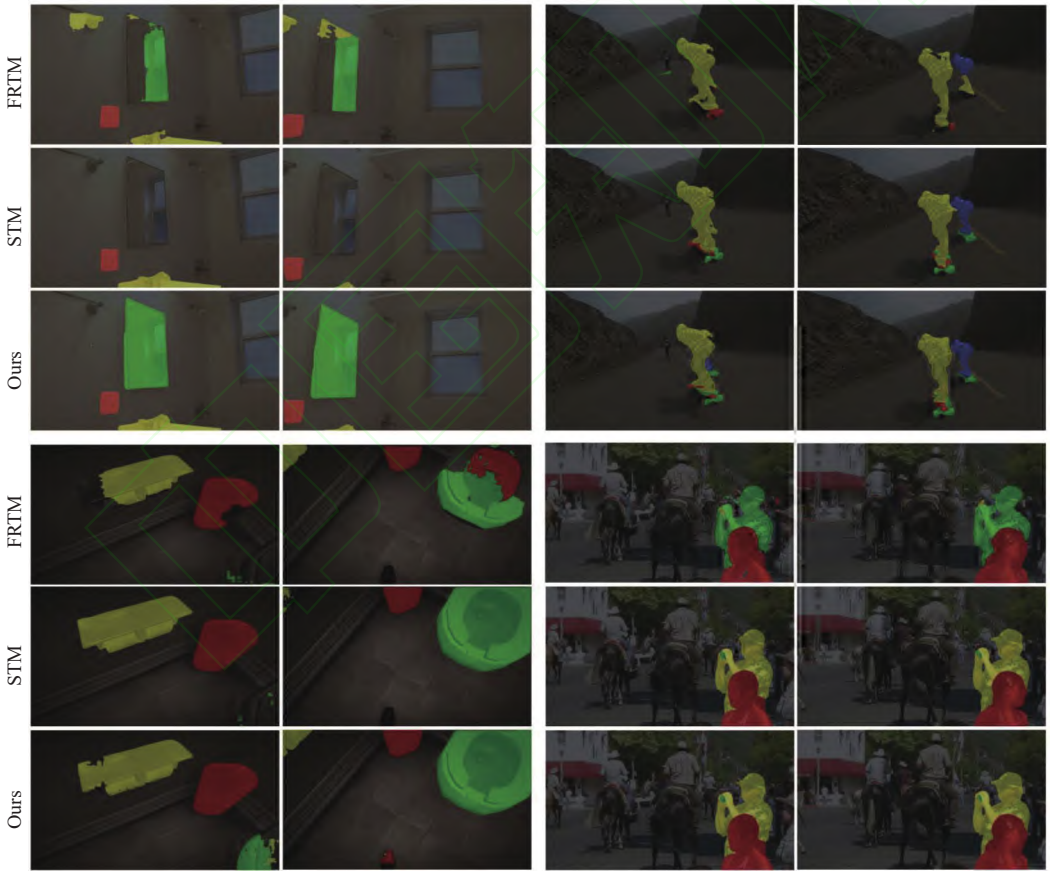


图 9 Youtube-2018 数据集的定性比较结果

图 9 下半部分为两个视频主要展示对小物体的分割效果, 仅有我们的方法能成功地分割出位于视频左下角的马桶 (绿色掩码, 图 9 下半部分第 1 列). 相比之下, FRTM 方法对于小物体的分割能力相对最弱, 而我们的方法和 STM 方法均能分割出拍照人手持的相机 (绿色掩码, 图 9 下半部分 3、4 列). 我们的方法对于小物体的辨别能力主要得益于我们构造的两条吸收马尔可夫链, 能较稳定有效地传播小物体的显著特征.

3.3 定量分析评估

本文在 3 个主流的视频分割数据集上, 即在 Youtube-2018 数据集, Davis 2016 数据集和 Segtrack-v2 数据集与当前主流的视频分割方法进行定量分析.

表 1 为本文算法和 9 种主流视频分割方法在 Youtube-2018 验证集的定量比较结果. Youtube-2018 验证集由 474 个视频序列组成, 且大部分视频序列包含多个目标. 该验证集的所有目标覆盖 97 种不同类别, 其中有 26 种类别从未在训练集中出现过. 在训练集中出现过的类别的区域相似度和轮廓精确度用 J_{Seen} 和 F_{Seen} 表示, 而训练集中未出现过的用 J_{Unseen} 和 F_{Unseen} 表示, 总评分 G 是 J_{Seen}、F_{Seen} 和 J_{Unseen}、F_{Unseen} 的平均值.

表 1 Youtube-2018 验证集的定量比较

方法	OL	合成数据	G	J_{Seen}	F_{Seen}	J_{Unseen}	F_{Unseen}	FPS
AGAME	×	×	66.1	67.8	69.5	60.8	66.2	14.3
STM−	×	×	68.2	—	—	—	—	6.25
FRTM	×	×	71.3	72.2	76.1	64.5	72.7	14.6
SwiftNet (ResNet18)	×	×	73.2	73.3	76.3	68.1	75.0	70
SST	×	×	81.8	80.9	—	76.0	—	6
Ours	×	×	74.6	73.2	78.1	71.8	76.1	0.9
OnAVOS	√	×	55.2	60.1	62.7	46.1	51.4	0.08
OSVOS	√	×	58.8	59.8	60.5	54.2	60.7	0.22
S2S	√	×	64.4	71.0	70.0	55.5	61.2	0.11
STM	×	√	79.4	84.2	72.8	72.8	80.9	6.25
LCM	×	√	82.0	82.2	86.7	75.7	83.4	—

我们对比的方法包括: AGAME[45]、STM[44]、FRTM[43]、SwiftNet[46]、SST[47]、OnAVOS[24]、OSVOS[23]、S2S[38]和 LCM[48]. 表 1 共分为 3 部分, 上半部分包括我们的方法和其他 5 种方法, 这些方法既不需在线训练又不需要使用合成的图像辅助训练, 在这些方法中, 只有 SST 方法的分割效果优于本文方法, 该方法基于时空 Transformer 和注意力机制实现了端到端的视频分割, 但和本文方法相比, 该网络结构更加复杂, 且需要更长的训练时间和更高的显存配置. STM− 与 STM 的区别: −代表与 STM 相比, STM−未使用合成的数据. 表 1 中部的 3 种方法需要在线训练但不需要使用合成数据训练网络, 这类方法不能较好地适应目标外观剧烈改变的场景, 且每帧分割的耗时较长, 不能实现实时分割. 表 1 下半部分的 2 种方法需要使用合成的数据进行训练, 这两种方法均属于基于记忆网络的分割方法, 达到了当前最佳的分割效果, 但基于记忆网络的方法时间复杂度相对较高, 且 LCM 算法还需要额外的微调步骤, 以减小训练集和测试集的差距.

表 2 描述了与当前主流 11 种分割方法, 在 Davis 2016 数据集的定量分析结果. 其中, FPS 代表每秒分割视频帧的数量; M_J 代表区域相似度, 用于计算每帧分割结果中, 正确分割像素的数量; M_F 代表轮廓精确度, 用于评价分割结果边界的准确性; $M_J\&M_F$ 代表 M_J 和 M_F 的均值; M_J 和 M_F 的值越高, 分割效果越好. 表 2 中, 按照是否需要在线训练分为两类: (1) 需要在线训练, 包括: OSVOS[23]、MSK[49]、OnAVOS[24]、RANet[50]; (2) 不需要在线训练, 包括 RGMP[51]、RMNet[16]、FEELVOS[25]、FAVOS[8]、SiamMask[9]、FRTM[43]、AMCT[14].

表 2 的前 4 种方法需要在线训练, 它们使用第 1 帧增广数据, 以训练网络权重, 可获取更精准分割结果, 但需要额外的模型训练和大量的时间消耗. 本文方法和后 7 种方法均不需要在线训练, 这类方法的分割速度通常快于在线训练方法. 表 2 中多数方法主要基于卷积神经网络提取目标的深度特征, 通过解码特征图输出视频分割结果,

常缺乏对目标底层特征和边缘特征的表达. RMNet−和 RMNet 相比, 未使用 Youtube-2018 数据集进行训练, 但这两种方法在预训练阶段, 均需使用合成的图像数据训练网络. 本文方法和 AMCT 方法通过支持向量机计算超像素间的相似度, 并通过吸收马尔可夫链获取超像素标签, 而 AMCT+CNN 方法则通过卷积神经网络计算超像素间相似度. 由表 2 可知, 本文方法的分割精确度远超上述两种基于吸收马尔可夫链的方法, 并具有和其他主流方法类似, 甚至优于其他主流方法的区域相似度和轮廓精确度.

表 2 Davis 2016 数据集中的定量比较

方法	*OL*	M_J	M_F	$M_J\&M_F$	*FPS*
OSVOS	√	79.8	80.6	80.2	0.25
MSK	√	79.7	75.4	77.5	0.08
OnAVOS	√	86.1	84.9	85.5	0.07
RANet	√	86.6	87.6	87.1	0.07
RGMP	×	81.5	82.0	81.8	7.7
FEELVOS	×	81.1	82.2	81.7	2.2
FAVOS	×	82.4	79.5	80.8	0.56
FRTM	×	—	—	80.3	35.2
SiamMask	×	71.7	67.8	69.8	35
RMNet−	×	80.6	82.3	81.5	11.9
RMNet	×	88.9	88.7	88.8	11.9
AMCT	×	60.9	—	—	2.3
AMCT+CNN	×	73.2	—	—	0.27
本文方法	×	81.7	80.9	81.1	1.2

表 3 描述了本文算法和 9 种主流视频分割方法在 Segtrack-v2 数据集定量分析评估的结果. 和 Davis 2016 单目标数据集相比, Segtrack-v2 部分视频有多个相似目标, 且有帧数大于 200 帧的长视频, 该数据集能反应分割模型在多种挑战下的鲁棒性. 对比方法包括: OFL[1]、MSK[49]、MoNet[52]、RGMP[51]、CVOS-Decoder[53]、AGUnet[54]、FCVOS[41]、CINN[55], 对比结果见表 3. 其中, OFL 和 FCVOS 为非深度学习方法, 它们分别基于光流和双边格网判断目标的显著区域, 但因缺乏对目标高级语义和整体特征的表达, 易将突变背景或相似物体分割成目标, 并造成分割错误的逐帧累积. 其余 7 种均通过构造深度学习模型进行视频分割. 在目标被长期遮挡或剧烈形变时, 上一帧的目标特征难以引导网络推断出当前帧的分割掩码. 表 3 中的多个分割网络缺乏对目标长期稳定特征和运动趋势的学习, 导致其在部分视频中出现严重的分割错误. 本文分割结果的区域相似度在所有方法中处于最优水平, 轮廓精确度为次优.

表 3 Segtrack-v2 数据集中的定量比较

方法	OFL	MSK	MoNet	RGMP	CVOS-Decoder	AGUnet	FCVOS	CINN	Ours
M_J	67.5	70.3	72.4	76.8	69.5	74.0	75.5	77.1	76.3
M_F	74.5	—	76.5	74.0	—	78.9	—	—	80.3

3.4 消融实验

本文在 Davis 2016 数据集中进行消融实验, 以评估算法不同模块. 本文共包含 5 个可移除的优化策略, 分别为: 目标跟踪获取感兴趣区域 (T-ROI)、第 1 帧和当前帧建立吸收马尔可夫链 (FN-AMC)、基于短期时空线索优化超像素标签 (SOL)、基于长期时空线索优化超像素标签 (LOL)、基于骨架映射网络优化分割 (SNR). 消融实验以 AMCT 方法为基准, AMCT 方法通过 EPPC 光流获取相邻帧超像素的位移量, 然后将当前帧和前一帧建立吸收马尔可夫链, 根据当前帧超像素的吸收时间, 判断超像素标签, 进而获取分割结果. 本文基于此基准线和优化策略进行的 6 组消融实验, 具体见表 4.

表 4 在 Davis 2016 数据集中的消融实验

组别	T-ROI	FN-AMC	SOL	LOL	SNR	M_J	分割速率 (FPS)
基线	×	×	×	×	×	60.9	4
实验1	√	×	×	×	×	66.2	3.32
实验2	√	√	×	×	×	72.4	1.92
实验3	√	√	√	×	×	72.8	1.85
实验4	√	√	√	√	×	73.6	1.69
实验5	√	√	√	√	√	81.7	1.2

通过表 4 可知, 目标跟踪获取感兴趣区域 (T-ROI)、第 1 帧和当前帧建立吸收马尔可夫链 (FN-AMC)、基于骨架映射网络优化分割 (SNR) 的分割优化效果明显, 区域相似度分别提升了 5.3%、6.2%、8.1%, 对大部分视频均有显著的提升. 而短期时空线索优化超像素标签 (SOL) 和长期时空线索优化超像素标签 (LOL) 仅对部分视频序列有明显提升, 也会造成个别视频的分割精度轻微下降, 在整个数据集中区域相似度平均提升 0.6% 和 0.4%. 本文算法的平均分割速率为 1.2 FPS, T-ROI、FN-AMC、SOL、LOL、SNR 的平均耗时分别为: 0.06 s、0.21s、0.02 s、0.05 s、0.24 s.

4 总 结

本文针对复杂场景下的视频分割展开研究, 提出了联合吸收马尔可夫链和骨架映射的视频分割方法. 该方法实现了“预分割—后优化—再提升”三阶段递进处理, 可得到精准目标分割结果. 预分割阶段: 本文以超像素为线索描述目标, 构建有效传播目标特征的吸收马尔可夫链, 初步得到超像素的前景标签和背景标签, 该阶段在目标剧烈形变和被遮挡时仍能准确定位目标. 后优化阶段: 本文提出基于长期和短期时空线索的超像素标签优化, 充分挖掘目标的整体变化规律和运动一致性约束, 降低了突变背景和相似物体的干扰. 再提升阶段: 本文基于超像素标签优化结果, 提取了表达目标位置与形态的前景骨架和表达干扰物的背景骨架, 并提出即能挖掘视频高级语义特征又融合视频与骨架特征的骨架映射网络, 进一步提升目标分割精度. 大量实验结果表明, 与当前主流视频分割方法相比, 在剧烈形变、遮挡、相似背景等多种挑战下, 本文方法具有更高的区域相似度和轮廓精确度. 未来工作中, 拟通过图神经网络优化超像素标签的分类, 并探索骨架在计算机视觉其他领域的应用, 如人体姿态评估、语义分割等.

References:

[1] Tsai YH, Yang MH, Black MJ. Video segmentation via object flow. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3899–3908. [doi: 10.1109/CVPR.2016.423]

[2] Ding MY, Wang Z, Zhou BL, Shi JP, Lu ZW, Luo P. Every frame counts: Joint learning of video segmentation and optical flow. In: Proc. of the 2020 AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 10713–10720. [doi: 10.1609/aaai.v34i07.6699]

[3] Li XX, Loy CC. Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 93–110. [doi: 10.1007/978-3-030-01219-9_6]

[4] Zheng Y, Chen YD, Hao CY. Video object segmentation algorithm based on consistent features. Journal of Image and Graphics, 2020, 25(8): 1558–1566 (in Chinese with English abstract). [doi: 10.11834/jig.190571]

[5] Hao CY, Chen YD, Yang ZX, Wu EH. Higher-order potentials for video object segmentation in bilateral space. Neurocomputing, 2020, 401: 28–35. [doi: 10.1016/j.neucom.2020.03.020]

[6] Liu ZY, Wang L, Hua G, Zhang QL, Niu ZX, Wu Y, Zheng NN. Joint video object discovery and segmentation by coupled dynamic Markov networks. IEEE Trans. on Image Processing, 2018, 27(12): 5840–5853. [doi: 10.1109/TIP.2018.2859622]

[7] Chen X, Li ZX, Yuan Y, Yu G, Shen JX, Qi DL. State-aware tracker for real-time video object segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9381–9390. [doi: 10.1109/CVPR42600.2020.00940]

[8] Cheng JC, Tsai YH, Hung WC, Wang SJ, Yang MH. Fast and accurate online video object segmentation via tracking parts. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7415–7424. [doi: 10.1109/CVPR.

2018.00774]

[9] Wang Q, Zhang L, Bertinetto L, Hu WM, Torr PHS. Fast online object tracking and segmentation: A unifying approach. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1328−1338. [doi: 10.1109/CVPR.2019.00142]

[10] Liu MH, Wang CS, Hu Q, Wang CX, Cui XH. Part-based object tracking based on multi collaborative model. Ruan Jian Xue Bao/Journal of Software, 2020, 31(2): 511−530 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5656.htm [doi: 10.13328/j.cnki.jos.005656]

[11] Yeo D, Son J, Han B, Han JH. Superpixel-based tracking-by-segmentation using Markov chains. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1812−1821. [doi: 10.1109/CVPR.2017.62]

[12] Chen YD, Hao CY, Liu AX, Wu EH. Multilevel model for video object segmentation based on supervision optimization. IEEE Trans. on Multimedia, 2019, 21(8): 1934−1945. [doi: 10.1109/TMM.2018.2890361]

[13] Liu LW, Xing JL, Ai HZ, Lao SH. Semantic superpixel based vehicle tracking. In: Proc. of the 21st Int'l Conf. on Pattern Recognition. Tsukuba: IEEE, 2012. 2222−2225.

[14] Milan A, Leal-Taixé L, Schindler K, Reid I. Joint tracking and segmentation of multiple targets. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5397−5406. [doi: 10.1109/CVPR.2015.7299178]

[15] Li Y, Liu Y, Liu GJ, Guo MZ. Weakly supervised semantic segmentation by iterative superpixel-CRF refinement with initial clues guiding. Neurocomputing, 2020, 391: 25−41. [doi: 10.1016/j.neucom.2020.01.054]

[16] Xie HZ, Yao HX, Zhou SC, *et al*. Efficient regional memory network for video object segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1286−1295. [doi: 10.1109/CVPR46437.2021.00134]

[17] Oh SW, Lee JY, Xu N, Kim SJ. Fast user-guided video object segmentation by interaction-and-propagation networks. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5242−5251. [doi: 10.1109/CVPR.2019.00539]

[18] Miao JX, Wei YC, Yang Y. Memory aggregation networks for efficient interactive video object segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10363−10372. [doi: 10.1109/CVPR42600.2020.01038]

[19] Wang WG, Song HM, Zhao SY, Shen JB, Zhao SY, Hoi SCH, Ling HB. Learning unsupervised video object segmentation through visual attention. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3059−3069. [doi: 10.1109/CVPR.2019.00318]

[20] Li SY, Seybold B, Vorobyov A, Fathi A, Huang Q, Kuo CCJ. Instance embedding transfer to unsupervised video object segmentation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6526−6535. [doi: 10.1109/CVPR.2018.00683]

[21] Hu YT, Huang JB, Schwing AG. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 813−830. [doi: 10.1007/978-3-030-01246-5_48]

[22] Li SY, Seybold B, Vorobyov A, Lei XJ, Kuo CCJ. Unsupervised video object segmentation with motion-based bilateral networks. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 215−231. [doi: 10.1007/978-3-030-01219-9_13]

[23] Caelles S, Maninis KK, Pont-Tuset J, Leal-Taixe L, Cremers D, van Gool L. One-shot video object segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5320−5329. [doi: 10.1109/CVPR.2017.565]

[24] Maninis KK, Caelles S, Chen Y, Pont-Tuset J, Leal-Taixe L, Cremers D, Van Gool L. Video object segmentation without temporal information. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1515−1530. [doi: 10.1109/TPAMI.2018.2838670]

[25] Voigtlaender P, Chai YN, Schroff F, Adam H, Leibe B, Chen LC. FEELVOS: Fast end-to-end embedding learning for video object segmentation. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9473−9482. [doi: 10.1109/CVPR.2019.00971]

[26] Oh SW, Lee JY, Xu N, Kim SJ. Space-time memory networks for video object segmentation with user guidance. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(1): 442−455. [doi: 10.1109/TPAMI.2020.3008917]

[27] Zhou Q, Huang ZL, Huang LC, Gong YC, Shen H, Huang C, Liu WY, Wang XG. Proposal, tracking and segmentation (PTS): A cascaded network for video object segmentation. arXiv:1907.01203, 2019.

[28] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4293−4302. [doi: 10.1109/CVPR.2016.465]

[29] Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully-convolutional Siamese networks for object tracking. In: Proc. of the 2016 European Conf. on Computer Vision. Amsterdam: Springer, 2016. 850−865. [doi: 10.1007/978-3-319-48881-3_56]

[30] Wen LY, Du DW, Lei Z, Li SZ, Yang MH. JOTS: Joint online tracking and segmentation. In: Proc. of the 2015 IEEE Int'l Conf. on

Computer Vision. Boston: IEEE, 2015. 2226–2234. [doi: 10.1109/CVPR.2015.7298835]

[31] Wang WG, Shen JB, Xie JW, Porikli F. Super-trajectory for video segmentation. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1680–1688. [doi: 10.1109/ICCV.2017.185]

[32] Wang LJ, Lu HC, Yang MH. Constrained superpixel tracking. IEEE Trans. on Cybernetics, 2018, 48(3): 1030–1041. [doi: 10.1109/TCYB.2017.2675910]

[33] Heo Y, Jun KY, Kim CS. Interactive video object segmentation using global and local transfer modules. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 297–313. [doi: 10.1007/978-3-030-58520-4_18]

[34] Li B, Yan JJ, Wu W, Zhu Z, Hu XL. High performance visual tracking with Siamese region proposal network. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8971–8980. [doi: 10.1109/CVPR.2018.00935]

[35] Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274–2282. [doi: 10.1109/TPAMI.2012.120]

[36] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[37] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[38] Xu N, Yang LJ, Fan YC, *et al*. YouTube-VOS: Sequence-to-sequence video object segmentation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 603–619. [doi: 10.1007/978-3-030-01228-1_36]

[39] Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A, van Gool L. The 2017 DAVIS challenge on video object segmentation. arXiv:1704.00675, 2017.

[40] Li FX, Kim T, Humayun A, Tsai D, Rehg JM. Video segmentation by tracking many figure-ground segments. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 2192–2199. [doi: 10.1109/ICCV.2013.273]

[41] Gui Y, Tian Y, Zeng DJ, Xie ZF, Cai YY. Reliable and dynamic appearance modeling and label consistency enforcing for fast and coherent video object segmentation with the bilateral grid. IEEE Trans. on Circuits and Systems for Video Technology, 2020, 30(12): 4781–4795. [doi: 10.1109/TCSVT.2019.2961267]

[42] Sun MJ, Xiao JM, Lim EG, Zhang BF, Zhao Y. Fast template matching and update for video object tracking and segmentation. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10788–10796. [doi: 10.1109/CVPR42600.2020.01080]

[43] Robinson A, Lawin FJ, Danelljan M, Khan FS, Felsberg M. Learning fast and robust target models for video object segmentation. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7404–7413. [doi: 10.1109/CVPR42600.2020.00743]

[44] Oh SW, Lee JY, Xu N, Kim SJ. Video object segmentation using space-time memory networks. In: Proc. of the 2019 IEEE Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9225–9234. [doi: 10.1109/ICCV.2019.00932]

[45] Johnander J, Danelljan M, Brissman E, Khan FS, Felsberg M. A generative appearance model for end-to-end video object segmentation. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8945–8954. [doi: 10.1109/CVPR.2019.00916]

[46] Wang HC, Jiang XL, Ren HB, Hu Y, Bai S. SwiftNet: Real-time video object segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1296–1305. [doi: 10.1109/CVPR46437.2021.00135]

[47] Duke B, Ahmed A, Wolf C, Aarabi P, Taylor GW. SSTVOS: Sparse spatiotemporal Transformers for video object segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5908–5917. [doi: 10.1109/CVPR46437.2021.00585]

[48] Hu L, Zhang P, Zhang B, Pan P, Xu YH, Jin R. Learning position and target consistency for memory-based video object segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4142–4152. [doi: 10.1109/CVPR46437.2021.00413]

[49] Perazzi F, Khoreva A, Benenson R, Schiele B, Sorkine-Hornung A. Learning video object segmentation from static images. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3491–3500. [doi: 10.1109/CVPR.2017.372]

[50] Wang ZQ, Xu J, Liu L, Zhu F, Shao L. RANet: Ranking attention network for fast video object segmentation. In: Proc. of the 2019 IEEE Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3977–3986. [doi: 10.1109/ICCV.2019.00408]

[51] Oh SW, Lee JY, Sunkavalli K, Kim SJ. Fast video object segmentation by reference-guided mask propagation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7376–7385. [doi: 10.1109/CVPR.2018.

00770]

[52] Xiao HX, Feng JS, Lin GS, Liu Y, Zhang MJ. MoNet: Deep motion exploitation for video object segmentation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1140−1148. [doi: 10.1109/CVPR.2018.00125]

[53] Tan ZT, Liu B, Chu Q, Zhong HS, Wu Y, Li WH, Yu NH. Real time video object segmentation in compressed domain. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(1): 175−188. [doi: 10.1109/TCSVT.2020.2971641]

[54] Yin YJ, Xu D, Wang XG, Zhang L. AGUnet: Annotation-guided U-net for fast one-shot video object segmentation. Pattern Recognition, 2021, 110: 107580. [doi: 10.1016/j.patcog.2020.107580]

[55] Bao LC, Wu BY, Liu W. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5977−5986. [doi: 10.1109/CVPR.2018.00626]

附中文参考文献:

[4] 征煜, 陈亚当, 郝川艳. 特征一致性约束的视频目标分割. 中国图象图形学报, 2020, 25(8): 1558−1566. [doi: 10.11834/jig.190571]

[10] 刘明华, 汪传生, 胡强, 王传旭, 崔雪红. 多模型协作的分块目标跟踪. 软件学报, 2020, 31(2): 511−530. http://www.jos.org.cn/1000-9825/5656.htm [doi: 10.13328/j.cnki.jos.005656]

梁云(1981−), 女, 博士, 教授, CCF 专业会员, 主要研究领域为计算机视觉, 计算机图形学, 智慧农业.

郑晋图(1999−), 男, 本科生, 主要研究领域为目标跟踪, 目标检测, 视频分割.

张宇晴(1999−), 女, 硕士生, 主要研究领域为机器学习, 目标跟踪, 视频分割.

张勇(1979−), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为大数据处理, 人工智能, 图形图像.