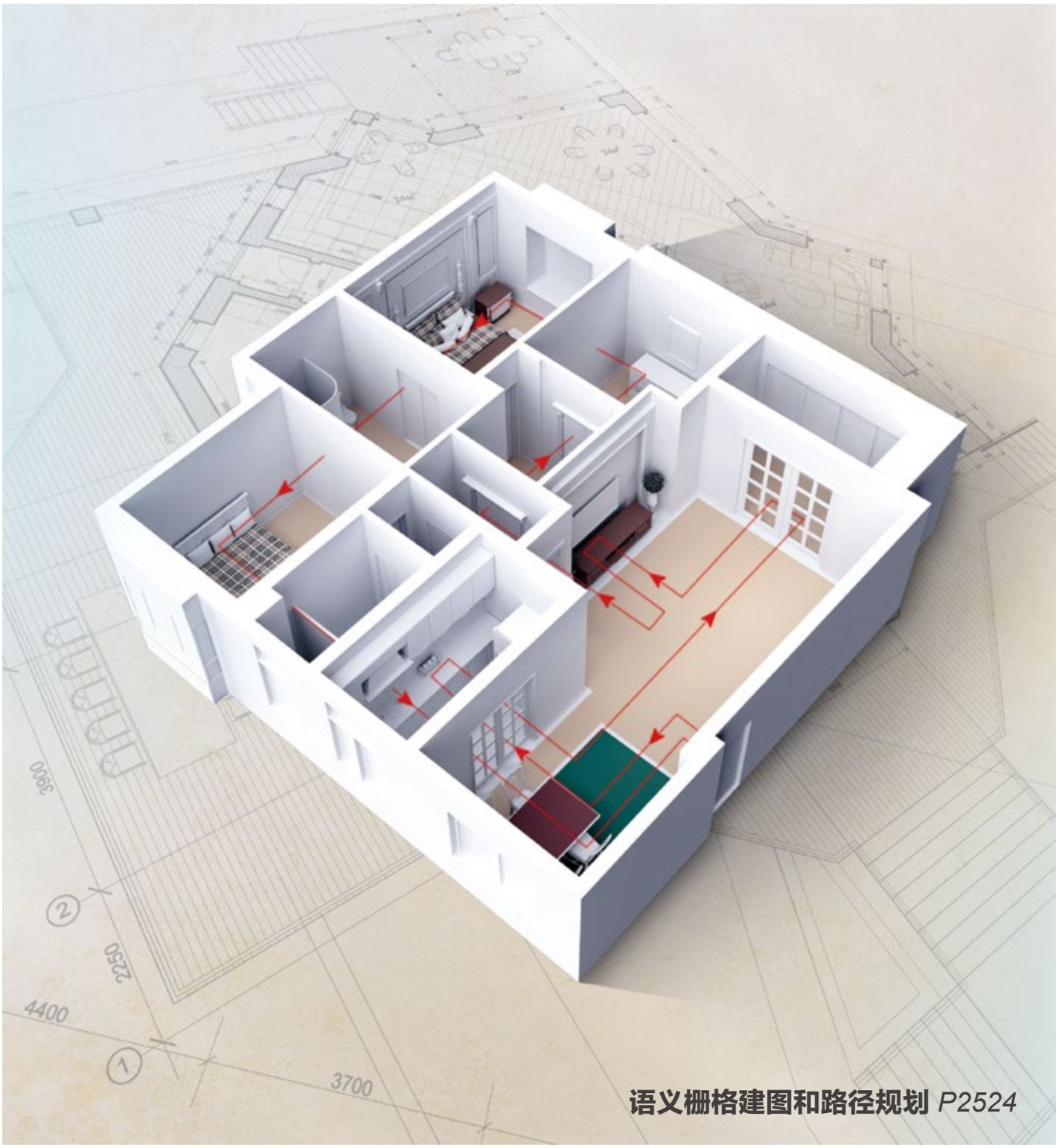
JOURNAL OF IMAGE AND GRAPHICS

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

中国图象图形学报

2021
10
VOL.26

ISSN1006-8961
CN11-3758/TB



语义栅格建图和路径规划 *P2524*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第10期（总第306期）
2021年10月16日

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

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

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

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

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

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

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

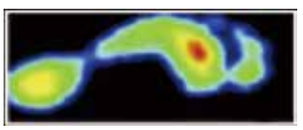
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

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

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

2021年10月 第26卷 第10期（总第306期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



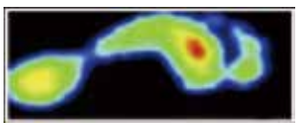
穿鞋足迹序列的足迹能量图组表达与识别(第2357页)

时空图卷积网络与注意机制的视频目标分割(第2376页)

双光流网络指导的视频目标检测(第2473页)

Volume 26, Number 10
Published October 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Shoeprint sequence representation and recognition using shoeprint energy map set (P2357)

Spatial-temporal video object segmentation with graph convolutional network and attention mechanism(P2376)

Dual optical flow network-guided video object detection(P2473)

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

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

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2021)10-2376-12

论文引用格式:Yao R, Xia S X, Zhou Y, Zhao J Q and Hu F Y. 2021. Spatial-temporal video object segmentation with graph convolutional network and attention mechanism. Journal of Image and Graphics,26(10):2376-2387(姚睿,夏士雄,周勇,赵佳琦,胡伏原. 2021. 时空图卷积网络与注意机制的视频目标分割. 中国图象图形学报,26(10):2376-2387)[DOI:10.11834/jig.200357]

时空图卷积网络与注意机制的视频目标分割

姚睿[1],夏士雄[1*],周勇[1],赵佳琦[1],胡伏原[2]

1. 中国矿业大学计算机科学与技术学院, 徐州　221116; 2. 苏州科技大学电子与信息工程学院, 苏州　215009

摘　要:**目的**　从大量数据中学习时空目标模型对于半监督视频目标分割任务至关重要,现有方法主要依赖第1帧的参考掩膜(通过光流或先前的掩膜进行辅助)估计目标分割掩膜。但由于这些模型在对空间和时域建模方面的局限性,在快速的外观变化或遮挡下很容易失效。因此,提出一种时空部件图卷积网络模型生成鲁棒的时空目标特征。**方法**　首先,使用孪生编码模型,该模型包括两个分支:一个分支输入历史帧和掩膜捕获序列的动态特征,另一个分支输入当前帧图像和前一帧的分割掩膜。其次,构建时空部件图,使用图卷积网络学习时空特征,增强目标的外观和运动模型,并引入通道注意模块,将鲁棒的时空目标模型输出到解码模块。最后,结合相邻阶段的多尺度图像特征,从时空信息中分割出目标。**结果**　在DAVIS(densely annotated video segmentation)-2016和DAVIS-2017两个数据集上与最新的12种方法进行比较,在DAVIS-2016数据集上获得了良好性能,Jacccard相似度平均值(Jaccard similarity-mean, *J*-*M*)和F度量平均值(F measure-mean, *F*-*M*)得分达到了85.3%,比性能最高的对比方法提高了1.7%;在DAVIS-2017数据集上,*J*-*M*和*F*-*M*得分达到了68.6%,比性能最高的对比方法提高了1.2%。同时,在DAVIS-2016数据集上,进行了网络输入与后处理的对比实验,结果证明本文方法改善了多帧时空特征的效果。**结论**　本文方法不需要在线微调和后处理,时空部件图模型可缓解因目标外观变化导致的视觉目标漂移问题,同时平滑精细模块增加了目标边缘细节信息,提高了视频目标分割的性能。

关键词:视频目标分割(VOS);图卷积网络;时空特征;注意机制;深度神经网络

Spatial-temporal video object segmentation with graph convolutional network and attention mechanism

Yao Rui[1], Xia Shixiong[1*], Zhou Yong[1], Zhao Jiaqi[1], Hu Fuyuan[2]

1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou　221116, China;

2. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou　215009, China

Abstract: **Objective** The task of video object segmentation (VOS) is to track and segment a single object or multiple objects in a video sequence. VOS is an important issue in the field of computer vision. Its goal is to manually or automatically provide specific object masks on the first frame or reference frame and then segment these specific objects in the entire video sequence. VOS plays an important role in video understanding. According to the types of video object labels, VOS methods can be divided into four categories: unsupervised, interactive, semi-supervised, and weakly supervised. In this

收稿日期:2020-07-16;**修回日期**:2020-08-26;**预印本日期**:2020-09-02

*通信作者:夏士雄　xiasx@cumt.edu.cn

基金项目:国家自然科学基金项目(62172417,61772530,61806206,61876121);江苏省自然科学基金项目(BK20180639);江苏省六大人才高峰项目(2018-XYDXX-044)

Supported by: National Natural Science Foundation of China (62172417, 61772530, 61806206, 61876121); Natural Science Foundation of Jiangsu Province of China (BK20180639); the Six Talent Peaks Project in Jiangsu Province (2018-XYDXX-044)

study, we deal with the problem of semi-supervised VOS; that is, the ground truth of object mask is only given in the first frame, the segmented object is arbitrary, and no further assumptions are made about the object category. Currently, semi-supervised VOS methods are mostly based on deep learning. These methods can be divided into two types: detection-based methods and matching-based or motion propagation methods. Without using temporal information, detection-based methods learn the appearance model to perform pixel-level detection and object segmentation at each frame of the video. Matching-based or motion propagation methods utilize the temporal correlation of object motion to propagate from the first frame or a given mask frame to the object mask of the subsequent frame. Matching-based methods first calculate the pixel-level matching between the features of the template frame and the current frame in the video and then directly divide each pixel of the current frame from the matching result. There are two types of methods based on motion propagation. One type of method is to introduce optical flow to train the VOS model. Another type of method learns deep object features from the object mask of the previous frame and refines the object mask of the current frame. Most existing methods mainly rely on the reference mask of the first frame (assisted by optical flow or previous mask) to estimate the object segmentation mask. However, due to the limitations of these models in modeling spatial and temporal domain, they easily fail under rapid appearance changes or occlusion. Therefore, a spatial-temporal part-based graph model is proposed to generate robust spatial-temporal object features. **Method** In this study, we propose an encode-decode-based VOS framework for spatial-temporal part-based graph. First, we use the Siamese architecture for the encode model. The input has two branches: the historical image frame branch stream and the current image frame branch stream. To simplify the model, we introduce a Markov hypothesis, that is, given the current frame and $K-1$ previous frames, and $K-1$ previously estimated segmentation masks. One branch inputs the dynamic features of the historical image frame and the mask, and the other branch inputs the current frame image and the segmentation mask of the previous frame. Both branches use ResNet50 as the base network, and the network weights are derived from the ImageNet pre-trained model. After obtaining the results of Res5 stage, we use the global convolution module to output image features, where the size of the convolution kernel is set to 7 and the number of channels of the feature is set to 512, which is the same as the other feature dimensions. Next, we design a structural graph representation model based on parts (nodes) and use the graph convolutional network to learn the object appearance model. To represent the spatial-temporal object model, we construct an undirected spatial-temporal part-based graph $\boldsymbol{G}_{\mathrm{ST}}$ on frames with dense grid parts (nodes) and K (i.e., $t-K$, $\cdots$, $t-1$), use a two-layer graph convolutional network to output feature matrix, and aggregate the target features of the spatial-temporal components through max pooling. In addition, we construct an undirected spatial part-based graph $\boldsymbol{G}_{\mathrm{S}}$ (similar to $\boldsymbol{G}_{\mathrm{ST}}$), which has the same processing steps as the above two-layer graph convolutional network, and then we obtain the spatial part-based object features. Next, the spatial-temporal part-based features and spatial part-based features are channel aligned to form a whole feature, and the channels are 256. The output functions of the spatial-temporal part-based feature model and the spatial part-based feature model have different characteristics, and we adopt an attention mechanism to assign different weights to all features. To optimize the feature map, we introduce a residual module to improve the edge details. Finally, in the decoding module, we construct a smooth refinement module, add an attention mechanism module, and merge features of adjacent stages in a multi-scale context. Specifically, the decoding module consists of three smooth and fine modules, plus a convolution layer and a Softmax layer, and then outputs the mask of the video object. The training process mainly includes two stages. First, we use the simulated images generated from the static images to pre-train the network. Second, we fine-tune this pre-trained model on the VOS dataset. The time window size K is set to 3. In the testing, the interval 3 is used to update the reference frame image and mask, so that the historical information can be effectively memorized. **Result** In the experimental section, the proposed method does not require online fine-tuning and post-processing, and it is compared with 12 latest methods on two datasets. On the DAVIS (densely annotated video segmentation)-2016 dataset, compared with the method with the highest performance, our Jaccard similarity-mean (*J-M*) & F measure-mean (*F-M*) score is 85.3% and increased by 1.7%. On the DAVIS-2017 dataset, compared with the method with the highest performance, our *J-M* & *F-M* score is 68.6% and is increased by 1.2%. At the same time, on the DAVIS-2016 dataset, a comparative experiment of network input and post-processing is carried out. **Conclusion** In this work, we studied the problem of robust spatial-temporal object model in VOS. A spatial-temporal VOS with part-based graph is proposed to alleviate the drift of visual object. The experimental results show that our model outperforms several state-of-the-art VOS approaches.

Key words: video object segmentation (VOS); graph convolutional network; spatial-temporal features; attention mechanism; deep neural network

0 引 言

视频目标分割(video object segmentation, VOS)的任务是在视频序列中分割单目标或多目标,是计算机视觉领域中的一个重要问题,目标是在第1帧或参考帧上手动或自动给出特定目标掩膜,而后在整个视频序列中分割这些特定目标(Yao 等,2020)。

根据不同的视频目标标签类型,VOS 分为无监督、交互式、半监督和弱监督视频目标分割。无监督和交互式 VOS 是用户与方法交互程度的两个极端:一方面,无监督 VOS 可以通过自下而上方式产生连贯的时空目标区域,而无需任何用户输入,即没有任何视频特定标签(Grundmann 等,2010)。另一方面,交互式 VOS 使用严格监督的交互方法,要求对第1帧进行像素级精确分割(人工配置非常耗时)(Maninis 等,2018)。在两个极端之间存在半监督 VOS 方法,需要手动标签确定前景目标,然后自动分割为序列的其余帧(Caelles 等,2017;Wug 等,2018)。而弱监督 VOS 方法在训练或/和测试过程中使用轻量级的标签形式,缓解了标记强烈的像素级分割掩膜问题,可克服监督方法的不足(Zhang 等,2015)。

给定视频中参考帧的目标掩膜,半监督 VOS 方法通常比无监督 VOS 方法具有更好的性能,受到计算机视觉领域的广泛关注。DAVIS(densely annotated video segmentation)-2016(Perazzi 等,2016)和 DAVIS-2017(Pont-Tuset 等,2017)等大型视频数据集的发布使得许多深度网络方法(Caelles 等,2017;Wug 等,2018;Newswanger 和 Xu,2017)极大改善了半监督方法的性能。但是,这些方法与实际应用还有很大距离。其中,目标遮挡、快速移动、外观变化以及不同实例之间的相似性仍然是主要障碍;繁重的后处理操作、人工干预和昂贵的模型微调也存在很多问题。构建一种鲁棒的视频目标分割模型成为关注的焦点。

半监督 VOS 大多基于深度学习的方法,分为基于检测的方法和基于匹配或运动传播的方法两种类型。

基于检测的方法在不使用时间信息的情况下,学习外观模型,以在视频每个帧处进行像素级检测和目标分割,依靠使用给定测试序列的第1帧目标掩膜微调深度网络(Caelles 等,2017;Voigtlaender 和 Leibe,2017)。Caelles 等人(2017)引入全卷积网络(Long 等,2015),提出单次视频目标分割(one-shot VOS,OSVOS)方法在静态图像上进行离线和在线训练,从而在目标视频的第1帧上微调。Voigtlaender 和 Leibe(2017)提出一种在线自适应视频目标分割(online adaptation VOS,OnAVOS)方法,对网络进行微调以适应外观变化。OSVOS-S 利用语义信息进行视频目标分割(Maninis 等,2019)。LucidTracker 引入一种用于在线微调的数据增强机制(Khoreva 等,2019)。Luiten 等人(2018)集成实例分割、光流、优化和重标识以及广泛的微调技术,获得了令人满意的视频目标分割性能。在线微调对 VOS 任务非常有效,视为提高 VOS 性能的常规技术,但在实际应用中耗时较多。

基于匹配或运动传播的方法利用目标运动的时间相关性,从第1帧或给定带标签帧开始传播到后续帧的目标掩膜。基于匹配的方法首先计算视频中模板帧与当前帧的特征之间的像素级匹配,然后直接从匹配结果中分割当前帧的每个像素。像素级度量学习通过像素空间中与模板帧的最近邻匹配来预测每个像素的分割结果(Yoon 等,2017)。Hu 等人(2018a)在 VideoMatch 中提出一种软匹配机制,对匹配特征的平均相似度评分图执行软分割,以生成平滑的预测。Voigtlaender 等人(2019)使用全局和局部匹配来获得更稳定的像素级匹配。基于运动传播的方法有两类。一类方法引入光流训练 VOS 模型。在视频描述的早期阶段,主要将光流应用于 VOS 以保持运动一致性,使用光流作为线索随时间跟踪像素以建立时间相关性。SegFlow(Cheng 等,2017)、MoNet(motion network)(Xiao 等,2018)和 VS-ReID(video object segmentation with re-identification)(Li 和 Change,2018)方法均由图像颜色分割和使用 FlowNet(Ilg 等,2017)光流两个分支组成,为了学习利用运动信息,会接收两个或多个输入,包括目标帧和多个相邻帧。Jampani 等人(2017)提出一种时间双向网络,使用光流作为附加功能以自适应方式传播视频帧。通过光流建立的时间依赖性,Bao 等人(2018)通过基于卷积神经网络(convolutional neural network, CNN)的时空马尔可夫随机场的推论提出一种 VOS 方法。此外,一些方法采用光流和递归神经网络(recurrent neural network, RNN)构建目

标掩膜传播(Hu 等,2018c)。另一类方法从前一帧的目标掩膜学习深度目标特征,而在当前帧的目标掩膜进行细化。MaskTrack 方法将经过精炼的前一帧掩膜训练为当前帧掩膜,直接从光流中推断出分割结果(Perazzi 等,2017)。与使用前一帧的精确前景掩膜的方法相比,Yang 等人(2018)在视觉和空间调制之前使用了非常粗糙的位置,提高分割速度。Wug 等人(2018)使用孪生网络框架,提出参考引导的掩膜传播(reference-guided mask propagation, RGMP)方法,将带有标签的参考帧和具有前一帧掩膜的当前帧同时用于深度网络,提升了视频目标分割性能。但是,这类方法使用第 1 帧中的初始目标掩膜匹配当前帧目标,由于视觉目标分割是变化场景的动态过程,在连续帧中目标外观之间存在很强的时空关系,且简单叠加参考帧图像和目标掩膜以及当前帧图像和前一帧掩膜,没有挖掘两帧图像上空间和时域信息,易导致视觉目标的漂移问题。

针对上述问题,本文提出一种基于时空目标模型的端到端编码—时空部件图—解码视频目标分割方法,利用了历史帧的时空部件特征结构信息。通过使用时空部件图卷积神经网络(Yan 等,2018),引入历史帧图像和掩膜,构造了一个时空目标模型以形成历史样本的结构化表示,生成时空目标特征。在解码模块,为了处理多尺度分割不一致问题,构建了平滑精细模块(smooth refinement module, SRM),细化目标分割的性能,并引入通道注意机制模块(channel attention block, CAB),实现鲁棒 VOS。本文算法框架如图 1 所示。

本文的创新之处在于:1)设计一种时空部件图卷积神经网络,构建时空部件图模型,利用通道注意机制,生成鲁棒的外观特征,缓解了外观变化导致的视觉目标漂移问题;2)提出一种端到端的编码—时空部件图—解码视频目标分割框架,利用平滑精细模块,增加目标边缘细节信息,提高视频目标分割的性能,不需要在线微调和后处理;3)在 DAVIS 2016 和 DAVIS 2017 数据集上进行实验,与主流方法进行对比,验证了本文方法的效率和有效性。

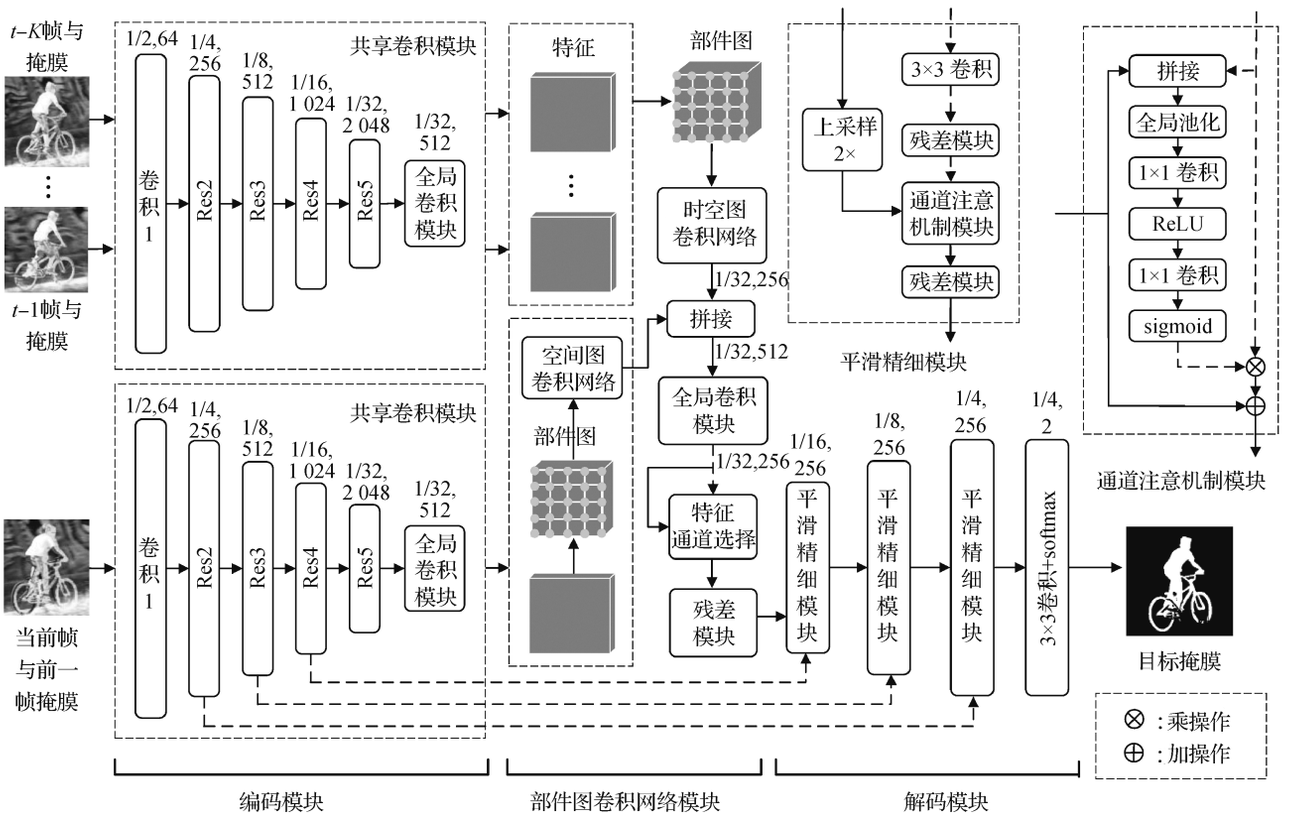


图 1　本文提出的时空部件图卷积网络的视频目标分割框架图

Fig. 1　Overview of our proposed method, spatial-temporal video object segmentation with part-based graph

1 时空图卷积网络的分割方法

本文方法的目标是设计一个深度学习模型,能够沿视频序列对目标进行分割,输出每个目标的分割掩膜。

在形式上,令 $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T)$ 是具有 T 帧的输入 RGB 视频序列,$\boldsymbol{x}_t\in\mathbf{R}^{H\times W\times 3}$ 表示第 t 帧图像,H 和 W 为图像帧的高度和宽度。目标是估计整个序列 $\boldsymbol{X}$ 上视频目标的掩膜 $\hat{\boldsymbol{Y}}$,即学习一个映射 $M:\boldsymbol{X}\rightarrow\hat{\boldsymbol{Y}}$,其中,$\hat{\boldsymbol{Y}}=(\hat{\boldsymbol{y}}_1,\cdots,\hat{\boldsymbol{y}}_T)$ 是序列的目标掩膜,$\hat{\boldsymbol{y}}_t\in\mathbf{R}^{H\times W}$ 是在第 t 个视频帧中跟踪的目标掩膜。本文将视频序列 $\boldsymbol{X}$ 的真实目标掩膜定义为 $\boldsymbol{Y}=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T)$。

1.1 模型概述

本文提出编码—时空部件图—解码的视频目标分割框架,关键思想是通过历史前 K 个帧的部件图像特征,学习时空目标外观模型,引入注意机制平滑精细机制,提升检测和分割视频目标的性能。具体来说,框架包括编码模块、时空部件图卷积网络模块和解码模块。编码模块使用孪生编码模型,分两个分支,一个分支输入历史帧和掩膜捕获序列的动态特征;另一个分支的输入是当前帧图像和前一帧的分割掩膜,以提取当前帧特征,其中前一帧目标掩膜指导当前帧分割结果。时空部件图卷积网络模块构建时空部件图(即编码模块中两个分支的图像特征),并使用图卷积网络,学习时空特征信息,增强目标的外观和运动模型,然后引入通道注意模块,将鲁棒的时空目标模型输出到解码模块。解码模块使用完全卷积网络,结合相邻阶段的图像特征,从时空信息中分割出视频中的目标。

1.2 编码模块

本文方法的输入包括历史图像帧和当前图像帧两个分支。大多数现有半监督视频目标分割方法的编码模块分为两类,第 1 类方法依赖第 1 帧的参考掩膜估计分割目标,即模型为 $p(\hat{\boldsymbol{y}}_t\mid\boldsymbol{x}_t,\boldsymbol{y}_0,\boldsymbol{x}_0)$,但在对时间组件建模方面具有局限性,在快速的外观变化或遮挡下容易失效。第 2 类方法学习多帧图像特征估计目标模型,即 $p(\hat{\boldsymbol{y}}_t\mid\boldsymbol{x}_t,\cdots,\boldsymbol{x}_1,\boldsymbol{y}_0,\boldsymbol{x}_0)$ 或 $p(\hat{\boldsymbol{y}}_t\mid\boldsymbol{x}_t,\hat{\boldsymbol{y}}_{t-1},\boldsymbol{x}_{t-1},\cdots,\hat{\boldsymbol{y}}_1,\boldsymbol{x}_1,\boldsymbol{y}_0,\boldsymbol{x}_0)$。

本文提出的目标模型属于第 2 类方法,为了简化模型,引入一个马尔可夫假设,其条件分布分解为

$$p(\hat{\boldsymbol{Y}}\mid\boldsymbol{X})=\prod_{t=1}^{T}p(\hat{\boldsymbol{y}}_t\mid\boldsymbol{X}_{t-K}^{t},\hat{\boldsymbol{Y}}_{t-K}^{t-1}) \tag{1}$$

式中,K 为序列中的帧号。给定当前帧和 $K-1$ 个先前帧 $\boldsymbol{X}_{t-K}^{t}=(\boldsymbol{x}_{t-K},\cdots,\boldsymbol{x}_t)$ 以及 $K-1$ 个先前估计的分割掩膜 $\hat{\boldsymbol{Y}}_{t-K}^{t-1}=(\hat{\boldsymbol{y}}_{t-K},\cdots,\hat{\boldsymbol{y}}_{t-1})$,可以跟踪和分割出视频目标。在训练中,设计一个回归函数 $f(\cdot)$ 通过对分布 $p(\hat{\boldsymbol{y}}_t\mid\boldsymbol{X}_{t-K}^{t},\hat{\boldsymbol{Y}}_{t-K}^{t})$ 建模来学习映射 M,即 $\hat{\boldsymbol{y}}_t=f(\boldsymbol{X}_{t-K}^{t},\hat{\boldsymbol{Y}}_{t-K}^{t-1})$。

如图 1 所示,给定当前帧 $\boldsymbol{x}_t$,在编码模块下半部分的输入为图像—掩膜对 $[\boldsymbol{x}_t,\hat{\boldsymbol{y}}_{t-1}]\in\mathbf{R}^{H\times W\times(3+1)}$,即当前 t 帧图像和前一帧视频目标的二进制掩膜的拼接;而上半部分的输入为 $t-1$ 到 $t-K$ 的帧图像与掩膜,即 $\{[\boldsymbol{x}_{t-K},\hat{\boldsymbol{y}}_{t-K}],\cdots,[\boldsymbol{x}_{t-1},\hat{\boldsymbol{y}}_{t-1}]\}$。两个分支都使用 ResNet50(He 等,2016)为基础网络,网络权重由 ImageNet(Deng 等,2009)预训练模型所得,权重在同一实例的每个网络之间共享。在得到 Res5 阶段的结果后,使用全局卷积模块(Peng 等,2017)输出图像特征 $\{\boldsymbol{z}_t,\boldsymbol{z}_{t-K},\cdots,\boldsymbol{z}_{t-1}\}$,提升了网络的分类和密集像素定位能力,且可提高处理效率。按 Peng 等人(2017)的方法,使用 $k=7$ 作为卷积核尺寸。$\boldsymbol{z}_t\in\mathbf{R}^{h\times w\times C_1}$ 中 $h\times w$ 是特征尺寸大小,C_1 为特征的通道数(此处 C_1 设为 512),其他特征维度相同。最后,将这些特征输入部件图卷积网络,处理后输出到解码模块,在图 1 中,每个模块上部显示了输出的维度,如卷积 1 上部的(1/2, 64),1/2 表示 $H\times W$ 尺寸的 1/2,64 为通道数。

1.3 时空部件的图卷积网络模块

为了利用大小为 K 的时间窗口内的所有信息,可以通过连接 K 个先前帧中的特征,简单拼接获得更多目标信息,并输出到解码模块分割目标掩膜,但这将严重影响模型的效率和适应性。因此,设计了时空部件图卷积网络,学习鲁棒的时空目标外观模型。

鲁棒的时空目标模型对视频目标分割至关重要。但是,大多数现有 VOS 方法都从全局角度描述目标外观而忽略历史信息,使得算法对图像的重大外观变化高度敏感,导致分割失败。为此,设计了一个基于部件(节点)的结构化图表示模型,并使用图卷积网络学习目标外观模型。具体来说,为了简化和提高效率,依照 Cui 等人(2019)的方法将特征图 $\{\boldsymbol{z}_t,\boldsymbol{z}_{t-K},\cdots,\boldsymbol{z}_{t-1}\}$ 的每个 $1\times1\times C_1$ 密集网格视为图像特征部件。

为了表示时空目标模型,在具有 $N=h\times w$ 个部

件(节点)和 K(即 $t-K$, …, $t-1$)帧上构建了一个无向时空部件图 $\boldsymbol{G}_{\mathrm{ST}}=(\boldsymbol{V},\boldsymbol{E})$,帧 K 具有帧内和帧间部件(节点)的关系(Yan 等,2018)。$\boldsymbol{V}$ 和 $\boldsymbol{E}$ 是无向图中的节点集和边集。其中,节点集 $\boldsymbol{V}=\{v_{kn}\mid k=t-1,\cdots,t-K;n=1,\cdots,N\}$ 包含所有 K 中的节点,其中 $\boldsymbol{F}(v_{kn})$ 为特征向量。边集 $\boldsymbol{E}$ 包含两类边。第 1 类是空间边 $\boldsymbol{E}_{\mathrm{S}}=\{v_{ki}v_{kj}\mid 1\leqslant i,j\leqslant N,i\neq j\}$,表示每一帧图像特征中帧内节点之间的关系,鉴于图像中目标部件具有各种变化,会出现各种相互关系,本文采用完全连接图来描述空间关系。第 2 类为时序边 $\boldsymbol{E}_{\mathrm{T}}$,表示帧间节点之间的关系,将连续帧中具有相同位置的部件(节点)连接,即 $\boldsymbol{E}_{\mathrm{T}}=\{v_{kj}v_{(k+1)j}\}$,可视为一个特定部件随时间的跟踪轨迹。

基于上述无向时空图,使用图卷积网络对其中节点之间关系进行处理,即采用与 Kipf 和 Welling(2016)类似的图卷积实现。首先,基于图 $\boldsymbol{G}_{\mathrm{ST}}$ 关系确定邻近矩阵 $\boldsymbol{A}$ 的权重;其次,将邻近矩阵和特征矩阵 $\boldsymbol{H}^{(0)}$ 表示为图卷积网络的输入,而图卷积网络的输出为更新 $\boldsymbol{H}^{(l+1)}$,具体为

$$\boldsymbol{H}^{(l+1)}=\delta(\boldsymbol{D}^{-1/2}\hat{\boldsymbol{A}}\boldsymbol{D}^{-1/2}\boldsymbol{H}^{(l)}\boldsymbol{\Theta}^{(l)})\quad l=0,1,\cdots,l-1 \tag{2}$$

$$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I},\quad \boldsymbol{D}_{ii}=\sum_{j}\hat{\boldsymbol{A}}_{ij} \tag{3}$$

式中,$\boldsymbol{\Theta}$ 是需要训练的特定层的权重矩阵,$\boldsymbol{I}$ 为单位矩阵,δ 为非线性激活函数 ReLU 。

随后,使用两层图卷积网络,输出矩阵为 $\{\hat{\boldsymbol{z}}_{t-K},\cdots,\hat{\boldsymbol{z}}_{t-1}\}$,其中 $\hat{\boldsymbol{z}}_{t-1}\in\mathbf{R}^{h\times w\times C_2}$,如图 1 所示,$C_2=256$。最后,通过最大池化(MaxPooling( ))聚合时空部件特征 $\boldsymbol{z}_{\mathrm{ST}}=MaxPooling([\hat{\boldsymbol{z}}_{t-K},\cdots,\hat{\boldsymbol{z}}_{t-1}])$。

时空部件的图卷积网络模块的输入有两个部分,已经生成了时空部件特征模型 $\boldsymbol{z}_{\mathrm{ST}}$,而另一个部分为当前帧图像特征模型。通过分析,可以看出当前帧内的特征部件(节点)也有关系。因此,构建一个无向空间部件图 $\boldsymbol{G}_{\mathrm{S}}$,$\boldsymbol{G}_{\mathrm{S}}$ 与 $\boldsymbol{G}_{\mathrm{ST}}$ 相似,不同之处只在于帧的数量,即 $\boldsymbol{G}_{\mathrm{S}}$ 的图像帧为 1,而 $\boldsymbol{G}_{\mathrm{ST}}$ 的图像帧为 K。然后,采用与上述两层图卷积网络相同的处理步骤,获得空间部件特征 $\boldsymbol{z}_{\mathrm{S}}\in\mathbf{R}^{h\times w\times C_2}$。接下来,将时空部件特征和空间部件特征进行通道对齐,拼接为一个整体特征。此时,使用全局卷积模块将这个特征的两个部分进行特征匹配。这一模块中所有卷积层产生的特征图的通道都为 256,输出特征为 $\boldsymbol{Z}$。

时空部件特征模型和空间部件特征模型的输出特征具有不同的特性,采用注意机制(Hu 等,2018b)为所有的特征分配不同的权重,即特征通道选择。形式上,通过非线性变换将特征 $\boldsymbol{Z}$ 变换为 $\hat{\boldsymbol{Z}}$,具体为

$$\hat{\boldsymbol{Z}}=\boldsymbol{Z}\otimes\boldsymbol{W},\quad \boldsymbol{W}=\varphi(\boldsymbol{\theta}_2\psi(\boldsymbol{\theta}_1 f_{\mathrm{GAP}}(\boldsymbol{Z}))) \tag{4}$$

式中,$\otimes$,φ,ψ,f_{GAP} 分别表示逐通道乘法、sigmoid 激活函数、ReLU 激活函数和全局平均池化,$\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 为卷积层权重。为了优化特征图,引入残差模块(Wug 等,2018;Peng 等,2017)提升边缘细节。

1.4 解码模块

解码模块将时空图像特征做为输入,并与编码模型中的当前帧中生成的特征进行连接(图 1 中下部连接虚线),产生图像帧中目标掩码输出。根据编码模型中 ResNet50 特征图的尺寸,可以分为 5 个阶段,不同的阶段具有不同的识别能力,从而导致不同的一致性表现。在较低阶段,网络对较精细的空间信息进行编码,但是没有空间上下文指导,且处理的视野较小,使得语义一致性较差。而较高阶段时,处理视野较大,具有很强语义一致性,但预测的空间像素比较粗糙,这样可以结合其优势,使用平滑精细模块 SRM(图 1 中右上部),加入注意机制模块 CAB(图 1 中右上部(Yu 等,2018)),合并多尺度上下文中相邻阶段的特征。

本文方法的解码模块由 3 个 SRM、1 个卷积层和 1 个 softmax 层组成。其中,SRM 有两个输入,一个是从上一阶段特征,进行了两倍上采样;另一个是从编码模块相同阶段特征,第一步使用 3 × 3 卷积层,作用是将通道数统一为 256。中间使用两个残差模块优化特征图,并通过 CAB 合并两个特征图。而 CAB 与上节特征通道选择的操作相似,不同之处在于两个特征图拼接之后再进行加操作。接下来,与 Wug 等人(2018)的方法类似,softmax 层之后掩膜输出的尺寸为输入图像的 0.25 倍,每一个目标都生成一个两通道掩膜图。

1.5 模型训练与推理

本文模型的编码模块是 ResNet50,该图像在 ImageNet 上进行了图像标注任务的预训练。在时空部件的图卷积网络模块操作之后,将特征压缩为具有 256 个通道并输入到解码模块。训练过程主要包括两个阶段(Wug 等,2018)。首先使用从静态图像生成的仿真图像对进行网络预训练。将真实图像和目标掩膜作为编码模块的 K 帧图像($K=1$),将真

实图像的仿真图像和目标掩膜作为编码模块的下部分输入。之后,在视频目标分割数据集上微调此预训练模型,即采用 DAVIS-2016 和 DAVIS-2017 的训练数据集,分辨率为 720 × 480 像素。为了更好地估计训练中在测试时发生的掩膜错误传播,将时间窗口大小 K 设置为 3,即使用来自视频的随机时间索引的 $K+1$ 个连续目标帧,最后一个图像帧作为分割的当前帧。此外,使用最小化交叉熵损失,采用 Adam 优化器(Kingma 和 Ba,2014),以 1E-5 的学习率训练模型,训练与测试在单个 NVIDIA GeForce 1080 Ti GPU 上进行。

推理目标分割中,用半监督方式给出第 1 帧的真实掩膜,依次估计其余帧的掩膜。初始化时,将第 1 帧重复 K 次做为参考帧与掩膜,在实验中,K 设置为 3。在视频目标分割过程中,使用间隔 3 来更新参考帧图像与掩膜,可以有效记忆历史信息。此外,对于每个间隔帧,删除一个样本,再添加新的样本。这样可减少编码模块特征计算内存和时间,使得推理更加高效。

2 实验结果与分析

为了评估本文方法的有效性,首先对方法的组件进行分析,定量结果如表 1 所示;其次在两个具有挑战性的数据集 DAVIS-2016 (Perazzi 等,2016)和 DAVIS-2017 (Pont-Tuset 等,2017)上进行实验验证,并与最近相关视频目标分割方法进行比较。

表 1 本文方法不同组件的定量结果

Table 1 The quantitative results of different components of the proposed method

方法	*J-M* 和 *F-M*/%	*J-M*/%	*F-M*/%	速度/(s/帧)
参考帧	82.3	81.9	82.7	**0.23**
$K=1$	82.0	81.6	82.3	0.28
$K=3$	85.3	84.6	85.9	0.39
$K=3$(w/o PG)	82.6	82.2	82.9	0.32
$K=5$	**85.6**	**84.9**	**86.2**	0.86
+CRF	83.8	84.7	82.9	3.13

注:加粗字体表示各列最优结果;+CRF 表示使用密集条件随机场。

2.1 评估指标

对视频目标分割的效果采用区域相似度和轮廓相似度指标进行验证(Perazzi 等,2016)。

区域相似度用于测量错误分类的图像像素数和与分割算法匹配的像素。使用 Jaccard 相似度进行量化,其值为预测目标分割掩膜 $\boldsymbol{R}_{\mathrm{M}}$ 与真实掩膜 $\boldsymbol{R}_{\mathrm{G}}$ 之间的并交比(intersection over union,IoU),即

$$J=\frac{\boldsymbol{R}_{\mathrm{M}}\cap\boldsymbol{R}_{\mathrm{G}}}{\boldsymbol{R}_{\mathrm{M}}\cup\boldsymbol{R}_{\mathrm{G}}} \tag{5}$$

轮廓相似度以 F 度量表示,F 度量由精度和召回率计算。假设分割掩膜 $\boldsymbol{R}_{\mathrm{M}}$ 表示一组闭合轮廓 $c(\boldsymbol{R}_{\mathrm{M}})$,则基于预测目标分割掩膜轮廓 $c(\boldsymbol{R}_{\mathrm{M}})$ 和真实掩膜轮廓 $c(\boldsymbol{R}_{\mathrm{G}})$ 可以计算轮廓精度和轮廓召回率,进而获得轮廓相似度的 F 度量。具体为

$$F=\frac{2P_CR_C}{P_C+R_C} \tag{6}$$

式中,P_C 表示轮廓精度,R_C 表示轮廓召回率。

在两个数据集上的实验中,在区域相似度和轮廓相似度上计算了 6 个统计数据:Jaccard 相似度平均值(Jaccard similarity—mean,*J-M*),Jaccard 相似度召回率(Jaccard similarity—recall,*J-R*),Jaccard 相似度衰减率(Jaccard similarity—decay,*J-D*);F 度量平均值(F-measure—mean,*F-M*),F 度量召回率(F-measure—recall,*F-R*)和 F 度量衰减率(F-measure—decay,*F-D*)。平均值是视频序列中所有帧的平均值,为了更好地进行定量比较,实验中计算了"*J-M* 和 *F-M*"的平均值"*J-M&F-M*";召回率计算帧的比例,其分割得分高于 0.5;衰减率首先将所有视频帧分为 4 个片段,然后计算最后一个视频片段与第一个视频片段之间的得分差异。

2.2 组件分析与结果

为了验证本文方法在训练和测试过程中的性能,通过实验对本文方法进行组件研究,分析本文方法中的网络输入和后处理组件,实验在 DAVIS-2016 数据集上进行,实验结果如表 1 所示。可以看出:1)在网络输入方面,将网络中编码模块的上部分输入分别使用参考帧(第 1 帧)、前 $K=1$ 帧、前 $K=3$ 帧(本文方法)和前 $K=5$ 帧进行代替。结果表明,由于使用了时空部件图模型与平滑精细模块,与 RGMP 方法相比,视频目标分割性能均有所提升,且随着前 K 帧的增加,性能也会提升,但处理时间会变长。因此,本文使用 $K=3$ 作为参数与其他方法进行比较。为了分析部件图模型对整个方法的作用,分别进行采用部件图模型($K=3$)和不采用部

件图模型(K = 3 w/o PG(part-based graph))实验。w/o PG 的模型与本文方法相似,只是删除了图 1 中的部件图模型,而在图 1 左上部分的共享卷积模块之后,将 $t-1$、$t-2$ 和 $t-K$ 的特征图以通道方向进行拼接,然后形成另外(1/32, 256)的特征图;而在图 1 左下部分的共享卷积模块中,将全局卷积模块的输出为(1/32, 256)的特征图;这样就可以将两个时空特征图进行拼接,之后的模块不变。结果表示,不采用部件图模型(K = 3 w/o PG)的方法在处理速度方面会有一些提升,但是 *J-M* & *F-M* 降低 2.7%,*J-M* 和 *F-M* 均受到影响。进一步分析可以看出,简单的多帧特征融合并不能有效利用时空信息,如仅使用参考帧的性能与 K = 3 w/o PG 的性能相似,但前者的处理速度更优,而部件图模型可生成鲁棒的外观特征。2)在后处理组件方面,使用密集条件随机场(conditional random field, CRF)作为后处理部分细化分割结果(使用 K = 3)时,*J-M*提高 0.1%,*F-M*下降 3%,可能是由于 CRF 对精细细节过于敏感。因此,本文不使用 CRF 进行后处理。

2.3 DAVIS-2016 数据集与结果分析

DAVIS-2016 是视频单目标分割中广泛使用的基准数据集,包含 50 个视频,其中 30 个用于训练,20 个用于测试。共 3 455 个视频帧,每个帧带有一个单一的目标掩膜。由于视频处理的计算复杂性,数据集中的序列的时间范围比较短(约 2～4 s),但包括了视频序列中遇到的所有主要挑战:例如背景杂波、目标运动、边缘模糊、相机抖动和视线范围之外等问题。本文方法在视频 480 p 分辨率下进行训练与测试,与 SegFlow(Cheng 等,2017)、MaskTrack(Perazzi 等, 2017)、OSVOS(Caelles 等, 2017)、OSVOS-S(Maninis 等,2019)、OnAVOS(Newswanger 和 Xu,2017)、PLM(pixel-level matching)(Yoon 等,2017)、VPN(video propagation networks)(Jampani 等,2017)、OSMN(object segmentation via network modulation)(Yang 等,2018)、RGMP(Wug 等,2018)和 FEELVOS(fast end-to-end embedding learning for video object segmentation)(Voigtlaender 等,2019)等 12 种主流的半监督方法进行比较。其中,SegFlow、MaskTrack、OSVOS、OSVOS-S 和 OnAVOS 使用在线微调方法。为了显示各算法的最佳性能,直接引用基准网站或论文中发布的数字,结果如表 2 所示。

表 2　不同方法在 DAVIS 2016 验证集中的定量指标比较

Table 2　Comparison of quantitative indexes in DAVIS 2016 validation set among different methods

方法	在线微调	*J-M* & *F-M*/%	区域相似度			轮廓准确度			速度/(s/帧)
			J-M/%	*J-R*/%	*J-D*/%	*F-M*/%	*F-R*/%	*F-D*/%	
SegFlow	√	76.1	76.1	90.6	12.1	76	85.5	10.4	7.9
MaskTrack	√	77.6	79.7	93.1	8.9	75.4	87.1	9.0	12.0
OSVOS	√	80.2	79.8	93.6	14.9	80.6	92.6	15.0	9.0
OSVOS-S	√	**86.6**	85.6	**96.8**	5.5	**87.5**	**95.9**	8.2	4.5
OnAVOS	√	85.5	**86.1**	96.1	5.2	84.9	89.7	**5.8**	13.0
STCNN	√	83.8	83.8	96.1	**4.9**	83.8	91.5	6.4	3.9
PLM	×	66.4	70.2	86.3	11.2	62.5	73.2	14.7	0.28
VPN	×	67.9	70.2	82.3	12.4	65.5	69.0	14.4	0.63
OSMN	×	73.5	74	87.6	9.0	72.9	84.0	10.6	0.13
RGMP	×	81.8	81.5	91.7	10.9	82.0	90.8	10.1	0.14
FEELVOS	×	81.7	81.1	90.5	13.7	82.2	86.6	14.1	0.55
DTN	×	83.6	83.7	–	–	83.5	–	–	**0.07**
OnAVOS	×	–	72.7	–	–	–	–	–	–
本文	×	85.3	84.6	92.1	9.7	85.9	92.5	9.5	0.39

注:加粗字体表示各列最优结果,“–”表示原论文没有提供结果。

从表2可以看出,在不使用在线微调的方法中,本文方法明显优于其他方法。与在线微调的方法相比,本文方法无需通过在线微调或后处理获得较好的准确性。本文方法的 *J-M & F-M* 得分为85.3%,高于使用在线微调的SegFlow、MaskTrack和OSVOS方法。OnAVOS方法使用基于网络的置信度和空间配置选择的训练样本在线更新网络,需要大量时间和计算资源。而本文方法效率更高,在训练和测试阶段均不需要光流。VOS方法通常以准确性或速度为重点、以牺牲另一种方法的性能为代价。本文方法既保持尽可能高的准确性,同时尝试专注于速度。在不使用在线微调的方法中,RGMP的 *J-M & F-M* 得分为81.8%,比本文方法低3.5%。在区域相似度和轮廓准确度方面,与效果最好的对比方法相比,本文方法的 *J-M* 提高3.1%,*J-D* 提高3.7%。不使用在线微调的OnAVOS的 *J-M* 只有72.7%,而本文方法的 *J-M* 为84.6%。为了显示时空部件图模型的性能,与STCNN(spatio-temporal convolutional neural network)(Xu等,2019)和DTN(dynamic targeting network)(Zhang等,2019)两种时空CNN模型进行比较,其中STCNN方法进行了在线微调,结果如表2所示。可以看出,本文方法在 *J-M* 和 *F-M* 方面都好于STCNN和DTN方法,具有较好的时空特性。

在测试时间方面,由于在不同平台上开发和评估不同算法,加之GPU类型多有不同,很难公平地比较运行时间效率,OSVOS、OnAVOS、RGMP、OSMN和FEELVOS方法使用原文献代码的运行结果,其他方法的运行时间来自RGMP(Wug等,2018)。从表2可以看出,本文方法处理速度较高,每一帧处理时间为0.39 s,且 *J-M & F-M* 性能比其他方法最高的提升了1.7%。

2.4 DAVIS-2017数据集与结果分析

DAVIS 2017是DAVIS 2016的扩展,视频序列中每帧带有多个目标,有60个训练视频和30个测试视频。在训练中,只单独训练每个目标,生成二进制掩膜。在测试中,使用网络分别获得每个目标的软概率图,并对帧中所有目标的图像使用softmax操作进行后处理,以生成多目标分割掩膜。

为进一步验证本文方法在多目标数据集上的性能,与OSVOS、OnAVOS、OSVOS-S、OSMN、STCNN、DTN、VideoMatch(Hu等,2018a)、RANet(ranking attention network)(Wang等,2019)、MaskRNN(Hu等,2018c)和RGMP等10种主流方法进行比较,各算法的区域相似度和轮廓准确度的平均值如表3所示。本文方法利用长期(long-term)时空特性,不需在线微调,获得了良好性能,有效证明了长期时空信息对视频目标分割的重要性。可以看出,OSVOS、OnAVOS和OSVOS-S在使用在线微调的情况下,各项指标仍低于本文方法。本文方法的 *J-M & F-M* 得分达到了68.6%,比RGMP和对比算法中性能最高的DTN方法分别提高了1.9%和1.2%。*J-M* 指标为65.6,与不需在线微调的OSVOS和OnAVOS方法相比,提高了26%。

表3 不同方法在DAVIS 2017验证集中的定量指标比较

Table 3 Comparison of quantitative indexes in DAVIS 2017 validation set among different methods

/%

方法	在线微调	*J-M & F-M*	*J-M*	*F-M*
OSVOS	√	60.3	56.6	63.9
OnAVOS	√	65.3	61.6	69.1
OSVOS-S	√	67.8	64.5	71.1
STCNN	√	61.7	58.7	64.6
OnAVOS	×	–	39.5	–
OSMN	×	54.8	52.5	57.1
VideoMatch	×	62.4	56.5	68.2
RANet	×	65.7	63.2	68.2
MaskRNN	×	–	60.5	–
DTN	×	67.4	64.2	70.6
RGMP	×	66.7	64.8	68.6
本文	×	**68.6**	**65.6**	**71.5**

注:加粗字体表示各列最优结果,“–”表示原论文没有提供结果。

2.5 定性分割结果

本文方法在DAVIS 2016和DAVIS 2017数据集中的定性分割结果如图2和图3所示。在图2显示了DAVIS-2016验证集中breakdance、drift-chicane、bmx-trees和motocross-jump等4个代表性视频序列。其中,红色掩膜表示像素级的分割结果。结果表明,RGMP会丢失部分目标,而本文方法能够在遮挡,变形,快速运动和背景杂波等多种挑战情况下分割目标,获得了较好性能。图3显示了DAVIS-2017验证集中india(3个目标)、horsejump-high(2个目

图 2　本文方法与 RGMP 在 DAVIS-2016 验证集上的定性分割结果比较

Fig. 2　Comparison of the qualitative segmentation results in DAVIS-2016 validation set between RGMP and ours ((a) RGMP; (b) ours)

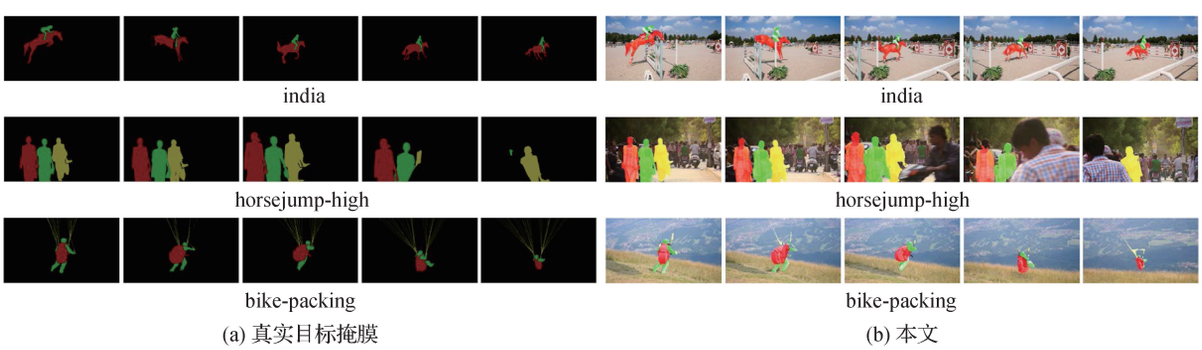


图 3　本文方法在 DAVIS-2017 验证集上的定性分割结果

Fig. 3　The qualitative segmentation results in DAVIS-2017 validation set of our method((a) ground truth; (b) ours)

标)和 bike-packing(3 个目标)等 3 个代表性视频序列,图像中多个目标以不同的颜色突出显示。可以看出,本文方法在许多挑战性场景(例如外观更改)中具有良好的鲁棒性。

3　结　论

本文研究了视频目标分割中鲁棒时空目标模型问题,提出时空部件图卷积网络的视频目标分割方法缓解视觉目标漂移情况,且不需在线微调和后处理。构建了编码—时空部件图—解码 VOS 框架,在 $K+1$ 帧图像与掩膜编码之后,为了解决视频序列场景变化问题,设计了时空部件图卷积网络,利用历史帧信息生成时空部件特征,并借助注意机制构建更好的特征表示。在解码模型中使用平滑精细模块处理不同尺度目标的分割。

在两个流行的数据集 DAVIS 2016 和 DAVIS 2017 上进行实验。与其他 VOS 方法相比,本文方法的性能有明显提升。与使用在线微调的方法相比,本文方法无需进一步在线微调或后处理即可获得较好的准确性,效率更高,在训练和测试阶段都不需要光流,且在目标的部件遮挡和细节边缘问题上达到良好效果。与最近的时空 CNN 模型进行对比,也获得了较好性能。此外,对本文方法进行输入、后处理和不采用部件图模型的组件分析,验证了本文方法有效性。

本文主要在视频目标分割方面进行研究,构建时空图卷积网络,有效生成时空特征图。但在多目标协同处理方面仍存在不足。下一步工作将探索多目标时空特征模型,充分挖掘多目标之间的关系,提升多目标视频目标分割性能。

参考文献(References)

Bao L C, Wu B Y and Liu W. 2018. CNN in MRF: video object seg-

mentation via inference in a CNN-based higher-order spatio-temporal MRF//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5977-5986 [DOI: 10.1109/CVPR.2018.00626]

Caelles S, Maninis K K, Pont-Tuset J, Leal-Taixé L, Cremers D and Van Gool L. 2017. One-shot video object segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5320-5329 [DOI: 10.1109/CVPR.2017.565]

Cheng J C, Tsai Y H, Wang A J and Yang M H. 2017. SegFlow: joint learning for video object segmentation and optical flow//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 686-695 [DOI: 10.1109/ICCV.2017.81]

Cui Z, Cai Y Y, Zheng W M, Xu C Y and Yang J. 2019. Spectral filter tracking. IEEE Transactions on Image Processing, 28(5): 2479-2489 [DOI: 10.1109/TIP.2018.2886788]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Grundmann M, Kwatra V, Han M and Essa I. 2010. Efficient hierarchical graph-based video segmentation//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 2141-2148 [DOI: 10.1109/CVPR.2010.5539893]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu J, Shen L and Sun G. 2018a. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Hu Y T, Huang J B and Schwing A G. 2018b. VideoMatch: matching based video object segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 56-73 [DOI: 10.1007/978-3-030-01237-3_4]

Hu Y T, Huang J B and Schwing A G. 2018c. Maskrnn: instance level video object segmentation [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1803.11187.pdf

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A and Brox T. 2017. FlowNet 2.0: evolution of optical flow estimation with deep networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1647-1655 [DOI: 10.1109/CVPR.2017.179]

Jampani V, Gadde R and Gehler P V. 2017. Video propagation networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3154-3164 [DOI: 10.1109/CVPR.2017.336]

Khoreva A, Benenson R, Ilg E, Brox T and Schiele B. 2019. Lucid data dreaming for video object segmentation. International Journal of Computer Vision, 127(9): 1175-1197 [DOI: 10.1007/s11263-019-01164-6]

Kingma D P and Ba J L. 2014. Adam: a method for stochastic optimization [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1412.6980.pdf

Kipf T N and Welling M. 2016. Semi-supervised classification with graph convolutional networks [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1609.02907.pdf

Li X X and Change L C. 2018. Video object segmentation with joint reidentification and attention-aware mask propagation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 93-110 [DOI: 10.1007/978-3-030-01219-9_6]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Luiten J, Voigtlaender P and Leibe B. 2018. PReMVOS: Proposal-Generation, Refinement and Merging for Video Object Segmentation//Proceedings of Asian Conference on Computer Vision (ACCV). Perth, Australia: Springer: 565 – 580 [doi:10.1007/978-3-030-20870-7_35]

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

Maninis K K, Caelles S, Pont-Tuset J and Van Gool L. 2018. Deep extreme cut: from extreme points to object segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 616-625 [DOI: 10.1109/CVPR.2018.00071]

Newswanger A and Xu C L. 2017. One-shot video object segmentation with iterative online fine-tuning [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1706.09364.pdf

Peng C, Zhang X Y, Yu G, Luo G M and Sun J. 2017. Large kernel matters – improve semantic segmentation by global convolutional network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1743-1751 [DOI: 10.1109/CVPR.2017.189]

Perazzi F, Khoreva A, Benenson R, Schiele B and Sorkine-Hornung A. 2017. Learning video object segmentation from static images//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3491-3500 [DOI: 10.1109/CVPR.2017.372]

Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M and

Sorkine-Hornung A. 2016. A benchmark dataset and evaluation methodology for video object segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 724-732 [DOI: 10.1109/CVPR.2016.85]

Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A and Van Gool L. 2017. The 2017 davis challenge on video object segmentation [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1704.00675.pdf

Voigtlaender P, Chai Y, Schroff F, Adam H, Leibe B and Chen L C. 2019. FEELVOS: fast end-to-end embedding learning for video object segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 9473-9482 [DOI: 10.1109/CVPR.2019.00971]

Voigtlaender P and Leibe B. 2017. Online adaptation of convolutional neural networks for video object segmentation//Proceedings of the British Machine Vision Conference (BMVC). London, UK: BMVA Press: 116.1-116.13 [DOI: 10.5244/C.31.116]

Wang Z Q, Xu J, Liu L, Zhu F and Shao L. 2019. Ranet: ranking attention network for fast video object segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3977-3986 [DOI: 10.1109/ICCV.2019.00408]

Wug O S, Lee J Y, Sunkavalli K and Kim S J. 2018. Fast video object segmentation by reference-guided mask propagation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7376-7385 [DOI: 10.1109/CVPR.2018.00770]

Xiao H X, Feng J S, Lin G S, Liu Y and Zhang M J. 2018. MoNet: deep motion exploitation for video object segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1140-1148 [DOI: 10.1109/CVPR.2018.00125]

Xu K, Wen L Y, Li G R, BO L F and Huang Q M. 2019. Spatiotemporal CNN for video object segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1379-1388 [DOI: 10.1109/CVPR.2019.00147]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition [EB/OL]. [2020-07-08]. https://arxiv.org/pdf/1801.07455.pdf

Yang L J, Wang Y R, Xiong X H, Yang J C and Katsaggelos A K. 2018. Efficient video object segmentation via network modulation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6499-6507 [DOI: 10.1109/CVPR.2018.00680]

Yao R, Lin G S, Xia S X, Zhao J Q and Zhou Y. 2020. Video object segmentation and tracking: a survey. ACM Transactions on Intelligent Systems and Technology, 11(4): #36 [DOI: 10.1145/3391743]

Yoon J S, Rameau F, Kim J, Lee S, Shin S and Kweon I S. 2017. Pixel-level matching for video object segmentation using convolutional neural networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2186-2195 [DOI: 10.1109/ICCV.2017.238]

Yu C Q, Wang J B, Peng C, Gao C X, Yu G and Sang N. 2018. Learning a discriminative feature network for semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1857-1866 [DOI: 10.1109/CVPR.2018.00199]

Zhang L, Lin Z, Zhang J M, Lu H C and He Y. 2019. Fast video object segmentation via dynamic targeting network//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 5581-5590 [DOI: 10.1109/ICCV.2019.00568]

Zhang Y, Chen X W, Li J, Wang C and Xia C Q. 2015. Semantic object segmentation via detection in weakly labeled video//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3641-3649 [DOI: 10.1109/CVPR.2015.7298987]

作者简介

姚睿，1982 年生，男，副教授，主要研究方向为计算机视觉、模式识别。
E-mail: ruiyao@cumt.edu.cn

夏士雄，通信作者，男，教授，主要研究方向为智能信息处理。
E-mail: xiasx@cumt.edu.cn

周勇，男，教授，主要研究方向为智能信息处理。
E-mail: yzhou@cumt.edu.cn
赵佳琦，男，副教授，主要研究方向为计算机视觉。
E-mail: jiaqizhao@cumt.edu.cn
胡伏原，男，教授，主要研究方向为计算机视觉。
E-mail: fuyuanhu@mail.usts.edu.cn