

时序特征引导的视频目标分割方法

朱琳琳, 李家勋, 刘珈溪, 梁 宵
(沈阳航空航天大学 自动化学院, 沈阳 110136)

摘要: 为了解决现有目标分割方法对光流图的过度依赖问题,并充分利用视频目标特有的时间相关性和空间相关性,提出一种先利用目标与背景的时序特征信息提取感兴趣区域,后利用空间显著性特征分割目标的两阶段方法。首先,设计光流异常区域检测器,以序列图像间像素位置变化产生的光流场作为时序特征,学习目标与背景在时序特征上的差异,从而检测出目标位置,并将分割范围引导至目标所在感兴趣区域(region of interest, ROI)。然后,使用提出的DC-U²Net作为显著目标检测器分割出ROI中的显著目标,得到掩码后,完成目标分割。其中,DC-U²Net通过引入双通道RSU模块与残差连接路径,进一步提高目标掩码的精度。该方法在两个公共数据集上进行了性能测试,取得了有竞争力的结果。在DAVIS2016数据集上,J分数、F分数与G分数分别达到83.7、86.7、85.2。结果表明,针对视频目标时序特征和显著性特征的特点,在不影响实时性的情况下提出的基于时序特征引导的两阶段视频目标分割方法,提升了目标分割精确度。

关键词: 时序特征;视频目标分割;深度学习;掩码;两阶段方法

中图分类号: TP391

文献标志码: A

DOI: 10.3969/j.issn.2095-1248.2026.04.010

Temporal feature-guided method for video object segmentation

ZHU Linlin, LI Jiaxun, LIU Jiayi, LIANG Xiao

(College of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: To address the excessive reliance of existing object segmentation methods on optical flow maps and fully leverage the temporal and spatial correlations inherent in video objects, a two-stage method was proposed that first extracts regions of interest (ROI) using the temporal feature information of objects and backgrounds, and then segments objects using spatial saliency features. Firstly, an optical flow anomaly region detector was designed, which used the optical flow field generated by pixel position changes between sequential images as temporal features to learn the differences in temporal features between objects and backgrounds, thereby detecting object positions and guiding the segmentation scope to the ROI where objects were located. Then, the proposed DC-U²Net was used as a salient object detector to segment salient objects in the ROI, obtain masks, and complete object segmentation. DC-U²Net further improved the accuracy of object masks by introducing a dual-

收稿日期: 2025-06-09 修回日期: 2025-06-29 接受日期: 2025-06-30

基金项目: 国家自然科学基金(项目编号: 61973222); 辽宁省高校基本科研业务费专项资金资助项目(项目编号: LJ232410143058)。

作者简介: 朱琳琳(1982—), 女, 山东聊城人, 副教授, 博士, 主要研究方向为图像分割、目标检测与跟踪, E-mail: zhulinlin82@126.com。

channel RSU module and residual connection paths. The proposed method was tested on two public datasets and achieved competitive results. On the DAVIS2016 dataset, the J-score, F-score, and G-score reached 83.7, 86.7, and 85.2 respectively. Aiming at the characteristics of temporal features and saliency features of video objects, the proposed two-stage video object segmentation method guided by temporal features-guided improves object segmentation accuracy without compromising real-time performance.

Key words: temporal feature; video object segmentation; deep learning; masks; two-stage method

视频目标分割是视觉感知领域的重点研究方向,其本质是通过对视频内容的理解^[1]从图像序列中检测出处于运动状态的目标。基于视觉的运动目标检测在工业、农业、交通安防及军事等领域被广泛应用,同时也对目标跟踪^[2]、行为识别^[3]、视频生成^[4]等其他计算机视觉任务的研究起到促进作用。

传统视频目标分割方法的核心思想是从连续的视频帧中找出发生明显变化的像素,主要包括帧差法^[5]、光流法^[6]、背景减除法^[7]。帧差法具有简单高效的特点, Saleemi等^[8]提出的高效帧差法以最小帧数实现运动目标检测; Barnich等^[9]提出了一种基于颜色特征的视觉背景提取器(visual background extractor, ViBe)背景减除算法; Bilodeau等^[10]提出了一种基于纹理特征的局部二值相似模式(local binary similarity patterns, LBSP)算法,实现了比 ViBe 更好的实时性。光流法对两种相机状态(运动和静止)均有较强的适应能力,其中霍恩-舒克(Horn-Schunck, HS)光流法^[11]与卢卡斯-卡纳德(Lucas-Kanade, LK)光流法^[12]均为运动目标检测领域的经典方法。

深度学习技术的快速发展使完全基于深度神经网络的运动目标分割方法逐渐成为主流。Jain等^[13]提出第一个端到端的可训练框架,实现视频前景目标的分割。该框架通过一个双流深度全卷积网络将外观与运动信息融合在一起,在多个数据集上实现了超过单纯应用光流估计方法的效果。Cheng等^[14]将 ResNet-101 与 FlowNetS 相结合提出 SegFlow 网络结构,通过在两个分支之间进行多尺度特征映射与双向传播,同时实现了像素级的目标

分割和光流估计,并在实验中证明了光流对前景目标分割的优化效果。Yang等^[15]基于上下文信息分离的思想,将生成的对抗框架引入到运动目标分割任务中,让一个深度神经网络利用运动目标的上下文信息来预测光流,同时训练另一个网络,试图准确分割出运动目标,使图像中可用的上下文信息尽可能减少,通过两个网络的对抗学习,它们能捕捉到运动目标与背景之间不同的运动特性。

尽管上述方法相对于传统方法已经取得了很大的进步,但还存在一些不足。传统算法的帧差法存在明显的鬼影效应;受限于背景建模的特征表达能力,背景减除法在复杂的场景中表现不佳;光流计算极易受到像素亮度变化的影响,且在不符合时间连续性假设和空间一致性假设的场景中检测效果较差;在相机静止状态下,现有的深度学习方法已经展现了强大的性能,在相机运动的状态下,结合光流估计网络的深度学习方法也可以有效利用光流图蕴含的运动信息完成检测,但现有方法的外观与运动信息使用策略过度依赖光流图。本文提出了一种两阶段方法来解决上述问题。

1 实验方法

本文主要包含两部分内容:1)提出一种基于运动信息引导的两阶段方法,先后通过光流异常区域检测器与显著目标检测器获得像素级的检测结果;2)以 DC-U²Net 作为显著目标分割网络,通过在 U²-Net 基础上引入双通道 RSU 模块与残差连接路径,解决 ROI 内像素点是否属于当前运动目标的决策问题。本文实

验流程如图 1 所示。

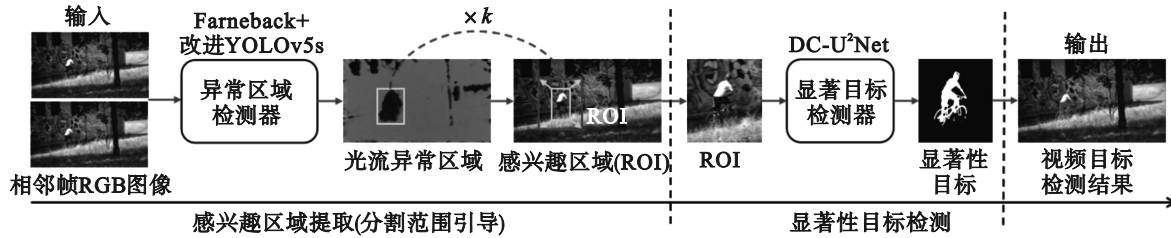


图 1 实验流程

1.1 基于光流图的感兴趣区域提取

通常情况下,视频中的显著性目标为运动目标,区分序列图像中目标与背景的运动状态是找到目标的关键,而光流图可以通过像素颜色与亮度体现出目标与背景的运动差异,因此,利用光流图中的运动信息能够确定运动目标的大致位置。实验第一阶段通过对光流异常区域的检测提取感兴趣区域,实现对视频显著目标的初步定位。

图 2 为相邻帧光流计算结果图,展示了目标被部分遮挡及运动不明显情况下的光流计算方法得到的结果,除目标边缘模糊问题外,还存在目标本身特征不一致及背景区分度低的问题。在这些情况下,传统图像处理方法不能有效地区分出目标所在的光流异常区域,所以本文在计算光流图的基础上结合深度学习网络来进行异常区域的检测。

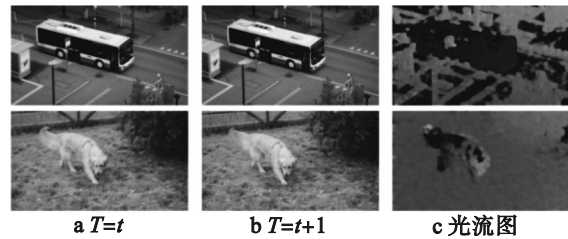


图 2 相邻帧光流计算结果图

因为 YOLOv5s 及其改进的模型具有较强的特征提取能力,故本文设计了一个以 YOLOv5s 为主体的光流异常区域检测器。光流异常区域检测器如图 3 所示。相邻两帧 RGB 图像输入到检测器后,首先通过 Farneback 算法计算得到光流图,再由改进 YOLOv5s 网络从光流图中检测出光流异常区域。本文使用轻量级的 RepViT block^[16]代替原始 YOLOv5s 主干特征提取网络中的 CBS 层,加强网络的全局特征提取能力,进一步提升检测器的性能。

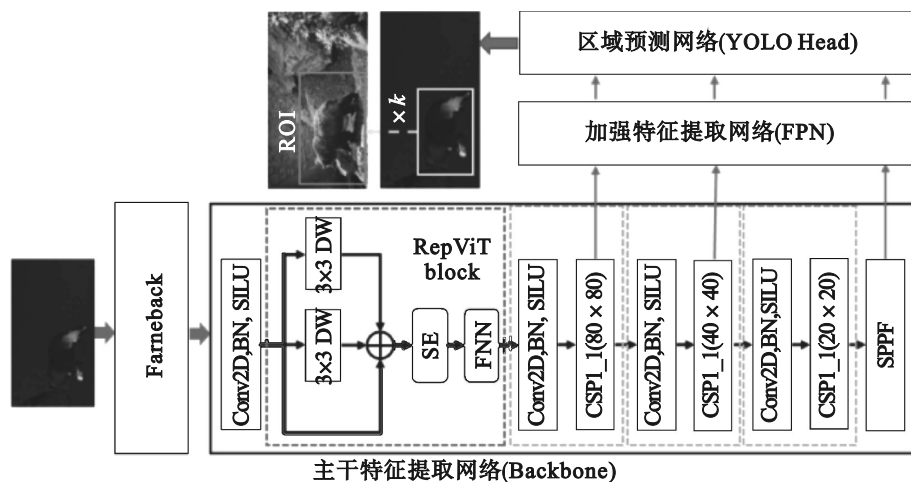


图 3 光流异常区域检测器

因为光流计算的不稳定性,所以光流图中的异常区域与真实运动目标之间存在一定的差异。为了实现对运动目标的完整覆盖,本文将检测到的异常区域进行扩大,得到运动目标所在的感兴趣区域(ROI),其中扩大系数计为 k 。

1.2 基于DC-U²Net的显著性目标分割

感兴趣区域内的显著性目标分割可以看作一个将运动目标作为前景,非运动目标区域作为背景的像素级二分类任务。在本任务中除像素颜色与亮度等特征外,空间特征也起到

了十分关键的作用。

U²-Net为显著性目标分割任务而设计,能够充分学习图像中局部细节与全局的对比信息,但在空间特征提取方面仍存在不足,因此本文在U²-Net结构基础上提出改进双通道U²-Net网络(DC-U²Net)。首先,将双通道RSU模块(DC-RSU)应用到U²-Net网络的编解码器中,增强空间特征提取能力;然后,在每组编码器与解码器的连接路径之间加入残差结构,构成残差连接路径(Res Path),优化特征融合过程。DC-U²Net结构如图4所示。

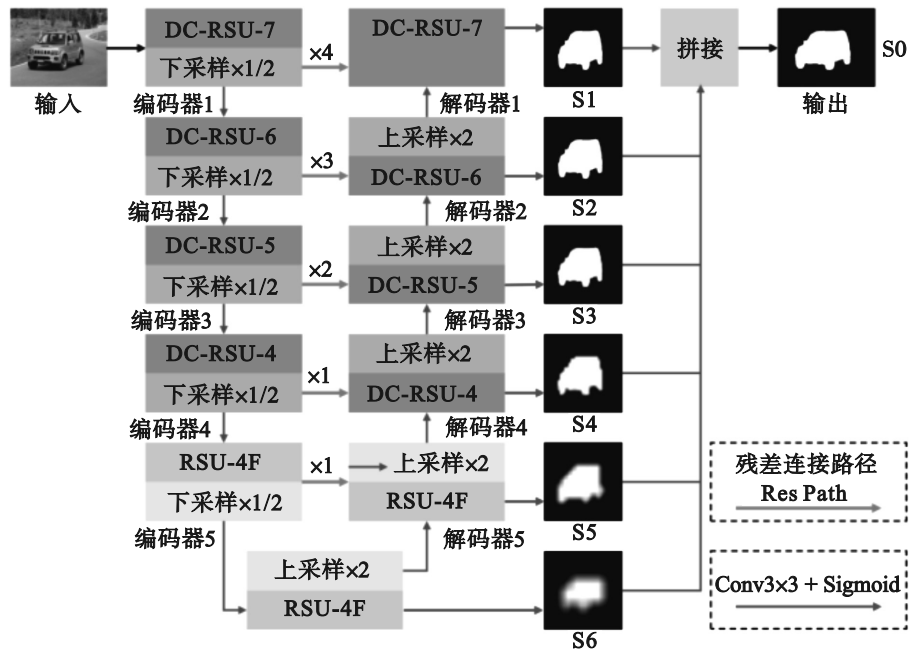


图4 DC-U²Net结构

双通道RSU模块与残差连接路径说明如下:

1) 双通道RSU模块

U²-Net网络包含5组对称的编码器与解码器,每个编解码器的主体为一个U型残差RSU模块。本文使用双通道模块(dual-channel block)^[17]替换RSU模块中用于特征提取的普通卷积层,得到双通道RSU(DC-RSU)模块,并将模块设置为不同的采样深度 L ,应用到DC-U²Net不同位置的编码器与解码器中,在保证网络能够充分提取多尺度特征的同时补充更

多的空间特征。DC-RSU模块结构如图5所示。双通道模块结构如图5a所示, C_{in} 与 C_{out} 分别为输入模块和输出模块的通道数,模块中由通道系数 α 控制输入到左右两个通道的通道数 C_{mid} ,输入的图像在通道内分别经过1~3个 3×3 卷积层的3个特征层进行拼接,并将两通道的输出特征相加后作为双通道模块的最终输出。

DC-RSU模块以卷积核大小为 3×3 、膨胀系数为2的膨胀卷积为轴,形成对称的U型结构,图像输入到编码器后,先经双通道模块进

行初步特征提取,再进行卷积下采样与上采样操作,最后将双通道模块输出的特征与上采样后输出的特征相加作为 DC-RSU 模块的最终

输出,采样层深度 $L=7$ 的 DC-RSU-7 模块结构如图 5b 所示。

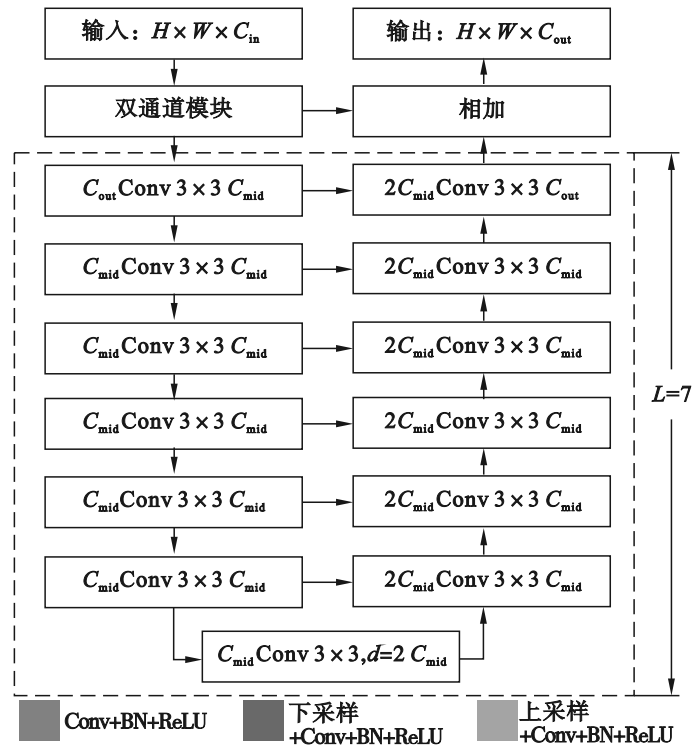
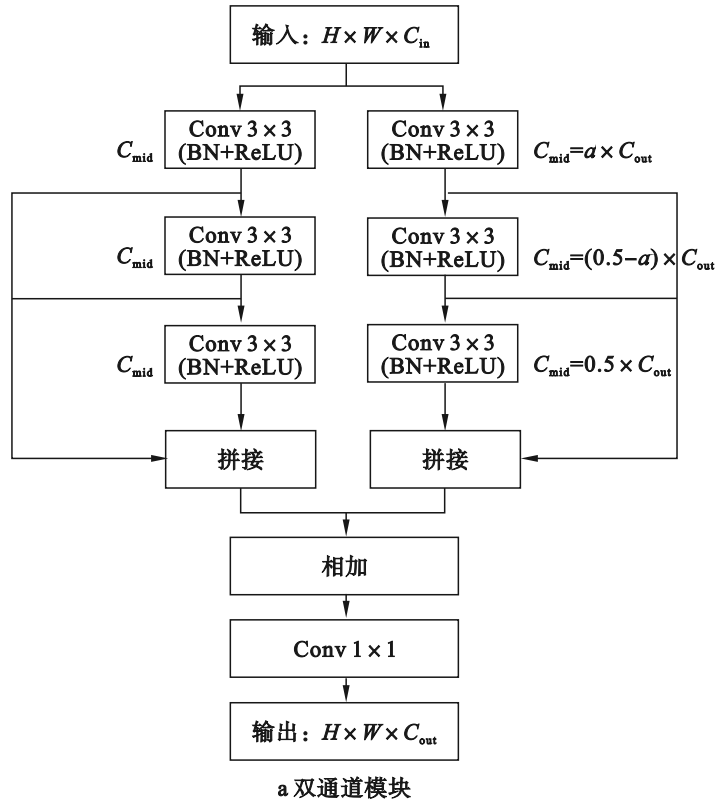


图 5 DC-RSU 模块结构

2)残差连接路径

U²-Net 网络在编码器与解码器之间应用跳跃连接结构,以保留在池化操作中损失掉的特征信息,但原始的跳跃连接存在一些缺陷。两组存在语义差异的特征直接融合,可能会影响到特征图预测的准确性。本文在 DC-U²Net 中将原始的连接路径改进为由多个残差结构

串联组成残差连接路径,以减小编码器与解码器特征连接时的语义差异。包含 4 个残差结构的残差连接路径结构如图 6 所示,每个残差结构由一个带有批量归一化层(batch normalization)与 ReLU 激活函数的 3×3 卷积层。如图 6 所示,从上至下的 5 个连接路径中分别使用 4、3、2、1、1 个残差结构。

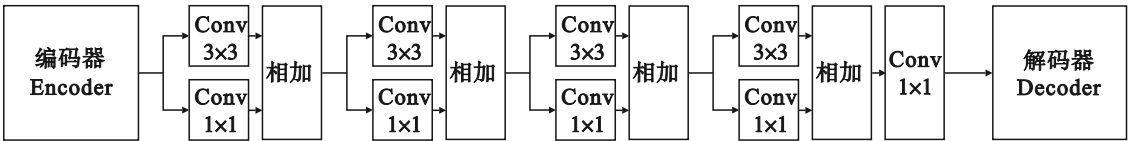


图 6 残差连接路径结构

2 实验与结果分析

2.1 实验设置

本文使用 DAVIS 2016 公共数据集制作光流异常区域检测数据集,并将其应用于光流异常区域检测器的模型训练。使用 DUTS 数据集作为 DC-U²Net 的训练集,其中 10 553 张图像用于模型训练,5 019 张图像用于测试。本文通过随机裁剪与随机水平翻转进行数据增强,并且使用混合精度训练(mixed precision training)策略进行加速。

实验选择 SegTrack-v2、DAVIS 2016 及 FBMS 作为主要测试数据集,并根据实验具体需求选择 J 分数(区域相似度,Jaccard index)、F 分数(轮廓准确度,boundary F-score)和 G 分数(J&F,J 分数与 F 分数的平均数)作为模型的定量评价指标。

2.2 感兴趣区域提取实验

本方法的第一阶段将提取出感兴趣区域

(ROI),有效的 ROI 提取会对后续的显著性目标分割起到积极的作用。本文在光流异常区域检测数据集上,对以 YOLOv5s 为主体的光流检测器进行了改进前后的对比实验。为了客观体现光流异常区域检测的精度,实验选择将检测结果与标签的交并比(intersection over union, IoU)作为定量评价指标。同时,在 $k=1.8$ 时,找到运动目标所在的 ROI,并评估 ROI 对运动目标的覆盖率(coverage)的关系,覆盖率与 ROI 的有效性成正相关,计算方法如式(1)所示。

coverage = \frac{ROI \cap R_L}{R_L} \tag{1}

式中: R_L 为目标掩码的外接矩形,即光流异常区域的标签。光流异常区域检测对比试验结果如表 1 所示,从表 1 可以看出,在主干网络中使用 RepViT block 代替原始的 CBS 层后,光流异常区域检测器的检测覆盖率提高 0.5%,并且对运动目标的平均覆盖率达到 98.2%,能够更好地检测出光流图中的异常区域,可以满足检测器的功能需求。

表 1 光流异常区域检测对比试验结果

方法	IoU	覆盖率/%
光流异常区域检测器(CBS)	0.86	97.7
本文光流异常区域检测器(RepViT block)	0.90	98.2

2.3 显著性目标分割消融实验

ROI内的显著性目标分割是本文方法的重要阶段,为了检验DC-U²Net的先进性,并验证使用双通道RSU模块及添加残差连接路径对原始U²-Net的改进效果。本实验使用J分数、F分数及G分数作为掩码精度的评价指标。本文在SegTrack-v2数据集上对模型设置了4组消融实验。U²-Net加入不同模块的消融实验结果如表2所示,由表2可以看出,在加入DC-RSU模块和残差连接路径后,模型的各项指标均有所提升。相较于原始U²-Net,DC-U²Net的整体评价指标更好,其中J分数提高3.2,F分数提高1.4,网络的综合性能得到较大改进。

表2 U²-Net加入不同模块的消融实验结果

模型	J分数	F分数	G分数
U ² -Net	76.8	83.9	80.4
U ² -Net + DC-RSU	79.7	85.2	82.5
U ² -Net + Res Path	78.9	84.5	81.8
本文DC-U ² Net	80.0	85.3	82.7

2.4 本文方法与其他方法对比

为了充分验证本文方法的表现,将本文方法与PDB^[18]、COSNet^[19]等针对无监督视频目标分割任务而设计的先进方法在DAVIS 2016与FBMS两个基准测试集中进行对比实验。

1)定量分析实验

本文方法与其他方法在DAVIS 2016测试集上定量评价结果如表3所示,本文方法通过两阶段渐进的方法来解耦目标的外观与运动信息,J分数与F分数分别达到83.7、86.7,综合表现接近IMP^[20]。但IMP在网络推理时需要将完整的视频作为输入,并从全局角度理解视频的内容,该算法为非实时分

割算法,所以本文方法在可以实时分割的方法中表现最优。

表3 本文方法与其他方法在DAVIS 2016测试集上定量评价结果

方法	发表	J分数	F分数	G分数	实时分割
PDB	ECCV 2018	77.2	74.5	75.9	
COSNet	CVPR 2019	80.5	79.4	80.0	√
A-GAME	CVPR 2019	82.0	82.2	82.1	√
FEELVOS	CVPR 2019	81.5	82.1	81.8	
STCNN	CVPR 2019	83.8	83.8	83.8	
SAT	CVPR 2020	82.6	83.6	83.1	√
WCS-Net	ECCV 2020	82.2	80.7	81.5	√
FRTM	CVPR 2020	83.6	83.4	83.5	√
F2Net	AAAI 2021	83.1	84.4	83.7	√
G-FRTM	CVPR 2021	-	-	84.3	√
FSNet	ICCV 2021	83.4	83.1	83.3	-
Z-VOS	CVPR 2021	84.5	84.6	84.6	-
IMP	AAAI 2022	84.5	86.7	85.6	
LSTA	PR 24	82.7	84.8	83.8	√
本文	-	83.7	86.7	85.2	√

注:“√”代表实时性算法,空白代表非实时性算法

2)定性分析实验

本文以各方法在FBMS测试集上的部分对比结果为例进行定性分析,其定性分析对比如图7所示。从图7可以看出,本文方法相较其他方法的细节保留更完整,得益于显著目标分割网络对图像空间特征的充分利用,DC-U²Net可以更好地把握目标与背景在图像中的相对位置关系,实现准确的目标分割。

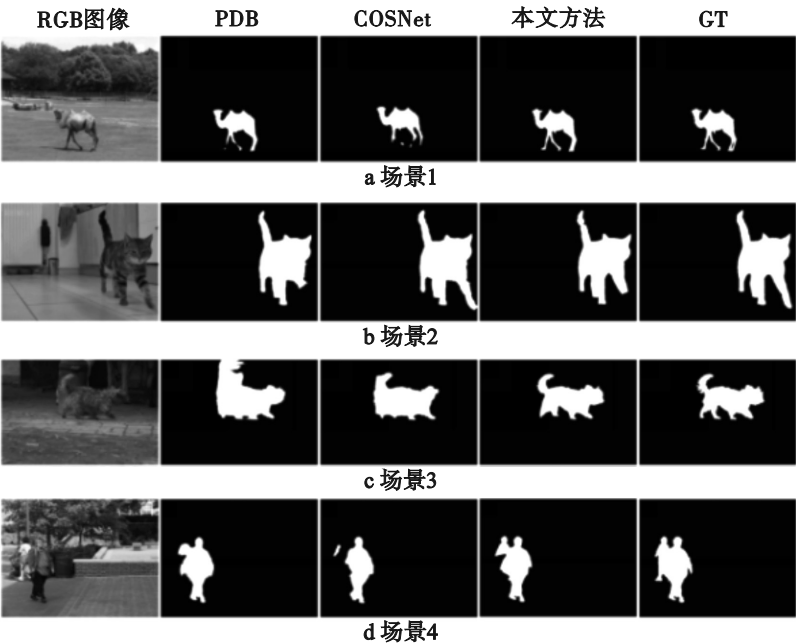


图7 本文方法与其他方法定性分析对比

3 结论

为了解决现有目标分割方法对光流图的过度依赖问题,本文提出一种两阶段渐进精炼的运动目标分割方法。该方法使用光流异常区域检测器找到光流图中的异常区域,并由此提取运动目标所在的感兴趣区域(ROI),实现对目标的初步定位,再通过显著目标分割网络对ROI中的显著性目标进行分割,最终完成运动目标分割。基于显著性的目标分割可以摆脱传统分割网络对训练数据中类别先验的高度依赖,为了使显著目标分割网络有更好的表现,本文基于U²-Net进行改进,提出了带有双通道RSU模块与残差连接路径的DC-U²Net,经消融实验验证网络性能有明显提升。本方法在DAVIS数据集上J分数、F分数和G分数为83.7、86.7和85.2,与其他方法的对比可以看出,本方法综合指标良好,能够准确找到运动目标并得到精细的目标掩码,实现了运动目标分割。

参考文献(References):

[1] Hutchinson M S, Gadepally V N. Video action understanding [J]. IEEE Access, 2021, 9: 134611–134637.

[2] Chen F, Wang X D, Zhao Y X, et al. Visual object tracking: a survey [J]. Computer Vision and Image Understanding, 2022, 222: 103508.

[3] Rahmani H, Bennamoun M, Ke Q. Human Action Recognition from Various Data Modalities: a review [J]. Cornell University, 2021, 45: 3200–3225.

[4] Chai W H, Guo X, Wang G A, et al. Stablevideo: text-driven consistency-aware diffusion video editing [C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22983–22993.

[5] Li T, Jiang B, Wu D H, et al. Tracking multiple target cows' ruminant mouth areas using optical flow and inter-frame difference methods [J]. IEEE Access, 2019, 7: 185520–185531.

[6] Amerini I, Galteri L, Caldelli R, et al. Deepfake video detection through optical flow based CNN [C]//2019 IEEE/CVF International Conference on

- Computer Vision Workshop. Seoul: IEEE, 2019: 1205–1207.
- [7] Bouwmans T, Javed S, Sultana M, et al. Deep neural network concepts for background subtraction; a systematic review and comparative evaluation [J]. Neural Networks, 2019, 117: 8–66.
- [8] Saleemi I, Shah M. Multiframe many-many point correspondence for vehicle tracking in high density wide area aerial videos [J]. International Journal of Computer Vision, 2013, 104(2): 198–219.
- [9] Barnich O, Van Droogenbroeck M. ViBe: a universal background subtraction algorithm for video sequences [J]. IEEE Transactions on Image Processing, 2011, 20(6): 1709–1724.
- [10] Bilodeau G A, Jodoin J P, Saunier N. Change detection in feature space using local binary similarity patterns [C]//2013 International Conference on Computer and Robot Vision. Regina: IEEE, 2013: 106–112.
- [11] Bruhn A, Weickert J, Schnörr C. Lucas/kanade meets horn/schunck: combining local and global optic flow methods [J]. International Journal of Computer Vision, 2005, 61(3): 211–231.
- [12] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision [C]//Proceedings of the 7th international joint conference on Artificial intelligence, Vancouver: IJCAI, 1981: 674–679.
- [13] Jain S D, Xiong B, Grauman K. FusionSeg: learning to combine motion and appearance for fully automatic segmentation of generic objects in videos [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2117–2126.
- [14] Cheng J C, Tsai Y H, Wang S J, et al. SegFlow: joint learning for video object segmentation and optical flow [C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 686–695.
- [15] Yang Y C, Loquercio A, Scaramuzza D, et al. Unsupervised moving object detection via contextual information separation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 879–888.
- [16] Wang A, Chen H, Lin Z, et al. RepViT: revisiting Mobile CNN From ViT Perspective [C]//CVPR 2024 Camera-ready Version Seattle: IEEE, 2023.
- [17] Lou A G, Guan S Y, Loew M H. DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation [C]//Medical Imaging 2021: Image Processing. Online Only: SPIE, 2021.
- [18] Song H M, Wang W G, Zhao S Y, et al. Pyramid dilated deeper ConvLSTM for video salient object detection [C]//Computer Vision-ECCV 2018. Munich: ChamSpringer International Publishing, 2018: 744–760.
- [19] Lu X K, Wang W G, Ma C, et al. See more, know more: unsupervised video object segmentation with co-attention Siamese networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3618–3627.
- [20] Lee Y, Seong H, Kim E. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1245–1253.

(责任编辑:刘划 英文审校:郑学东)