反馈学习高斯表观网络的视频目标分割

王 龙 [1] 宋慧慧 [1] 张开华 [1] 刘青山 [1]

摘 要 大量基于深度学习的视频目标分割方法存在两方面局限性: 1) 单帧编码特征直接输入网络解码器, 未能充分利用多帧特征, 导致解码器输出的目标表观特征难以自适应复杂场景变化; 2) 常采用前馈网络结构, 阻止了后层特征反馈前层进行补充学习, 导致学习到的表观特征判别力受限. 为此, 本文提出了反馈高斯表观网络, 通过建立在线高斯模型并反馈后层特征到前层来充分利用多帧、多尺度特征, 学习鲁棒的视频目标分割表观模型. 网络结构包括引导、查询与分割三个分支. 其中, 引导与查询分支通过共享权重来提取引导与查询帧的特征, 而分割分支则由多尺度高斯表观特征提取模块与反馈多核融合模块构成. 前一个模块通过建立在线高斯模型融合多帧、多尺度特征来增强对外观的表征力, 后一个模块则通过引入反馈机制进一步增强模型的判别力. 最后, 本文在三个标准数据集上进行了大量评测, 充分证明了本方法的优越性能.

关键词 视频目标分割, 表观建模, 反馈机制, 深度学习

引用格式 王龙, 宋慧慧, 张开华, 刘青山. 反馈学习高斯表观网络的视频目标分割. 自动化学报, 2022, **48**(3): 834−842

DOI 10.16383/j.aas.c200288

Feedback Learning Gaussian Appearance Network for Video Object Segmentation

WANG Long[1] SONG Hui-Hui[1] ZHANG Kai-Hua[1] LIU Qing-Shan[1]

Abstract There are two limitations in existing deep learning based video object segmentation methods: 1) the single frame encoding features are directly input into the network decoder, which fails to make full use of the multi-frame features, resulting in the difficulty in adapting complex scene changes of the target appearance features of the decoded output; 2) the feedforward network structure is adopted to prevent the feature feedback of the latter layer from the former layer for complementary learning. Therefore, this paper proposes a feedback Gaussian appearance network. By building an online Gaussian model and feedback the features of the back layer to the front layer, we can make full use of the multi-frame and multi-scale features to learn a robust video object segmentation model. Network structure includes three branches: guidance, query and segmentation branches. The guidance and the query branches extract the features of the guidance frame and the query frame by sharing the weights of the network, while the segmentation branch is composed of the multi-scale Gaussian appearance feature extraction module and the feedback multi-kernel fusion module. The former module enhances the representation of the appearance by building an online Gaussian model to fuse the multi-frame and multi-scale features, and the second module further enhances the discriminative capability of the model by introducing a feedback mechanism. Finally, experiments are carried out on three benchmark datasets, which fully proves the superiority of this method.

Key words Video object segmentation, appearance model, feedback mechanism, deep learning

Citation Wang Long, Song Hui-Hui, Zhang Kai-Hua, Liu Qing-Shan. Feedback learning gaussian appearance network for video object segmentation. *Acta Automatica Sinica*, 2022, **48**(3): 834−842

视频目标分割[1−6] 通常被建模为半监督学习任务, 即在给定初始帧目标掩模标注的前提下, 精确分割出后续帧中特定目标区域. 视频目标分割在众多计算机视觉任务中具有重要的应用价值, 包括视频编辑[7−8]、目标追踪[9−10] 和动作识别[11−12] 等. 近年来, 随着深度学习的兴起, 视频目标分割也取得了突破性进展. 但是, 精度高且速度快的算法仍然非常匮乏, 其原因在于所学深度模型仍难以有效应对复杂视频场景的变化, 如严重遮挡、快速运动、相似目标干扰等.

为此, 一些视频目标分割算法在不同方面进行了尝试. 其中, 文献 [1−2, 13−14] 中的算法在测试阶段用第 1 帧及其标注在线微调网络; 另外, 文献 [2,

收稿日期 2020-05-08 录用日期 2020-07-21
Manuscript received May 8, 2020; accepted July 21, 2020
国家新一代人工智能重大项目 (2018AAA0100400), 国家自然科学基金 (61872189, 61876088, 61532009), 江苏省自然科学基金 (BK20191397, BK20170040) 资助
Supported by National Major Project of China for New Generation of Artificial Intelligence (2018AAA0100400), National Natural Science Foundation of China (61872189, 61876088, 61532009), and Natural Science Foundation of Jiangsu Province (BK20191397, BK20170040)
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming
1. 南京信息工程大学, 江苏省大数据分析技术重点实验室, 大气环境与装备技术协同创新中心 南京 210044
1. Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044

15−17] 中的算法将视频分割任务视为掩膜逐帧传播过程. 但是, 由于未充分考虑复杂场景的表观建模, 这些方法在一些复杂场景下表现不佳. 为此, 一些算法试图通过增强目标与背景的特征表征力来提升表观模型的判别力. 譬如, 文献 [3] 通过利用匹配 (Matching) 操作与排序注意力模块学习查询帧中的每个像素与引导帧中所有像素之间的相似程度来构建鲁棒的表观模型; 文献 [4] 设计了两个新颖的子网络调制器, 将视觉和空间信息通过网络调制构建表观模型, 并嵌入分割子网络进行学习. 但是, 这些方法只利用单帧的特征学习表观建模, 未能充分捕获视频的时域上下文信息, 难以自适应复杂场景的变化. 针对该问题, 本文设计出一种在线多帧、多尺度高斯表观网络模块, 通过在线学习目标与背景的特征分布来提升表观模型的判别力.

除此之外, 大量掩膜传播类算法[2−3, 15−18] 将前一帧的预测结果作为当前帧的额外输入进行处理. 这种结构可视为时间维度的循环结构. 但是, 鲜有算法在空间维度也构建循环结构, 而空间维循环可将高层特征反馈到低层, 从而充分利用前、后层特征信息学习更加鲁棒的表观模型. 鉴于此, 本文通过引入这种反馈机制[19−20] 设计出反馈多核融合模块, 用于引导学习更加鲁棒的表观模型.

本文的主要贡献总结如下:

1) 提出一种在线多帧、多尺度高斯表观模型, 充分学习多尺度特征的统计信息, 增强对目标与背景表观的判别力;

2) 将信息反馈的思想引入视频目标分割, 设计出一种反馈多核融合模块, 允许前层特征捕捉后层的有用信息;

3) 本文算法与当前最先进的方法相比, 在多个标准数据集上达到领先水平, 证明了本文算法的优越性.

1 相关工作

1.1 基于在线微调的视频目标分割

一些视频目标分割算法严重依赖在线学习. 文献 [1] 及其扩展算法[21] 预先训练一个语义分割网络, 然后利用初始帧微调该网络, 使其关注分割目标; 文献 [13] 在文献 [1] 的基础上引入了在线自适应机制以学习跨视频目标表观变化. 这类方法将视频简单地视为无关图片的集合, 忽略了视频序列的时间相关性, 严重影响建模精度. 为此, 一些方法开始考虑采用简单的时序信息建模, 通过传播上一帧掩膜来建模时序信息. 文献 [2] 利用了光流算法传播掩膜, 首创了掩膜传播类视频目标分割方法; 文献 [22] 将 4 个不同功能的子网络组合为一体进行微调, 获得 2018 DAVIS[23] 挑战赛的冠军. 尽管在线微调能够大幅提升视频目标分割的精度, 但是严重影响运行效率, 导致其难以应用于对实时性要求较高的实际任务之中.

1.2 基于离线学习的视频目标分割

为降低运算成本并达到精度与速度之间的平衡, 最近提出的一些视频目标分割算法抛弃了在线微调过程, 转而只依赖于离线学习. 文献 [16] 提出了基于孪生网络的视频目标分割模型, 其中, 子网络 1 对初始帧及其掩模标注进行编码, 子网络 2 对当前帧和上一帧预测结果进行编码. 两者的输出结果再通过全局卷积 (Global convolution) 进行融合; 另外, 文献 [18] 在时空域引入非局部 (Non-local) 注意力机制来充分利用视频中丰富的时序信息, 在多个标准数据集上都表现出优异的性能.

1.3 基于表观建模的视频目标分割

表观建模对视频目标分割至关重要. 文献 [24] 设计了软匹配层来计算相似得分图; 文献 [25] 同时进行全局匹配和局部匹配, 并结合了参考帧和上一帧的信息学习鲁棒的表观模型; 文献 [3] 则将掩膜传播与特征匹配结合, 优势互补, 性能表现出色;文献 [15] 和文献 [26] 分别设计了专门的目标表观模型来自适应学习目标和背景区域之间的差异.

1.4 反馈机制

近年来, 反馈机制[19−20] 在视觉任务中得到了广泛应用, 如图像超分[27]、显著目标检测[28]、人群计数[29] 等. 文献 [27] 利用反馈结构, 以高层特征补充学习浅层表征, 取得不错效果;文献 [28] 在解码器中应用多阶段反馈机制, 进一步纠正显著图估计偏差, 提升了显著性检测的精度; 文献 [29] 设计了一种通用架构, 将自顶向下的信息以反馈的形式传递给自底向上的网络进行特征学习, 在多个数据集上表现出优异性能.

2 本文方法

如图 1 所示, 本文网络主要包含引导、查询与分割三个分支. 其中, 引导与查询分支为在 ImageNet 数据集上预训练的 ResNet101 网络, 通过共享网络权重分别用于提取引导帧与查询帧的深度特征. 深度卷积网络各层特征的特性不同: 低层富含纹理细节信息, 高层富含语义信息, 而中层则介于两者之间. 本文将利用多层特征构建多尺度表观网络, 以充分利用它们之间的互补优势.

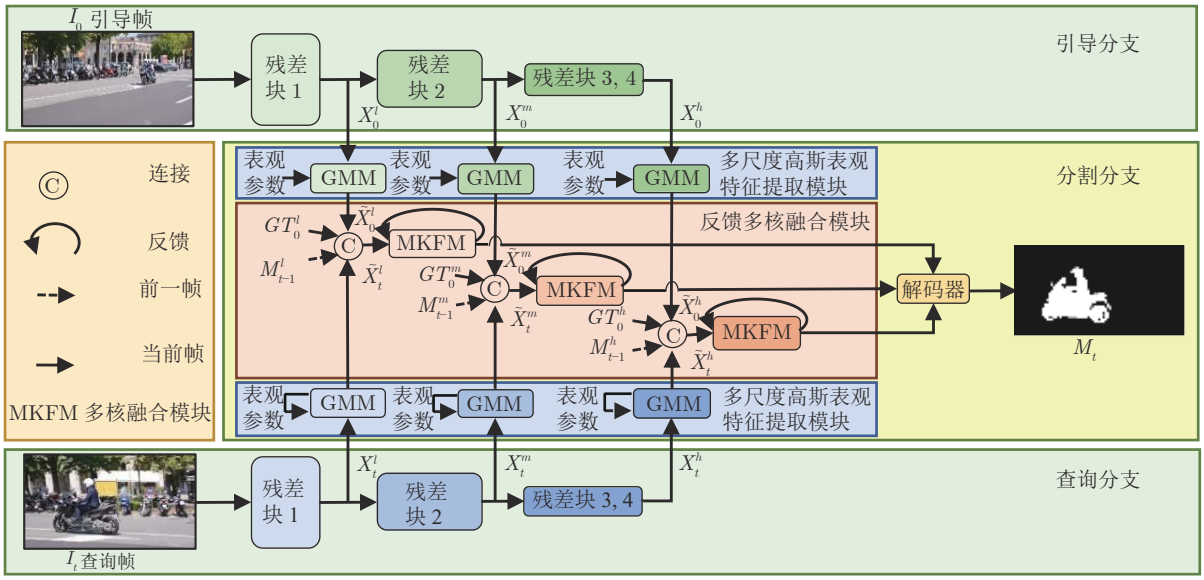


图 1 网络结构图

Fig. 1 Network structure diagram

首先, 利用多层特征构建多尺度高斯表观模型, 并通过在线更新来捕获多帧信息. 通过图 1 中的多尺度高斯表观特征提取模块生成目标和背景区域的概率密度分布图作为目标表观特征 $\tilde{X}$, 其能够有效凸显目标而抑制背景干扰. 之后, 将引导帧的高斯表观特征 $\tilde{X}_0$、引导帧真实标注 GT_0、查询帧的高斯表观特征 $\tilde{X}_t$ 和上一帧掩膜预测 M_{t-1} 一同输入反馈多核融合模块, 该模块同时考虑时空双重反馈, 不仅沿时间维反馈传播掩膜, 而且将后层特征反馈至前层以融合两者优势, 达到丰富自身信息容量与提升判别力的效果. 最后, 融合后的多级特征通过一个简单的类似文献 [30] 所提出的 U 形网络 (UNet) 结构的解码器网络, 输出最终分割预测结果 M_t.

本文的主要创新点在于所设计的多尺度高斯表观特征提取模块与反馈多核融合模块, 并将在后续内容中进行详细介绍.

2.1 多尺度高斯表观特征提取模块

鉴于视频的多帧、多尺度特征的统计信息在表观建模中的重要作用, 本文在低、中、高三种特征尺度层面上, 通过设计混合高斯模型 (Gaussian mixed model, GMM) 在线学习多帧的表观统计信息, 以获取各个位置的像素属于前景目标的概率, 来突出目标并抑制背景干扰, 为后续模块精确预测提供有效支持.

本文的 GMM 模型在整个视频上在线更新目标的均值 μ 与方差 σ. 具体来讲, 本文利用当前帧 I_t 对应的多尺度特征 $\{X_t^i\}_{i=l,\,m,\,h}$ 与分割结果 M_t^i, 通过掩膜平均池化操作估计目标和背景的均值与方差参数, 对应公式化描述为

$$\hat{\mu}_t^i = \frac{\Sigma_{x,\,y} 1\{M_t^i(x,\,y)=c\} \odot X_t^i(x,\,y)}{\Sigma_{x,\,y} 1\{M_t^i(x,\,y)=c\}} \tag{1}$$

$$\hat{\sigma}_t^i = \frac{\Sigma_{x,\,y} 1\{M_t^i(x,\,y)=c\} \odot \sigma(x,\,y)}{\Sigma_{x,\,y} 1\{M_t^i(x,\,y)=c\}} \tag{2}$$

其中, $\sigma = (X_t^i - \mu_t^i)(X_t^i - \mu_t^i)^{\mathrm{T}}$, $1\{\cdot\}$ 代表指示函数, $\odot$ 代表按位相乘, $(x,\,y)$ 表示空间位置坐标, c 则表示目标或者背景类别. 为了能够自适应目标表观的复杂变化, 如图 2 所示, 本文利用一种在线更新机制来融合多帧信息[31], 增强模型在复杂变化场景下的鲁棒性. 在线更新公式为

$$\mu_{t-1}^i \leftarrow (1-\gamma)\mu_{t-2}^i + \gamma\hat{\mu}_{t-1}^i \tag{3}$$

$$\sigma_{t-1}^i \leftarrow (1-\gamma)\sigma_{t-2}^i + \gamma\hat{\sigma}_{t-1}^i + \gamma(1-\gamma)(\mu_{t-2}^i - \hat{\mu}_{t-1}^i)(\mu_{t-2}^i - \hat{\mu}_{t-1}^i)^{\mathrm{T}} \tag{4}$$

其中, γ 是超参, $\hat{\mu}_{t-1}^i$ 与 $\hat{\sigma}_{t-1}^i$ 分别为式 (1) 与式 (2) 在时间为 $t-1$ 时计算所得. 最后, 更新后的参数代入 GMM 公式, 并忽略掉无关的常数项, 可输出高斯表观特征如下:

$$\tilde{X}_t^i = \ln|\sigma_{t-1}^i| + (X_t^i - \mu_{t-1}^i)^{\mathrm{T}}(\sigma_{t-1}^i)^{-1}(X_t^i - \mu_{t-1}^i) \tag{5}$$

2.2 反馈多核融合模块

由于视频相邻帧间目标的表观变化比较平稳, 故大量算法将视频目标分割视为掩膜从初始帧逐帧传播的过程, 把上帧预测掩膜作为处理当前帧的额

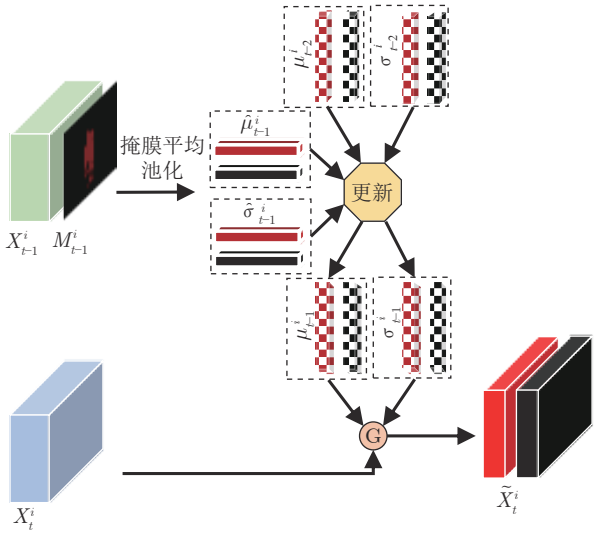
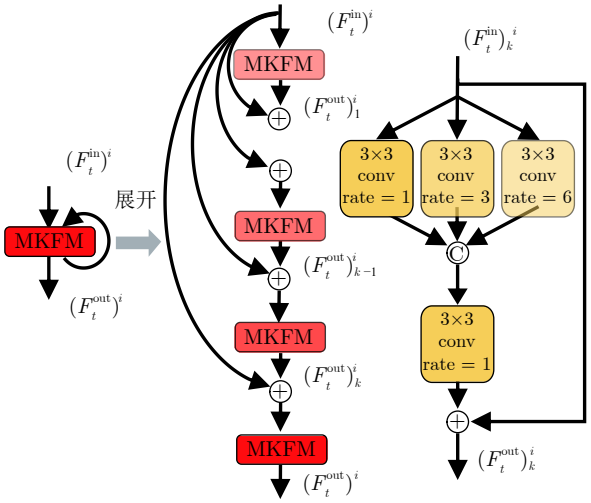


图 2　高斯表观特征提取模块 (G 表示高斯模型)

Fig. 2　Gaussian appearance feature extraction module (G denotes Gaussian model)



(a) 反馈多核融合模块　(b) 多核融合模块 (MKFM)

(a) Feedback multiple kernels fusion module　(b) Multiple kernels fusion module

图 3　反馈结构

Fig. 3　Feedback structure

外输入, 通过这种时域循环结构捕捉时序信息. 但是, 这类方法忽视了与时域循环结构相对应的空域循环结构. 而这种结构允许靠近监督约束的特征回流到网络浅层, 能够进一步增强所学特征的判别力. 所以, 为了同时充分利用这两种结构, 本文将反馈机制和掩膜传播相结合, 构建了反馈多核融合模块, 分别在低、中、高三种特征尺度上进一步增强多尺度高斯表观特征的信息容量.

图 3 展示了反馈多核融合模块的结构. 在视频目标分割任务中, 首帧的掩膜标注提供全局引导信息, 而前一帧预测出的掩膜则富含局部引导信息. 反馈多核融合模块融合这两种引导信息以及第 2.1 节所述的引导帧与查询帧的高斯表观特征:

$$(F_t^{\text{in}})^i = f_{\text{cat}}(X_t^i,\, X_0^i,\, M_{t-1}^i,\, GT_0^i) \tag{6}$$

其中, f_{cat} 代表沿通道维度连接. 不同于简单的前馈结构, 反馈结构中的高层信息可通过反馈连接向前流动. 如图 3 所示, 多核融合模块在第 $n = 1, \cdots, k$ 次循环接收前一次循环的输出 $(F_t^{\text{out}})_{n-1}^i$ 以及式 (6) 定义的 $(F_t^{\text{in}})^i$:

$$(F_t^{\text{in}})_n^i = (F_t^{\text{out}})_{n-1}^i \oplus (F_t^{\text{in}})^i \tag{7}$$

$$(F_t^{\text{out}})_n^i = f_{\text{MKFM}}((F_t^{\text{in}})_n^i) \tag{8}$$

其中, f_{MKFM} 为多核融合模块, $\oplus$ 为按位加.

图 3(b) 所示多核融合模块 (Multiple kernels fusion module, MKFM) 为反馈多核融合模块的基本单元, 除了传播掩膜的功能之外, 还通过并行多个不同扩张率的卷积操作构成空洞金字塔[32], 扩大感受野以捕捉更丰富的上下文信息. 首先, $(F_t^{\text{in}})_n^i$ 输入三个平行的扩张率分别为 $r = 1, 3, 6$ 的带孔卷积:

$$(F_t^{\text{out}'})_n^i = f_{\text{cat}}(CONV_{r=1,\,3,\,6}((F_t^{\text{in}})_n^i)) \tag{9}$$

其中, $CONV$ 表示卷积操作. 然后, 式 (9) 的结果输入一个普通卷积以融合三路特征融合并还原通道数; 最后, 利用残差跳跃连接来防止梯度消失, 以上操作公式化为

$$(F_t^{\text{out}})^i = CONV_{r=1}((F_t^{\text{out}'})_k^i) + (F_t^{\text{in}})^i \tag{10}$$

最后, 式 (10) 的输出 $(F_t^{\text{out}})^i$ 输入到解码器, 输出最终预测结果

$$M_t = f_{\text{dec}}(\{(F_t^{\text{out}})^i\}_{i=l,\,m,\,h}) \tag{11}$$

其中, f_{dec} 由多个上采样层和卷积层组成, 还原到原始尺寸的同时逐级融合多层特征.

3　实验设置与结果分析

3.1　训练细节

为了公平起见, 本文借鉴文献 [15] 采用两阶段训练:

第 1 阶段: 以尺寸调整至 240×480 像素大小的 DAVIS 2017[33] 和 YouTube-VOS[34] 训练集为训练数据, 采用 Adam[35] 优化器, 训练 80 个周期. 其中, 每批训练数据包括 4 段视频, 每段视频随机选取连续 8 帧, 学习率及其衰减率分别设置为 10^{-4} 和 0.95, 权重衰减率设置为 10^{-5}.

第 2 阶段: 对第 1 阶段训练的预训练模型进行进一步微调. 首先, 以尺寸调整为 480×864 像素大小的 DAVIS 2017[33] 和 YouTube-VOS[34] 的训练集

为训练数据, 采用 Adam[35] 优化器, 训练 100 个周期. 其中, 每批训练数据包括 2 段视频, 每段视频随机选取连续 4 帧, 学习率及其衰减率分别设置为 10^{-5} 和 0.985, 权重衰减率设置为 10^{-6}.

3.2 评价指标

本文使用 DAVIS 2017[33] 的标准评价指标, 包括区域相似度 J 和轮廓精度 F. 其中, J 为分割结果和标注真值掩膜的交并比, 即

$$J = \frac{|M \cap GT|}{|M \cup GT|} \tag{12}$$

其中, M 表示预测的分割结果, GT 表示分割真值掩膜. F 将掩膜视为系列闭合轮廓的集合, 计算基于轮廓的 F 度量, 即

$$F = \frac{2PR}{P+R} \tag{13}$$

其中, P 为准确率, R 为召回率. 另外, 本文还采用综合指标 $J\&F$, 其表示为两者的均值, 即

$$J\&F = \frac{J+F}{2} \tag{14}$$

3.3 单目标数据集上的比较结果

DAVIS 2016[36] 是密集标注的单目标视频目标分割数据集, 包括 30 段训练视频和 20 段验证视频. 表 1 中对比了本文算法与其他 18 种最先进的算法, 包括 10 种基于在线学习的算法和 8 种基于离线学习的算法. 本文算法的综合指标 $J\&F = 85\%$, 在基于离线的对比方法中排名第 2, 仅低于排名第 1 的 RANet[3] 0.5 个百分点, 与性能最先进的在线方法 MHP[14] 相比, 结果仅相差 1.9%. 但是, 本文算法的运行速度达到 0.1 s/帧, 远快于对比的离线方法, 如 MHP[14] 分割每帧用时超过 14 s. 此外, 虽然 RANet[3] 几乎在所有指标上都略高于本文算法, 但是, 其在验证单目标与多目标分割任务前, 需分别在单目标数据集 DAVIS 2016[36] 和多目标数据集 DAVIS 2017[33] 各自的训练集上进行网络微调, 即针对不同数据集使用更有针对性的不同模型参数. 而本文算法则无需这一过程, 在验证不同数据集时使用同样模型参数, 因而更具普适性.

3.4 多目标数据集上的比较结果

1) 数据集 DAVIS 2017 上的结果

DAVIS 2017[33] 是 DAVIS 2016[36] 针对多目标视频分割任务的扩展, 其包括 60 段训练视频、30 段验证视频、30 段测试视频以及 30 段竞赛视频. 表 2 比较了本文算法与 9 种基于离线学习算法和 8 种基

表 1 不同方法在 DAVIS 2016 验证集的评估结果

Table 1 Evaluation results of different methods on DAVIS 2016 validation dataset

方法	在线	$J\&F$	J_{Mean}	J_{Recall}	J_{Decay}	F_{Mean}	F_{Recall}	F_{Decay}	T (s)
MSK[2]	√	77.6	79.7	93.1	8.9	75.4	87.1	9.0	12
LIP[37]	√	78.5	78.0	88.6	5.0	79.0	86.8	6.0	—
OSVOS[1]	√	80.2	79.8	93.6	14.9	80.6	92.6	15.0	9
Lucid[17]	√	83.6	84.8		—	82.3	—	—	> 30
STCNN[38]	√	83.8	83.8	96.1	4.9	83.8	91.5	6.4	3.9
CINM[39]	√	84.2	83.4	94.9	12.3	85.0	92.1	14.7	> 30
OnAVOS[13]	√	85.5	86.1	96.1	5.2	84.9	89.7	5.8	13
OSVOS_{S}[21]	√	86.6	85.6	96.8	5.5	87.5	95.9	8.2	4.5
PReMVOS[22]	√	86.8	84.9	96.1	8.8	88.6	94.7	9.8	> 30
MHP[14]	√	86.9	85.7	96.6	—	88.1	94.8	—	> 14
VPN[40]	×	67.9	70.2	82.3	12.4	65.5	69.0	14.4	0.63
OSMN[4]	×	73.5	74.0	87.6	9.0	72.9	84.0	10.6	0.14
VM[24]	×	—	81.0	—	—	—	—	—	0.32
FAVOS[41]	×	81.0	82.4	96.5	4.5	79.5	89.4	5.5	1.8
FEELVOS[25]	×	81.7	81.1	90.5	13.7	82.2	86.6	14.1	0.45
RGMP[16]	×	81.8	81.5	91.7	10.9	82.0	90.8	10.1	0.13
AGAM[15]	×	81.8	81.4	93.6	9.4	82.1	90.2	9.8	0.07
RANet[3]	×	85.5	85.5	97.2	6.2	85.4	94.9	5.1	0.03
本文算法	×	85.0	84.6	97.1	5.8	85.3	93.3	7.2	0.1

表 2　不同方法在 DAVIS 2017 验证集的评估结果
Table 2　Evaluation results of different methods on DAVIS 2017 validation dataset

方法	在线	J	F	T (s)
MSK[2]	√	51.2	57.3	15
OSVOS[1]	√	56.6	63.9	11
LIP[37]	√	59.0	63.2	—
STCNN[38]	√	58.7	64.6	6
OnAVOS[13]	√	61.6	69.1	26
OSVOS_S[21]	√	64.7	71.3	8
CINM[39]	√	67.2	74.0	50
MHP[14]	√	**71.8**	**78.8**	20
OSMN[4]	×	52.5	57.1	0.28
FAVOS[41]	×	54.6	61.8	1.2
VM[24]	×	56.6	68.2	0.35
RANet[3]	×	63.2	68.2	—
RGMP[16]	×	64.8	68.6	0.28
AGSS[42]	×	64.9	69.9	—
AGAM[15]	×	67.2	72.7	—
DMMNet[43]	×	68.1	73.3	0.13
FEELVOS[25]	×	69.1	74.0	0.51
本文算法	×	**70.7**	**76.2**	0.14

表 3　不同方法在 DAVIS 2017 测试集的评估结果
Table 3　Evaluation results of different methods on DAVIS 2017 test-dev dataset

方法	在线	J	F
OSVOS[1]	√	47.0	54.8
OnAVOS[13]	√	49.9	55.7
OSVOS_S[21]	√	52.9	62.1
CINM[39]	√	64.5	70.5
MHP[14]	√	66.4	72.7
PReMVOS[22]	√	**67.5**	**75.7**
OSMN[4]	×	37.7	44.9
FAVOS[41]	×	42.9	44.3
Capsule[44]	×	47.4	55.2
RGMP[16]	×	51.4	54.4
RANet[3]	×	53.4	57.2
AGAM[15]	×	53.3	58.8
AGSS[42]	×	54.8	59.7
FEELVOS[25]	×	55.2	60.5
本文算法	×	**58.3**	**63.5**

表 4　不同方法在 YouTube-VOS 验证集的评估结果
Table 4　Evaluation results of different methods on YouTube-VOS validation dataset

方法	在线	G	J_S	F_s	J_u	F_u
MSK[2]	√	53.1	59.9	59.5	45.0	47.9
OnAVOS[13]	√	55.2	60.1	62.7	46.6	51.4
OSVOS[1]	√	58.8	59.8	60.5	54.2	60.7
S2S[45]	√	**64.4**	**71.0**	**70.0**	**55.5**	**61.2**
OSMN[4]	×	51.2	60.0	60.1	40.6	44.0
DMMNet[43]	×	51.7	58.3	60.7	41.6	46.3
RGMP[16]	×	53.8	59.5	—	45.2	—
RVOS[46]	×	56.8	63.6	67.2	45.5	51.0
S2S[45]	×	57.6	66.7	—	48.2	—
Capsule[44]	×	62.3	67.3	68.1	53.7	59.9
PTSNet[47]	×	63.2	69.1	—	53.5	—
AGAM[15]	×	66.0	66.9	—	61.2	—
本文算法	×	**68.1**	**69.9**	**72.3**	**62.1**	**68.3**

于在线学习算法在 DAVIS 2017[33] 验证集上的结果. 本文算法以 $J = 70.7\%$ 和 $F = 76.2\%$ 的结果在所有离线方法中排名第 1, 非常接近最优在线方法 MHP[14] 的性能 $J = 71.8\%$ 和 $F = 78.8\%$. 但是, 本文算法运行速度达到 0.14 s/帧, 而 MHP[14] 则为 20 s/帧.

表 3 是各算法在 DAVIS 2017[33] 测试集上的表现. 本文算法在离线算法中仍表现最优, 且与排名第 2 的 FEELVOS[25] 相比, J 和 F 指标分别高出 3.1% 和 3%. 此外, 本文算法精度不及最优离线方法 PReMVOS[22], 但是其网络模型是由 4 个不同功能的子网络组成, 结构异常复杂, 并且其缓慢的在线学习过程导致其推理速度 (> 30 s/帧) 远慢于本文算法. 另外, DAVIS 2017[33] 测试集中平均每段视频包含的目标物体数目多于验证集, 导致离线算法与在线算法之间的精度差距要比在验证集上的更大.

2) 数据集 YouTube-VOS 上的结果

YouTube-VOS[34] 是第一个大规模视频目标分割数据集, 包含 3 471 段训练视频和 474 段验证视频. 验证集又分为 65 类可见类别和 26 类未见类别. 评估指标为分别计算可见和未见的 J 和 F: J_s, J_u, F_s 和 F_u. 综合指标 G 为 4 项指标均值. 如表 4 所示, 本文算法 $G = 68.1\%$, 排名第 1, 超越第 2 名 AGAME[15] 2.1%, 甚至比在线学习的 S2S (Sequence-to-sequence)[45] 高 3.7%. 尤其, 本文算法对未见类别取得了 $J_u = 62.1\%$ 和 $F_u = 68.3\%$ 的出色性能, 充分体现了本文模型良好的泛化性能.

3.5　消融实验

表 5 展示了本文算法在 DAVIS 2017[33] 验证集上的消融实验结果. 三个算法变体分别用于验证各组成部分的作用. 不考虑高斯表观建模即去除多尺度高斯表观特征提取模块, $J = 62.2\%$, 与原模型相

表 5 消融实验 (M, F 和 f 分别代表多尺度高斯表观特征提取模块、反馈多核融合模块和反馈机制)

Table 5 Ablative experiments (M, F, f, denotes the multi-level Gaussian feature module, feedback multi-kernel fusion module and feedback mechanism, respectively)

算法变体	本文算法	$-M$	$-F$	$-f$	$-M-F$
J (%)	70.7	62.2	66.6	69.1	59.8

比下降了 8.5%, 证明了高斯表观建模的重要作用. 另外, 将反馈多核融合模块替换为几层简单的卷积后, $J=66.6\%$; 只去除反馈连接后 $J=69.1\%$. 从这两种模型变体的结果指标可见, 多核融合模块和反馈机制的贡献分别为 2.5% 与 1.6%. 最后, 将上述两模块都去除, J 仅为 59.8%. 以上消融实验充分证明了本文算法各部分的重要作用.

表 6 展示了反馈次数 k 对本算法精度、速度的影响. 当反馈次数为 0 时, 意味着只有前馈没有反馈; 当 k 由 0 变为 1 时, J 提升了 0.8%; 进一步地, 当 k 取 2 和 3 时, J 也随之继续提升; 最后, 当 k 再进一步增加时, J 不再改变. 而另一方面, 随着 k 值的增加, 算法运行速度逐渐变慢, 这是多核融合模块 (MKFM) 被重复调用导致计算量上升的结果. 但是, 由于多核融合模块结构简单、计算量小, k 每加 1, 速度仅变慢 2 ~ 3 ms/帧. 相较于 J 的大幅提升, 此数量级的速度变慢和计算量增加是微乎其微的. 综上, 本文将反馈机制和多核融合模块相结合, 能够以较少的计算代价换来精度的大幅提升.

表 6 不同反馈次数对比

Table 6 Comparisons with different numbers of feedback

反馈次数 k	0	1	2	3	4
J (%)	69.1	69.9	70.3	70.7	70.7
T (ms)	132	135	137	140	142

3.6 分割结果展示

图 4 展示了本文算法在各数据集上的分割结果. 可见本文算法在多种挑战场景下性能出色. 前两行中, 跳舞女孩和街舞男孩被几乎无错地分割, 展示了本算法对单个目标的强大分割能力. 第 3 行金鱼和第 8 行斑马视频中, 算法未被多个相似物体误导, 未发生混淆和丢失, 体现了本算法对于相似物体的良好区分能力. 另外, 第 4 行中, 两个进行柔道比赛的男士之间相互遮挡与交互不断, 但是本算法仍可准确分割, 表现出很强的鲁棒性. 最后, 第 5 行和第 6 行出现的自拍杆和小提琴琴弓都被准确分割, 充分展示了本算法对小物体出色的分割能力.

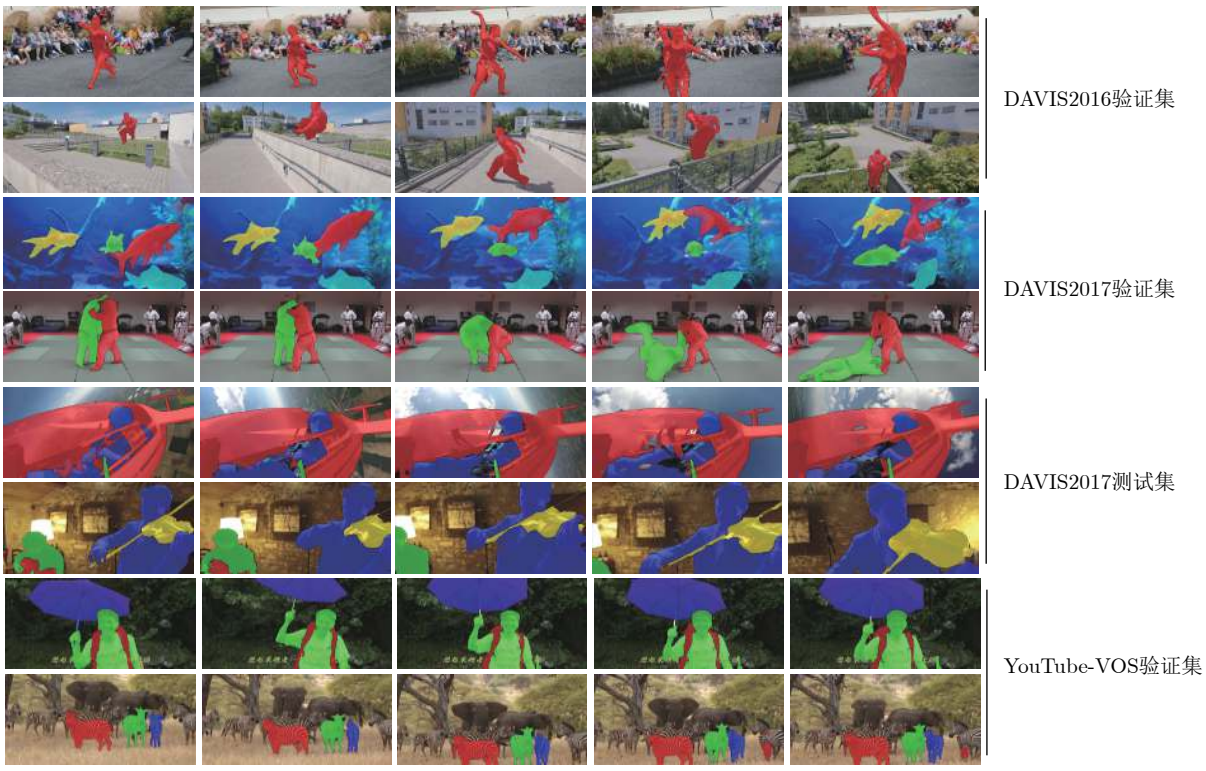


图 4 分割结果展示

Fig. 4 Display of segmentation results

4 结束语

本文提出了一种新颖的反馈学习高斯表观网络的视频目标分割算法, 集成了多尺度高斯表观特征提取模块与反馈多核融合模块. 前者通过高斯混合模型在线建模跨多帧和多尺度的目标和背景稳定表观特征, 生成粗糙但鲁棒的中间结果, 方便后续模块进一步处理. 而后者结合反馈机制和掩模传播, 通过时空双重循环结构更好地利用上下文信息, 增强模型的判别力. 在多个标准评测数据集上的实验结果都充分验证了本文所提出算法的优越性.

References

1 Caelles S, Maninis K, Ponttuset J, Lealtaixe L, Cremers D, Van Gool L. One-shot video object segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 221−230

2 Perazzi F, Khoreva A, Benenson R, Schiele B, Sorkinehornung A. Learning video object segmentation from static images. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 2663−2672

3 Wang Z Q, Xu J, Liu L, Zhu F, Shao L. Ranet: Ranking attention network for fast video object segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 3978−3987

4 Yang L J, Wang Y R, Xiong X H, Yang J C, Katsaggelos A K. Efficient video object segmentation via network modulation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6499−6507

5 Bo Yi-Hang, Hao Jiang. A rotation- and scale-invariant human parts segmentation in videos. *Acta Automatica Sinica*, 2017, **43**(10): 1799−1809
(薄一航, Hao Jiang. 视频中旋转与尺度不变的人体分割方法. 自动化学报, 2017, **43**(10): 1799−1809)

6 Chu Yi-Ping, Zhang Yin, Ye Xiu-Zi, Zhang San-Yuan. Adaptive video segmentation algorithm using hidden conditional random fields. *Acta Automatica Sinica*, 2007, **33**(12): 1252−1258
(褚一平, 张引, 叶修梓, 张三元. 基于隐条件随机场的自适应视频分割算法. 自动化学报, 2007, **33**(12): 1252−1258)

7 Li Y, Sun J, Shum H Y. Video object cut and paste. In: Proceedings of the 2005 Special Interest Group on Computer Graphics and Interactive Techniques Conference. Los Angeles, USA: ACM, 2005. 595−600

8 Wang W G, Shen J B, Porikli F. Selective video object cutout. *IEEE Transactions on Image Processing*, 2017, **26**(12): 5645−5655

9 Wang Q, Zhang L, Bertinetto L, Hu W M, Torr P H. Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 1328−1338

10 Yeo D H, Son J, Han B Y, Hee Han J. Superpixelbased tracking-by-segmentation using markov chains. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 1812−1821

11 Guo J M, Li Z W, Cheong L F, Zhou S Z. Video co-segmentation for meaningful action extraction. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 2232−2239

12 Qian Yin-Zhong, Shen Yi-Fan. Hybrid of pose feature and depth feature for action recognition in static image. *Acta Automatica Sinica*, 2019, **45**(3): 626−636
(钱银中, 沈一帆. 姿态特征与深度特征在图像动作识别中的混合应用. 自动化学报, 2019, **45**(3): 626−636)

13 Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation, ArXiv Preprint, ArXiv: 1706.09364, 2017.

14 Xu S J, Liu D Z, Bao L C, Liu W, Zhou P. Mhpvos: Multiple hypotheses propagation for video object segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 314−323

15 Johnander J, Danelljan M, Brissman E, Khan F S, Felsberg M. A generative appearance model for end to-end video object segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 8953−8962

16 Oh S W, Lee J, Sunkavalli K, Kim S J. Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7376−7385

17 Khoreva A, Benenson R, Ilg E, Brox T, Schiele B. Lucid data dreaming for object tracking. *DAVIS Challenge on Video Object Segmentation*, 2017.

18 Oh S W, Lee J, Xu N, Kim S J. Video object segmentation using space-time memory networks. In: Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 9226−9235

19 Cao C S, Liu X M, Yang Y, Yu Y, Wang J, Wang Z L, Huang Y Z, Wang L, Huang C, Xu W, Ramanan D, Huang T S. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 2956−2964

20 Zamir A R, Wu T L, Sun L, Shen W B, Shi B E, Malik J, Savarese S. Feedback networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 1308−1317

21 Maninis K, Caelles S, Chen Y H, Ponttuset J, Lealtaixe L, Cremers D, Van Gool L. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **41**(6): 1515−1530

22 Luiten J, Voigtlaender P, Leibe B. Premvos: Proposal-generation, refinement and merging for video object segmentation. In: Proceedings of the 2018 Asian Conference on Computer Vision. Perth Australia: Springer, 2018. 565−580

23 Caelles S, Montes A, Maninis K, Chen Y H, Van Gool L, Perazzi F, Ponttuset J. The 2018 DAVIS challenge on video object segmentation. ArXiv preprint: ArXiv: 1803.00557, 2018.

24 Hu Y T, Huang J B, Schwing A G. Videomatch: Matching based video object segmentation. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Spring, 2018. 54−70

25 Voigtlaender P, Chai Y N, Schroff F, Adam H, Leibe B, Chen L C. Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 9481−9490

26 Robinson A, Lawin F J, Danelljan M, Khan F S, Felsberg M. Discriminative online learning for fast video object segmentation, ArXiv preprint: ArXiv: 1904.08630, 2019.

27 Li Z, Yang J L, Liu Z, Yang X M, Jeon G, Wu W. Feedback network for image super-resolution. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 3867−3876

28 Wei J, Wang S H, Huang Q M. F3net: Fusion, feedback and focus for salient object detection. In: Proceedings of the 2019 AAAI Conference on Artificial Intelligence. Hilton Hawaiian Village, Honolulu, Hawaii, USA: Spring, 2019.

29 Sam D B, Babu R V. Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the 2018 AAAI

Conference on Artificial Intelligence. New Orleans, USA: Spring, 2018.

30 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 2015 International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany: Springer, Cham, 2015. 234−241

31 Zhang K H, Song H H. Real-time visual tracking via online weighted multiple instance learning. *Pattern Recognition*, 2013, **46**(1): 397−411

32 Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(4): 834−848

33 Ponttuset J, Perazzi F, Caelles S, Arbelaez P, Sorkinehornung A, Van G L. The 2017 DAVIS challenge on video object segmentation. ArXiv Preprint, ArXiv: 1704.00675, 2017.

34 Xu N, Yang L J, Fan Y C, Yue D C, Liang Y C, Yang J C, Huang T S. Youtube-vos: A large-scale video object segmentation benchmark. ArXiv preprint: ArXiv: 1809.03327, 2018.

35 Kingma D P, Ba J. Adam: A method for stochastic optimization. ArXiv preprint ArXiv: 1412.6980, 2014.

36 Perazzi F, Ponttuset J, Mcwilliams B, Van Gool L, Gross M, Sorkinehornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, IEEE: 2016. 724−732

37 Lv Y, Vosselman G, Xia G S, Ying Y M. Lip: Learning instance propagation for video object segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops. Seoul, Korea (South): IEEE, 2019.

38 Xu K, Wen L Y, Li G R, Bo L F, Huang Q M. Spatiotemporal CNN for video object segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 1379−1388

39 Bao L C, Wu B Y, Liu W. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 5977−5986

40 Jampani V, Gadde R, Gehler P V. Video propagation networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 451−461

41 Cheng J C, Tsai Y H, Hung W C, Wang S J, Yang M H. Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7415−7424

42 Lin H J, Qi X J, Jia J Y. AGSS-VOS: Attention guided single-shot video object segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 3948−3956

43 Zeng X H, Liao R J, Gu L, Xiong Y W, Fidler S, Urtasun R. DMM-net: Differentiable mask-matching network for video object segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 3929−3938

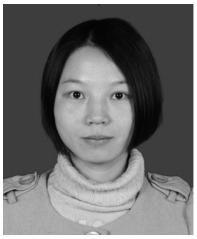
44 Duarte K, Rawat Y S, Shah M. Capsulevos: Semisupervised video object segmentation using capsule routing. In: Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 8480−8489

45 Xu N, Yang L J, Fan Y C, Yang J C, Yue D C, Liang Y C, Cohen S, Huang T. Youtubevos: Sequence-to-sequence video object segmentation. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Spring, 2018. 585−601

46 Ventura C, Bellver M, Girbau A, Salvador A, Marques F, Giroinieto X. Rvos: End-to-end recurrent network for video object segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 5277−5286

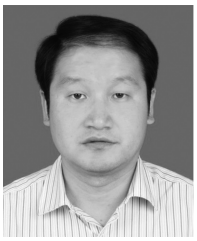
47 Zhou Q, Huang Z L, Huang L C, Gong Y C, Shen H, Huang C, Liu W Y, Wang X. Proposal, tracking and segmentation (PTS): A cascaded network for video object segmentation. ArXiv Preprint ArXiv: 1907.01203, 2019.

王 龙 南京信息工程大学自动化学院硕士研究生. 主要研究方向为视频目标分割, 深度学习.
E-mail: nj-wl@foxmail.com
(**WANG Long** Master student at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video object segmentation and deep learning.)



宋慧慧 南京信息工程大学自动化学院教授. 主要研究方向为视频目标分割, 图像超分. 本文通信作者.
E-mail: songhuihui@nuist.edu.cn
(**SONG Hui-Hui** Professor at the School of Automation, Nanjing University of Information Science and Technology. Her research interest covers video object segmentation and image super-resolution. Corresponding author of this paper.)

张开华 南京信息工程大学自动化学院教授. 主要研究方向为视频目标分割, 视觉追踪.
E-mail: zhkhua@gmail.com
(**ZHANG Kai-Hua** Professor at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video object segmentation and visual tracking.)



刘青山 南京信息工程大学自动化学院教授. 主要研究方向为视频内容分析与理解.
E-mail: qsliu@nuist.edu.cn
(**LIU Qing-Shan** Professor at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video content analysis and understanding.)