Opto-Electronic Engineering
光 电 工 程

Article
2021 年，第 48 卷，第 10 期

DOI: 10.12086/oee.2021.210193

基于自适应模板更新与多特征融合的视频目标分割算法

汪水源[1,2]，侯志强[1,2*]，王　囡[1,2]，
李富成[1,2]，蒲　磊[3]，马素刚[1,2]

[1]西安邮电大学计算机学院，陕西 西安 710121；
[2]西安邮电大学陕西省网络数据分析与智能处理重点实验室，陕西 西安 710121；
[3]火箭军工程大学作战保障学院，陕西 西安 710025

摘要：针对 SiamMask 不能很好地适应目标外观变化，特征信息利用不足导致生成掩码较为粗糙等问题，本文提出一种基于自适应模板更新与多特征融合的视频目标分割算法。首先，算法利用每一帧的分割结果对模板进行自适应更新；其次，使用混合池化模块对主干网络第四阶段提取的特征进行增强，将增强后的特征与粗略掩码进行融合；最后，使用特征融合模块对粗略掩码进行逐阶段细化，该模块能够对拼接后的特征进行有效的加权组合。实验结果表明，与 SiamMask 相比，本文算法性能有明显提升。在 DAVIS2016 数据集上，本文算法的区域相似度和轮廓相似度分别为 0.727 和 0.696，比基准算法提升了 1.0%和 1.8%，速度达到 40.2 f/s；在 DAVIS2017 数据集上，本文算法的区域相似度和轮廓相似度分别为 0.567 和 0.615，比基准算法提升了 2.4%和 3.0%，速度达到 42.6 f/s。

关键词：视频目标分割；模板更新；特征融合；掩码细化
中图分类号：TP391　　**文献标志码：**A

汪水源，侯志强，王囡，等. 基于自适应模板更新与多特征融合的视频目标分割算法[J]. 光电工程，2021，**48**(10): 210193
Wang S Y, Hou Z Q, Wang N, *et al*. Video object segmentation algorithm based on adaptive template updating and multi-feature fusion[J]. *Opto-Electron Eng*, 2021, **48**(10): 210193

Video object segmentation algorithm based on adaptive template updating and multi-feature fusion

Wang Shuiyuan[1,2], Hou Zhiqiang[1,2*], Wang Nan[1,2], Li Fucheng[1,2], Pu Lei[3], Ma Sugang[1,2]

[1]Institute of Computer, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[3]Rocket Force University of Engineering, Operational Support School, Xi'an, Shaanxi 710025, China

Abstract: In order to solve the problem that SiamMask cannot adapt to the change of target appearance and the lack of use of feature information leads to rough mask generation, this paper proposes a video object segmentation

收稿日期：2021-06-06；　收到修改稿日期：2021-09-09
基金项目：国家自然科学基金资助项目(62072370)
作者简介：汪水源(1996-)，男，硕士研究生，主要从事计算机视觉、视频目标分割的研究。E-mail：wsy_wang1@163.com
通信作者：侯志强(1973-)，男，博士，教授，博士生导师，主要从事图像处理、计算机视觉和信息融合的研究。E-mail：hzq@xupt.edu.cn

algorithm based on the adaptive template update and the multi-feature fusion. First of all, the algorithm adaptively updates the template using the segmentation results of each frame; secondly, the hybrid pooling module is used to enhance the features extracted in the fourth stage of the backbone network, and the enhanced features are fused with the rough mask; finally, the feature fusion module is used to refine the rough mask stage by stage, which can effectively combine the spliced features. Experimental results show that, compared with SiamMask, the performance of the proposed algorithm is significantly improved. On the DAVIS2016 data-set, the region similarity and contour similarity of this algorithm are 0.727 and 0.696, respectively, which is 1.0% and 1.8% higher than that of the benchmark algorithm, and the speed reaches 40.2 f/s. On the DAVIS2017 data-set, the region similarity and contour similarity of this algorithm are 0.567 and 0.615, respectively, which is 2.4% and 3.0% higher than that of the benchmark algorithm, and the speed reaches 42.6 f/s.

Keywords: video object segmentation; template update; feature fusion; mask thinning

1 引 言

近年来，视频目标分割(video object segmentation，VOS)在视频监控、自动驾驶和智能机器人等领域具有广泛的应用，受到了越来越多研究人员的关注。

按照人工参与程度的不同，可以将视频目标分割分为交互式视频目标分割、无监督视频目标分割和半监督视频目标分割。交互式 VOS 根据用户的迭代输入来分割感兴趣目标，它主要用于获取高精度的分割结果[1]。无监督 VOS 旨在使用显著特征、独立运动或已知类别标签自动估计目标掩码[2]，它不需要用户给出任何输入，通常用来自动分割视频中最关键、最显著的目标。半监督 VOS 是视频目标分割领域中最受关注的任务，也是本文的研究方向。半监督 VOS 给出了视频第一帧中目标的真实掩码，它的目的是在剩余帧中自动分割出目标掩码，然而，在整个视频序列中，待分割目标可能会经历较大的外观变化、遮挡和快速运动等情况，因此，想要在视频序列中鲁棒地分割出目标是一项极具挑战性的任务。

早期的半监督视频目标分割相关工作以 OSVOS[3]，MaskTrack[4]等算法为代表。OSVOS 利用视频首帧掩码独立地处理视频的每一帧，虽然有效地解决了遮挡问题，但它忽略了视频中隐含的时序信息。MaskTrack 使用光流将分割掩码从当前帧传播到下一帧。OnAVOS[5]通过在线自适应机制扩展了第一帧微调。PReMVOS[6]通过使用广泛的微调和合并算法组合了包括光流网络在内的四个不同的神经网络。尽管这些方法取得了不错的分割效果，但它们所采用的在线微调技术严重影响了分割速度。DyeNet[7]将模板匹配引入到重识别网络中，并抛弃了在线微调，但利用光流和循环神经网络使其训练复杂且计算量大。之后的一些工作旨在避免微调和使用光流，从而实现更快的分割速度。FAVOS[8]提出了一种基于部分区域的跟踪方法来跟踪目标对象的局部区域。PML[9]使用最近邻分类器学习像素方式的嵌入。VideoMatch[10]使用软匹配层，将当前帧的像素映射到学习嵌入空间中的第一帧。以上方法仅使用视频的前一帧或第一帧掩码作为当前帧的参考，利用前一帧掩码可以更好地处理外观的变化，但同时会牺牲对遮挡和误差漂移的鲁棒性，而利用第一帧掩码与此相反。

后续工作更注重前一帧和第一帧相结合。FEELVOS[11]扩展了 MaskTrack，它采用语义像素级嵌入以及全局和局部匹配机制将目标信息从视频的第一帧和前一帧传输到当前帧。与微调方法相比，FEELVOS 实现了更快的运行速度，但容易产生累积误差。AGAME[12]提出了一种概率生成模型来预测目标和背景的特征分布。OSMN[13]使用两个网络分别提取第一帧和前一帧的实例级信息，从而对当前帧进行分割预测。RGMP[14]采用在多个阶段中训练的编码器-解码器孪生网络架构来捕捉搜索图像和模板图像之间的局部相似性。STMVOS[15]利用存储网络从当前帧之前的包括第一帧和上一帧在内的更多帧中存储和读取信息，其性能优于之前所有的方法，但是，STMVOS 的训练过程较为繁琐，对硬件需求较高。

SiamMask[16]通过在 SiamRPN[17]的基础上增加 Mask 分支，形成了一种多分支的孪生网络框架。在视频目标分割领域，SiamMask 在 DAVIS2016[18]和 DAVIS2017[19]数据集上取得具有竞争性分割精度的同时，速度比同时期的方法快了近一个数量级。对比经典的 OSVOS，SiamMask 快了两个数量级，使得视频目标分割可以得到实际应用。但是，由于缺少模板更新，在复杂视频中，SiamMask 容易出现跟踪漂移现象；此外，在掩码生成过程中，SiamMask 所使用的特征信

息损失较多，融合过程较为粗糙，没有采用主干网络全阶段的特征图对掩码进行细化。为了解决以上问题，本文提出一种基于自适应模板更新与多特征融合的视频目标分割算法。首先，所提算法使用自适应更新策略对模板进行处理，该策略可以利用每一帧的分割结果对模板进行更新；其次，为了使用更多的特征信息对掩码进行细化，本文算法使用混合池化模块对主干网络第四阶段提取的特征进行增强，将增强后的特征与粗略掩码进行融合；最后，为了生成更为精细的掩码，本文算法使用特征融合模块将主干网络各个阶段具有更丰富空间信息的中间特征参与到掩码细化过程中。实验结果表明，本文算法显著改善了基准算法因遮挡、相似背景干扰等原因导致的跟踪漂移现象，在DAVIS2016 和 DAVIS2017 数据集上的性能得到明显提升，运行速度满足实时性要求。

2 本文算法

本文提出一种基于自适应模板更新与多特征融合的视频目标分割算法。算法在 SiamMask[6]基础上，利用自适应更新策略对模板进行处理，使用混合池化模块对主干网络第四阶段提取的特征进行增强，并采用特征融合模块对粗略掩码进行逐阶段细化。

2.1 SiamMask 算法简介

SiamMask 包括提取特征的 ResNet-50 主干网络、RPN 分支和掩码生成模块(mask generation module)三个部分，算法整体框架如图 1 所示。算法首先需要人工在视频的第一帧(模板帧)中选定待跟踪目标；接着，将选定目标与视频当前帧(搜索帧)同时输入主干网络，分别得到目标模板和提取到的当前帧的特征图，对二者进行互相关得到响应图；随后，根据 RPN 分支的指导在响应图的对应位置选取部分区域，并上采样得到粗略的初始掩码；最后，利用主干网络所提取的当前帧每阶段的特征图，对粗略掩码进行逐阶段的逐点相加并上采样，得到精细掩码，以此精细掩码作为对应视频每一帧的最终分割结果。

2.2 模板更新模块

基于孪生网络的视频目标分割和视觉目标跟踪算法大多使用视频第一帧中已标记的目标作为模板，在后续帧的搜索区域中与该模板进行匹配，从而得到目标在该帧对应的位置。虽然保持目标模板不变可以提升算法对遮挡和误差漂移的鲁棒性，但在整个视频中，目标的外观和姿态通常改变很大，如不更新模板，跟踪过程会受到目标漂移的影响，进而导致跟踪失败且无法恢复。

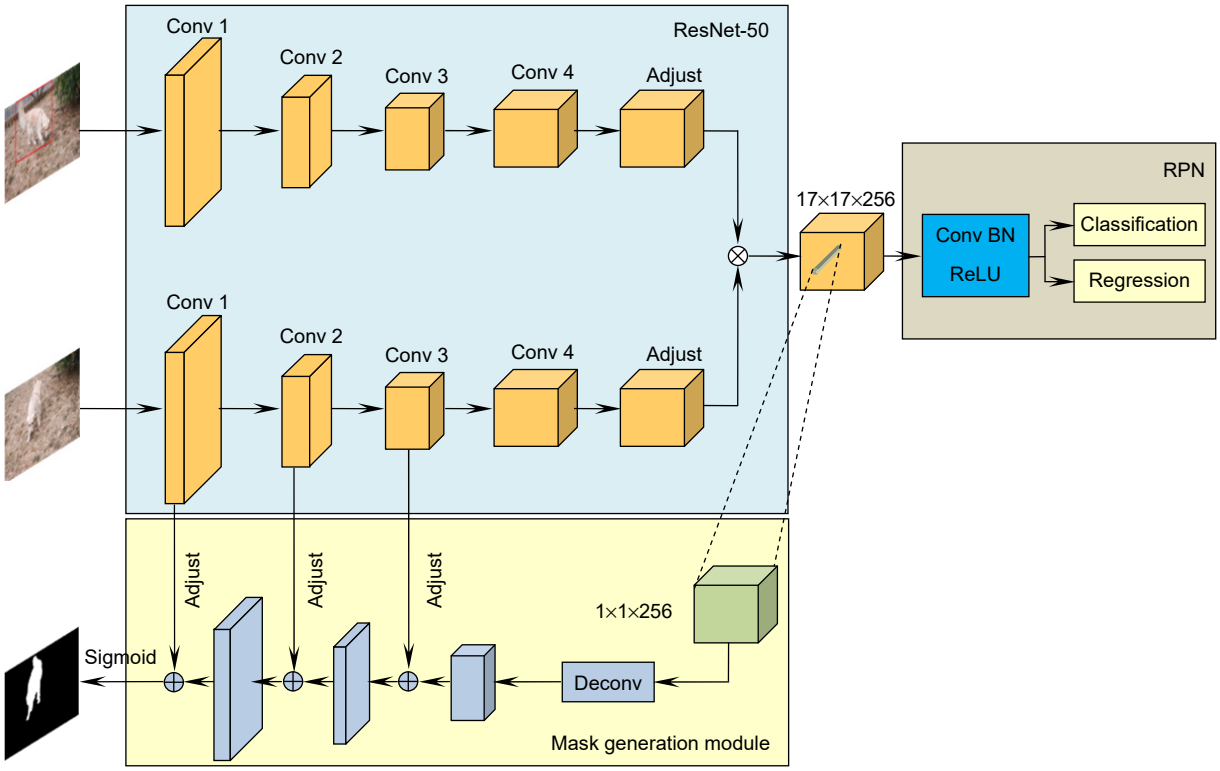

图 1 SiamMask 算法整体框架

Fig. 1 Overall framework of siamMask algorithm

受 UpdateNet[20]的启发，本文在 SiamMask 中引入模板更新模块(template update module)，该模块采用自适应更新策略对模板进行在线更新，该策略可用以下的表达式表示：

$$\tilde{T}_i = F(T_0^{\mathrm{GT}}, \tilde{T}_{i-1}, T_i) \quad , \tag{1}$$

式中：T_0^{GT} 表示从视频第一帧中目标的对应区域提取到的特征信息，T_i 表示根据本次跟踪结果从当前帧的对应区域中提取到的特征信息，$\tilde{T}_{i-1}$ 表示上一帧累积的模板，函数 F 以 T_0^{GT}、T_i 和 $\tilde{T}_{i-1}$ 为输入并输出下一帧所需的参考模板 $\tilde{T}_i$。基于当前帧的跟踪结果和累积模板之间的差异，F 可以适应每一帧的特定更新需求。此外，它还融入了初始模板 T_0^{GT}，不同于 T_i 和 $\tilde{T}_{i-1}$，T_0^{GT} 是第一帧的真实标注，这给生成模板提供了高度可靠的目标信息，并增强对目标漂移的鲁棒性。此外，类似残差学习的策略，模板更新模块中还添加了 T_0^{GT} 到输出的跳过连接，从而确保模板更新以最准确的样本为中心。

本文使用两层卷积和一层激活函数来实现模板更新模块的功能。首先，将 T_0^{GT}、$\tilde{T}_{i-1}$ 和 T_i 按通道拼接并输入模板更新模块，将它的输出和 T_0^{GT} 相加之后即可得到更新之后的下一帧模板。如图 2 所示，在第 i 帧跟踪过程中，模板更新模块(Update)的三个输入分别为从第 $i-1$ 帧目标对应区域所提取的特征 T_{i-1}、第 $i-1$ 帧对应的累积模板 $\tilde{T}_{i-2}$ 和从视频第一帧中目标对应区域提取到的特征 T_0^{GT}，将输出与 T_0^{GT} 进行跳过连接后，得到第 i 帧所需模板 $\tilde{T}_{i-1}$，第 i+1 帧对应模板 $\tilde{T}_i$ 的生成与此类似。

2.3 混合池化模块

由于 SiamMask 只使用 ResNet-50 的前四个阶段作为主干网络，且末层提取特征仅下采样到原图尺寸的 1/8，这就导致深层特征既没有足够的感受野，又缺少丰富的上下文信息。此外，在掩码生成模块中，SiamMask 只使用了主干网络前三个阶段的特征图对掩码进行细化，这使掩码又进一步损失了多尺度的语义信息。为了解决这些问题，综合速度与性能的考虑，本文算法在保持原始算法主干网络结构不变的基础上，继续使用第四阶段的特征对掩码进行细化。前人工作已经证明，金字塔池化模块[21]是增强场景解析网络的有效方法，它可以有效地捕捉长程上下文信息。受 SPNet[22]的启发，本文算法引入混合池化模块(mixed pooling module，MPM)对主干网络第四阶段的特征进行增强，该模块可以同时收集特征图的长程和短程依赖关系，增强特征图的感受野。

本文所使用的混合池化模块(MPM)如图 3 所示。设输入特征图形状为 $H\times W\times C$，其中 H，W，C 分别代表特征图的高度，宽度和通道数。为了降低计算复杂度，MPM 首先将特征图通道数调整为原来的 1/4；随后，将调整后的特征图同时送入上下四个并行分支。其中，Pool_W、Pool_H 分别对特征图的水平和垂直方向进行条状池化，得到 $H\times 1$ 和 $1\times W$ 的特征图；接

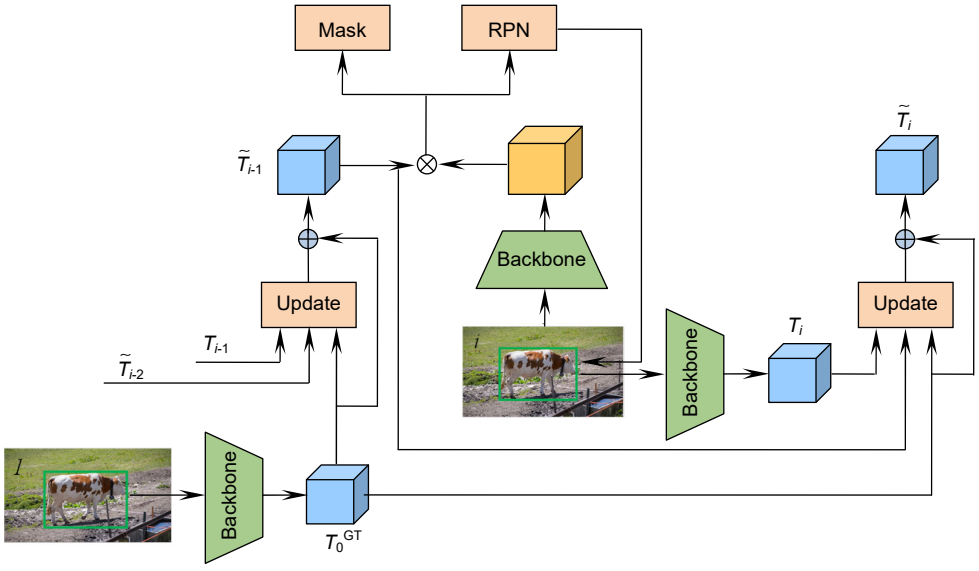

图 2 模板更新模块与模板更新流程

Fig. 2 Template update module and template update process

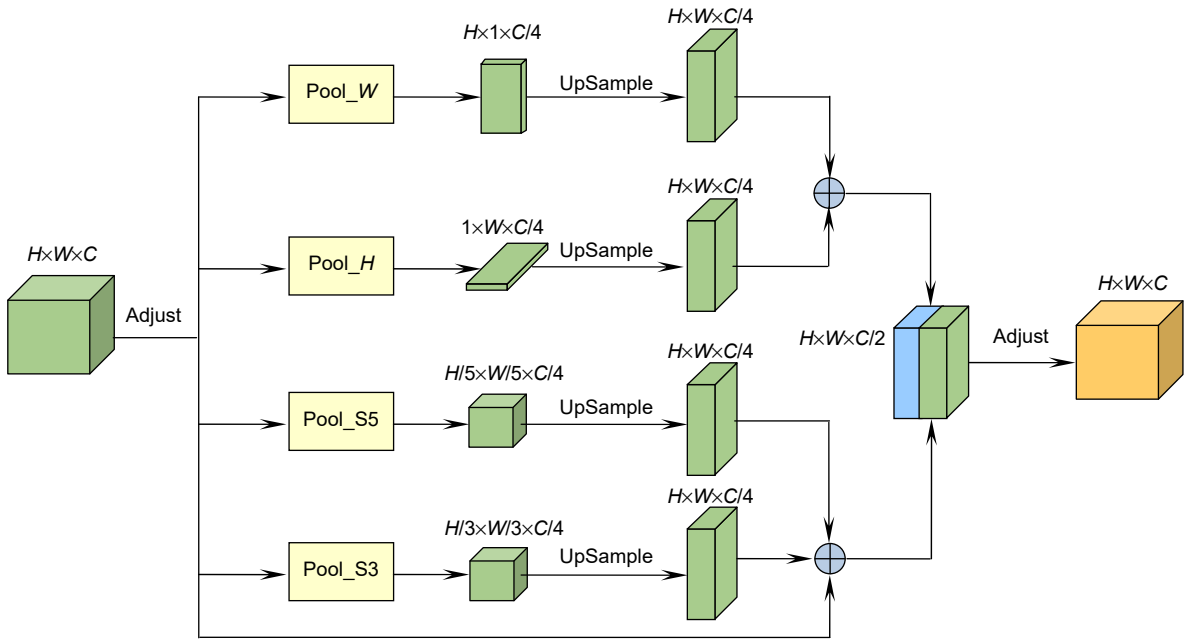

图 3 混合池化模块(MPM)

Fig. 3 Mixed pooling module (MPM)

着，将两个特征图同时扩张至 $H\times W$ 后进行相加，得到具有充足远程上下文信息的融合特征图。Pool_S5 和 Pool_S3 首先对调整后的特征图进行不同比例的池化，分别得到尺寸为原特征图 1/5 和 1/3 的两个特征图；然后，对这两个输出进行上采样并与原始尺寸的特征图相加，得到具备充足短程上下文信息的融合特征图。最后，将两个融合特征图按通道拼接并调整，得到最终融合特征图。

2.4 多尺度掩码细化与特征融合模块

在粗略掩码细化过程中，SiamMask 首先将主干网络前三阶段所提取的中间特征通道数调整为原来的 1/16，随后再分别与粗略掩码逐点相加并上采样。整个过程没有充分利用浅层网络的特征信息，这会导致掩码丢失更多的空间与语义信息。因此，除继续使用第四阶段的特征外，本文只将浅层网络所提取的特征通道数调整为原来的 1/4。受 BiseNet[23]的启发，本文算法使用特征融合模块(feature fusion module，FFM)将调整后的每一阶段特征与粗略掩码进行通道拼接。如图 4 所示，此特征融合模块可以对融合特征重新加权组合，自适应地选择需要关注的通道信息。

2.5 算法整体框架

如图 5 所示，本文算法在 SiamMask 的基础上增加了模板更新模块(template update module)，改进了掩码生成模块(mask generation module)，主干网络和

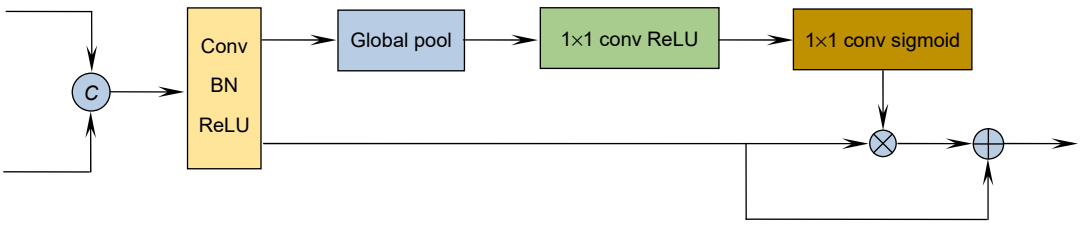

图 4 特征融合模块(FFM)

Fig. 4 Feature fusion module (FFM)

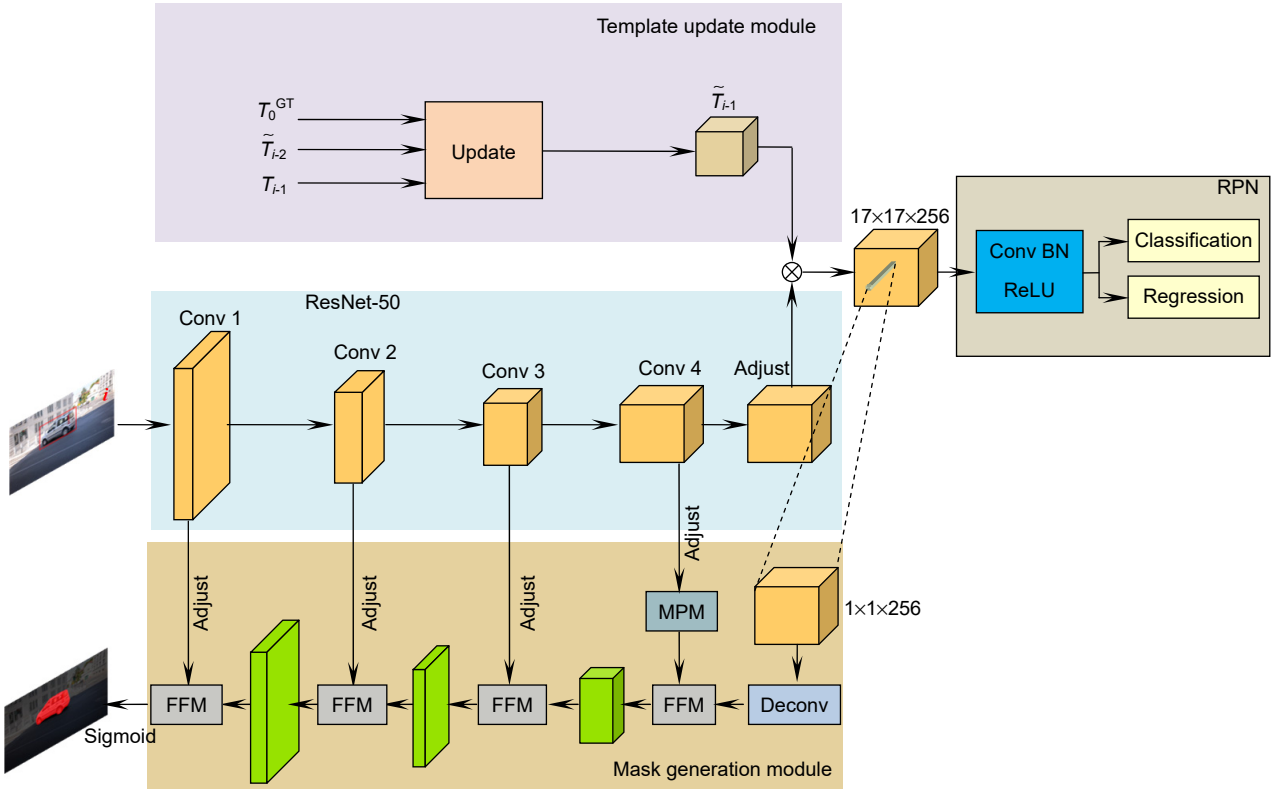

图 5 本文算法整体框架

Fig. 5 The overall framework of the algorithm in this paper

RPN 分支的设置皆与原文保持一致。本文算法依然需要人工在视频的第一帧中选定待跟踪目标，不同于 SiamMask，与当前帧特征进行互相关的不再是从第一帧中所提取的目标模板，而是经过模板更新模块逐帧更新的特定模板；采用与 SiamMask 相同的方式获取粗略掩码后，本文算法使用改进后的掩码生成模块对粗略掩码进行处理，从而得到视频每一帧的精细分割结果。

3 实验结果

为验证所提算法的有效性，本文采用 DAVIS2016 和 DAVIS2017 数据集对其进行评估。

DAVIS2016 和 DAVIS2017 是当前视频目标分割界常用的测试数据集，DAVIS2016 包含 50 个高质量视频，其中 30 个用于训练，20 个用于评估，每个视频序列只注释一个目标。DAVIS2017 是对 DAVIS2016 的扩展，包括 60 个用于训练的视频序列和 30 个用于评估的视频序列，它涵盖了视频目标分割任务中常见的多种挑战场景，如遮挡、运动模糊和外观变化，每个视频序列平均包含 2.03 个对象，单个视频序列最多包含 5 个要跟踪的对象。

本文的实验环境如下：操作系统为 64 位的 Ubuntu 16.04，PyTorch 版本为 0.4.1，16 G 内存，GPU 为 1 块 NVIDIA 1080Ti。

3.1 网络训练细节

本文算法采用两阶段方法完成整个网络的训练，其中，掩码生成模块仅在第二阶段进行训练。在第一阶段，网络首先加载在 ImageNet-1k 上预训练的 ResNet-50 权重模型；随后，使用随机梯度下降优化算法，在 Youtube-VOS、COCO、ImageNet-DET 和 ImageNet-VID 数据集上对网络进行训练，epoch 设置为 50。前五个 epoch 使用预热策略，学习率从 1×10^{-3} 逐渐增长到 5×10^{-3}，后 45 个 epoch 使用对数下降策略，学习率从 5×10^{-3} 逐渐降低到 2.5×10^{-3}。第二阶段仅使用带有掩码标注的 Youtube-VOS 和 COCO 数据集进行训练，epoch 设置为 20，整个第二阶段采用对数下降策略，学习率从 1×10^{-2} 逐渐下降到 2.5×10^{-3}。

3.2 定性分析

图 6 给出了本文算法与原始算法在 DAVIS2016 和 DAVIS2017 上的分割效果图，其中，前两列为 DAVIS2016 上的定性实验结果，第三和第四列为在多

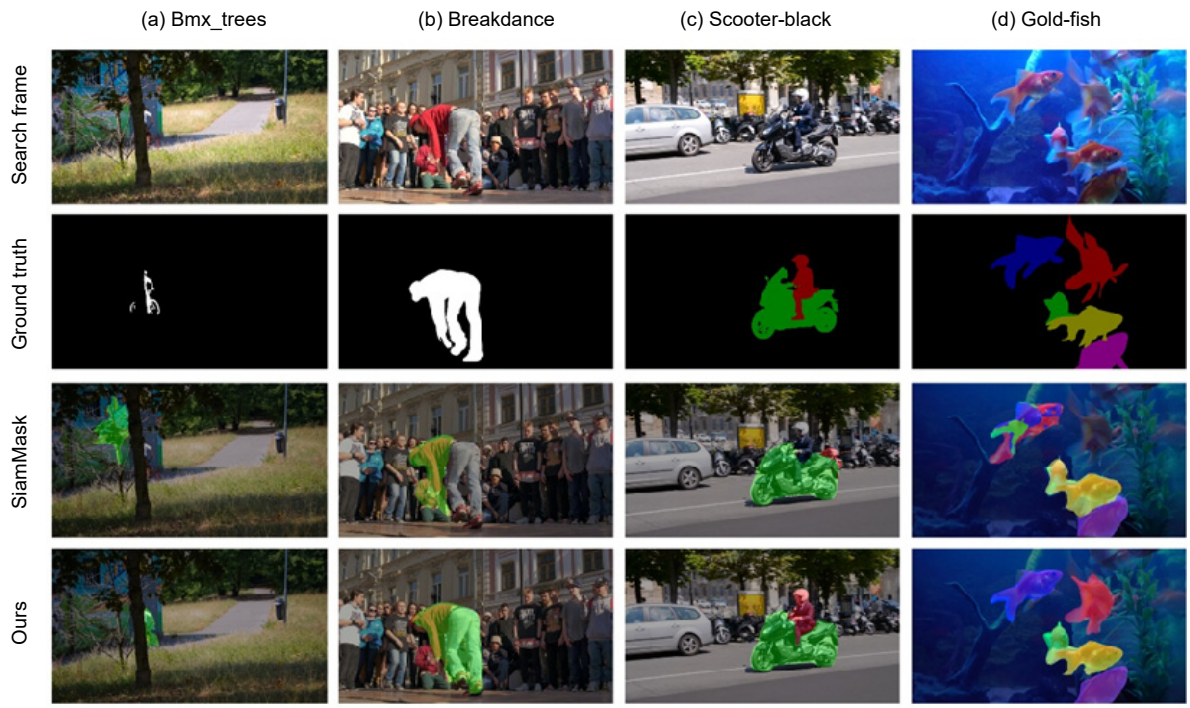

图 6 定性实验结果

Fig. 6 Qualitative experimental results

目标的 DAVIS2017 上的定性实验结果。

在第一列图中,原始算法在目标出现遮挡现象时,跟踪结果发生漂移,而本文算法可以较好地克服此种情景。

在第二列图中,在目标发生较大形变并伴有相似物干扰时,原始算法的分割效果并不理想,而本文算法则实现了对相似干扰背景的剔除,在目标形变的情况下依然可以给出较好的分割掩码。

第三列图中,当两个不同类别的目标同时出现快速运动时,原始算法出现了误判现象(误把摩托车尾部认为是骑摩托的人),而本文算法仍可以分别对两个目标给出较好的分割结果。

第四列图中出现了多个同类别的目标(五条金鱼),它们在图片中分布紧密并伴有目标之间的粘连和遮挡,其中,右上角的金鱼在该帧还出现了较大的形变。在面对此种极具挑战性的场景时,原始算法出现了明显的相似目标误判(左上角金鱼和中间两条粘连的金鱼)和目标的漏检(右上角金鱼),目标的掩码轮廓也比较粗糙。本文算法则精准地分割出全部目标,掩码质量也明显优于原算法。

3.3 定量分析

DAVIS 系列数据集的评价指标主要有 Jaccard index(J)和 F-Measure(F)。Jaccard index 是评价分割质量的常用指标,它被计算为预测掩码和掩码真值的交并比(IOU),用来衡量二者之间的区域相似度。F-Measure 基于准确率和召回率进行计算,它衡量的是预测掩码与掩码真值之间的轮廓相似度。

表 1 给出了本文算法和其他五种对比算法(VPN[24]、BVS[25]、PLM[26]、MuG-W[2]、SiamMask[16])在 DAVIS2016 上的性能指标,从中可以看出,本文算法的区域相似度(J)为 0.727,轮廓相似度(F)为 0.696,超越了所有的对比算法。相比于 SiamMask,J 和 F 分别提升 1.0%和 1.8%的同时,速度满足实时性要求,达到 40.2 f/s。

表 2 给出了本文算法和其他五种对比算法(OSVOS[3]、FAVOS[8]、OSMN[13]、MuG-W[2]、SiamMask[16])在 DAVIS2017 上的性能指标,从中可以看出,本文算法的区域相似度(J)为 0.567,优于其他五种对比算法,比原算法提升了 2.4%,轮廓相似度(F)为 0.615,比原算法提升了 3.0%,虽然略低于 OSVOS 和 FAVOS,但本文算法的速度比它们快了一个甚至是两个数量级,达到 42.6 f/s,依然满足实时性要求。SiamMask_R 为按 SiamMask 开源代码进行复现的测试结果,由于硬件设备存在差异及测试参数的影响,本文在 DAVIS2016 上的复现结果略低于 SiamMask,

表 1 DAVIS2016 验证集上不同算法之间的性能对比

Table 1 Performance comparison between different algorithms on the DAVIS2016 verification set

Method	J-mean	F-mean	Speed/(f/s)
VPN[24]	0.702	0.655	1.6
BVS[25]	0.600	0.588	2.7
PLM[26]	0.702	0.625	6.7
MuG-W[2]	0.657	0.636	1.4
SiamMask[16]	0.717	0.678	**55.0**
SiamMask_R	0.692	0.665	**55.0**
Ours	**0.727**	**0.696**	40.2

表 2 DAVIS2017 验证集上不同算法之间的性能对比

Table 2 Performance comparison between different algorithms on the DAVIS2017 verification set

Method	J-mean	F-mean	Speed/(f/s)
OSVOS[3]	0.566	**0.639**	0.1
FAVOS[8]	0.546	0.618	0.8
OSMN[13]	0.525	0.571	8.0
MuG-W[2]	0.541	0.580	1.4
SiamMask[16]	0.543	0.585	**55.0**
SiamMask_R	0.543	0.585	**55.0**
Ours	**0.567**	0.615	42.6

DAVIS2017 上的复现结果与 SiamMask 相同，本文所有工作皆在此复现基础上进行。

3.4 消融实验

为了验证所提模块的有效性，本文算法采用 DAVIS2017 数据集进行消融实验。如表 3 所示，MPM 表示是否使用混合池化模块处理过后的主干网络第四阶段特征进行掩码细化；FFM 表示在逐阶段掩码细化过程中，是否采用特征融合模块；Update 表示是否使用模板更新模块。结果表明，使用多尺度特征可以给粗略掩码提供更多的语义信息，一定程度上提升了分割精度。利用数量更为丰富的通道信息，以通道拼接并自适应选择的特征融合方式代替逐点相加，使得掩码细化过程更为合理，进一步优化了分割效果。Update 模块则利用了视频中潜在的时序信息，对每一帧的分割做出更加准确的指导，再次将本文算法的分割精度提升到新的高度。

表 3 本文算法在 DAVIS2017 上的消融实验

Table 3 Ablation experiment of this algorithm on the DAVIS2017

SiamMask	MPM	FFM	Update	J	F
√				0.543	0.585
√	√			0.546	0.591
√		√		0.550	0.597
√			√	0.559	0.602
√	√	√		0.552	0.598
√	√		√	0.562	0.608
√		√	√	0.565	0.612
√	√	√	√	**0.567**	**0.615**

4 结 论

本文提出了一种基于自适应模板更新与多特征融合的视频目标分割算法。首先，算法利用每一帧的分割结果对模板进行自适应更新；其次，在掩码生成过程中，使用特征信息更为丰富的中间特征和更为合理的融合过程对掩码进行细化。与 SiamMask 相比，所提算法性能得到明显提升的同时，速度达到实时。但是，本文算法需要对每一个数据集分别进行参数调试，过程较为繁琐，这也是 Siamese 系列算法难以复现的原因之一。它们需要使用先验知识和多个后处理来辅助跟踪和分割结果的选择，后处理过程会引入对应的超参数，而 Siamese 系列算法对超参数的选择非常敏感，如果没有合适的超参数，算法性能会受到比较大的影响。最近的一些工作将 Transformer 的思想融入到跟踪算法中[27]，单个参数即可适用于所有数据集，显著降低了后处理操作对算法性能的影响。因此，本文后续工作将考虑对此进行探索，以进一步优化视频目标分割算法的后处理过程。

参考文献

[1] Miao J X, Wei Y C, Yang Y. Memory aggregation networks for efficient interactive video object segmentation[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10366–10375.

[2] Lu X K, Wang W G, Shen J B, *et al*. Learning video object segmentation from unlabeled videos[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8957–8967.

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

[4] Perazzi F, Khoreva A, Benenson R, *et al*. Learning video object segmentation from static images[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3491–3500.

[5] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation[Z]. arXiv: 1706.09364, 2017.

[6] Luiten J, Voigtlaender P, Leibe B. PReMVOS: proposal-generation, refinement and merging for video object segmentation[C]//*Proceedings of the 14th Asian Conference on Computer Vision*, 2018: 565–580.

[7] Li X X, Loy C C. Video object segmentation with joint re-identification and attention-aware mask propagation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 93–110.

[8] Cheng J C, Tsai Y H, Hung W C, *et al*. Fast and accurate online video object segmentation via tracking parts[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7415–7424.

[9] Chen Y H, Pont-Tuset J, Montes A, *et al*. Blazingly fast video object segmentation with pixel-wise metric learning[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1189–1198.

[10] Hu Y T, Huang J B, Schwing A G. VideoMatch: matching based video object segmentation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 56–73.

[11] Voigtlaender P, Chai Y N, Schroff F, *et al*. FEELVOS: fast end-to-end embedding learning for video object segmentation[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 9473–9482.

[12] Johnander J, Danelljan M, Brissman E, *et al*. A generative appearance model for end-to-end video object segmentation[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 8945–8954.

[13] Yang L J, Wang Y R, Xiong X H, *et al*. Efficient video object segmentation via network modulation[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6499–6507.

[14] Oh S W, Lee J Y, Sunkavalli K, *et al*. Fast video object segmentation by reference-guided mask propagation[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7376–7385.

[15] Oh S W, Lee J Y, Xu N, *et al*. Video object segmentation using space-time memory networks[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 9225–9234.

[16] Wang Q, Zhang L, Bertinetto L, *et al*. Fast online object tracking and segmentation: a unifying approach[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 1328–1338.

[17] Li B, Yan J J, Wu W, *et al*. High performance visual tracking with Siamese region proposal network[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8971–8980.

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

[19] Pont-Tuset J, Perazzi F, Caelles S, *et al*. The 2017 DAVIS challenge on video object segmentation[Z]. arXiv: 1704.00675, 2018.

[20] Zhang L C, Gonzalez-Garcia A, Van De Weijer J, *et al*. Learning the model update for Siamese trackers[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 4009–4018.

[21] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C]//*Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6230–6239.

[22] Hou Q B, Zhang L, Cheng M M, *et al*. Strip pooling: rethinking spatial pooling for scene parsing[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 4002–4011.

[23] Yu C Q, Wang J B, Peng C, *et al*. BiSeNet: bilateral segmentation network for real-time semantic segmentation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 334–349.

[24] Jampani V, Gadde R, Gehler P V. Video propagation networks[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3154–3164.

[25] Märki N, Perazzi F, Wang O, *et al*. Bilateral space video segmentation[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 743–751.

[26] Yoon J S, Rameau F, Kim J, *et al*. Pixel-level matching for video object segmentation using convolutional neural networks[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2186–2195.

[27] Chen X, Yan B, Zhu J W, *et al*. Transformer tracking[Z]. arXiv: 2103.15436, 2021.

Video object segmentation algorithm based on adaptive template updating and multi-feature fusion

Wang Shuiyuan[1,2], Hou Zhiqiang[1,2*], Wang Nan[1,2], Li Fucheng[1,2], Pu Lei[3], Ma Sugang[1,2]

[1]Institute of Computer, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[3]Rocket Force University of Engineering, Operational Support School, Xi'an, Shaanxi 710025, China

Experimental results

Overview: In recent years, video object segmentation (VOS) has been widely used in video surveillance, autopilot, intelligent robot, and other fields, and it has attracted more and more researchers' attention. According to the degree of human participation, video object segmentation can be divided into interactive video object segmentation, unsupervised video object segmentation, and semi-supervised video object segmentation. Semi-supervised VOS is the most concerned task in the field of video object segmentation, and it is also the research direction of this paper. Semi-supervised VOS gives the real mask of the target in the first frame of the video, and its purpose is to segment the target mask automatically in the remaining frames. However, in the whole video sequence, the target to be segmented may experience great appearance changes, occlusion, and fast movement, so it is a very challenging task to segment the target robust in the video sequence.

SiamMask forms is a multi-branch twin network framework by adding Mask branches to SiamRPN. In the field of video object segmentation, SiamMask achieves competitive segmentation accuracy on DAVIS2016 and DAVIS2017 data-sets. At the same time, the speed is nearly an order of magnitude faster than the method in the same period. Compared with the classical OSVOS, SiamMask is two orders of magnitude faster, so the video object segmentation can be applied in practice. However, due to the lack of template update, SiamMask is prone to tracking drift in complex videos. In addition, in the process of mask generation, SiamMask uses a lot of feature information loss, the fusion process is relatively rough, and does not use the feature map of the whole stage of the backbone network to refine the mask. In order to solve the above problems, this paper proposes a video object segmentation algorithm based on the adaptive template update and the multi-feature fusion. First of all, the proposed algorithm uses an adaptive update strategy to process the template, which can update the template using the segmentation results of each frame. Secondly, in order to use more feature information to refine the mask, this algorithm uses the hybrid pooling module to enhance the features extracted in the fourth stage of the backbone network, and fuses the enhanced features with the rough mask. Finally, in order to generate a more fine mask, this algorithm uses the feature fusion module to participate in the mask thinning process of intermediate features with richer spatial information in each stage of the backbone network. The experimental results show that the proposed algorithm significantly improves the tracking drift caused by occlusion and similar background interference, the performances on DAVIS2016 and DAVIS2017 data-sets are significantly improved, and the running speed meets the real-time requirements.

Wang S Y, Hou Z Q, Wang N, *et al.* Video object segmentation algorithm based on adaptive template updating and multi-feature fusion[J]. *Opto-Electron Eng*, 2021, **48**(10): 210193; DOI: 10.12086/oee.2021.210193

Foundation item: National Natural Science Foundation of China (62072370)
* E-mail: hzq@xupt.edu.cn