基于前景感知视觉注意的半监督视频目标分割

付利华[1],赵 宇[1,2],姜涵煦[1],赵 茹[1],吴会贤[1],闫绍兴[1]
(1. 北京工业大学信息学部,北京 100124;2. 北京航空航天大学计算机学院,北京 100191)

摘 要: 半监督视频目标分割是计算机视觉领域中的一个研究热点. 传统半监督视频目标分割方法的网络模型缺乏对相似目标的判别力,且传统的掩码传播方式对模型的指导能力较弱. 本文提出一种基于前景感知视觉注意的半监督视频目标分割方法. 三流孪生编码器将输入图像映射到同一特征空间,使得相同目标具有相似的特征. 基于前景感知的视觉注意将编码器输出的特征进行相似度匹配,并利用分割掩码突显前景特征,形成前景感知视觉注意,以此关注给定的分割目标,提升模型对待分割目标的判别力. 基于残差细化的解码器采用残差学习的思想,融合当前帧图像的低阶特征,逐步改善分割细节. 在公开基准数据集上的实验结果表明,本文方法可以较好地解决相似目标容易产生混淆等问题,并能较为准确地跟踪给定的分割目标.

关键词: 视频目标分割; 孪生网络; 特征空间; 前景感知; 视觉注意力
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2022)01-0195-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201256

Semi-Supervised Video Object Segmentation Based on Foreground Perception Visual Attention

FU Li-hua[1], ZHAO Yu[1,2], JIANG Han-xu[1], ZHAO Ru[1], WU Hui-xian[1], YAN Shao-xing[1]
(1. *Faculty of Information Technology, Beijing University of Technology, Beijing* 100124, *China*;
2. *School of Computer Science and Engineering, Beihang University, Beijing* 100191, *China*)

Abstract: Semi-superised video object segmentation(SVOS) is a research hotspot in the field of computer vision. Most semi-supervised video object segmentation methods lack the ability to discriminate similar object, and the traditional mask propagation method is weak in guiding the model. This paper proposes a semi-supervised video object segmentation method based on foreground perception visual attention. The three-stream Siamese encoder maps the input frame to the same feature space, so that the same objects have similar features. Visual attention based on foreground perception calculates the similarity of encoder features and highlights the foreground through the mask, so as to focus on the given object and improve the model discrimination. The decoder based on residual refinement fuses the low-level features of the current frame to gradually improve the segmentation details. Experiments on public benchmark datasets show that the proposed method can deal with the similar confusion of the object and track the given object accurately.

Key words: video object segmentation; siamese network; feature space; foreground perception; visual attention

1 引言

半监督视频目标分割技术(Semi-Supervised Video Object Segmentation,SVOS)是视频分析的基础,其主要目标是根据第一帧给定的目标分割掩码,在视频后续帧中分割出特定的目标对象,是当前计算机视觉的研究热点,其被广泛应用于基于视频理解的精确目标跟踪、智能监控、视频检索和增强现实等领域.

在深度学习的驱动下,半监督视频目标分割方法主要依赖 3 种策略:在线学习策略(Online Learning-Based Methods)、基于掩码传播策略(Propagation-Based Methods)以及基于特征匹配策略(Matching-Based Methods). 对于处理一段新的视频,基于在线学习策略的半监督视频目标分割方法需要在父网络上多次迭代训练,具有很好的域适应能力[1,2],但却大大增加了时间成本. 并且这类方法大多数从静态图分割的角度来对待视频帧,较少地利用视频帧间的时序信息,网络模型

收稿日期:2020-11-06;修回日期:2021-01-20;责任编辑:王天慧
基金项目:北京市自然科学基金(No.4173072)

难以适应由于目标对象长期运动所带来的形变.

基于掩码传播策略的半监督视频目标分割方法,主要是通过传播前一帧的目标分割掩码,对给定目标进行跟踪[3,4],从而实现视频目标的分割.但是由于这类方法依赖于视频中目标的连续性,掩码传播过程会受到目标遮挡、多个相似目标重叠和目标快速运动的影响,容易造成跟踪漂移,导致分割性能下降.

基于特征匹配策略的半监督视频目标分割方法主要通过计算第一帧与当前帧的像素级相似度,判断当前帧中每个像素是否属于前景目标对象[5],其主要优势在于分割速度快于基于在线学习策略的半监督视频目标分割方法,同时在一定程度上解决了跟踪漂移等问题,但当待分割目标出现新的外观特征时,会出现误匹配等现象.

半监督视频目标分割任务的本质是一个对比验证过程,模型对于不同目标的判别力是提升分割精度的关键.尽管现有的半监督视频目标分割方法无论在精度上还是运行速度上都取得了较大的进步,但仍存在以下几点问题:(1)大多数模型内部将高层特征进行简单的特征拼接,这种方式过于通用化,难以有效提升模型对于不同目标的判别力,导致视频的多目标分割精度下降;(2)现有的分割方法大多将前一帧预测的分割掩码与当前帧图像合并为四通道图像,以此进行掩码传播,然而,这种掩码传播方式对于模型的指导意义并不明显,容易造成跟踪漂移,导致分割性能下降;(3)现有模型大多关注编码阶段以及特征处理,往往忽视解码器的重要性,导致模型不能有效利用低阶特征,丢失边缘细节等信息.

为了解决上述问题,本文主要贡献有以下3点:

(1)设计全局前景感知的视觉注意,将第一帧特征与当前帧特征进行逐元素全局匹配,并利用第一帧掩码突显前景特征,然后将其加权到当前帧特征,得到全局前景感知视觉注意特征图,提升模型对待分割目标的重识别能力,增强模型对于不同目标的判别力;

(2)设计局部前景感知的视觉注意,将当前帧特征与前一帧对应局部邻域的特征进行特征匹配,并利用前一帧掩码突显前景特征,然后将其加权到当前帧特征,得到局部前景感知视觉注意特征图,提升模型对待分割目标的跟踪能力,能有效处理目标的外观变化,代替传统的掩码传播方法;

(3)设计一种基于残差细化的解码器,利用残差学习的思想进行特征还原,在解码过程中融入当前帧的低阶特征,逐步改善分割细节.

2 相关工作

2.1 基于在线学习的半监督视频目标分割

基于在线学习的半监督视频目标分割是利用给定的分割掩码,在线微调网络模型,使其适用于给定的目标对象[1].OSVOS[2]预先离线训练一个通用的前景-背景分割网络,即父网络(Parent Network),用于通用目标的前景和背景分割,然后使用视频第一帧和给定目标的分割掩码,在线微调网络参数.OSVOS-S[6]基于OSVOS的思想,利用语义实例分割(Semantic Instance Segmentation),引入语义先验知识指导分割过程,传播实例分割掩码.OnAVOS[7]将在线学习的思想扩展到整个视频,通过边框约束筛选出高质量的分割结果,并利用这些分割结果再次训练模型,以实现不断提升分割的效果.Lucid Tracker[8]通过对第一帧进行大量的数据增强,扩展了第一帧与给定的分割掩码,以此模拟目标物体在后续帧中可能发生的变形,获得了较高的分割精度.DyeNet[9]通过重识别(Re-ID)技术匹配同一目标,并利用具有较高置信度的分割结果更新网络模型,以此提高分割精度.PReMVOS[10]将在线学习、实例分割[11]、光流(Optical Flow)[12]、细化(Refinement)和Re-ID[13]等技术融合在一起,目前取得了最高的精度.

2.2 基于掩码传播的半监督视频目标分割

基于掩码传播的半监督视频目标分割利用帧间时序信息,将前一帧的分割结果传播到当前帧,增加相邻两帧的时序一致性约束,提升分割精度.MaskTrack[3]通过将前一帧分割结果与当前帧RGB图像进行合并,形成四通道图像,输入网络模型,以此提供时序信息.VPN[14]提出一个可学习的双边滤波网络,用于传播视频帧之间的结构化信息.FAVOS[15]首先将第一帧所标注的目标拆分为多个部分,然后在后续帧中跟踪目标的各部分,并对跟踪结果进行分割,最后将目标各部分的分割结果进行合并,得到最终的目标分割掩码.RGMP[4]提出使用孪生编码器结构(Siamese Network),将第一帧和当前帧的高层特征堆叠在一起,进而获得当前帧的分割掩码.OSMN[16]提出使用网络调制技术,将第一帧特征作为视觉调制参数,将前一帧分割掩码的高斯分布作为空间调制参数,并将两个调制参数融合到主分割网络中的每一层,从而实现给定目标的分割.

除了利用前一帧预测的分割掩码,将光流作为运动指导信息也是非常有效的处理方式.MPN[17]利用光流变换前一帧的分割结果,并将其与当前帧的RGB图像合并,形成四通道图像作为网络输入.CTN[18]提出一个三端全卷积神经网络,输出分割概率图、确定性前景概率图和确定性背景概率图,然后使用马尔科夫随机场进行优化得到最终结果.CRN[19]首先在光流上应用运动轮廓模型,提取粗糙的目标分割掩码,然后提出级联优化网络,将粗糙的目标分割掩码作为输入,以级联优化的方式生成最

终的目标分割掩码.

2.3 基于特征匹配的半监督视频目标分割

基于特征匹配的半监督视频目标分割通过计算第一帧与当前帧的像素级相似度,判断当前帧中每个像素是否属于前景目标对象[20],其主要优势在于分割速度快于基于在线学习的半监督视频目标分割方法,同时在一定程度上解决了跟踪漂移等问题. PML[5]利用三元组损失函数(Triplet Loss),计算每一个像素点的嵌入向量(Embedding Vector),基于最邻近分类算法判断当前像素点是否属于前景目标. VideoMatch[21]类似于PML,但其直接使用交叉熵损失函数优化分割概率图,并将前景像素和背景像素分开匹配,最后进行合并优化. FEELVOS[22]使用全局匹配和局部匹配进行更鲁棒的像素级匹配. MTN[23]采用全局匹配的方式匹配待分割目标,同时提出一种新颖的掩码转换层代替原有的掩码传播方法,为了进一步提升分割速度,MTN极大地缩小特征图大小以及对应的通道数,在保证基本分割精度的同时,具有较高的分割速度.

3 基于前景感知视觉注意的半监督视频目标分割

本文提出一种基于前景感知视觉注意的半监督视频目标分割方法,整体网络模型主要包含4个部分:基于三流孪生网络的编码器、基于全局前景感知的视觉注意、基于局部前景感知的视觉注意和基于残差细化的解码器. 其主体结构如图1所示.

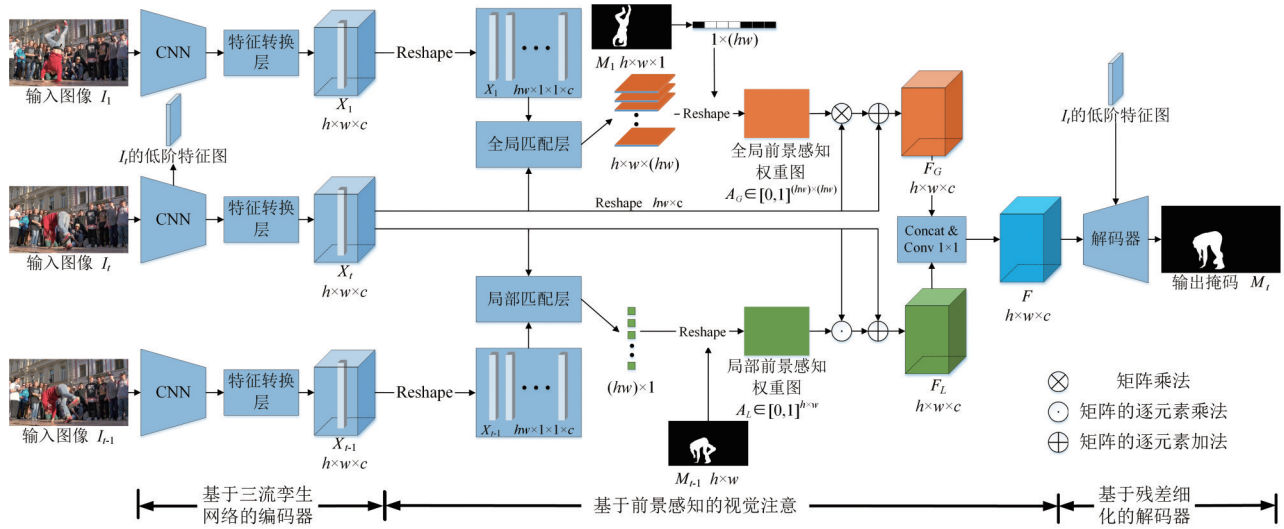

图1　基于前景感知视觉注意的半监督视频目标分割网络结构图

首先,基于三流孪生网络的编码器将第一帧、当前帧和前一帧共同映射到同一特征空间,使得相同目标具有相似特征;然后,通过全局逐元素地匹配第一帧特征与当前帧特征的相似性,并利用第一帧掩码突显前景特征,获得基于全局前景感知的视觉注意,提升模型对于不同目标的判别力;其次,通过局部地匹配当前帧特征和前一帧对应局部邻域特征的相似性,并利用前一帧掩码突显前景特征,形成获得基于局部前景感知的视觉注意,提升模型对待分割目标的跟踪能力,有效处理目标的外观变化,代替传统的掩码传播方法;最后,基于残差细化的解码器利用残差学习的思想,融合当前帧图像的低阶特征,逐步改善分割细节.

3.1 基于三流孪生网络的编码器

孪生编码器由三流孪生网络和特征转换层共同组成,其目的是将输入图像映射到同一特征空间,使同一实例目标的像素点特征不断接近,不同实例目标的像素点特征不断远离.

3.1.1 三流孪生网络

三流孪生网络基于ResNet-50,其输入分别为第一帧图像、当前帧图像以及前一帧图像. 原始的ResNet-50网络具有较大的步长和较深的通道数,相对于输入图像,最终输出的特征图将被下采样32倍. 但较低的特征图分辨率会丢失大量空间信息,不利于后续的特征匹配.

由于扩张卷积可以在不缩小特征图分辨率的前提下,增加卷积操作的感受野,因此本文将ResNet-50的最后两个残差块的下采样卷积采用扩张卷积代替,最终输出的特征图相对于输入图像下采样8倍,使编码器输出的特征保留更丰富的空间信息. 然后,将输出特征图利用1×1卷积调整到256通道,目的是降低后续特征匹配操作的计算量,提高分割速度. 编码器通过ImageNet[24]进行预训练获得初始化权值.

为了保证分割速度,本文方法每帧只需前向传播一次,并将其编码器输出特征进行保存,以便后续帧使

用. 通过这种前后帧特征复用的方式,三流孪生编码器不会增加分割网络复杂度,同样具有较高的分割速度.

3.1.2 特征转换层

对于编码器输出的特征图,本文利用一个特征转换层提取更广泛的上下文信息,其结构如图2所示. 特征转换层由一系列不同扩张率的扩张卷积和全局平均池化共同组成. 通过这种方式获得具有不同感受野的特征图,然后以多尺度的方式融合不同感受野的特征图,提取更广泛的上下文信息. 最后将特征图中每个特征点都视为一个特征向量,利用一个1×1逐点卷积整合每个特征向量,使其在特征空间中,属于同一实例目标的像素点特征不断接近,不同实例目标的像素点特征不断远离.

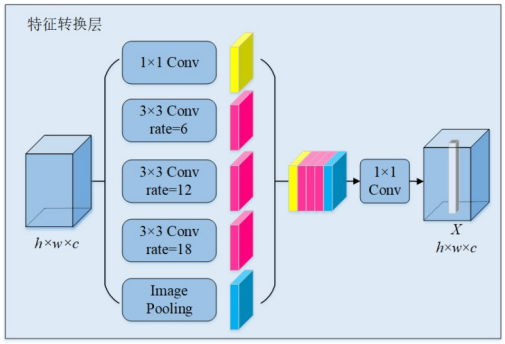

图2 特征转换层结构图

3.2 基于全局前景感知的视觉注意

基于全局前景感知的视觉注意目标是提升模型对于特定目标的重识别能力,增强模型对不同目标的判别力. 首先,利用全局匹配层将第一帧特征和当前帧特征进行逐元素匹配,获得全局相似度矩阵;然后,利用第一帧给定的分割掩码提取出全局相似度矩阵中的前景信息,忽略背景信息,生成全局前景感知权重图;最后,将获得的全局前景感知权重图加权到当前帧特征图,突显当前帧特征中与前景相似的特征,抑制与背景相似的特征,获得全局前景感知视觉注意特征图.

3.2.1 全局匹配

全局匹配目的是计算第一帧和当前帧的任意两个位置之间的空间依赖性,具有相似特征的任何两个位置均可以相互促进,且不受到空间维度的距离限制. 全局匹配过程如图3所示. 假设第一帧I_1的特征图为$X_1 \in \mathbb{R}^{h\times w\times c}$,当前帧$I_t$的特征图为$X_t \in \mathbb{R}^{h\times w\times c}$,并将特征图上的每个特征点都视为一个1×1×$c$的特征向量,其中$h$和$w$是特征图的大小,$c$为特征图深度. 为了方便后续匹配,将第一帧特征$X_1$重塑为特征集合$X_1=\left\{X_1^j \in \mathbb{R}^{1\times c} | j=1,2,\cdots,hw\right\}$. 通过式(1),计算特征向量$\boldsymbol{X}_1^j$与$\boldsymbol{X}_t^i$之间的相似度$s_{ij}$.

$$s_{ij}=\frac{X_t^i \cdot X_1^j}{\left\|X_t^i\right\| \cdot \left\|X_1^j\right\|} \tag{1}$$

其中,$\boldsymbol{X}_1^j$表示第一帧的特征图中第j个特征点的特征向量,$\boldsymbol{X}_t^i$表示当前帧的特征图中第i个特征点的特征向量,两个特征向量越相似,s_{ij}的值越大.

如图3所示,通过将$\boldsymbol{X}_1^j$与当前帧特征进行逐元素的相似度计算,获得$\boldsymbol{X}_1^j$与当前帧特征图X_t的相似度矩阵$\boldsymbol{G}_j$.

$$\boldsymbol{G}_j=\left\{s_{ij} | i=1,2,\cdots,hw\right\} \in [0,1]^{h\times w} \tag{2}$$

$\boldsymbol{G}_j$刻画了第一帧特征图中第j个特征点的特征向量与当前帧特征图中所有特征向量的相似度. 计算第一帧特征图中每个特征点的特征向量与当前帧特征图的相似度矩阵,获得矩阵集合$\left\{\boldsymbol{G}_j \in [0,1]^{h\times w} | j=1,2,\cdots,hw\right\}$,即为全局相似度矩阵$\boldsymbol{G}$.

$$\boldsymbol{G}=(s_{ij})_{(hw)\times(hw)} \in [0,1]^{(hw)\times(hw)} \tag{3}$$

全局相似度矩阵$\boldsymbol{G}$中的每个元素s_{ij}表示第j个位置的第一帧特征对第i个位置的当前帧特征的影响,两个位置的特征越相似,则这个值越大.

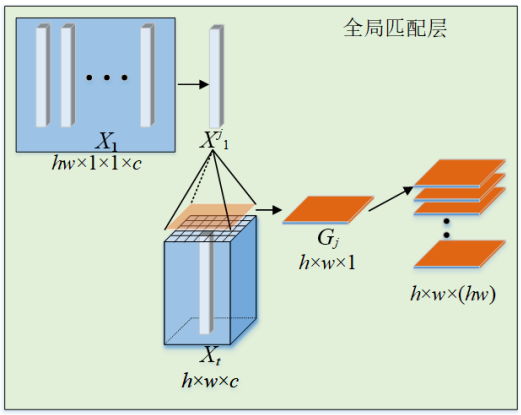

图3 全局匹配示意图

3.2.2 基于全局前景感知的视觉注意

全局相似度矩阵$\boldsymbol{G}^{(hw)\times(hw)}$中行表示当前帧特征的信息,列表示第一帧特征的信息. 为了使模型关注特定分割目标,利用第一帧掩码关注前景目标,忽略背景信息. 首先将第一帧掩码M_1大小调整至$h\times w\times 1$,并将其转置并重构成一维行向量$\boldsymbol{M}_1^{1\times hw}$,其中前景像素点的元素为1,背景像素点的元素为0;其次,利用逐元素乘法$\odot$,将$\boldsymbol{M}_1^{1\times hw}$按行加权到矩阵$\boldsymbol{G}$的每一行,生成全局前景感知权重图$A_G=(a_{ij})_{(hw)\times(hw)} \in [0,1]^{(hw)\times(hw)}$;然后,将全局前景感知权重图按照矩阵乘法的方式加权到当前帧特征图;最后,为了使得最终被关注的特征充分有效,利用残差学习的思想,使用矩阵加法补充可能被错误抑制的特征,获得全局前景感知视觉注意特征图F_G.

$$F_G=\alpha(A_G \cdot X_t)+X_t \tag{4}$$

其中,α为可学习参数,由反向传播时自动调整.

全局前景感知视觉注意特征图F_G是当前帧特征与第一帧特征所有位置的加权和,并使用第一帧的分割掩码突显当前帧与前景目标相似的特征,抑制与背景相似的特征. 因此,F_G具有全局上下文信息,并使当前帧特征充分关注给定的待分割目标,从而提升模型对于特定目标的重识别能力,增强模型对于不同目标的判别力.

3.3　基于局部前景感知的视觉注意

FlowNet[12]利用局部互相关操作提取连续两帧间的运动信息. 受到这种运动提取方式的启发,本文设计一种基于局部前景感知的视觉注意,目的是提升对待分割目标的跟踪能力,有效处理目标的外观变化,代替传统的掩码传播方法. 首先,利用局部匹配层将当前帧的特征与前一帧对应局部邻域的特征进行特征匹配;其次,根据最近邻思想,选取邻域匹配的最大值作为当前帧特征与前一帧对应局部邻域的相似度,获得局部相似度矩阵;然后,利用前一帧预测的分割掩码,提取局部相似度矩阵中的前景信息,忽略背景信息,生成局部前景感知权重图;最后,将获得的局部前景感知权重图加权到当前帧特征图,有效传播前景信息,获得局部前景感知视觉注意特征图.

3.3.1　局部匹配

局部匹配目的是计算前后两帧的局部依赖性,将前一帧的前景信息有效传播到当前帧,进一步提升网络模型对不同目标的判别能力. 局部匹配过程如图4所示. 假设当前帧的前一帧I_{t-1}的特征图为$X_{t-1}\in\mathbb{R}^{h\times w\times c}$,并将其整理为特征集合$X_{t-1}=\{X_{t-1}^j\in\mathbb{R}^{1\times c}|j=1,2,\cdots,hw\}$. 局部匹配与全局匹配类似,主要区别在于匹配范围不同. 因为目标对象可能会随着时间而产生较大的位移,所以全局匹配需要当前帧特征图与第一帧特征图的每一个特征向量都进行匹配. 然而,视频中的连续两帧满足时空一致性,即每个目标对象外观和位置均不会发生太大变化. 因此对于当前帧特征图中每个特征向量$\boldsymbol{X}_t^i$,只需要考虑前一帧特征图对应位置的局部邻域.

如图4所示,假设当前帧第i个位置的特征点对应前一帧特征图的局部邻域为$n(i)$,其窗口大小为d,假设$N(\boldsymbol{X}_t^i)\subseteq X_{t-1}$为$\boldsymbol{X}_t^i$在前一帧特征图中对应局部邻域的特征集合. 特征集合$N(\boldsymbol{X}_t^i)$中的元素是横纵方向上距离$\boldsymbol{X}_t^i$最多d个位置的前一帧特征向量,因此$N(\boldsymbol{X}_t^i)$包含D个特征向量,其中$D=(2d+1)^2$. 窗口大小d根据特征图大小自适应改变,变化公式为$d=(h/5+w/5)/2$. 根据式(1)计算$\boldsymbol{X}_t^i$与$N(\boldsymbol{X}_t^i)$内所有特征向量之间的相似度,由于连续两帧的外观信息差别较小,根据最近邻思想,选取其中的最大值作为$\boldsymbol{X}_t^i$与前一帧特征图中对应局部邻域的相似度l_i.

$$l_i=\max_{j\in n(i)}(s_{ij})\in[0,1] \tag{5}$$

l_i刻画了当前帧特征图中第i个特征点的特征向量与前一帧特征图中对应局部邻域的相似度. 计算当前帧特征图中每一个特征向量与其前一帧特征图中对应局部邻域的相似度,获得局部相似度矩阵$\boldsymbol{L}$.

$$\boldsymbol{L}=(l_i)_{h\times w}\in[0,1]^{h\times w} \tag{6}$$

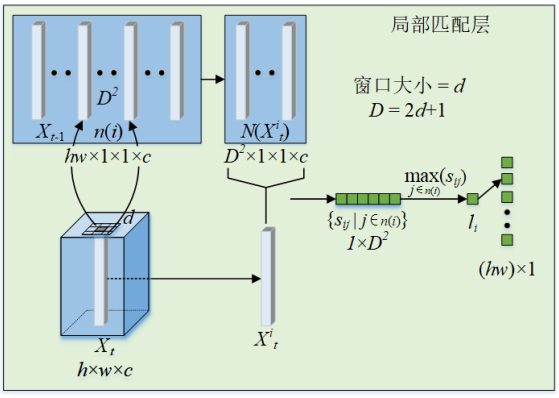

图4　局部匹配示意图

3.3.2　基于局部前景感知的视觉注意

局部前景感知的视觉注意与全局前景感知的视觉注意的方式相同,首先利用前一帧的分割掩码M_{t-1}提取局部相似度矩阵的前景信息,抑制背景信息,生成局部前景感知权重图$A_L=(a_{ij})_{h\times w}\in[0,1]^{h\times w}$;然后,利用矩阵的逐元素乘法$\odot$加权到当前帧特征,同样利用矩阵加法补充可能被错误抑制的特征,获得局部前景感知视觉注意特征图F_L.

$$F_L=\beta(A_L\odot X_t)+X_t \tag{7}$$

其中,β为可学习参数,由反向传播时自动调整.

局部前景感知视觉注意特征图F_L是当前帧特征图与前一帧特征图中对应局部邻域特征的加权和,并使用前一帧分割掩码传播前景信息,使得当前帧特征更加关注给定的分割目标.

通过全局前景感知的视觉注意和局部前景感知的视觉注意,分别获得全局前景感知视觉注意特征图F_G和局部前景感知视觉注意特征图F_L,将两者按照通道维度进行特征拼接,并利用1×1卷积调整通道维度,最终生成充分关注待分割目标的高阶特征图$F\in\mathbb{R}^{h\times w\times c}$,并将其输入残差细化解码器,获得最终的分割结果.

3.4　基于残差细化的解码器

基于残差细化的解码器主要目的是将抽象的高阶特征逐步还原,并通过连接当前帧的低阶特征,融合当前帧细节信息,最终输出当前帧的分割掩码,实现目标分割. 基于残差细化的解码器结构如图5所示.

如图5所示,在训练过程中,使用双重损失监督的

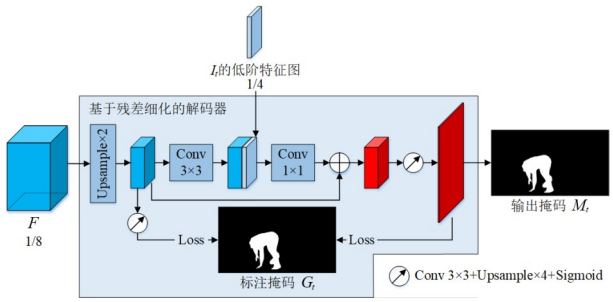

图5 基于残差细化的解码器结构图

方式逐步细化分割结果. 首先将高阶特征图F进行2倍上采样,以便融合当前帧的低阶特征;其次,利用Sigmoid函数直接将上采样后的特征图进行二分类,并与当前帧的分割掩码计算损失值L_1;再次,利用残差学习的思想,融合当前帧的低阶特征图,残差结构具有保留有效信息去除冗余信息的优点,在残差结构中加入低阶特征图,可以增强当前高阶特征缺失的细节特征,提升最终的输出结果;然后,将输出的特征图同样利用Sigmoid函数进行二分类,并与当前帧的分割掩码计算损失值L_2;最后,将两个损失值相加,对网络反向传播,进行端到端的训练.

3.5 训练细节

对于训练数据的预处理,首先,将训练图像分辨率调整为240×432,并对其进行归一化和标准化操作;然后,采用随机翻折、旋转作为数据增强策略;最后,利用随机擦除策略模拟图像可能出现待分割、目标被遮挡等情况,擦除的部分使用三通道均值进行填充.

训练过程中,在同一个视频中随机选择三帧作为视频的第一帧、当前帧以及前一帧. 为了模拟在分割过程中出现前一帧分割结果错误的情况,本文将前一帧分割掩码进行随机膨胀、腐蚀、随机擦除以及置为空图等操作,增强网络模型的鲁棒性. 本文方法的损失函数由加权二分类交叉熵[25]损失函数和Lovász-Softmax[26]损失函数共同组成,基于PyTorch开源框架,采用自适应矩估计(Adaptive Moment Estimation, Adam)优化算法,batch大小为4,初始学习率为10^{-4},权重衰减率为10^{-5},循环训练数据集100次.

4 实验结果与分析

本文的半监督视频目标分割方法目的是更好地跟踪并分割给定的目标对象,解决目标的相似混淆等问题. 为评价提出方法的有效性,本文在3个公开的大型基准数据集DAVIS-2016[27],DAVIS-2017[28]和YouTube-VOS[29]进行实验. YouTube-VOS数据集是2018年9月ECCV(European Conference on Computer Vision)最新推出的公开基准数据集,包含4000多个来自YouTube网站的高分辨率视频数据,其中训练集包含3471个视频,验证集包含474个视频,比DAVIS-2017数据集大30倍.

在训练过程中,由于YouTube-VOS数据集的数据量较大,本文分割方法直接在YouTube-VOS数据集上进行训练,在DAVIS-2016和DAVIS-2017数据集上进行测试. 实验环境为Intel(R) Xeon(R) E5-2620 v3 2.40GHz CPU、两块NVIDIA GeForce GTX 1080 Ti GPU和Linux 64位操作系统.

在测试过程中,本文分割方法在线处理视频的每一帧. 每帧只需前向传播一次,并将其编码器输出特征进行保存,以便后续帧使用,因此本文方法具有较高的分割速度.

4.1 主流方法对比

本文使用DAVIS数据集[21,22]提供的基准代码计算预测的目标分割掩码与正确标注的目标分割掩码之间的区域相似度J(Region Similarity)、轮廓精确度F(Contour Accuracy)以及每帧的处理时间Time. 区域相似度J用于评估目标分割结果与正确标注的目标分割掩码之间的区域覆盖率. 轮廓精确度F用于评估目标分割结果与正确标注的目标分割掩码之间分割边界的相似程度. 由于视频帧具有不同的分辨率,因此本文在处理之前将视频帧统一缩放到320×576.

将本文的视频目标分割方法与当前几种较流行的视频目标分割方法进行比较. 其中基于在线微调的方法有OSVOS[2],OnAVOS[7],MSK[3],STCNN[30];基于掩码传播的方法有OSMN[16],FAVOS[15],RGMP[4],RVOS[31];基于特征匹配的方法有PLM[20],PML[5],VM[21],FEELVOS[22],MTN[23],AGUnet[32],MRARnet[33].

4.1.1 DAVIS-2016数据集上的实验结果及分析

DAVIS-2016数据集用于视频单目标分割. 在DAVIS-2016数据集上,本文的视频目标分割方法与对比方法的性能评估结果如表1所示. 在表1中,将基于在线微调、掩码传播、特征匹配等方法分开列出,其中"–"表示未公开源码以及分割结果,表中的数据结果根据原论文获得.

(1)OSVOS,OnAVOS,MSK,STCNN等方法都采用了在线微调方式,基于在线微调的视频目标分割方法对每一个测试视频均在线微调分割网络,因此可以取得较好的目标分割效果,但是,在线微调非常耗时,且不能很好地适应场景的快速变化. 本文分割网络采用孪生网络结构,代替在线微调的方式,从而有效地减少了目标分割的时间.

(2)OSMN,FAVOS,RGMP等方法均采用了传统掩码传播的方式,然而,这种掩码传播方式对于模型的指导意义并不明显,且依赖视频中目标的连续性,容易受

表1　不同视频目标分割方法在DAVIS-2016数据集的定量评估结果

方法		*J&F* ↑（%）	*J* Mean ↑（%）	*F* Mean ↑（%）	Time（s/帧）↓
在线微调	OSVOS	80.2	79.8	80.6	9
	OnAVOS	85.5	86.1	84.9	13
	MSK	77.6	79.7	75.4	12
	STCNN	83.8	83.8	83.8	3.9
掩码传播	OSMN	73.5	74	72.9	0.13
	FAVOS	81	82.4	79.5	1.8
	RGMP	81.8	81.5	82.0	0.13
特征匹配	PLM	66.4	70.2	62.5	0.5
	PML	77.4	75.5	79.3	0.28
	VM	81.0	–	–	0.32
	FEELVOS	81.7	81.1	82.2	0.51
	MTN	75.7	75.3	76.1	0.027
	AGUnet	80.9	80.7	81.0	0.09
	MRARnet	83.9	83.9	83.8	0.62
	本文方法	81.1	80.5	81.6	0.11

到目标遮挡、多个相似目标重叠等影响，造成跟踪漂移，导致分割性能下降. 本文提出了一种基于局部前景感知的视觉注意，提升了模型对待分割目标的跟踪能力，能有效处理目标的外观变化，代替传统的掩码传播方法.

(3) PLM, PML, VM, FEELVOS, MTN, AGUnet 和 MRARnet 等方法采用特征匹配的方式. MTN 方法仅利用全局匹配，并提出一种掩码转换层代替原有掩码传播方法. 同时，MTN 大幅度缩小特征图大小以及通道数量，因此分割速度较快. 但 MTN 只将第一帧与当前帧做相似度匹配，同时掩码转换层的输入为第一帧分割掩码，这导致 MTN 方法本身难以适应目标的外观变化，当待分割目标的外观信息相对于第一帧发生巨大改变时，网络整体分割精度明显下降. PLM, PML, VM, FEELVOS 等方法均将匹配的距离图直接进行解码输出，但是逐元素匹配容易产生较大噪声，当出现新的外观特征时，会出现误匹配等问题. AGUnet 模型基于全卷积孪生网络对前景和背景区域进行自动标注，并将这些标注信息融合到视频目标分割的 U-net 网络中. 从表1可以看出，该模型可以获得较快的分割速度，但是模型的分割精度依赖全卷积孪生网络自动标注的精度. MRARnet 模型通过感知的运动趋势，预测当前帧的目标感兴趣区域作为模型输入，并基于帧间的光流信息，动态更新参考帧，以适应待分割目标的变化. 从表1可以看出，该模型可以获得较好的分割精度，但是由于光流信息的引入，模型不能获得较快的分割速度.

(4)本文提出的基于前景感知视觉注意的半监督视频目标分割方法，利用相同目标特征相似的特点关注前景目标特征，设计的全局前景感知和局部前景感知的视觉注意，可增强模型对待分割目标的重识别能力和跟踪能力，因此本文的分割方法具有较高的分割精度，分割精度达到81.1(*J&F*)，并且本文方法每帧只需前向传播一次，在分割过程中保存每一帧的编码器输出特征，以便后续帧使用，因此模型具有较高的分割速度，分割速度为每帧0.11s.

4.1.2　DAVIS-2017数据集上的实验结果及分析

DAVIS-2017 主要用于视频多目标分割，验证集包含30个高清视频. 本文分割方法与对比方法在 DAVIS-2017 数据集上的性能评估结果如表2所示，其中“–”表示未公开源码以及分割结果，表中的数据结果根据原论文获得.

表2　不同视频目标分割方法在DAVIS-2017数据集的定量评估结果

方法		*J&F* ↑（%）	*J* Mean ↑（%）	*F* Mean ↑（%）	Time（s/帧）↓
在线微调	OSVOS	60.3	56.6	63.9	9
	OnAVOS	65.4	61.6	69.1	13
	STCNN	61.7	58.7	64.6	3.9
掩码传播	OSMN	54.8	52.5	57.1	0.13
	FAVOS	58.2	54.6	61.8	1.8
	RGMP	66.7	64.8	68.6	0.13
	RVOS	60.6	57.5	63.6	–
特征匹配	PML	57.2	–	–	0.28
	VM	56.6	–	–	0.32
	MTN	54.2	49.4	59.0	0.048
	AGUnet	64.1	60.9	67.2	0.18
	MRARnet	63.4	61.3	65.4	0.63
	本文方法	62.1	61.5	62.8	0.11

由表2可知，本文方法的分割结果达到62.1%(*J&F*)，分割速度为每帧0.11 s. 本文的网络模型仅在 YouTube-VOS 训练集进行训练，在 DAIVS-2017 验证集上进行测试，因此分割精度略低于如 RGMP 和 OnAVOS 等直接在 DAVIS-2017 数据集上训练的方法.

大多数模型内部将高层特征进行简单地特征拼接，这种方式过于通用化，难以提升模型对不同目标的判别力，导致多目标分割精度下降. 相比于其他分割方法，本文提出一种基于前景感知视觉注意的半监督视频目标分割方法，利用特征匹配的思想使模型关注前景目标，因此在多目标分割中具有较好的分割精度. 本文方法与对比方法在 DAVIS-2017 数据集上的部分分割结果如图6所示.

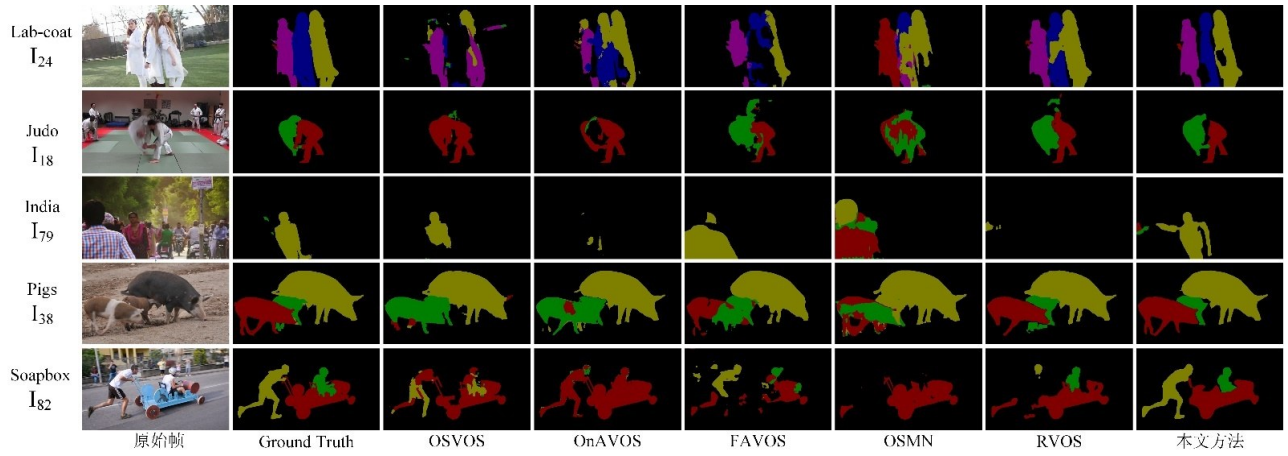

图6 本文方法与对比方法在DAVIS-2017数据集上的部分分割结果比较

4.1.3 YouTube-VOS数据集上的实验结果及分析

YouTube-VOS的官方验证集包含474个带有第一帧目标分割掩码的视频序列,其中具有91个目标类别.为了评估算法对特定分割目标的泛化能力,验证集中有65个是训练集中包含的目标类别,称为已知类别(seen),有26个是训练集中不包含的目标类别,称为未知类别(unseen).

对于YouTube-VOS数据集,同样采用区域相似度J、轮廓精确度F作为评估指标,并将J和F分成已知类别分割精度和未知类别分割精度.G overall代表四个评估指标的平均值.本文分割方法与对比方法在YouTube-VOS验证集上的性能评估结果如表3所示,其中"–"表示未公开源码以及分割结果,表中的数据结果根据原论文获得.

表3 不同视频目标分割方法在YouTube-VOS验证集的定量评估结果(%)

方法	Overall	Seen		Unseen	
	$G\uparrow$	$J\uparrow$	$F\uparrow$	$J\uparrow$	$F\uparrow$
OSVOS	58.8	59.8	60.5	54.2	60.7
OnAVOS	55.2	60.1	62.7	46.6	51.4
RGMP	53.8	59.5	45.2	–	–
OSMN	51.2	60.0	60.1	40.6	44.0
RVOS	56.8	63.6	67.2	45.5	51.0
本文方法	64.2	65.4	58.8	67.4	65.2

从表3中可以看出,本文方法在YouTube-VOS验证集上,无论在已知类别还是未知类别上都具有稳定的分割结果,总体分割精度达到64.2%(G overall).

(1)OSVOS和OnAVOS均采用在线微调技术.在已知类别和未知类别的视频中,在线微调的方法均可达到较高的分割精度.虽然在离线训练阶段没有预先学习过未知类别的目标对象,但是这些方法在测试阶段会基于未知类别视频的第一帧和对应的分割掩码对主分割网络进行在线微调,使得其分割网络能学习到当前待分割目标的外观信息,从而实现对未知类别目标对象的有效分割.但由于在线微调需要对主网络进行多次迭代训练,这会大大增加视频目标分割的时间.

本文利用孪生网络将输入图像映射到高维特征空间,并利用全局匹配的方式形成全局前景感知的视觉注意.通过这种方法可以提升模型对于给定目标的重识别能力,增强模型对于不同目标的判别力,代替在线微调过程,在不损失精度的前提下,提升分割速度.

(2)RGMP,OSMN,RVOS均采用传统掩码传播的方式.从在表3可以看出,现有的掩码传播方式在YouTube-VOS数据集上表现不好,因为该数据集中存在大量遮挡、目标丢失等情况,使得传统掩码传播方式容易造成跟踪漂移,导致最终的分割效果下降;此外,这些方法不能很好地处理未知类别,这是由于掩码传播方式通常以前一帧预测掩码作为额外输入,但这种方式对于模型的指导意义并不明显.本文提出利用局部匹配的方式形成局部前景感知的视觉注意,通过这种方式可以提升模型对待分割目标的跟踪能力,有效处理目标外观变化,代替传统的掩码传播方法.

本文分割方法在YouTube-VOS验证集上部分的分割结果如图7所示.从图7可以看出,无论对于单目标还是多目标分割,本文分割方法均能较为准确地分割目标对象,并且随着视频序列的播放,分割效果可以保持较高的鲁棒性.

4.2 方法分阶段的效果对比

4.2.1 定量分析

为了验证本文分割方法各阶段的有效性,分别从基于全局前景感知的视觉注意、基于局部前景感知的视觉注意、特征转换层和基于残差细化的解码器4个方面,在DAVIS-2017数据集上进行实验分析,其有效性评估实验结果如表4所示.

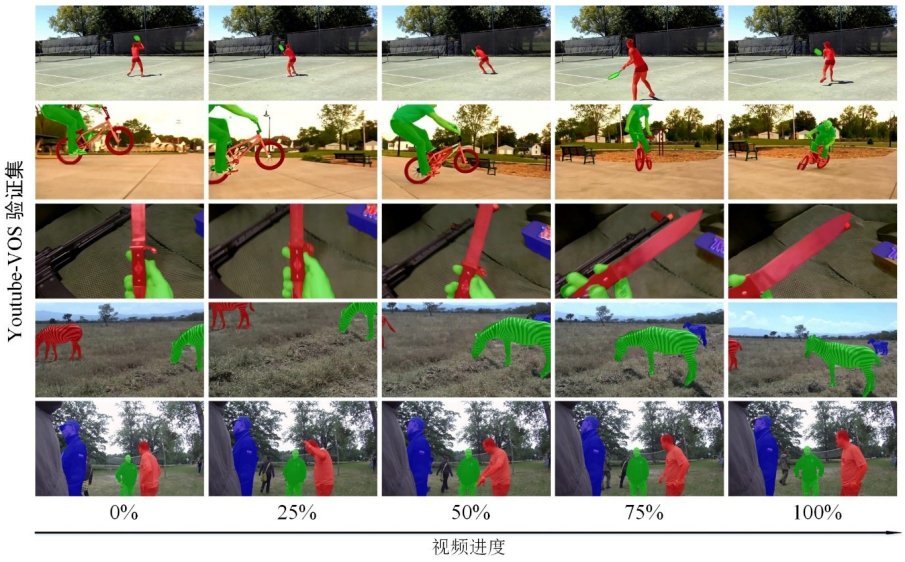

图7　本文方法在YouTube-VOS验证集上的部分定性结果展示

表4　本文方法分阶段效果的定量分析(%)

注意力机制	J	ΔJ
- Global	52.4	-9.1
- Local	45.2	-16.3
- ASPP	53.8	-7.7
- ReDecoder	55.5	-6.0
完整算法	61.5	-

为了评估基于全局前景感知的视觉注意的有效性,保持原有网络结构,删去基于全局前景感知的视觉注意部分,直接将局部前景感知视觉注意特征图F_L作为解码器的输入,本文将此网络命名为"- Global". 此时整体网络失去对待分割目标的重识别能力,仅通过连续帧之间的运动关系跟踪待分割目标. 从表4可以看出,仅依靠基于局部前景感知的视觉注意,分割精度将降低9.1%. 由此可以看出,基于全局前景感知的视觉注意可以使得网络充分关注待分割目标,提升分割精度.

为了评估基于局部前景感知的视觉注意的有效性,保持原有网络结构,删去基于局部前景感知的视觉注意部分,直接将全局前景感知视觉注意特征图F_G作为解码器的输入,本文将此网络结构命名为"- Local". 此时网络在没有任何时序信息的情况下,根据第一帧给定目标的外观特征匹配后续帧的待分割目标,失去对特定目标的跟踪能力. 从表4可以看出,仅依靠基于全局前景感知的视觉注意,分割精度将降低16.3%. 以此可以看出,基于局部前景感知的视觉注意可以增强模型对待分割目标的跟踪能力,有效提升分割精度. 同时,基于局部前景感知的视觉注意的性能提升要明显高于基于全局前景感知的视觉注意,说明在半监督视频目标分割任务中,有效的掩码传播方式对分割精度的影响较大.

为了评估编码器中的特征转换层的有效性,保持原有网络结构,删去基于三流孪生网络的编码器中的特征转换层,得到对应的网络为"- ASPP". 此时网络"- ASPP"提取的特征没有包含更丰富的上下文信息,从表4可以看出,其分割精度将降低7.7%. 由此可以看出,在编码器中加入特征转换层,通过融合不同感受野的特征,可提取具有更广泛上下文信息的特征,能有效提升分割精度.

为了评估基于残差细化的解码器的有效性,保持原有网络结构,删去基于残差细化的解码器中的低阶特征部分,得到对应的网络为"- ReDecoder". 此时网络"- ReDecoder"由于去除了基于残差细化的解码器中的低阶特征,在解码过程中没有融合当前帧的细节信息,从表4可以看出,网络"- ReDecoder"的分割精度将降低6.0%. 由此可以看出,在解码器中,利用残差学习的思想,融合当前帧的低阶特征图,可以增强当前高阶特征缺失的细节特征,进一步提升分割精度.

4.2.2　定性分析

本文对全局前景感知权重图A_G和局部前景感知权重图A_L进行可视化,以直观的方式分析本文所提出的前景感知视觉注意的有效性,可视化结果如图8所示.

基于全局前景感知的视觉注意首先将第一帧特征图和当前帧特征图进行逐元素匹配,然后利用第一帧给定的分割掩码提取对应的前景信息,抑制背景信息,进而获得全局前景感知权重图A_G. 全局前景感知权重图中每一个点代表当前帧与前景目标的相似度. 颜色越接近黄色,表示相似度越高;颜色越接近紫色,表示相似度越低. 基于全局前景感知的视觉注意目的是提升模型对特定目标的重识别能力,增强模型对不同目标的判别力.

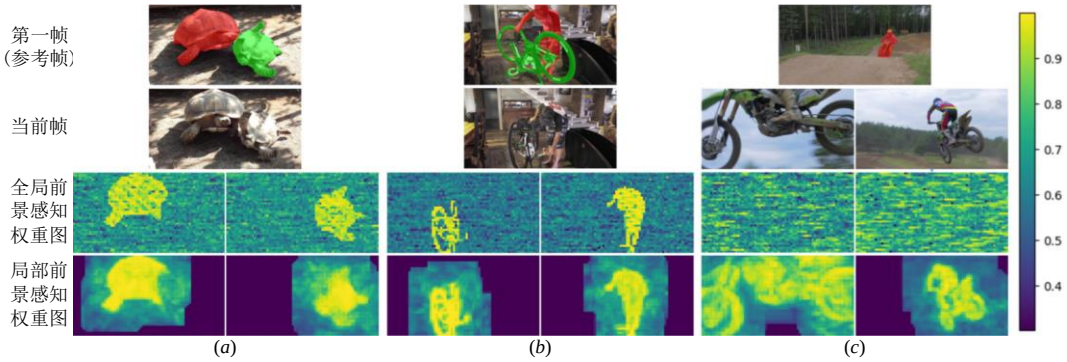

图8 全局前景感知权重图与局部前景感知权重图的可视化示意图

如图8(*a*)、图8(*b*)所示,基于全局前景感知的视觉注意可以较好地将两个不同的目标进行分离.当目标对象的运动相对平缓且外观变化相对稳定时,基于全局前景感知的视觉注意具有良好的指导意义,使得模型更加关注特定的分割目标,提升模型对于不同目标的判别能力;但当目标对象外观变化剧烈时,如图8(*c*)所示,基于全局前景感知的视觉注意的精确度则会大幅度下降,并产生大量噪声.因此在分割过程中,单纯利用基于全局前景感知的视觉注意不能较好地处理分割目标的外观变化,使得最终的分割精度下降.

基于局部前景感知的视觉注意首先将当前帧的特征与前一帧对应局部邻域的特征进行特征匹配,并选取邻域匹配的最大值作为当前帧特征与前一帧对应局部邻域特征的相似度,然后利用前一帧预测的分割掩码提取前景信息,忽略背景信息,生成局部前景感知权重图A_L.局部前景感知权重图中每一个点同样代表当前帧与前景目标的相似度.基于局部前景感知的视觉注意目的是提升模型对待分割目标的跟踪能力,有效处理目标的外观变化,代替传统的掩码传播方法.

如图8(*c*)所示,基于局部前景感知的视觉注意可以有效地跟踪给定的分割目标.当目标对象发生较大的外观变化时,由于出现大量新的外观特征,基于全局前景感知的视觉注意可能会失效.但基于局部前景感知的视觉注意是一种逐步跟踪的过程,由于视频帧之间存在时空一致性,相邻两帧在外观上不会出现剧烈变化,因此基于局部前景感知的视觉注意可以较为准确地捕获到给定的分割目标,且不受其外观变化的影响.但由于基于局部前景感知的视觉注意依赖视频的连贯性,若出现长时间的遮挡,则不能进行目标跟踪,此时只能依赖基于全局前景感知的视觉注意重新识别待分割目标.

5 总结

针对大多数半监督视频目标分割网络模型缺乏对相似目标的判别力,且简单的掩码传播对网络模型的指导意义不明显,本文通过利用基于全局前景感知的视觉注意,提升模型对不同目标的判别力;其次,利用基于局部前景感知的视觉注意,提升模型对待分割目标的跟踪能力,有效处理目标的外观变化,代替传统的掩码传播方法;最后,基于残差细化的解码器利用残差学习的思想,融合当前帧图像的低阶特征,并使用多级损失监督,逐步提升分割细节.实验结果表明,本文的目标分割方法能有效地解决目标的相似混淆等问题,能快速、有效地分割出视频中的给定目标.

参考文献

[1] 李瀚, 刘坤华, 刘嘉杰, 等. 实时视觉目标跟踪与视频对象分割多任务框架[J]. 中国图象图形学报, 2021, 26(1): 101-112.
LI H, LIU K H, LIU J J, et al. Multitask framework for video object tracking and segmentation combined with multi-scale interframe information[J]. Journal of Image and Graphics, 2021, 26(1): 101-112. (in Chinese)

[2] 付利华, 赵宇, 孙晓威, 等. 基于孪生网络的快速视频目标分割[J]. 电子学报, 2020, 48(4): 625-630.
FU L H, ZHAO Y, SUN X W, et al. Fast video object segmentation based on siamese networks[J]. Acta Electronica Sinica, 2020, 48(4): 625-630. (in Chinese)

[3] PERAZZI F, KHOREVA A, BENENSON R, et al. Learning video object segmentation from static images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA:IEEE, 2017: 2663-2672.

[4] WUG OH S, LEE J Y, SUNKAVALLI K, et al. Fast video object segmentation by reference-guided mask propagation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7376-7385.

[5] CHEN Y, PONT-TUSET J, MONTES A, et al. Blazingly fast video object segmentation with pixel-wise metric learning[C]//Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition. Salt Lake City, USA:IEEE, 2018: 1189-1198.

[6] MANINIS K K, CAELLES S, CHEN Y, et al. Video object segmentation without temporal information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1515-1530.

[7] VOIGTLAENDER P, LEIBE B. Online adaptation of convolutional neural networks for the 2017 DAVIS challenge on video object segmentation[C]//The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops. Hawaii, USA: BMVC 2017.

[8] KHOREVA A, BENENSON R, ILG E, et al. Lucid data dreaming for object tracking[C]//The DAVIS Challenge on Video Object Segmentation-CVPR Workshops. Hawaii, USA: IEEE, 2017: .

[9] LI X, CHANGE LOY C. Video object segmentation with joint re-identification and attention-aware mask propagation [C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, Germany, 2018: 90-105.

[10] LUITEN J, VOIGTLAENDER P, LEIBE B. PReMVOS: Proposal-generation, refinement and merging for video object segmentation[C]//Asian Conference on Computer Vision. Perth, Australia: ACCV, 2018: 565-580.

[11] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2961-2969.

[12] ILG E, MAYER N, SAIKIA T, et al. Flownet 2.0: Evolution of optical flow estimation with deep networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 2462-2470.

[13] XIAO T, LI S, WANG B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 3415-3424.

[14] JAMPANI V, GADDE R, GEHLER P V. Video propagation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 451-461.

[15] CHENG J, TSAI Y H, HUNG W C, et al. Fast and accurate online video object segmentation via tracking parts [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7415-7424.

[16] YANG L, WANG Y, XIONG X, et al. Efficient video object segmentation via network modulation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA:IEEE, 2018: 6499-6507.

[17] SUN J, YU D, LI Y, et al. Mask propagation network for video object segmentation[C]//The 2018 DAVIS Challenge on Video Object Segmentation-CVPR Workshops. Salt Lake City, USA: IEEE, 2018: 1-4.

[18] JANG W D, KIM C S. Online video object segmentation via convolutional trident network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 5849-5858.

[19] HU P, WANG G, KONG X, et al. Motion-guided cascaded refinement network for video object segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 1400-1409.

[20] SHIN YOON J, RAMEAU F, KIM J, et al. Pixel-level matching for video object segmentation using convolutional neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2167-2176.

[21] HU Y T, HUANG J B, SCHWING A G. Videomatch: Matching based video object segmentation[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 54-70.

[22] VOIGTLAENDER P, CHAI Y, SCHROFF F, et al. Feelvos: fast end-to-end embedding learning for video object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 9481-9490.

[23] ZHUO T, CHENG Z, KANKANHALLI M. Fast video object segmentation via mask transfer network[J]. (2019-08-28)[2021]. https://arxiv.org/abs/1908.10717.

[24] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Florida, USA: IEEE, 2009: 20-25.

[25] XIE S, TU Z. Holistically-nested edge detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1395-1403.

[26] BERMAN M, RANNEN TRIKI A, BLASCHKO M B. The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4413-4421.

[27] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video

object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 724-732.

[28] PONT-TUSET J, PERAZZI F, CAELLES S, et al. The 2017 davis challenge on video object segmentation[OL]. (2018-05-27)[2021]. http://arXiv:1704.00675.

[29] XU N, YANG L, FAN Y, ET AL. Youtube-vos: Sequence-to-sequence video object segmentation[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 585-601.

[30] XU K, WEN L, LI G, et al. Spatiotemporal cnn for video object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 1379-1388.

[31] VENTURA C, BELLVER M, GIRBAU A, et al. Rvos: End-to-end recurrent network for video object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 5277-5286.

[32] YIN Y, XU D, WANG X, et al. AGUnet: annotation-guided U-net for fast one-shot video object segmentation[J]. Pattern Recognition, 2021, 110: 107580.

[33] FU L, ZHAO Y, SUN X, et al. Video object segmentation based on motion-aware ROI prediction and adaptive reference updating[J]. Expert Systems with Applications, 2020, 167(4): 114153.

作者简介

付利华 女. 1976年9月出生,四川安岳人. 2005年在西北工业大学计算机学院获得工学博士学位. 现为北京工业大学信息学部副教授. 主要研究方向为智能信息处理、图像处理和计算机视觉.

E-mail: fulh@bjut.edu.cn

赵 宇(通信作者) 男. 1994年8月出生,河北唐山人. 2020年在北京工业大学信息学部获得工学硕士学位. 现为北京航空航天大学计算机学院博士研究生. 主要研究方向为图像处理和计算机视觉.

E-mail: yzhao0812@foxmail.com