基于运动跟踪与特征融合的视频实例分割方法

周 震,李 莹,柳德云,吉根林
(南京师范大学 计算机与电子信息学院/人工智能学院,江苏 南京 210023)

摘 要:视频实例分割(VIS)提供了对视频更深层次的理解,是智能监控、自动驾驶、机器人等领域高级任务的前置任务之一。目前对于图像实例分割已经有很多研究,但是对于视频实例分割的研究却相对较少,而将图像分割方法直接应用到视频领域也存在很多问题,其中实例被遮挡、实例成像差以及高速运动引起实例模糊等异常情况导致的追踪和分割效果差是主要问题。针对该问题,提出一种基于运动跟踪与注意力特征融合的视频实例分割方法(MTFA)。该方法利用运动跟踪头依据运动和特征信息在整个视频中跟踪实例并分配实例标签,然后按照实例标签对每一帧中实例去其他帧提取同一实例的特征信息,通过注意力机制融合这些特征信息用以增强当前帧的特征并生成分割掩码。该方法在 Youtube-VIS 数据集测试中最佳 AP 为 38.3%(ResNet-50)和 41.2%(ResNet-101)。

关键词:视频实例分割;图像实例分割;运动跟踪;特征融合;注意力机制
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2022)11-0043-07
doi:10.3969/j.issn.1673-629X.2022.11.007

Video Instance Segmentation Method Based on Motion Tracker and Feature Aggregation

ZHOU Zhen,LI Ying,LIU De-yun,JI Gen-lin
(School of Computer and Electronic Information/Artificial Intelligence,Nanjing Normal University,
Nanjing 210023,China)

Abstract:Video instance segmentation (VIS) provides a deep understanding of video and it is a pre-task for advanced tasks such as intelligent surveillance,autonomous driving and robotics. Many works focus on the image segmentation,but there is relatively few research on the video instance segmentation. There are many problems in applying image segmentation to video,the main problem is the poor segmentation and tracking result caused by instance occlusion,image blurring and so on. To solve the above problem,we propose a video instance segmentation method based on motion tracker and feature aggregation (MTFA). This method uses motion tracker to track instances across frames and assign labels to instances. According to these labels,the feature information of the same instance is extracted from other frames by instances in current frame,then the feature information of the current frame is enhanced by fused features from attentional feature aggregation module and segmentation masks are generated with enhanced feature. The best AP of the proposed method in the Youtube-VIS dataset test is 38.3% (ResNet-50) and 41.2% (ResNet-101).

Key words:video instance segmentation;image instance segmentation;motion tracker;feature aggregation;attention mechanism

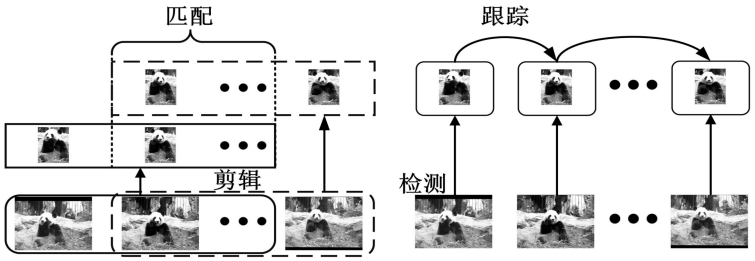
0 引 言

视频实例分割(VIS)的研究正变得越来越重要,它是计算机视觉中一项具有挑战性的研究内容。在图像领域中实例分割需要同时检测和分割对象实例[1],而在视频领域中,实例分割更具挑战性[2],因为它还需要准确跟踪和分类整个视频中的对象。

现有的 VIS 方法通常采用两种不同的思路来处理实例分割任务:第一种思路是“剪辑-匹配”,基于分而治之的思想。它将整个视频分成多个重叠的短片段(剪辑),并获得每个剪辑的 VIS 结果,最后合并生成具有逐个剪辑匹配的实例序列[3-4],如图 1(a)所示。另一种思路是“检测-跟踪”,利用跟踪头扩展现有的图像实例分割网络,首先使用图像实例分割网络逐帧进行目标检测和分割,然后通过跟踪头以分类或重识别的方法将这些目标进行关联以生成实例序列[2-5],如图 1(b)所示。这两种思路都需要从视频中生成多

收稿日期:2022-06-13 **修回日期**:2022-08-15
基金项目:国家自然科学基金资助项目(41971343,62102186)
作者简介:周 震(1996-),男,硕士,研究方向为大数据分析与挖掘;通讯作者:吉根林,博士,教授,博导,CCF 高级会员(09027S),研究方向为大数据分析与挖掘。

个不完整的序列(帧或者剪辑),然后通过跟踪/匹配来合并它们。



(a)剪辑-匹配 (b)检测-跟踪

图 1 视频实例分割的两种范式

现有的方法在合并序列的过程中很容易受到因目标被遮挡或快速运动导致的误检的影响,累积之后可能造成更大的分割误差。此外,现有方法普遍有一个缺点,即它们忽略了帧间实例关系。帧间实例关系是指不同帧实例之间的关系,这种帧间关系通常包含丰富的时空信息,对处理 VIS 任务很有用。最近的一些方法[4-5]已经注意到了这个问题,但是它们直接在帧级别融合了来自相邻帧的特征,这可能会导致目标特征信息传播不精确,从而对准确性产生负面影响,并且这些方法仅利用这些信息进行检测和分割,没有用于跟踪。

针对上述问题,在“检测-跟踪”思路的基础上提出了一种新的视频实例分割方法(MTFA)。具体而言,在目前图像分割网络的基础上添加一个新的运动跟踪头和自注意力特征融合模块,运动跟踪头借助运动模拟的位置信息和检测分割的特征信息来跨帧关联实例,特征融合模块借助运动跟踪头跟踪结果提取相应支持帧上的实例特征与目标帧实例特征进行基于注意力的融合,并将融合后实例特征增强的原特征图传入图像实例分割网络以生成效果更好的实例分割掩码。该方法实现了跟踪与分割任务之间的信息共享与相互帮助,提升了检测分割结果,有效解决了遮挡、快速运动和成像质量差导致追踪和分割效果差的问题。

1 MTFA 视频实例分割方法

文中的网络 MTFA 基于 QueryInst[6],包含一个特征提取骨干网络,一个基于实例查询的检测框 & 掩码生成器和一个运动跟踪头,还包括一个基于注意力的特征融合模块。MTFA 网络处理 VIS 任务的整体流程如图 2 所示。

图中上部实线箭头部分代表一阶段完成内容,通过图像实例分割网络实现帧级别的目标检测任务,然后通过运动跟踪头跨帧关联实例,并为每个检测框分配一个实例标签,这与先前的“检测-跟踪”网络相似。二阶段以目标帧 t 为例,以 $t \pm n$ 帧为支持帧,对于 t 帧中的每一个实例,提取 $t \pm n$ 帧中相应实例的特征图,通过一个基于注意力的特征融合模块得到融合的实例特征用以增强 t 帧特征图,并传入图像实例分割网络生成实例掩码。

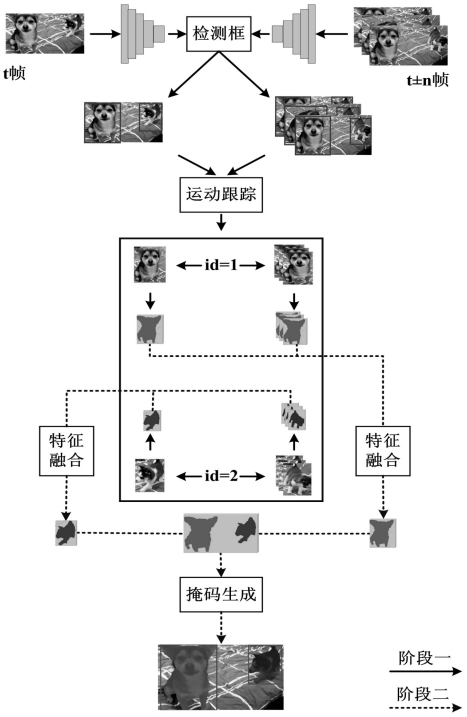


图 2 MTFA 视频实例分割处理流程

在本节中,首先介绍生成图像中实例检测框和分割掩码的 QueryInst 网络架构。然后详细介绍 MTFA 的网络结构以及各个模块的细节。

1.1 图像中实例的检测分割

QueryInst[6]是两阶段图像实例分割网络,它将图片中的实例作为一组查询来驱动整个网络,整个网络分为 6 个阶段,每个阶段以前一阶段输出为输入。以第三阶段 s_3 为例,流程如图 3 所示。

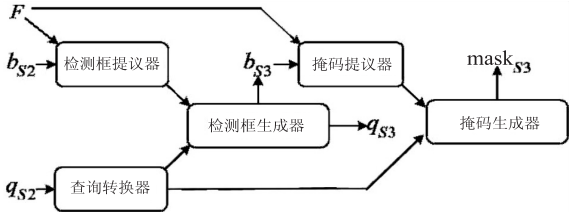


图 3 实例检测与分割流程

实例检测:对于检测框预测,处理顺序如下:在阶段 s_3 处理时,一个检测框提议器在前一阶段检测框预测结果 b_{s_2} 的指导下,从特征图 F (feature map)中提取当前阶段检测框特征。同时,将前一阶段查询对象 q_{s_2} 输入基于注意力的查询转换器中以获得转换后的查询对象。然后将当前阶段检测框特征信息和转换后查询对象输入检测框生成器用以生成当前阶段的检测框预测 b_{s_3},并为下一阶段生成查询对象 q_{s_3}。

实例分割:对于实例掩码预测,处理顺序如下:在当前阶段检测框预测 b_{s_3} 的指导下,掩码提议器从特征图 F 中提取当前阶段掩码特征。将当前阶段掩码特征和转换后查询对象输入掩码生成器用以生成实例级别掩码预测 mask_{s_3}

1.2　基于运动和特征的实例跟踪

MTFA 的运动跟踪头将实例检测框的中心点作为实例本身,然后根据视频的时间序列来跟踪中心点。该跟踪头能够计算前一帧到目前帧的中心点偏移量,并用运动信息和特征信息来关联相关实例。将时刻 t 的帧设置为目标帧,则帧中的一个追踪实例 i_t 可以用它的检测框中心点位置 c_t^i 和实例标识 $\text{id}_t^i \in [0, N]$ 来表示,其中 N 是到目前为止的不同实例总量。如果 i_t 是先前实例之一,则分配 N 中的一个实例标识。如果它是新的实例,则分配新的实例标识,并将 N 扩充。为了实现上述过程,跟踪头预测一个二维运动过程 $M \in R^{H\times W\times 2}$ 并计算特征相似度 $i_t^F \otimes i_{t-1}^F$,对实例的运动和匹配的损失进行如下定义:

$$L_{\text{track}} = \frac{1}{n}\sum_{i\in n} |M_t^i - (c_t^i - c_{t-1}^i)| - \sum_j \log(p^j(y_{t-1}^i)) \tag{1}$$

其中,M_t^i 描述了实例 i 在 t 和 $t-1$ 帧之间的中心点运动 $c_t^i - c_{t-1}^i$,n 为当前帧里的实例总量,y_{t-1}^i 是 $t-1$ 帧搜索范围内的 j 个实例标签,p^j 是对应标签的归一化相似度分数。

具体来说,对于 t 帧中心点在位置 c_t^i 的实例对象 i,运动跟踪头在 $t-1$ 帧中心点位置在 $c_t^i - M_t^i$ 处半径为 r 圆内搜索同类实例,如果在半径 r 内仅有一个同类实例则直接与其相关联。如果有多个同类实例,则将其实例特征与区域内同类实例特征进行内积,得到得分矩阵,将实例身份与最高得分的实例相关联。如果在半径 r 内没有找到匹配的候选者,则会生成一个新的追踪对象 id_i。将半径 r 定义为当前帧跟踪对象检测框宽度和高度的平均值。如图 4 所示,图中方框内箭头表示模拟运动 $c_t^i - M_t^i$ 的过程,⊗表示内积计算特征相似度。

MTFA 的运动跟踪头结合了运动信息和外观特征信息,使得跟踪性能超越了之前的运动跟踪头,并克服例如目标交错、遮挡导致的跟踪错误。相较于对全图特征进行跟踪的工作计算量更小,且忽略了背景信息和其他不可见的干扰信息,使得追踪更具效率。

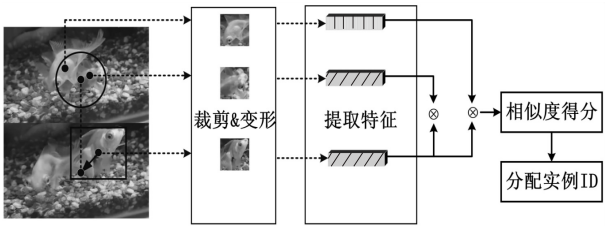


图 4　运动跟踪头示意图

1.3　基于注意力的特征融合

在上述运动跟踪头的作用下,MTFA 能够在整个视频中追踪不同的实例。为了让来自不同帧的实例信息帮助 MTFA 更好完成视频实例分割任务,提出了基于注意力的特征融合模块,借助该模块就能够对成像质量较差(遮挡、模糊等)的帧中实例进行更好的分割掩码生成。同样的设定时刻 t 的帧为目标帧,时间段内的其他 T 帧为支持帧。下面的关键是如何有效地聚合这些特征并生成质量更好的分割掩码。由于实例在某些帧中可能是模糊的,而在其他帧中可能是清晰的,因此很自然地想到学习一组注意力权重来聚合它们。目前的多头自注意力网络[7]可以通过不同的通道关注来自不同子空间的信息。因此 MTFA 构建了一个多头注意力模块来处理特征聚合中的不同模式,如图 5 所示。

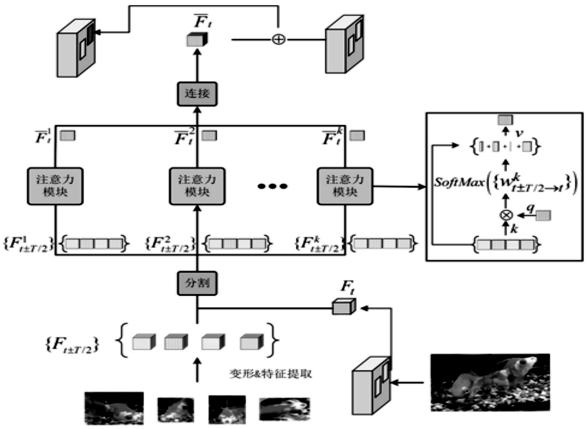


图 5　特征融合示意图

输入是一组以目标帧为中心支持帧总数为 T 的帧中某一实例特征 $F_{t\pm T/2}$。目标帧利用原图检测框坐标点映射到特征图进行定位提取检测框中的 $H \times W \times c$ 的特征图[8],支持帧的特征通过对 T 帧内同实例检测框裁剪 & 变形提取得到 $T \times H \times W \times c$ 的特征图,这里使用的是经过 FPN 提取的 4 层 256 通道的特征图,对每层进行上述操作。用 K 个注意力模块从不同维度聚合这些特征。

首先,将 $F_{t\pm T/2}$ 中的目标特征沿通道维度分成 K 组:

$$F^k_{t\pm T/2} = F_{t\pm T/2}[:,:,(k-1)\frac{c}{K}:k\frac{c}{K}] \tag{2}$$

其中, $F_{t\pm T/2} \in R^{H\times W\times(c/K)}$ 并且 $k \in \{1,2,\cdots,K\}$, 每个 $F_{t\pm T/2}$ 用于生成一个注意力权重图:

$$w^k_{t\pm T/2\to t} = \mathrm{softmax}(u^k(F^k_{t\pm T/2}) \otimes u^k(F^k_t)) \tag{3}$$

其中, $u^k(\cdot)$ 是一个微型的嵌入网络,并且 $u^k(F_{t\pm T/2}) \in R^{H\times W\times(c/K)}$ 。其中 $w^k_{t\pm T/2\to t}$ 表示第 k 组特征图从 $t\pm T/2$ 帧到 t 帧的归一化注意力权重。基于归一化的第 k 组时间注意力权重图 $w^k_{t\pm T/2\to t}$,第 k 组目标特征 $F^k_{t\pm T/2}$ 加权求和如下:

$$\bar{F}^k_t = \sum w^k_{t\pm T/2\to t} \bullet F^k_{t\pm T/2} \tag{4}$$

最终的跨帧融合特征 $\overline{F_t}$ 可以通过沿通道维度连接所有 $\overline{F^k_t}$ 来获得并且 $\overline{F^k_t} \in R^{H\times W\times(c/K)}$, $\overline{F_t}$ 与 F_t 大小及通道数相同为 $H\times W\times c$,但它包含视频中同一对象实例不同时间的信息,然后将 $\overline{F_t}$ 与初始目标帧中相应区域特征图作相加操作,增强目标区域特征,最后作为目标帧的特征传入图像实例分割网络生成最佳实例掩码。在某些情况下,例如某对象实例在某些支持帧中消失,借助运动跟踪头的优秀设计并不将这些帧纳入特征融合的计算范围,仅计算从其他支持帧中提取的实例特征。运动跟踪头的设计保证了 MTFA 的特征融合模块可以适用于大多数情况。注意力特征融合模块在目标帧中的对象被部分遮挡或模糊时,仍然可以从支持帧中提取对象特征,由于大多数空间位置而没有遮挡。因此,实例对象的可见部分在整个视频中仍占主导地位,目标帧这些实例的特征仍然可以得到增强,从而得到更好的实例分割效果。

1.4 损失函数

MTFA 总的损失函数包含目标检测、目标分割和目标追踪这三个方面的损失,具体的损失项如下式所示:

$$L_{\mathrm{all}} = L_{\mathrm{det}} + L_{\mathrm{mask}} + L_{\mathrm{track}} \tag{5}$$

其中, L_{det} 是目标检测损失函数, L_{mask} 是目标分割损失函数, L_{track} 是目标追踪损失函数。L_{track} 的定义已在上文中给出, L_{mask} 是 Dice loss [9],如下式所示:

$$L_{\mathrm{mask}} = 1 - \frac{2|m^i \cap m^j|}{|m^i| + |m^j|} \tag{6}$$

其中, m^i 是预测掩码, m^j 是真实掩码, L_{det} 包含类鉴别损失和检测框损失,其中类鉴别损失是 Focal loss[10]计算多分类的准确性,检测框损失是 L1 loss 计算检测框真实坐标与预测坐标的平均差值。

$$L_{\mathrm{det}} = \frac{\sum_{q=1}^{n}|y_q - \overline{y_q}|}{n} - \frac{(1-p_t)^2\log(p_t)}{4} \tag{7}$$

其中, y_q 指的是真实坐标, $\overline{y_q}$ 是预测坐标, n 为坐标点个数,此部分是 L1 loss。p_t 为归一化的标签预测值, $-\log(p_t)$ 为交叉熵损失, $(1-p_t)^2/4$ 为调节因子,两部分组合为 Focal loss。

2 实验及结果分析

2.1 实验数据和评价指标

该文的实验数据集是公开的 VIS 数据集 Youtube-VIS 2019[2]和 Youtube-VIS 2021,遵循大多数以前的工作[2,5,11]在测试集上评估提出的方法。

评估指标是平均精度(AP)和平均召回率(AR),以视频预测目标掩码序列与真实掩码序列的交集/并集(IoU)为阈值[2]。具体来说,该文的 AP 是按照多个 IoU 为阈值得到的精度(precision)取的平均值,AR 定义是视频里固定数量(该文为 1 和 10)检测结果最大的召回率(recall)。这两个指标都先在每个类别内求平均,再在所有类别上求平均,计算公式如下:

$$\begin{cases}\mathrm{precision} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\\ \mathrm{recall} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\end{cases} \tag{8}$$

其中,TP 代表正样本归为正类的数量,FP 代表负样本归为正类的数量,FN 代表正样本归为负类的数量。对于 IoU 的计算,预测掩码 m^i 和真实掩码 m^j 的交并集为:

$$\mathrm{IoU}(i,j) = \frac{\sum_{t=1}^{T}|m^i_t \cap m^j_t|}{\sum_{t=1}^{T}|m^i_t \cup m^j_t|} \tag{9}$$

根据定义,如果模型仅成功检测和分割实例但未成功关联实例,它仍然得到很低的 IoU。因此实例的准确跨帧关联对于实现模型高性能至关重要。

2.2 实验设定

该方法基本的训练设置主要遵循 QueryInst[6]。检测头包含 6 个阶段,查询总数设置为 100。采用 ResNet-50 和 ResNet-101 作为骨干网络,并使用 COCO 数据集预训练的权重进行参数初始化,运动跟踪头为 3 层卷积神经网络,每层包含一个卷积层、一个归一化层和一个 ReLU 的激活函数层。注意力融合模块的注意力块个数 K 设置为 4,微型嵌入网络 $u^k(\cdot)$ 为 3×3 的卷积层。代码在训练和测试阶段均使用了基于 Pytorch 的 MMDection[12]和 MMTracking[13]提供的开发框架。

对于训练,在 8 个 12G 显存的 GPU 上执行了总共 36 轮迭代训练,对于每次迭代,batch size 设置为 5,使用 SGD 作为优化器。初始学习率为 1.25×10^{-4},在第

27 和第 32 轮迭代,学习率除 10。使用 Youtube-VIS 数据集进行训练,输入为同一个视频的 5 帧,每帧为原视频间隔 5 帧的关键帧,遵循之前的工作[2-3,14-15]调整输入图像的大小,使输入尺寸为 640×360。

对于测试,使用一个 12G 显存的 TiTan XP 进行评测,来自同一视频的 4 帧(支持帧)与目标帧一起被采样。如果支持帧超出视频开始/结束,复制视频的第一帧/最后一帧作为支持帧。跟踪头用于关联实例,实例掩码是从最后阶段图像实例分割网络中生成的。评测阶段的所有输入图像都被调整大小,使输入尺寸为 640×360。

2.3　实验结果对比

在 Youtube-VIS 2019 的实验结果以及与现有方法的对比见表 1,表中列出了不同方法所用的骨干网络,处理视频的分辨率均为 640×360。

文中方法在所有评价指标上都取得了相当有竞争力的结果,在 ResNet-50 骨干网络下取得了 38.3% AP,在 ResNet-101 骨干网络下取得了 41.2% AP,表中其他方法数据均来自于原论文。具体来说,比此前最通用的方法,同样使用“检测-跟踪”方式的 MaskTrack R-CNN[2] AP 高出约 10 个百分点(在 ResNet-101 的情况下)。比起所对比的最好方法 VisTR[14]高出 2.1 个百分点,CrossVIS[11]和 SG-Net[15]高出 3.5 个百分点(在 ResNet-50 的情况下)。最后相比较文中的 baseline QueryInst[6]高出约 2.1 个百分点,充分说明了文中添加模块的作用。

表 1　Youtube-VIS 2019 数据集与现有方法性能比较

Backbone	Method	Venue	AP/%	AR/%
ResNet-50	MaskTrack R-CNN[2]	ICCV'19	30.3	31.0
	SipMask-VIS[5]	ECCV'20	33.7	35.4
	STEm-Seg[3]	ECCV'20	30.6	31.6
	VisTR[14]	CVPR'21	36.2	37.2
	CrossVIS[11]	ICCV'21	34.8	34.0
	SG-Net[15]	CVPR'21	34.8	35.8
	QueryInst[6]	ICCV'21	36.2	36.1
	Our Method	-	38.3	37.7
ResNet-101	MaskTrack R-CNN[2]	ICCV'19	31.9	33.5
	STEm-Seg[3]	ECCV'20	34.6	34.4
	VisTR[14]	CVPR'21	40.1	38.3
	SG-Net[15]	CVPR'21	36.3	35.9
	CrossVIS[11]	ICCV'21	36.6	36.0
	Our Method	-	41.2	39.6

在 Youtube-VIS 2021 的实验结果以及与现有方法的对比见表 2。由于对比方法未提供 ResNet-101 的相关数据,这里仅展示文中方法在 ResNet-101 骨干网络下的性能表现。

表 2　Youtube-VIS 2021 数据集与现有方法性能比较

Backbone	Method	Venue	AP/%	AR/%
ResNet-50	MaskTrack R-CNN[2]	ICCV'19	28.6	26.5
	SipMask-VIS[5]	ECCV'20	31.7	30.8
	CrossVIS[11]	ICCV'21	33.3	30.1
	Our Method	-	34.4	31.6
ResNet-101	Our Method	-	34.9	32.7

特意选择了与该文使用同样的“检测-跟踪”范式的方法,可以看到文中方法性能最好。其中 Mask Track R-CNN[2]仅使用特征信息关联实例,MTFA 结合了运动与特征来关联实例。Sip Mask-VIS[5]仅使用当前帧信息生成掩码,MTFA 让不同帧实例的特征信息协助生成掩码。与 Cross VIS[11]利用帧级别信息协助生成掩码不同,MTFA 利用实例级别的信息生成掩码,效果更好。

2.4　消融实验

为了验证该文添加模块的作用,在 ResNet-50 骨干网络下进行了消融实验,分别比较了仅添加运动跟踪头和添加了特征融合模块的网络性能,结果如表 3

所示。

表 3 消融实验

Method	AP/%	AR/%
Baseline	36.2	36.1
Baseline + Tracker	37.9	37.0
Baseline + Tracker + FA	38.3	37.7

表中 Tracker 代表运动跟踪头,FA 代表特征融合模块。通过消融实验可以看到在添加了运动跟踪头后,MTFA 相比 baseline(QueryInst)已经有了很大提升(约 1.7 个百分点),但是没有充分发挥出跟踪器对检测结果的作用,在添加了特征融合模块后,MTFA 比较添加跟踪器的方法也有了一定提升。这表明好的跟踪结果对于视频实例分割的作用是巨大的,因为 MTFA 特征融合的模块建立在运动跟踪的结果之上,所以不能单独作消融实验,可以看到 MTFA 两个模块在视频实例分割任务中是相辅相成的。

2.5 参数影响分析

MTFA 的网络主要包含两个超参数,注意力模块个数 K 和支持帧数量 T,为了充分探讨网络受超参数的影响,对于 K 在 $T=4$ 的情况下使用不同的 K 进行了实验,比较了不同参数下网络的性能,结果如表 4 所示。

表 4 不同数量的注意力模块对于网络性能的影响

#K	1	2	4	8
AP/%	37.9	38.1	38.3	38.3

表 4 显示了在 FA 中使用不同数量的时间注意块的效果。随着 K 从 1 增加到 4,AP 从 37.9% 增加到 38.3%。这表明使用更多的时间注意力块可以提高准确性。但当 K 大于 4 时性能饱和且消耗更多资源。因此,选择 $K=4$ 作为默认设置。

关于 T 的数量对性能的影响,也对此进行了实验,此时将 K 设置为 4,结果如表 5 所示。

表 5 不同数量的支持帧对网络性能的影响

#T	2	4	6	8
AP/%	38.0	38.3	38.3	38.1

表 5 显示了不同数量支持帧对网络性能的影响。可以看到,支持帧数量从 2 ~ 4 的提升很大,4 ~ 6 性能不变,6 ~ 8 性能下降,推断是此时融合了过多噪声信息并且最后直接对特征使用加和操作导致无法过滤无用信息导致的。因此将支持帧个数 T 设置为 4。

2.6 结果可视化

图 6 展示了 MTFA 在 Youtube-VIS 数据集和部分自己测试数据使用 ResNet-101 为骨干网络得到的可视化结果,以一个视频的 5 帧为例,前后帧中同一实例用相同颜色的掩码进行标记。

第一列展示了同一实例不同姿态以及部分遮挡的分割结果,总体来说分割效果是比较好的,但是在毛发边缘模糊的部分分割效果不是太好。第二列展示了多个同类实例高速运动和大量重叠时的分割结果,可以看到 MTFA 无论分割还是追踪效果都非常出色,图片中的物体微小的边角也被识别分割出来,分割边缘也比较清晰。第三列展示了两个同类物体靠近交错时的跟踪效果和分割结果,可以看到在实例交错时,分割掩码边缘仍然是清晰没有杂乱的,追踪也没有混乱。

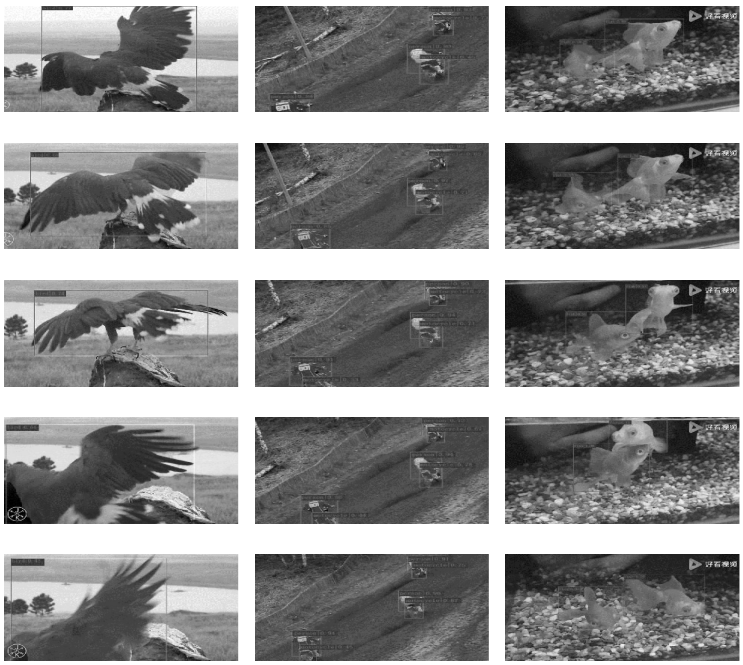


图 6 可视化结果

3　结束语

本文提出了一种基于运动跟踪和注意力特征融合的方法 MTFA。该方法充分利用运动信息和特征相似性加强了对实例身份的跟踪操作,并利用跟踪结果对模糊不清或被遮挡的帧中实例进行特征融合从而实现更好的分割。MTFA 在 Youtube-VIS 2019 和 Youtube-VIS 2021 测试中对比目前许多主流方法都取得了相当有竞争力的结果。此外,经过消融研究表明,MTFA 的运动跟踪和注意力特征融合结合的模块可以显著提高视频实例分割的性能。

参考文献:

[1] 梁新宇,林洗坤,权冀川,等. 基于深度学习的图像实例分割技术研究进展[J]. 电子学报,2020,48(12):2476-2486.

[2] YANG L,FAN Y,XU N,et al. Video instance segmentation [J]. arXiv:1905.04804,2019.

[3] ATHAR A,MAHADEVAN S,OSEP A,et al. Stem-seg: spatio-temporal embeddings for instance segmentation in videos[C]//European conference on computer vision. Glasgow:Springer,2020:158-177.

[4] BERTASIUS G, TORRESANI L. Classifying,segmenting, and tracking object instances in video with mask propagation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:9739-9748.

[5] CAO J,ANWER R M,CHOLAKKAL H,et al. Sipmask: spatial information preservation for fast image and video instance segmentation[C]//European conference on computer vision. Glasgow:Springer,2020:1-18.

[6] FANG Y,YANG S,WANG X,et al. Instances as queries [C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal:IEEE,2021:6910-6919.

[7] TAO Y,SUN Q,DU Q,et al. Nonlocal neural networks,nonlocal diffusion and nonlocal modeling [J]. arXiv: 1806.00681,2018.

[8] HE K,ZHANG X,REN S,et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,37(9):1904-1916.

[9] MILLETARI F,NAVAB N,AHMADI S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[J]. arXiv:1606.04797,2016.

[10] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,42(2):2999-3007.

[11] YANG S,FANG Y,WANG X,et al. Crossover learning for fast online video instance segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal:IEEE,2021:8043-8052.

[12] CHEN K,WANG J,PANG J,et al. MMDetection: open mmlab detection toolbox and benchmark[J]. arXiv: 1906.07155,2019.

[13] WU C,ZHANG F,WANG B,et al. mmTrack:passive multi-person localization using commodity millimeter wave radio [C]//IEEE INFOCOM 2020-IEEE conference on computer communications. Beijing:IEEE,2020:2400-2409.

[14] WANG Y,XU Z,WANG X,et al. End-to-end video instance segmentation with transformers [J]. arXiv: 2011.14503,2020.

[15] LIU D,CUI Y,TAN W,et al. SG-net: spatial granularity network for one-stage video instance segmentation[J]. arXiv:2103.10284,2021.