极化自注意力调控的情景式视频实例多尺度分割

黄 滢 何自芬 杨宏宽 赵崇任 张印辉
(昆明理工大学机电工程学院 昆明 650500)

摘 要 视频实例分割(Video Instance Segmentation)是开发智能机器人视觉系统的一项关键技术,部署视频实例分割算法的智能机器人能够精确地实现目标追踪、避障等高复杂度任务.机器人在特定情景下自主移动时的成像效果会受到自身速度、拍摄角度、距离远近及目标移动速度的影响,导致捕获的运动目标普遍存在拓扑形变和尺度缩放等随机性问题.对于在相同视频序列中跨静态帧的同一目标实例而言,模型所学习的可辨识特征往往具有多样性和不确定性.现有模型更多强调帧间掩膜传播或特征跟踪等时序交互方法,而忽略了对拓扑实例的深层语义解析和尺度目标的轮廓分辨,因此缺乏对高层细粒度特征的有效关注和低层空间信息的准确定位.本文提出一种极化自注意力调控的多尺度视频实例分割 PSAM-Net(Polarized Self-Attention Manipulation Network)模型.首先,在残差网络中嵌入单级式和级联式的极化自注意力机制,以建立任意空间位置的非线性关联信息,及其正交方向的通道特征依赖关系,克服高层特征图中细粒度特征分布弥散问题,增强模型的区域特征聚焦能力,完成对拓扑实例的深层语义解析;其次,由特征金字塔自上而下的特征流动方式,所导致的低层特征空间位置和实例边缘信息匮乏问题,对聚合了多粒度信息的空间定位分支模型进行构建,以适应不同尺度下前景目标定位检测和轮廓分割需求.最后,从 Youtube-VIS 中划分出多个适用于动物场景下的数据集.交叉验证结果表明,相较于 YolactEdge 基准模型,PSAM-Net 在平均检测和平均分割精度上分别提升 6.08%和 8.87%,达到 44.06%和 44.41%,测试速度高达 80FPS,表现出较好的鲁棒性与稳定性.本文方法实现了视频序列输入下的实时高精度分割,为智能移动机器人的自主环境感知提供了有效理论依据和一定参考价值.

关键词 视频实例分割;拓扑形变;尺度缩放;PSAM-Net;极化自注意力调控;空间定位分支
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02605

Multi-Scale Segmentation of Episodic Video Instance through Polarized Self-Attention Manipulation

HUANG Ying HE Zi-Fen YANG Hong-Kuan ZHAO Chong-Ren ZHANG Yin-Hui
(*Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming* 650500)

Abstract Video instance segmentation (VIS) is a key technology for developing vision systems of intelligent robots, and ones deployed with video instance segmentation algorithms can accurately perform highly complex robotic tasks, such as target tracking and obstacle avoidance. The imaging results that are acquired during environment perception when the robot moves autonomously in a specific scene is easily affected by its own motion speed, shooting angle, distance from the target position and the target relative motion speed, resulting in randomness problems such as topological deformation and scale scaling of the captured moving targets in general. For the same target instance exsting across the adjacent static frames in a series of video frames, it is generally diverse and uncertain in terms of the discernible feature representations which are learned by the model of common methods. Existing video instance segmentation models mostly emphasize more

收稿日期:2021-12-21;在线发布日期:2022-08-12.本课题得到国家自然科学基金(62061022,62171206)资助.**黄 滢**,博士研究生,主要研究方向为深度学习和机器视觉.E-mail:erying_h@163.com.**何自芬**,博士,副教授,主要研究方向为图像处理和机器视觉.**杨宏宽**,硕士研究生,主要研究方向为深度学习和机器视觉.**赵崇任**,硕士研究生,主要研究方向为深度学习和机器视觉.**张印辉**(通信作者),博士,教授,主要研究领域为图像处理和机器视觉.E-mail:zhangyinhui@kust.edu.cn.

on temporal interaction methods such as inter-frame mask propagation or multi-scale feature tracking, through which the deep semantic parsing of topological instances and contour discrimination of targets in multi-scale are neglected, thus the effective attention to high-level fine-grained features and accurate localization of low-level spatial information are seriously limited. To address the above mentioned issues, a model based on polarized self-attention manipulation for multi-scale video instance segmentation (named PSAM-Net) is proposed in this paper. First of all, in order to establish the positional correlation information between arbitrary non-linear spaces and dependences among orthogonal channels, we propose a single-stage and a cascaded polarized self-attention manipulation mechanisms, which are respectively embedded in the residual network after each residual block in an optimal form. The above mentioned measures benefit to overcome the dispersion of regression distribution for fine-grained features in the feature maps of deep levels and enhance the feature focusing ability of the model on key regions, so as to complete the task of deep semantic parsing of topological instances. Secondly, a multi-scale spatial location model of multi-granularity spatial information is established, which can make up for the lack of low-level feature space location and indistinct instance edge information caused by the feature flow from the top to the bottom of the feature pyramid networks. Through this can we achieve better requirements of target location detection and contour segmentation for foreground objects under different scales. Finally, we construct the episodic video dataset of animal instances extracted from Youtube-VIS, and extensive experiments are conducted to verify the competitive performance of our method. Compared with the YolactEdge benchmark model, the comprehensive testing results obtained by the PSAM-Net model show multiple improvements, where the average detection reaches 44.06% which is increased by 6.08%, and the average segmentation accuracy is increased by 8.87% to reach 44.41% respectively. In addition, the proposed PSAM-Net is capable of processing 80 video frames per second, which is well above the real-time requirements for video instance segmentation task. The model in this paper realizes real-time high-precision segmentation of video instances, and provides an effective theoretical basis and certain reference value for the autonomous environment perception of intelligent mobile robots.

Keywords video instance segmentation; topological deformation; scale scaling; PSAM-Net; polarized self-attention manipulation; spatial positioning branch

1 引 言

情景式移动机器人的视觉分割系统是开发智能机器人领域的一项难题. 机器人在特定场景下移动时视觉成像系统会受到自身速度、拍摄角度及距离目标拍摄位置远近的影响,加上捕捉的目标对象在运动过程中存在拓扑形变和尺度缩放的随机性,导致后续成像时目标实例边缘模糊、且同一目标尺度大小形态不一,对分割任务会造成极大挑战[1]. 因此,在不确定因素下的视频序列中如何准确稳定地分割目标实例是情景式机器人视觉系统一个亟待解决的问题.

围绕实例分割需要同步实现的多目标检测和分割任务,研究人员从双阶段和单阶段模型两条主线出发以分别改善现有模型的分割精度和速度. 在由感兴趣区域提取和掩膜分割构成的双阶段主线上,He 等人[2]在 Faster R-CNN[3]目标检测框架基础上提出 Mask R-CNN,通过增加一个用于预测实例掩膜的分支能在有效检测目标的同时为每个实例生成一个高质量的分割掩膜. 针对 MaskR-CNN 中利用边界框的分类置信度作为实例掩膜得分而导致的掩膜分割质量与最终得分不配准问题,Huang 等人[4]提出 Mask Scoring R-CNN,采用预测掩膜与真实掩膜的交并比作为新的掩膜质量评价策略,使得模型最终输出的掩膜轮廓信息更加贴合真实值以提高实例分割精度. 为进一步改进复杂遮挡背景下的掩膜分割质量,Ke 等人[5]提出双图层级联图神经网络

BCNet，通过将顶层和底层图卷积神经网络作为现有分割模型的掩膜预测分支，分别对被遮挡实例的形状位置及掩膜进行检测分割从而改善现有实例分割模型在处理多目标遮挡时的掩膜分割质量.双阶段模型虽然能取得较好的分割精度，但需要提取感兴趣区域导致模型计算复杂度增加分割效率低下.

单阶段模型摒弃了感兴趣区域提取以提高分割速度，YOLACT[6]同时生成系列原型掩膜和各个实例的掩膜系数后对生成结果执行对应线性加权组合操作，依据目标边界框裁剪得到每个实例对应的掩膜特征并提出快速非极大值抑制(Fast Non-Maximum Suppression)算法进一步提高模型的推理速度.Xie 等人[7]以单阶段无锚框目标检测模型 FCOS[8]为基础框架提出 PolarMask，将实例分割转化为极坐标下的实例中心分类和密集距离回归问题，以设定的目标先验角度和预测的中心坐标到目标轮廓射线距离来回归最终实例掩膜，通过优化实例分割任务建模步骤提升了分割速度.单阶段模型无需经过区域候选及位置修正等操作，相比于双阶段模型而言，分割速度更快，然而因其无法提取拓扑形变和尺度缩放目标候选区域而制约了分割精度.

双阶段和单阶段实例分割算法虽然在一定程度上能够提高模型的分割精度或速度，但在处理视频序列时还是把各帧作为单独的图像输入，只关注帧间独立的空间信息而忽略了时序一致性关系，本质上还是对图像独立分割.视频实例分割任务要求在各帧检测分割基础上，同时挖掘帧间目标实例的时空关联性并捕捉同一实例的全局上下文语义信息，在视频帧间完成实例跟踪的任务.然而不同运动实例在视频序列中会存在目标漂移、拓扑形变及尺度缩放等问题，即同一实例在不同帧中特征表示具有多样性.若先前静态帧特征学习结果不准确或错误，则后续帧在对其进行特征跟踪或掩膜传播时会引入噪声而干扰学习，因此视频分割任务对目标实例在静态帧中特征表述的准确度有很强的依赖性.针对上述问题，本文的主要贡献如下：

(1)提出一种基于单级式和级联式的极化自注意力(Polarized Self-Attention，PSA)[9]调控模型，其目的是激活高层特征的显著区域抽象语义信息，提高模型对可变形目标在静态帧中的可辨识特征表达能力.

(2)提出多尺度空间定位分支，与特征金字塔网络(Feature Pyramid Networks，FPN)[10]进行多粒度特征交互，丰富空间位置及实例轮廓边缘信息，在目标尺度动态变换的前提下增强模型学习的适应性和有效性.

(3)在视频实例分割数据集 Youtube-VIS[11]中构建适用于动物场景下的情景式数据集，并与基准模型[12]进行对比.实验结果表明，所提出的模型其平均检测精度提升了 6.35%，达到 44.33%，平均分割精度提升了 10.21%，达到 41.75%，在 RTX2080Ti 上推理速度高达 73FPS.通过 Bootstrap 交叉验证方式在所构建的五组数据集上综合验证结果表明，PSAM-Net 的平均检测精度提升了 6.08%，达到 44.06%；平均分割精度提升了 8.87%，达到 40.41%.

2　相关工作

为了充分利用视频序列的时序一致性关系及满足视频实例分割任务需求，现有视频实例分割方法主要从实例追踪和掩膜传播两个方向开展工作.

基于实例跟踪的方法主要是利用帧间的时序一致性特征，通过不同帧间实例跟踪的方式丰富每帧的全局上下文信息，关联各帧结果从而消除单帧信息带来的模糊性.Yang 等人[11]通过在 Mask-RCNN 检测和分割头部的基础上引入了一个追踪头部分支，将当前帧的实例追踪任务转变为一个多分类问题用于跨帧跟踪实例对象，然而各帧的特征提取过程没有利用其它帧的信息，依然是相互独立关系因此缺乏时序上的关联性.此外，Fu 等人[13]提出一种结合时间和空间的双重注意力综合特征聚合模型 CompFeat，利用时间和空间上下文信息分别完善序列帧级和对象级细化特征，提高模型的跟踪能力.然而建模当前帧与整个视频的时序关系时会反复计算时序注意力模块，造成一定的复杂度开销而影响模型推理速度.

基于掩膜传播的方法是利用视频序列中帧间的像素相似性或冗余性，将视频片段的帧级对象空间特征通过时序信息传播到其它帧中，通过特征交互完成实例追踪.Yang 等人[14]提出 MaskProp 以视频序列中各帧作为中心，将实例对象级掩膜特征传播给时间邻域的其它帧从而关联帧间对象级实例特征，虽然分割精度得到了提升然而其离线学习的方式导致了模型占用内存较大且训练周期较长.Liu 等人提出 YolactEdge[12]利用视频帧间的冗余性对关键帧与参考帧分别采用不同的特征提取方式，关

键帧的特征保持整体 YOLACT 的计算方式，而参考帧深层特征利用与其时间意义最近的前一关键帧传播方式获得，以确保减少不必要的时空复杂度开销并提高分割速度. YolactEdge 在时间意义上虽然可以通过特征传播减少冗余帧的计算复杂度从而加快模型测试推理速度，但对所定义关键帧的深层特征有很大依赖性，若关键帧的分割效果较差或参考帧与关键帧所包含的前景目标相比尺度变化较大且形变复杂时，这种特征传播方式会对参考帧的分割任务造成一定误导.

综上所述，相较于之前大多的视频实例分割算法更加关注帧间掩膜传播或特征跟踪等问题，本文更加注重同时伴随尺度变换的拓扑形变目标，在静态帧中的可辨识特征表达能力和掩膜分割能力.

3 PSAM-Net 模型

3.1 模型总体架构

PSAM-Net 模型总体架构主要包含特征提取和特征后处理两个部分. 特征提取部分由 ResNet50-PSA 调控模型、特征金字塔网络和空间定位分支构成；特征后处理部分由预测头分支、原型掩膜分支、非极大值抑制、裁减及阈值化五部分构成. 模型总体架构如图 1 所示.

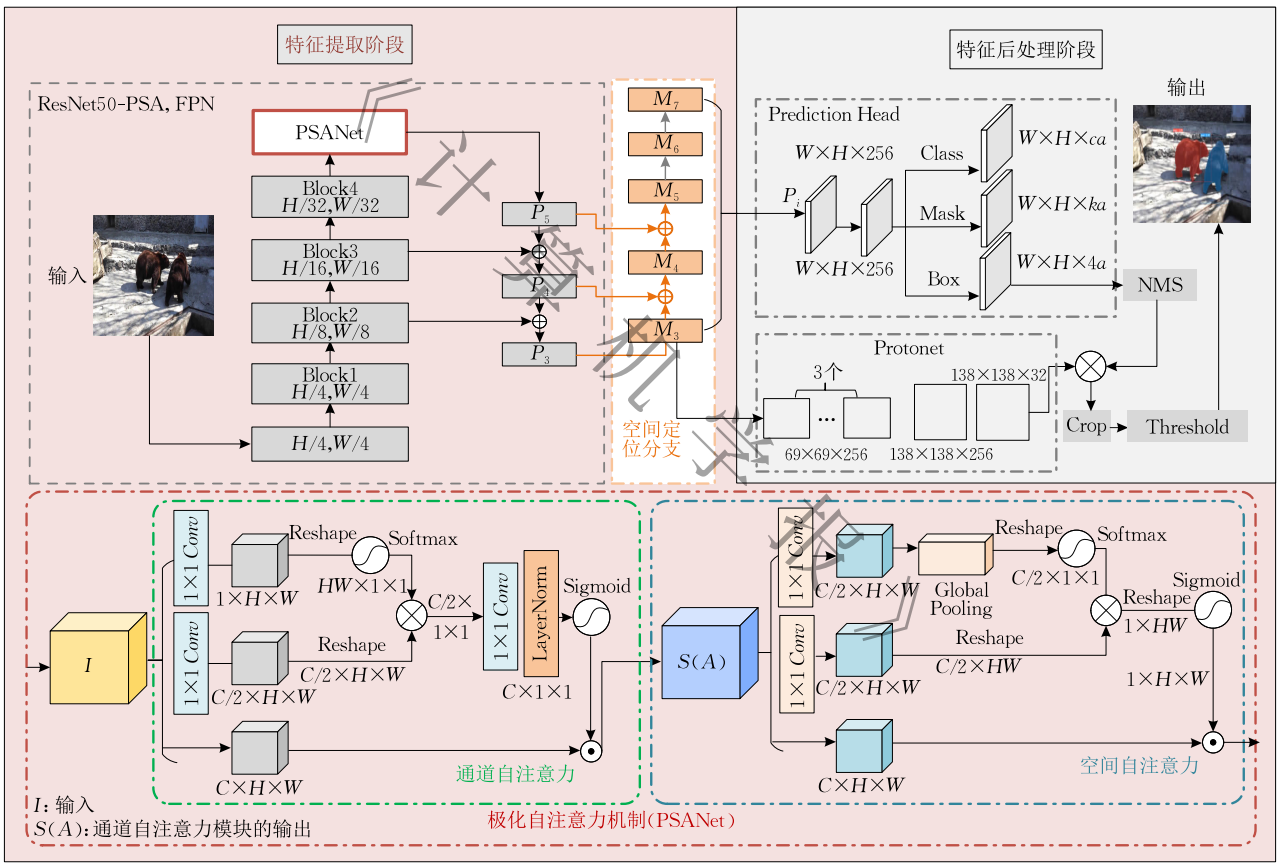


图 1 PSAM-Net 模型架构

在特征提取部分，首先将视频帧输入到 ResNet50-PSA 调控模型中提取语义特征，通过建立非线性的深层特征通道依赖关系和空间位置关联信息，聚焦重要区域及目标特征. 该调控模型的骨干网络主要由四个残差块(Residual Block)组成，将残差块简称为 Block，四个残差块依次简称为 Block1、Block2、Block3 和 Block4. 其次，通过特征金字塔自顶而下的融合方式为浅层特征引入更丰富的深层语义信息. 最后，对聚合了目标实例多粒度空间信息的多尺度空间定位分支模型进行构建，通过自下而上的方式将高分辨率、语义信息相对较弱但空间位置信息较强的低层特征与低分辨率但语义信息较强的高层特征逐级融合，丰富多尺度特征图的深层抽象特征和空间定位信息.

在特征后处理部分，将空间定位分支的多尺度特征图作为预测头网络输入，经过三个并行分支头得到输出结果. 其中，分类分支头获取当前图像的目标分类置信度；边界框分支头生成当前图像的目标

检测框位置偏移量，掩膜分支头对每一个锚框都预测一个掩膜系数；由于 M_3 层具有丰富的高层语义和低层空间特征，将其作为与预测头并行的原型掩膜分支输入，经过由五层卷积组成的全卷积网络生成通道维度为 32 的原型掩膜，该通道维度与掩膜系数个数相对应；最后对预测头的边界框分支应用快速非极大值抑制对同一目标位置上的重叠候选框完成筛选，并将原型掩膜与其对应的模板系数线性组合生成实例掩膜.

3.2　ResNet50-PSA 调控模型

启发于光学透镜在滤光时会对所有平行于主轴横向的光线进行折射和反射以抑制其传播，而允许正交于横向的光线透过这一原理. 极化自注意力机制将非线性关联信息的通道与空间分辨率两者抽象为正交关系以此建模为通道自注意力与空间自注意力两个子机制，对任一建模方向的子自注意力机制利用极化滤波思想将其中一个分支张量完全折叠，即对该建模维度完全压缩以过滤当前特征，而另外一个分支张量在该建模方向保持较低压缩水平以避免丢失较多信息，同时令其对应正交方向特征保持高维度张量以显著特征对比. 其次利用 Softmax 操作提升在子自注意力机制中由于压缩折叠所损失的特征信息，最后通过 Sigmoid 函数完成概率分布的动态映射，拟合细粒度回归结果的输出分布，聚焦显著区域特征信息.

由于本文所构建的情景式视频数据集背景环境复杂且目标实例运动时形状多变，原始残差网络难以在训练过程中为显著特征分配重要权重，因此会导致模型跨帧分割时出现实例丢失、错误分割或不完全分割等问题. 将极化自注意力嵌入至残差网络中可以增强模型对显著目标区域特征捕捉能力以此提升模型对细粒度任务的分割性能. 原始残差网络 ResNet50[15] 由浅层特征提取部分和深层特征提取部分组成. 在浅层网络中由于对输入样本不涉及过多特征映射变换操作可以保留较多原始输入的纹理特征，因此在浅层网络中嵌入极化自注意力机制有利于对目标实例的边缘纹理进行定位从而能更好分割；而在深层网络中需要对输入在原特征空间和新特征空间中进行逐层变换以生成丰富抽象的语义特征，因此在深层网络中嵌入极化自注意力机制可以更好把握目标实例的语义要素以指导模型的最终分割. 若直接将极化自注意力机制按照任意位置或数量对其嵌入，会存在使得模型对依赖于浅层或深层特征的目标实例分割结果较好而另外一者较差的问题. 因此，本文提出一种在残差网络基础上改进的极化自注意力调控模型结构 ResNet50-PSA，以丰富前景目标实例的语义信息并对其显著区域特征进行学习从而适应前景目标形状多变的视频动态场景.

在 ResNet50-PSA 模型中，残差网络的特征迭代计算过程需要使用 4 个 Block 模块，假设给定任一 Block 的输入为 $I_{\text{in}} \in R^{C_{\text{in}} \times H_{\text{in}} \times W_{\text{in}}}$，则其输出 $f_n^B \in R^{C_{\text{out}} \times H_{\text{out}} \times W_{\text{out}}}$ 计算公式可由式(1)表示：

$$f_n^B = Conv_{n,m}^{1\times1}(Conv_{n,m}^{3\times3}(Conv_{n,m}^{1\times1}(I_{\text{in}}))) + Conv_{n,m}^{1\times1,s}(I_{\text{in}}) \tag{1}$$

式中，C,H,W 分别表示特征图的通道数和尺寸大小，$Conv(\cdot)$ 表示卷积操作，s 表示卷积核步长，n 表示残差块序列号，m 表示当前第 n 个 Block 中对应卷积的输出通道数. 将极化自注意力中通道和空间自注意力计算方式分别简化为 $f_n^{ch}(\cdot)$ 和 $f_n^{sp}(\cdot)$，则 ResNet50-PSA 模型输出可由式(2)表示：

$$f_n^P = f_n^{sp}(f_n^{ch}(f_n^B(I_{\text{in}}))) \tag{2}$$

考虑到对极化自注意力不同嵌入位置和相异数量调控时会影响模型最终的检测分割精度，本文将 ResNet50-PSA 设计为单级式调控模型和级联式调控模型. 如图 2 所示，通过 8 个开关符号调控极化自注意的嵌入数量和位置. 使用序号为 1～4 的开关符号拟合调控原始残差网络中的特征流动情况，当开关闭合时表示按照基准模型中 ResNet50 原始方式连接残差块，此时上层 Block 的输出直接作为下层 Block 的特征输入，并不经过该原始调控开关所在开关对的另一 PSA 调控开关；使用序号为 5～8 的 PSA 开关符号模拟调控嵌入极化自注意力后的 Block 特征交互方式，每个 Block 后的两个开关并行连接且互斥即不能同时闭合或断开. 具体关系为开关 1 和 5 互斥，2 和 6 互斥，3 和 7 互斥，4 和 8 互斥.

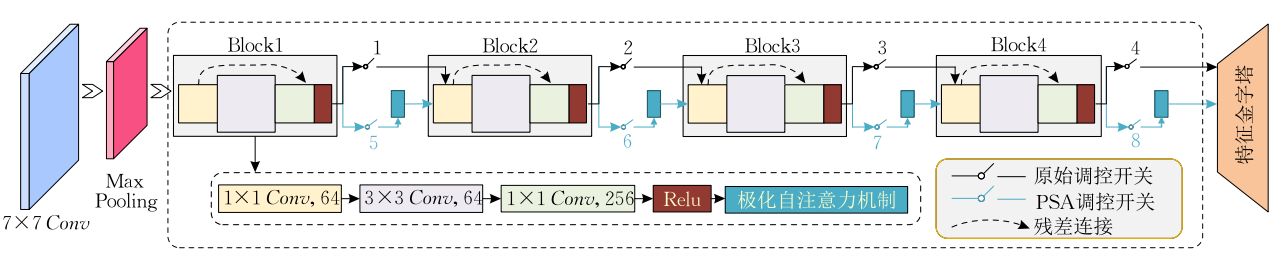


图 2　本文 ResNet50-PSA 调控模型结构

3.2.1 单级式调控模型

单级式调控模型考虑到极化自注意力机制嵌入在各 Block 位置的影响,将嵌入极化自注意力机制的残差块视为独立个体,具体工作情况如表 1 所示,用 ON 和 OFF 分别表示开关状态闭合和断开.将单级式调控模型的初始状态设定为原始调控开关全部闭合,而 PSA 调控开关全部打开即保持残差网络 ResNet50 的特征提取方式,依次在 Block1、Block2、Block3、Block4 之后调控闭合 PSA 开关同时默认其余残差块位置后原始调控开关状态闭合,PSA 调控开关状态断开.对应在表 1 的单级式调控模型中表现为原始调控开关 1～4 状态均为 ON,PSA 调控开关均为 OFF.

表 1 单级式调控模型

调控位置	原始调控开关				PSA 调控开关			
	1	2	3	4	5	6	7	8
初始状态	ON	ON	ON	ON	OFF	OFF	OFF	OFF
Block1	OFF	ON	ON	ON	ON	OFF	OFF	OFF
Block2	ON	OFF	ON	ON	OFF	ON	OFF	OFF
Block3	ON	ON	OFF	ON	OFF	OFF	ON	OFF
Block4	ON	ON	ON	OFF	OFF	OFF	OFF	ON

3.2.2 级联式调控模型

鉴于不同调控位置且相异调控数量的个体极化自注意力机制间交互作用对于模型最终分割效果的影响,在所设计单级式调控模型基础上将不同调控位置的极化自注意力机制按照相异数量依次组合得到级联式调控模型,具体情况如表 2 所示.同单级式调控模型相同将初始状态设定为原始调控开关全部闭合,依次将嵌入极化自注意力机制的 Block1、Block2、Block3、Block4 按照 2 和 4 的数量组合方式进行叠加调控.

表 2 级联式调控模型

调控位置	组合数量	原始调控开关				PSA 调控开关			
		1	2	3	4	5	6	7	8
初始状态	0	ON	ON	ON	ON	OFF	OFF	OFF	OFF
Block1,2	2	OFF	OFF	ON	ON	ON	ON	OFF	OFF
Block1,3	2	OFF	ON	OFF	ON	ON	OFF	ON	OFF
Block1,4	2	OFF	ON	ON	OFF	ON	OFF	OFF	ON
Block2,3	2	ON	OFF	OFF	ON	OFF	ON	ON	OFF
Block2,4	2	ON	OFF	ON	OFF	OFF	ON	OFF	ON
Block3,4	2	ON	ON	OFF	OFF	OFF	OFF	ON	ON
Block1,2,3,4	4	OFF	OFF	OFF	OFF	ON	ON	ON	ON

3.3 空间定位分支

视频实例分割任务在同一帧上区分不同实例以及跨帧跟踪相同实例时对个体目标的边缘纹理信息有很大依赖性,然而在真实场景中目标尺度和形状具有多变性.因此,在加强语义特征判别目标类别的同时,也要融合不同尺度特征来关联特征图的全局和局部特征.然而残差网络所提取的特征有着空间分辨率大小与语义信息深浅程度相矛盾而无法折衷的问题:低层特征图拥有细粒度空间分辨率但缺乏高层的抽象语义信息;高层特征图的语义特征相当丰富然而对细粒度空间边缘纹理信息定位较差.尽管特征金字塔网络通过自上而下的路径进行特征融合对深层特征逐步上采样与对应下部邻近特征相融合的方式弥补高分辨率特征图中深度语义信息不足的缺陷,然而上采样过程大多对高层语义信息进行特征映射,而对低级空间信息的传递能力较为薄弱,且残差网络中高层的语义特征是经过很长的信息路径从原始输入逐步下采样生成的,因此依然缺乏目标的空间定位信息和浅层纹理信息.为了弥补这一缺陷,YolactEdge 通过简单的横向映射操作对特征金字塔中各级别特征进行提取,以期生成纹理细节较为丰富的同级特征,但忽略了邻近特征交互作用因此缺乏对不同层级的尺度特征和语义要素的描述.

鉴于 PANet[16] 提出的自下而上的路径聚合方式可以达到与特征金字塔中低层纹理信息进行有效交互融合的目的,本文在 FPN 后建立空间定位分支(Multi-Scale Spatial Positioning,MSP)实现高低层级多粒度特征的有效融合并克服特征金字塔解析高级语义信息带来的空间定位问题,从不同层级更好描述目标实例的空间纹理细节,实现对尺度变换的

目标实例边缘定位和轮廓分割.

如图 3 所示，将特征金字塔中上采样解析后的高分辨率低级特征通过卷积操作提取其浅层空间信息，与邻近的相对高级抽象语义特征给与融合，缩小同一尺度特征中高层语义与低层空间信息分配不均的差距，使不同尺度拥有丰富的语义要素及空间纹理信息，充分发挥低级分辨率特征对最终分割的细节指导作用. 相较于 PANet 最低层分辨率特征通过特征金字塔中同级信息直接复制生成，本文空间定位分支进一步利用横向映射卷积操作加强了其低层信息的生成传播能力以此能够提取更多的细节纹理特征，然后通过相同分辨率特征间信息交互为不同尺度生成丰富的多粒度空间分辨率特征.

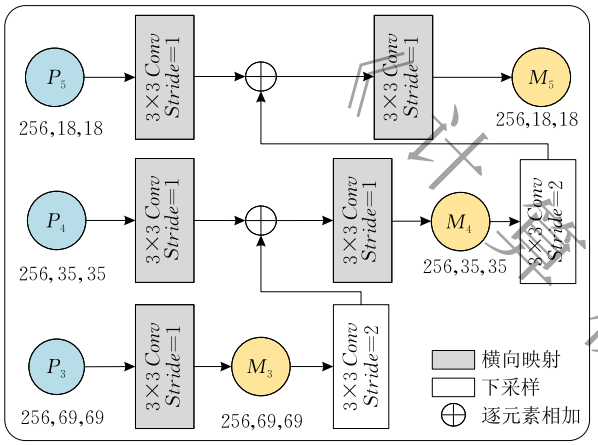


图 3　空间定位分支模型示意图

空间定位分支主要包括分辨率横向映射和下采样操作，分别由步长为 1 和步长为 2 的卷积表示. 首先对 P_3 层级特征经过横向映射得到最低层特征 M_3；然后将 M_3 下采样至 P_4 特征相同分辨率大小，并与其逐元素相加后再次通过横向映射进行特征整合操作得到 M_4，以克服直接相加的特征融合方式带来的背景信息干扰问题；最后利用与 M_4 相同的特征抽象融合方式生成 M_5 层级特征信息. 通过邻近层与恒定空间分辨率大小的特征生成内容丰富的高质量多粒度特征图. 空间定位分支中 M_3、M_4、M_5 的特征计算公式分别由式(3)、式(4)和式(5)表示.

$$M_3 = Conv_{3\times3,1}(P_3) \tag{3}$$

$$M_4 = Conv_{3\times3,1}(Add(Conv_{3\times3,2}(M_3), Conv_{3\times3,1}(P_4))) \tag{4}$$

$$M_5 = Conv_{3\times3,1}(Add(Conv_{3\times3,2}(M_4), Conv_{3\times3,1}(P_5))) \tag{5}$$

其中 $Conv_{k\times k,m}(\cdot)$ 代表尺寸大小为 $k\times k$，步长为 m 的卷积，$Add(\cdot)$ 代表逐元素相加.

4　实验结果与分析

4.1　数据集建立

本文实验数据来源于公开视频实例分割数据集 Youtube-VIS2019，该数据集包含 40 类对象，2883 个实例视频，其中训练集的视频数量有 2238 个，测试集为 338 个. YolactEdge 对该数据训练集按照 85% 和 15% 的比例重新随机划分了训练集和测试集分别为 1904 和 334 个视频. 为验证模型对情景式视频学习的鲁棒性，本文在该数据集基础上抽取了适用于情景式视频分割的动物类集合，包含大熊猫、猿猴、猴子、熊、长颈鹿、豹子、狐狸、鹿、斑马、老虎、大象、海狮共计 12 个类别，其中训练集 549 个视频中包含 829 个实例，测试集 108 个视频中包含 176 个实例. 动物情景数据集训练集和测试集中各类别实例数目分布情况如图 4 所示.

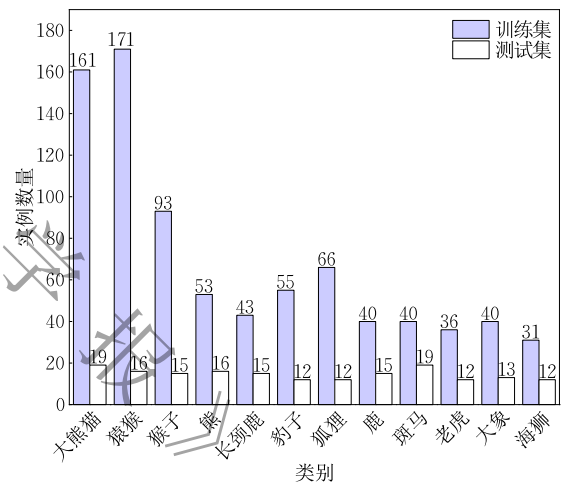


图 4　本文数据集中每个类别实例数分布情况

4.2　实验设置

本文使用操作系统为 Ubuntu18.04，CPU 为 Intel(R) Core(TM) i5-10400F，GPU 为 NVIDIA RTX 2080Ti，显存 11GB 的台式计算机. 深度学习框架为 Pytorch1.6.0、Python 版本为 3.6、采用 CUDA11.2、cuDNN8.0.5 加速模型训练以及使用 TensorRT 7.1.3.4[17] 加快模型在测试阶段的推理速度.

实验过程中将输入图像尺度统一为 550×550 并将批处理尺寸(Batch_Size)设置为 8. 在训练阶段设置迭代次数为 10 万步、初始学习率设置为 5e-4，在 0～3400 步使用 Warmup 学习率预热的方式从 0 逐步增加到预先设定的学习率大小提高模型损失收敛速度，3400 步以后使用余弦退火(Cosine Annealing LR)方法对学习率进行衰减以保证模型损失收敛的

稳定性，并每隔 5000 步保存一次模型权重共保存 20 个模型，对最后四个训练模型的精度和推理速度综合比较后选出最优模型.

4.3 评价指标

为客观评价情景式视频实例分割算法的性能，本文采用分割速度、运行时长和平均精度均值 mAP (mean Average Precision)作为评价指标. 采用模型每秒处理的帧数 FPS(Frames Per Second)指标衡量分割速度，采用模型处理每张图片所需时间来评估运行时长. 模型分割速度越快，即 FPS 越大，说明在一秒内分割的帧数越多，则所需运行时间越短；反之运行时间越长. 按照 0.05 的增量在 0.50 至 0.95 区间取值 10 个交并比(Intersecstion over Union, IoU)作为阈值，mAP 为这 10 个阈值下对应的平均精度 AP 的平均值且区间大小为$[0,1]$，mAP 越接近 1 代表检测和分割效果越好.

mAP 需要通过准确率(Precsion, P)与召回率(Recall, R)进一步求得，其中目标检测锚框的评价指标为 Box_mAP，分割任务的掩膜评价指标为 $Mask\_mAP$，相关计算公式如下：

$$P=\frac{TP}{TP+FP}\times 100\% \tag{6}$$

式中，TP(True Positive)为真正例，即真实值是目标，预测也判别为目标，FP(False Positive)为假正例，真实值是目标，预测判别为不是目标.

$$R=\frac{TP}{TP+FN}\times 100\% \tag{7}$$

式中，FN(False Negative)为假负例，即真实值不是目标，预测判别为不是目标. 对应的 mAP 计算公式为

$$mAP=\frac{1}{c}\sum_{a}^{c}\left(\frac{1}{|threshold|}\sum_{t}P(t)\right) \tag{8}$$

式中，a 为当前类别序号；c 为总类别数量；$threshold$ 为当前阈值，t 为 COCO 评价指标下设定的总阈值数，$P(t)$为当前阈值下所计算的准确率.

4.4 结果及分析

4.4.1 ResNet50-PSA 调控模型实验分析

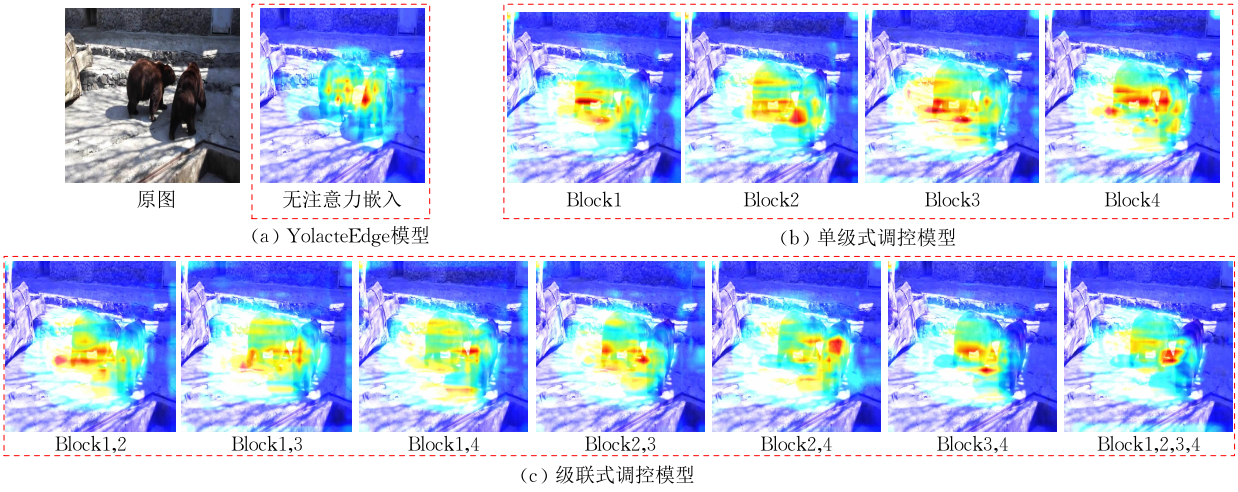
为对比本文设计的单级式和级联式调控模型对动物场景视频数据集分割任务的优越性，针对不同嵌入位置和相异嵌入数量的 ResNet50-PSA 调控模型进行实验分析，结果分别如表 3 和表 4 所示. 分析出相比于单级式而言，级联式调控模型的分割效果整体相对较弱，在不同位置嵌入个体自注意力机制时模型的分割精度会有所提升但按照设定嵌入数量调控组合时会导致模型分割精度下降. 表 3 中嵌入在 Block1 和 Block2 后的单级式调控模型分割精度分别为 34.52%和 35.16%，而在表 4 的级联式调控模型中两者调控组合后分割精度下降为 32.39%. 为清晰解释极化自注意力机制的内在工作机理对最终分割结果的有效性，在相同位置 Block4 后分别对单级式和级联式模型特征图在通道维度进行最大池化，并与原图叠加得到类激活热力图后进行可视化对比分析. 如图 5 所示，图中区域的颜色越接近红色说明模型对其激活程度越大且关注度越高. 可视化热图发现嵌入在 Block1 和 2 位置的级联式模型感兴趣区域扩散到背景信息上，对当前的分割任务有一定的干扰作用，因此分割精度相对于独立的单级式调控模型而言有所下降，但其总体分割精度较 Yolact-Edge 基准模型普遍得到了一定提升. 从图 5 中观察可以看出嵌入极化自注意力机制后的模型更能学到有效的感兴趣区域信息，对前景目标的重要特征予以关注.

表 3 单级式调控模型实验结果对比

调控嵌入位置	*Bbox* /%			*Mask* /%			FPS	运行时长/ms
	AP	*AP50*	*AP75*	*AP*	*AP50*	*AP75*		
Block1	32.89	55.21	34.77	34.52	54.05	39.08	90	11.1
Block2	35.20	35.20	38.15	35.16	55.24	39.46	90	11.1
Block3	37.35	60.27	41.37	36.89	58.32	39.76	98	10.2
Block4	**37.99**	**61.66**	**41.92**	**38.36**	**59.44**	**42.55**	**80**	**12.5**

表 4 级联式调控模型实验结果对比

调控嵌入位置	调控嵌入数量	*Bbox* /%			*Mask* /%			FPS	运行时长/ms
		AP	*AP50*	*AP75*	*AP*	*AP50*	*AP75*		
Block1+2	2	34.27	56.99	37.29	32.39	52.56	35.11	83	12.1
Block1+3	2	36.07	57.70	39.01	34.82	56.12	37.56	86	11.6
Block1+4	2	34.46	57.36	37.20	34.23	55.84	36.58	84	11.9
Block2+3	2	33.73	56.62	36.31	34.18	54.84	37.65	96	10.4
Block2+4	2	36.93	59.83	40.41	36.79	58.12	40.60	78	12.8
Block3+4	2	36.78	60.04	39.82	36.53	57.78	40.44	80	12.5
Block1+2+3+4	4	33.08	58.75	33.44	34.41	34.41	37.17	80	12.5



(a) YolacteEdge模型　(b) 单级式调控模型

(c) 级联式调控模型

图 5　ResNet50-PSA 调控模型与基准模型可视化热图对比

由表 3 可以看出单级式调控模型的检测分割精度随着极化自注意力嵌入位置加深呈现一种逐级递增的趋势，对应在图 5 发现热图区域越来越趋向于整个有效的前景目标信息. 这是因为模型的特征提取是一个语义形态逐渐抽象的过程，Block1 作为骨干网络的第一个残差块缺乏深层抽象的语义特征，对于有效信息的辨认还需要一个认知过程，因此于 Block1 后嵌入极化自注意力机制对于模型精度提升相对较低；而 Block4 作为骨干网络的尾部残差块，生成的深度抽象语义可以为注意力机制的显著目标特征提取任务提供指导作用，此时模型分割精度达到最优. 最优单级式调控模型与 YolactEdge 实验对比如表 5 所示，结果表明最优单级式调控模型分割精度达到 38.36%，相比于 YolactEdge 基准模型提升 6.82%，检测精度达到 37.99%，但速度从每秒分割 90 帧下降为 80 帧，这是因为自注意力机制的引入不可避免地会增加模型参数量从而影响测试速度，但模型总体还是达到实时分割的效果.

表 5　YolactEdge 基准与最优单级式调控模型实验结果对比

模型	*Bbox* /%			*Mask* /%			FPS	运行时长/ms
	AP	*AP*50	*AP*75	*AP*	*AP*50	*AP*75		
YolactEdge	37.98	57.15	41.53	31.54	52.73	30.55	90	11.1
最优单级式调控模型	**37.99**	**61.66**	**41.92**	**38.36**	**59.44**	**42.55**	**80**	**12.5**

综上所述，本文在 Block4 后嵌入极化自注意力机制的单级式调控模型分割精度达到最优，且不论是单级式还是级联式调控模型都为当前任务提高了分割精度，证明了 ResNet-PSA 模型的有效性. 然而极化自注意力机制倾向于通过对空间像素位置或特征通道关系的建立来捕捉同一语义内部特征相关性，更加注重像素回归问题，因此当多个目标实例距离较近时会过分将其他实例语义特征回归至当前实例，从而对检测任务造成错误引导降低检测精度.

4.4.2　空间定位分支实验

为验证聚合目标实例多粒度空间信息的多尺度空间定位分支相比于 PANet 路径聚合方式的优越性和对模型最终检测分割效果的有效性，选择嵌入于 Block4 位置后的最优单级式调控模型作为基准模型，在此基础上嵌入空间定位分支后的模型命名为 PSAM-Net. 模型改进前后的实验结果对比如表 6 所示，三组对比试验均在本文情景式数据集场景下进行重新训练测试. 可以看出添加 PANet 后的模型相比于最优单级式调控模型其平均检测和平均分割精度分别提升了 3.81%和 1.08%；嵌入空间定位分支后的 PSAM-Net 模型其平均检测和平均分割精度相较于 PANet 分别提升了 2.53%和 2.31%，相较于最优单级式调控模型分别提升了 6.34%和 3.39%. 说明空间定位分支能对高低层的语义特征

表 6　空间定位分支实验结果

模型	*Bbox_mAP* /%			*Mask_mAP* /%			FPS	运行时长/ms
	AP	*AP*50	*AP*75	*AP*	*AP*50	*AP*75		
最优单级式调控模型	37.99	61.66	41.92	38.36	59.44	42.55	80	12.5
最优单级式调控模型＋PANet	41.80	63.26	46.79	39.44	61.20	43.11	77	13.0
PSAM-Net	**44.33**	**65.52**	**47.53**	**41.75**	**62.84**	**43.45**	**73**	**13.6**

和分辨率信息通过多尺度交互实现目标实例的边缘轮廓特征优化和定位，从而提升模型的检测分割精度，然而空间定位分支的引入同时也为模型增加了一定的参数量导致分割速度有所下降.

为清晰解释空间定位分支的内在工作机理对最终分割结果的有效性，本文选择对空间维度为 35×35 的 M_4 层级特征图进行可视化分析. 考虑 M_4 层包含 256 维通道特征，分别提取最优单级式调控模型与 PSAM-Net 模型的第一层通道特征图进行可视化分析以保证对比条件的一致性，引入空间定位分支前后可视化对比结果如图 6 所示. 可以看出 PSAM-Net 模型对前景目标实例的边缘轮廓特征提取能力明显高于基准模型，增大了背景与前景信息的反差对比度，优化了模型对目标边缘的定位准确性. 结果表明空间定位分支可以通过多尺度邻近特征的聚合丰富当前尺度特征图的多粒度信息，弥补低级空间细粒度特征和高级语义粗粒度信息无法折衷的缺陷.

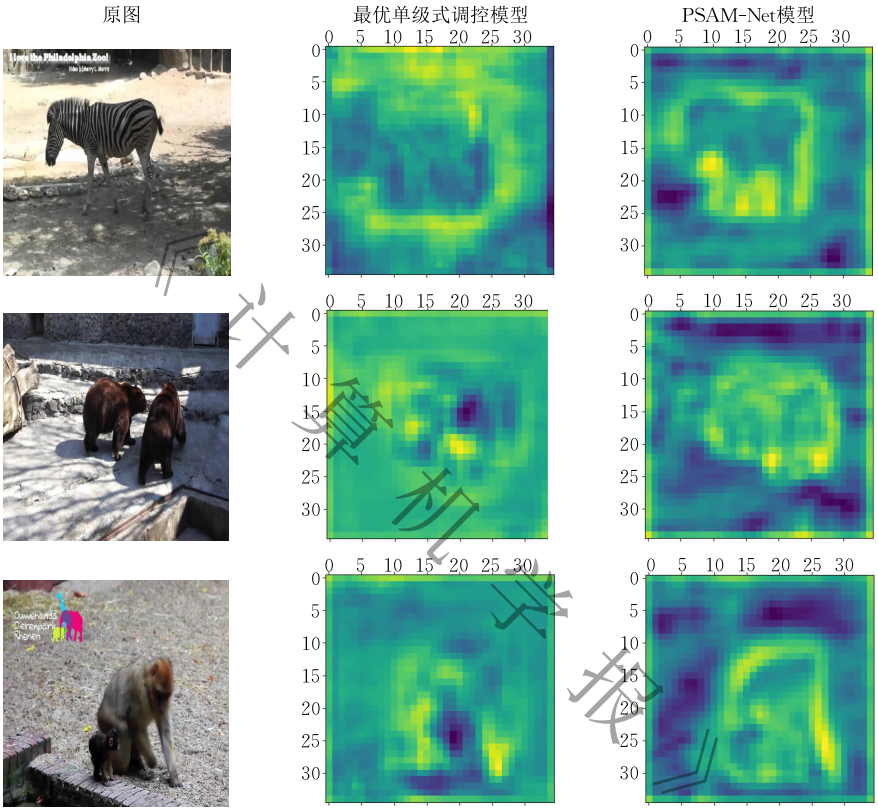


图 6 引入空间定位分支前后可视化结果对比

4.4.3 PSAM-Net 交叉验证

为评价本文模型在不同数据集上测试结果的鲁棒性与泛化能力，利用 Bootstrap 采样交叉验证法设计多组实验验证本文模型的检测分割性能. 在 3.1 节所构建数据集基础上在动物类数据集全体样本中进行 5 次可重复抽样. 在重复抽样过程中，保证训练集与验证集划分比例标准与原文相同为 85%∶15%，在此基础上组成 5 个新的样本空间以构建对应的数据集，且各类别在每组划分数据集中的实例总数相同，五组验证集具体实例分布信息如图 7 所示.

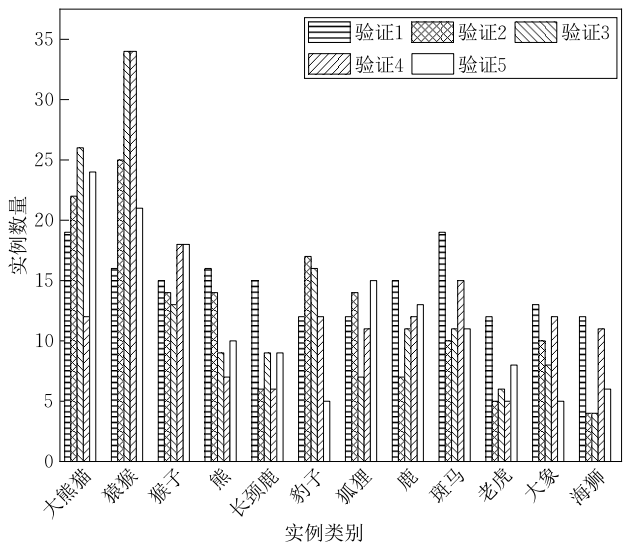


图 7 交叉验证组数据集实例样本分布信息

在图 7 所划分的数据基础上，将本文 PSAM-Net 模型分别对五组数据集作以验证并计算其均值与基准模型作以对比，实验结果分别如表 7 和表 8 所示，其中 N 表示重复抽样次数. 可以看出五组验证模型

表 7　Bootstrap 采样交叉验证实验结果对比

模型	N	*Bbox* /%			*Mask* /%			FPS	运行时长/ms
		AP	AP50	AP75	*AP*	*AP*50	*AP*75		
PSAM-Net	1	**44.33**	65.52	47.53	**41.75**	62.84	43.45	73	13.7
	2	**40.85**	64.57	44.83	**39.18**	61.93	42.22	80	12.5
	3	**44.70**	68.39	50.35	**38.87**	61.93	43.59	86	11.6
	4	**46.82**	69.40	54.66	**39.63**	65.60	43.00	81	12.3
	5	**43.62**	66.75	48.53	**42.60**	66.21	47.60	80	12.5

表 8　基准模型与采样交叉验证实验结果对比

模型	*Bbox_AP* /%	*Mask_AP* /%	FPS	运行时长/ms
基准模型	37.98	31.54	90	11.1
PSAM-Net	44.33	41.75	73	13.7
PSAM-Net(交叉验证)	44.06(±0.50)	40.41(±0.60)	80(±4.95)	12.5(±0.76)

的平均检测精度均值为 44.06%，方差为 0.50；平均分割精度的均值为 40.41%，方差为 0.60，证明模型在不同分布验证数据下结果波动较小，可以保持稳定的分割效果. 相比基准模型其平均检测精度和平均分割精度分别提高了 6.08%和 8.87%. 表明本文提出的情景式视频实例分割模型 PSAM-Net 有较好的稳定性与鲁棒性.

4.4.4　PSAM-Net 模型分割效果

为定性分析 PSAM-Net 模型对复杂背景下拓扑形变和尺度缩放目标的分割效果，将其与 YolactEdge 基准模型在同一测试集上的分割结果可视化并将部分分割细节信息进行局部放大以增强直观对比，第(a1)、(b1)、(c1)、(d1)行和第(a2)、(b2)、(c2)、(d2)行分别表示 YolactEdge 和 PSAM-Net 模型分割结果，黄色矩形框表示漏检目标实例区域，蓝色矩形框为分割细节放大区域，如图 8 所示. 在(a1)与(b1)视频序列中，大象与海狮运动时自身会产生拓扑形变使得 YolactEdge 产生误检、不完全分割和漏检等问题，对应在(a2)与(b2)中 PSAM-Net 可以通过极化自注意力机制中细粒度的特征回归任务关联整合空间像素和通道语义信息，激活映射显著目标区域信息且对其学习从而适应运动目标的拓扑形变特征，提高对目标实例的捕捉能力与分割精度. 在(c1)与(d1)视频序列中，由于相机拍摄距离发生变化造成目标实例尺度存在缩放问题，导致 YolactEdge 对目标实例轮廓分割时性能有所欠缺无法完整分割出整个实例特征，而在(c2)与(d2)中 PSAM-Net 可以通过多粒度空间定位分支为模型高层特征丰富边缘纹理细节信息从而改善目标尺度缩放导致的分割精度不足问题，实现对目标实例的准确定位与分割.

综合上述结果，本文模型相比于 YolactEdge 而言能更好地适应复杂背景下拓扑形变和尺度缩放目标的分割问题，有效降低基准模型的误检率和漏检率，有更高的检测分割精度和更好的鲁棒性.

4.4.5　对比实验

本文选择与基准模型和主流算法 SipMask[18]、STMask[19] 在分割速度、运行时长和精度上做以对比实验以客观评价 PSAM-Net 模型对视频实例分割任务的优越性. 为保证验证结果有效性与公平性将对比实验均在同一配置设备上进行重新训练且使用数据集均为本文所提取动物场景下的情景式视频数据，实验结果如表 9 所示.

结果表明，与其他三种模型相比，本文设计的 PSAM-Net 模型平均分割精度达到最高为 41.75%，比 STMask、Sipmask 和 YolactEdge 分别高出 4.45%、3.95%和 10.21%，平均分割速度大于 Sipmask 和 STMask，略小于 YolatEdge 达到 73FPS；处理一帧图片的运行时长仅需要 13.7ms. 综上所述，在同时权衡模型的平均分割精度、分割速度及运行时长的条件下，本文设计的 PSAM-Net 模型鲁棒性更好更适用于情景式的视频分割任务.

表 9　不同模型性能对比实验结果

模型	Box /%			Mask /%			FPS	运行时长/ms
	AP	AP50	AP75	AP	AP50	AP75		
YolactEdge	37.98	61.66	41.92	31.54	58.95	41.53	90	11.1
SipMask	—	—	—	37.80	56.90	41.10	18	71.4
STMask				37.30	53.70	40.50	31	32.5
PSAM-Net	**44.33**	**65.52**	**47.53**	**41.75**	**62.84**	**43.45**	**73**	**13.7**

图 8 PSAM-Net 与 YolactEdge 结果可视化对比

5 结　语

本文提出了一种极化自注意力调控的情景式视频实例多尺度分割模型 PSAM-Net. 首先相比于现有视频实例分割算法在非受限视频场景中作以工作，本文更加注重考虑在具体分割任务中模型的表现能力. 从非受限整体视频序列中构建出适用于动物场景下的特定数据集. 其次，在同时具有拓扑形变和尺度变换特点的目标数据上表明，所设计的极化自注意力调控模型和空间定位分支，提高了模型的可辨识特征表达能力和适应能力，达到了在复杂背景中对前景运动目标进行高性能分割的目的.

本文仅针对静态帧中目标实例的特征表达重要程度及算法进行了分析及优化，在未来研究工作中，将针对视频序列同一帧中多目标实例的重叠遮挡问题进行探索，并在此基础上，考虑定义时域一致性损失函数约束同一目标像素点使其在不同帧间具有相同的语义要素标签，完成动态场景下像素级时序关联任务的多目标实例准确分割.

参　考　文　献

[1] Chen Jia, Chen Ya-Song, Li Wei-Hao, et al. Application and prospect of deep learning in video object segmentation. Chinese Journal of Computers, 2021, 44(3): 609-631(in Chinese)
(陈加，陈亚松，李伟浩等. 深度学习在视频对象分割中的应用与展望. 计算机学报，2021，44(3)：609-631)

[2] He Kai-Ming, Gkioxari G, Dollar P, et al. Mask R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969

[3] Ren Shao-Qing, He Kai-Ming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[4] Huang Zhao-Jin, Huang Li-Chao, Gong Yong-Chao, et al. Mask scoring R-CNN//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6402-6411

[5] Ke Lei, Tai Yu-Wing, Tang Chi-Keung. Deep occlusion-aware instance segmentation with overlapping BiLayers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Kuala Lumpur, Malaysia, 2021: 4019-4028

[6] Bolya D, Zhou C, Xiao F, et al. YOLACT: Real-time instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9157-9166

[7] Xie En-Ze, Sun Pei-Ze, Song Xiao-Ge, et al. PolarMask: Single shot instance segmentation with polar representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12193-12202

[8] Tian Zhi, Shen Chun-Hua, Chen Hao, et al. FCOS: Fully convolutional one-stage object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9627-9636

[9] Liu Hua-Jun, Liu Fu-Qiang, Fan Xin-Yi, et al. Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv: 2107. 00782, 2021

[10] Lin Tsung-Yi, Dollar P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 936-944

[11] Yang Lin-Jie, Fan Yu-Chen, Xu Ning, et al. Video instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5188-5197

[12] Liu Hao-Tian, Soto R, Xiao Fan, et al. YolactEdge: Real-time instance segmentation on the edge//Proceedings of the IEEE International Conference on Robotics and Automation. Xi'an, China, 2021: 9579-9585

[13] Fu Yang, Yang Lin-Jie, Liu Ding, et al. CompFeat: Comprehensive feature aggregation for video instance segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Seattle, USA, 2021: 1361-1369

[14] Bertasius G, Torresani L, et al. Classifying, segmenting, and tracking object instances in video with mask propagation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9739-9748

[15] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[16] Liu Shu, Qi Lu, Qin Hai-Fang, et al. Path aggregation network for instance segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8759-8768

[17] Nvidia tensorrt. https://developer. nvidia. com/tensorrt. Accessed: 2020

[18] Cao Jia-Le, Anwer R M, Cholakkal H, et al. SipMask: Spatial information preservation for fast image and video instance segmentation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 1-18

[19] Li Ming-Han, Li Shu-Ai, Li Li-Da, et al. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Kuala Lumpur, Malaysia, 2021: 11215-11224

HUANG Ying, Ph. D. candidate. Her research interests include deep learning and machine vision.

HE Zi-Fen, Ph. D., associate professor. Her research interests include image processing and machine vision.

YANG Hong-Kuan, M. S. candidate. His research interests include deep learning and machine vision.

ZHAO Chong-Ren, M. S. candidate. His research interests include deep learning and machine vision.

ZHANG Yin-Hui, Ph. D., professor. Ph. D. supervisor. His research interests include image processing and machine vision.

Background

Video instance segmentation is one of the most difficult tasks in computer vision, which can be widely applied in autonomous scene perception tasks based on episodic intelligent robots. When a robot moves in a specific scene, the visual imaging system can be affected by its own speed, shooting angle and distances to the instances, as well as the randomness of topological deformation and scale scaling of the captured instances. As a result, the instances in the images have blurred edges and different scales and shapes across frames, which causes a great challenge to the segmentation tasks.

Recently, a host of researchers use the contextual information of video sequences to improve the effectiveness of the algorithms by means of inter-frame propagation or tracking. However, the algorithms still lack of effective attention to the high-level fine-grained features and the accurate positioning of low-level spatial information for the instances, which results in low segmentation accuracy for video instance segmentation.

Based on previous works, considering the adaptability of the algorithm to the variable shape of the instances, the focusing ability of important regional features and the blurring of edge information, a model based on polarized self-attention manipulation for multi-scale video instance segmentation is designed in this work. The experimental show that compared to YolactEdge benchmark, the comprehensive testing results obtained by the PSAM-Net model on the episodic video dataset of animals extracted from Youtube-VIS show that the average detection increased by 6.08% and reached to 44.06%, the average segmentation accuracy increased by 8.87% and reached to 44.41% respectively.

This work is supported by the National Natural Science Foundation of China under Grant No. 62171206, whose name are "High-order Polynomial Time-varying Memory Projection for Weakly Supervised Instance Segmentation of Spatio-temporal Sparse Scenes". In this project, we focus on the in-depth research on deep semantic parsing of topological instances and contour discrimination of scale targets. Besides, the work is supported by the National Natural Science Foundation of China under Grant No. 62061022.