# 锚框校准和空间位置信息补偿的街道场景视频实例分割

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

**摘 要:** 街道场景视频实例分割是无人驾驶技术研究中的关键问题之一,可为车辆在街道场景下的环境感知和路径规划提供决策依据.针对现有方法存在多纵横比锚框应用单一感受野采样导致边缘特征提取不充分以及高层特征金字塔空间细节位置信息匮乏的问题,本文提出锚框校准和空间位置信息补偿视频实例分割(Anchor frame calibration and Spatial position information compensation for Video Instance Segmentation,AS-VIS)网络.首先,在预测头3个分支中添加锚框校准模块实现同锚框纵横比匹配的多类型感受野采样,解决目标边缘提取不充分问题.其次,设计多感受野下采样模块将各种感受野采样后的特征融合,解决下采样信息缺失问题.最后,应用多感受野下采样模块将特征金字塔低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿,解决高层特征空间细节位置信息匮乏问题.在Youtube-VIS标准库中提取街道场景视频数据集,其中包括训练集329个视频和验证集53个视频.实验结果与YolactEdge检测和分割精度指标定量对比表明,锚框校准平均精度分别提升8.63%和5.09%,空间位置信息补偿特征金字塔平均精度分别提升7.76%和4.75%,AS-VIS总体平均精度分别提升9.26%和6.46%.本文方法实现了街道场景视频序列实例级同步检测、跟踪与分割,为无人驾驶车辆环境感知提供有效的理论依据.

**关键词:** 街道场景;视频实例分割;锚框校准;空间信息补偿;无人驾驶
**基金项目:** 国家自然科学基金(No.62061022,No.62171206)
**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)01-0094-13
**电子学报URL:**http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220885

# Anchor Frame Calibration and Spatial Position Information Compensation for Street Scene Video Instance Segmentation

ZHANG Yin-hui, ZHAO Chong-ren, HE Zi-fen*, YANG Hong-kuan, HUANG Ying
(*Department of Mechanical and Electrical Engineering*, *Kunming University of Science and Technology*, *Kunming*, *Yunnan* 650500, *China*)

**Abstract:** Due to the decision-making provision for vehicle environment perception and path planning, street scenes video instance segmentation as one of the key issues in research of self-driving technology has aroused wide concern. However, current researches focus on insufficient edge feature extraction, which is caused by utilization of single receptive field sampling for multi-aspect ratio anchor frames and deficiencies of spatial detailed position information in the high-level feature pyramid architecture. To alleviate these problems, we propose a network anchor frame calibration and spatial position information compensation for video instance segmentation (AS-VIS). Firstly, we conduct the anchor frame calibration module as additional branch in parallel with three prediction branches to align multi-type receptive field sampling with different aspect ratio of anchor frame. Secondly, a multi-receptive field subsampling module is designed to fuse the features of various receptive fields achieving less information missing compared with traditional down-sampling. Finally, for spatial location information compensation and detail location information dispersion in the higher-level feature space, we design multi-receptive field subsampling module embedded in higher level to map active feature of target region in lower level of the feature pyramid. The street scene video dataset is extracted from Youtube-VIS benchmark, including 329 videos in training set and 53 videos in validation set. Quantitative comparison of experimental results with Yolact-

收稿日期:2022-07-25;修回日期:2023-03-10;责任编辑:朱梅玉
*通讯作者:何自芬

Edge show that the average accuracy of anchor frame calibration is improved by 8.63% and 5.09%, spatial position information compensation feature pyramid network is improved by 7.76% and 4.75%, and the overall average accuracy of AS-VIS is improved by 9.26% and 6.46%. The proposed network AS-VIS realizes detection, tracking, and segmentation synchronously on instance-level street scene video sequences, and provides an effective theoretical basis for environment perception of self-driving vehicles.

Key words: street scene; video instance segmentation; anchor frame calibration; spatial information compensation; self-driving vehicle

Foundation Item(s): National Natural Science Foundation of China (No.62061022, No.62171206)

## 1 引言

环境感知是无人驾驶技术研究中的关键问题之一,可为车辆在交通场景下的路径规划和决策、控制执行提供重要依据. 车辆在行驶过程中需要实时获取环境信息并进行处理. 根据获取环境信息的传感器类型,目前常见的环境感知方法主要分为雷达[1~4]、多传感器信息融合[5~8]以及视觉[9~12]的方法. 雷达设备成本高并且只能识别深度信息而无法获取纹理和色彩;多传感器信息融合的方法成本更高并且技术难度大;而视觉的方法具有成本低、能获取纹理和色彩信息等优点,受到学者的广泛关注.

视频实例分割是由图像实例分割进一步扩展的计算机视觉任务. 图像传达的仅仅是空间信息,而视频是序列图像在时间上的连续播放,故同时具有空间信息和时间信息. 视频实例分割中需要对目标同时进行检测、跟踪和分割,从而获取整个视频分割掩膜序列,可以帮助智能车辆更好地做出驾驶策略.

视频实例分割网络主要通过跟踪或特征传播两类方法关联帧序列实例之间的关系. 基于跟踪的方法包括 MaskTrack R-CNN[13]在 Mask-RCNN[14]网络架构中添加跟踪分支,最先实现帧级实例同时检测、跟踪和分割,对视频实例分割有着开创性的贡献. 该网络没有利用其他帧的信息对目标帧的特征进行提取、检测和分割,而是独立完成之后再将实例通过跟踪分支关联起来,导致该网络计算复杂度较高. SipMask[15]在 Yolact[16,17]的基础上通过一个空间保存模块生成空间系数以及特征对齐,更好地描绘空间上相邻的对象,进而改善对象遮挡的预测. 为解决 Yolact 用一组系数来预测目标的整个掩膜导致边界框内空间信息匮乏的问题,SipMask 将掩膜预测分为多个子掩膜预测. 为弥补 SipMask 固有的特征表达不足,Liu 等人[18]提出 SG-Net,根据目标的大小和形状将目标动态划分为不同子区域进行分割以提高掩膜粒度,解决 SipMask 特征表达不足的问题. 在上述视频实例分割方法[13,15,18]中,检测、分割和跟踪等任务都是独立完成的,帧间时域一致连续信息并没有被用于提高分割质量. Fu 等人[19]提出 CompFeat 模型,通过改进空时记忆模型结合时间注意力机制,实现帧间信息建模,然而 CompFeat 建模当前帧与整个视频的时序关系时需要反复计算时序注意力模块,因此需要消耗较大计算资源. STMask[20]通过获取参考帧和当前帧之间的偏差并回归到当前帧,从而实现帧间信息传递. 但参考帧是根据当前帧随机选择的,因此在训练过程中易造成时域信息混淆.

特征传播方法早期用于视频分类[21]和视频目标检测[22,23]来提高推理速度和准确性. 在视频实例分割中,MaskProp[24]在 Mask-RCNN 网络架构中引入掩膜传播分支,将掩膜从前一帧传播至后一帧. 但 MaskProp 是一种离线学习方式,能够实现准确的预测,但训练时间长、内存消耗大. MSN[25]将 Swin[26]产生的分割掩码集和 CFBI[27]产生的传播掩码集,通过掩膜选择网络,确保只传播最优的掩膜,从而限制了噪声累积. 然而,Swin 和 CFBI 产生掩膜集以及掩膜选择网络需要消费较大的计算资源和内存. 与上述方法[24,25]不同,Liu 等人[28]在 Yolact 的基础上提出 YolacteEdge. 该网络通过计算帧间光流信息应用特征传播机制,将关键帧的深层特征传播至非关键帧,使非关键帧不需要计算深层特征,从而减少计算量. 然而,其对所定义关键帧的深层特征有较大依赖性,若关键帧分割效果较差,会对非关键帧的分割任务造成一定误导.

在特征图每个像素点上生成多个面积比例和纵横比不同的矩形框被定义为锚框(Anchor boxes,Abox),是卷积神经网络用于目标识别的重要概念之一. 锚框的应用能够提高目标检测模型的性能,然而上述实例分割模型[13~20,24,25]中存在每个空间位置的多个不同纵横比锚框共享同一卷积特征问题,导致预定义的锚框与真实边界框无法对齐,易使模型对目标边缘位置敏感不足,从而导致分割效果不佳. 此外,上述视频实例分割模型[13,15,18~20,24,25]中特征金字塔网络均采用自顶向下的方式,将高层特征融入到低层特征中,以丰富低层特征的语义信息,缺乏对高层特征经过多层卷积后,目标的纹理和边缘等空间细节信息缺失的思考. 针对上述问题,本文贡献如下:

(1)提出锚框校准模块,目的是应用同锚框纵横比匹配的多类型感受野采样,提高模型对目标边缘位置的敏感度,从而解决现有方法存在多纵横比锚框应用单一感受野采样导致边缘特征提取不充分的问题;

(2)提出多感受野下采样模块,目的是利用多种感受野找到目标区域并在空间上重复,从而解决下采样信息缺失的问题;

(3)提出空间位置信息补偿特征金字塔,应用多感受野下采样模块将特征金字塔低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿,解决高层特征空间细节位置信息匮乏的问题.

## 2 AS-VIS网络架构

YolactEdge整体架构设计十分轻量在速度和精度上都能获得较好成绩,故而选择其作为AS-VIS的基础网络.AS-VIS使用ResNet50作为特征提取网络,针对高层特征图空间位置信息匮乏的问题,将特征金字塔(Feature Pyramid Network,FPN)改为空间位置信息补偿特征金字塔(Spatial Location information compensation Feature Pyramid Network,SL-FPN),将低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿.AS-VIS预测头沿用YolactEdge中的预测头,针对多纵横比锚框应用单一感受野卷积采样导致特征提取不充分的问题,在预测头中引入锚框校准模块,实现在纵横比不同的锚框采用同锚框匹配的感受野提取特征,提高网络边缘特征提取性能.原型掩膜(Protonet)分支生成原型掩膜与预测头的掩膜因子进行线性组合,组合结果激活后得到分割掩膜,使AS-VIS在获得较好的精度的同时也能达到实时分割的效果.AS-VIS总体模型架构如图1所示.

### 2.1 锚框校准

基于锚框的一级检测器需要对输入特征图中的大部分区域进行采样,判断这些区域是否存在感兴趣的目标并调整区域边缘,从而能更准确地预测出边界框(Bounding box,Bbox).对于纵横比不同的锚框,应采用同锚框匹配的感受野提取特征,较大的锚框采用大的感受野,较小的锚框采用较小的感受野[20].针对上述问题,引用锚框校准(Anchor Frame Calibration,AFC),以提高网络对目标检测和分割的性能.常用一级检测器中直接在锚框的中心点上采用感受野为3×3的卷积核来提取特征,导致特征提取不充分.为解决该问题,本文将采用3×3,3×5和5×3这3个不同感受野的卷积核提取特征.特征图先通过这3个卷积核采样后得到3张

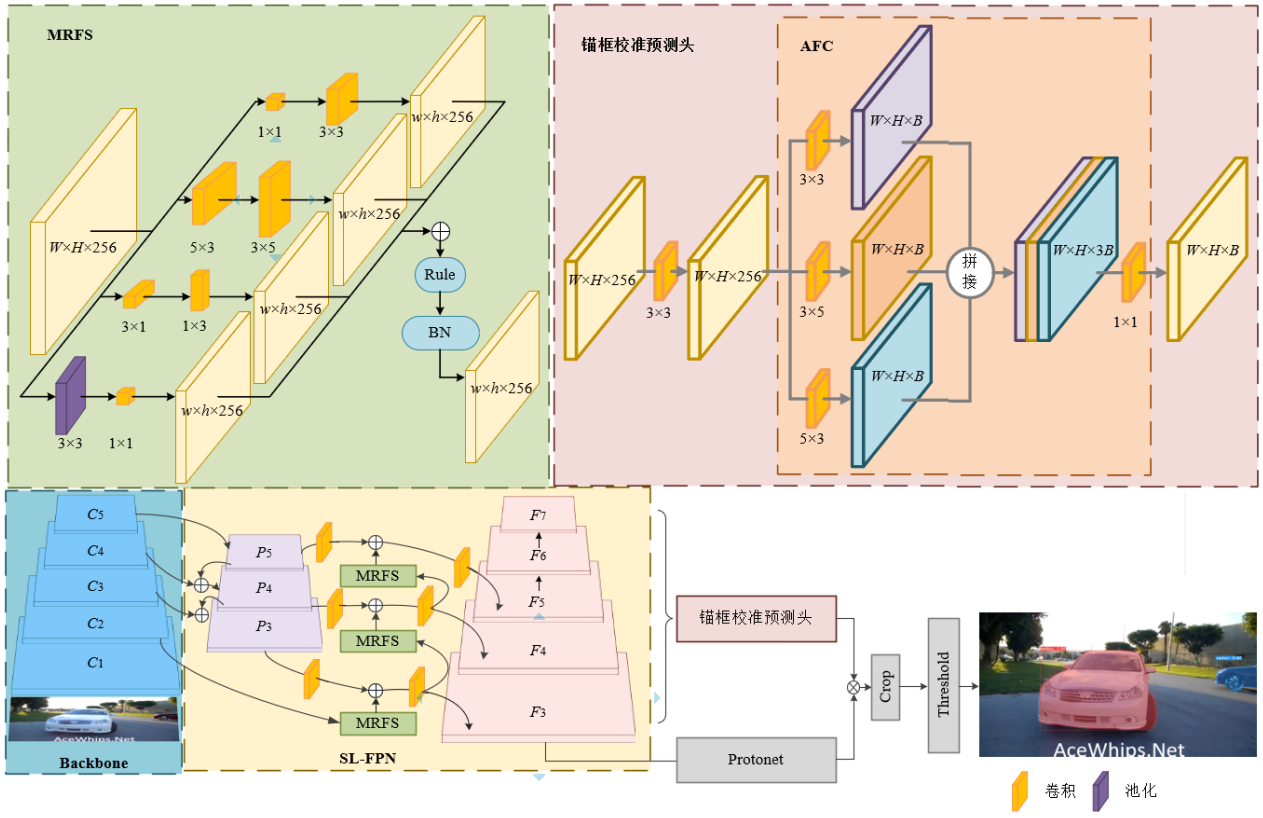



图1 AS-VIS总体模型架构

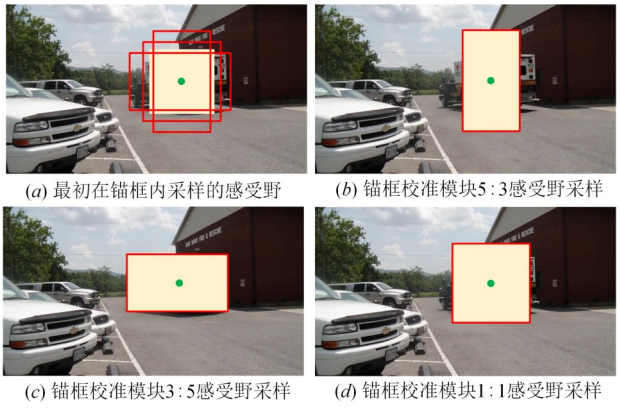
新特征图,再将输出的3张新特征图拼接一起后,通过一个1×1的卷积核将拼接后通道维度降回到拼接前通道维度.因为3个卷积核分别为3×3,3×5和5×3,因此对应的锚框纵横比为1∶1,3∶5和5∶3,卷积核纵横比与锚框纵横比相同,以实现锚框校准,从而有效解决不同纵横比锚框内应用单一感受野卷积采样导致的特征提取不充分问题.

锚框校准原理如图2所示,红色框为锚框,绿色点为锚点,卷积特征的感受野为淡黄色区域.图2(*a*)为最初在锚框内采样的方法,即无论锚框纵横比如何都采用相同感受野提取特征;而图2(*b*)~(*d*)则是对不同纵横比的锚框应用与锚框匹配的感受野提取特征.该部分用于预测头的Bbox,Mask和Class的3个分支,网络结构如图1中AFC部分.在Bbox分支上使用锚框校准时通道数$B$值为$4a$,在Class分支上添加锚框校准时通道数$B$值为$c$,在Mask分支上使用锚框校准时通道数$B$值为$m$.其中,$a$为每个像素点生成锚框数,取值为3;$c$为类别数;$m$为掩膜系数.锚框校准模块表达式为

$$F_{\text{out}}=\text{Conv}_{1\times1}\left(\text{Cat}\left(\text{Conv}_{5\times3}\left(F_{\text{in}}\right),\text{Conv}_{3\times5}\left(F_{\text{in}}\right),\text{Conv}_{3\times3}\left(F_{\text{in}}\right)\right)\right) \tag{1}$$

其中,$\text{Conv}_{i\times j}(\cdot)$表示卷积核为$i\times j$的卷积操作;$\text{Cat}(\cdot)$表示将特征图拼接;$F_{\text{in}}$表示输入的特征图;$F_{\text{out}}$表示输出的特征图.

验证时预测头中3个分支的特征图第一个通道进行可视化如图3所示.从图中红圈部分可以看出,经过锚框校准的Bbox和Mask分支上特征图中目标轮廓更为清晰,在Class分支上的特征图背景抑制效果好,目标显示明确.



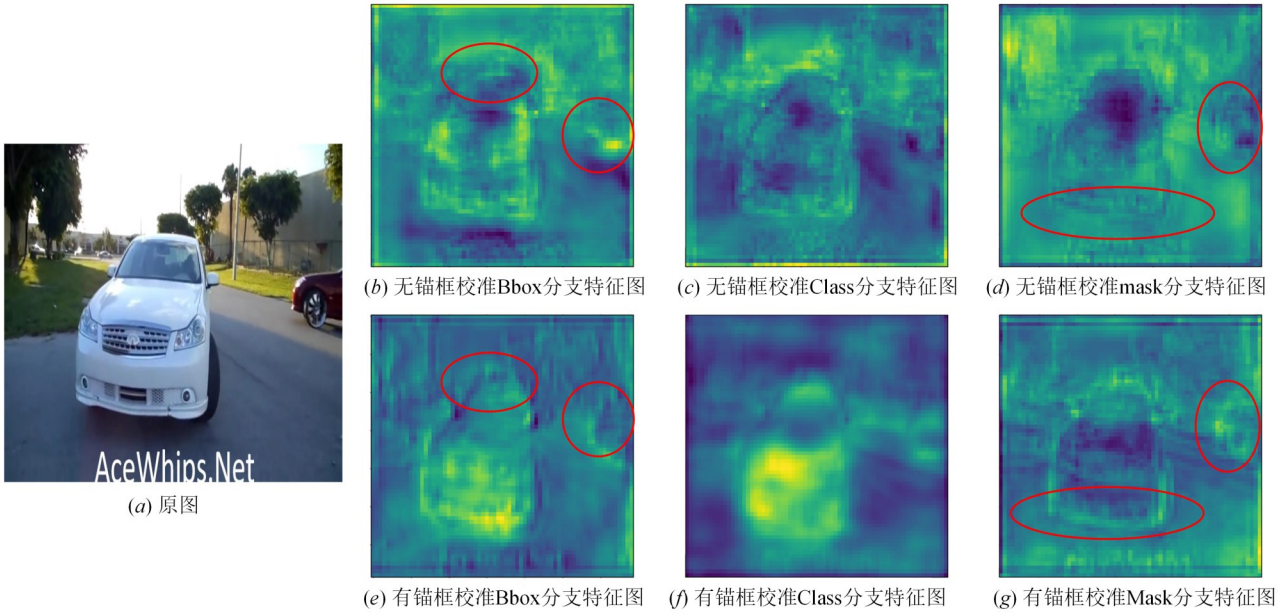
(*a*) 最初在锚框内采样的感受野　(*b*) 锚框校准模块5∶3感受野采样

(*c*) 锚框校准模块3∶5感受野采样　(*d*) 锚框校准模块1∶1感受野采样

图2　锚框校准原理

## 2.2　多感受野下采样模块

在深度学习中常用的下采样方式为池化操作和卷积操作.池化操作一般采用最大值池化、平均池化、随机池化以及求和区域池化等方法,池化操作具有平移、旋转以及尺度不变性,在保留特征图主要信息的同时能减小特征图参数量,还能防止过拟合提高网络泛化能力等,但池化操作会将特征图中的细节信息滤除掉,对分割精度有较大影响.卷积操作则通过步长为2的卷积实现下采样操作,其将卷积核区域内的像素组合成新的特征像素得到分辨率更低的特征图.与池化下采样相比,卷积下采样能够保留的信息更多,但步长为2的卷积操作会缺失一半特征信息.为此,设计一种多分支并行多种感受野下采样模块对特征图进行下采样操作,解决下采样后信息缺失问题.先应用多种感受野的卷积下采样操作和一个池化操作对特征图进行下采




(*a*) 原图　(*b*) 无锚框校准Bbox分支特征图　(*c*) 无锚框校准Class分支特征图　(*d*) 无锚框校准mask分支特征图

(*e*) 有锚框校准Bbox分支特征图　(*f*) 有锚框校准Class分支特征图　(*g*) 有锚框校准Mask分支特征图

图3　预测头各分支有无锚框校准特征图可视化对比

样,然后将多种采样后的特征融合.本文采用4个分支(包括3个卷积下采样分支和1个池化操作分支),卷积下采样分支的区别在于卷积核的尺寸不一样,即感受野不一样.第1个分支先采用1×1的卷积核对特征图进行卷积操作,其主要作用是实现跨通道的信息组合增加非线性特征,然后再通过一个3×3的卷积核步长为2的卷积操作对图像进行下采样.第2个分支先使用5×3步长为(2,1)的卷积对特征图的宽($w$)方向进行采样操作,然后再使用一个3×5步长为(1,2)的卷积对特征图的高($h$)方向进行采样操作.第3个分支先使用3×1步长为(2,1)的卷积对特征图的宽($w$)方向进行采样操作,然后再使用一个1×3步长为(1,2)的卷积对特征图的高($h$)方向进行采样操作.第4个分支先用核为3×3步长为2的平均池化操作,然后使用1×1卷积增加特征图的非线性特征.最后,将4个分支所得到的特征进行融合、激活和归一化后输出下采样最终结果,该法称为多感受野下采样(Multi-Receptive Field Subsampling,MRFS),网络结构如图1中MRFS部分,其表达式为

$$P_{\mathrm{cc}}=\mathrm{Conv}_{i\times j}\left(\mathrm{Conv}_{i\times j}\left(F_{\mathrm{in}}\right)\right) \tag{2}$$

$$P_{\mathrm{pc}}=\mathrm{Conv}_{i\times j}\left(\mathrm{Pool}_{i\times j}\left(F_{\mathrm{in}}\right)\right) \tag{3}$$

$$\mathrm{MRFS}=\mathrm{BN}\left(\mathrm{ReLU}\left(\sum_{n=1}^{3}P_{\mathrm{cc}}+P_{\mathrm{pc}}\right)\right) \tag{4}$$

其中,$P_{\mathrm{cc}}$表示为两个串联的卷积;$P_{\mathrm{pc}}$表示为先池化再卷积;$\mathrm{Conv}_{i\times j}(\cdot)$表示卷积核为$i\times j$的卷积操作;$\mathrm{Pool}_{i\times j}(\cdot)$表示池化核为$i\times j$的池化操作;$F_{in}$表示输入的特征图;MRFS表示最终输出;ReLU为激活函数;BN为归一化.

根据常规的感受野计算方式,当卷积核为$k$,步长为1时,感受野计算方法为

$$r=r+(k-1) \tag{5}$$

当卷积核为$k$,步长为2时,感受野计算方法为

$$r=(r\times 2)+(k-2) \tag{6}$$

综上所述,当卷积核为$k$,步长为$s$时,感受野的计算公式为

$$r=(r\times s)+(k-s) \tag{7}$$

因为本文中使用的卷积核和步长的方法是非对称的,因此将卷积核与步长写成向量的形式,计算方式为

$$\begin{bmatrix}R_{i+1}\\C_{i+1}\end{bmatrix}=\begin{bmatrix}R_i\\C_i\end{bmatrix}\odot\begin{bmatrix}S_R\\S_C\end{bmatrix}+\begin{bmatrix}k_R\\k_C\end{bmatrix}-\begin{bmatrix}S_R\\S_C\end{bmatrix} \tag{8}$$

其中,$R_i$为行感受野;$C_i$列感受野;$S_R$为行步长;$S_C$为列步长;$k_R$为卷积核的行大小;$k_C$为卷积核的列大小;$\odot$为Hadamard乘积.

进行池化操作时,感受野计算公式为

$$r=r\times 2 \tag{9}$$

MRFS中,第1,2和3分支应用式(8)计算每个分支的感受野,第4个分支的池化操作应用式(9)计算感受野,得到每个分支上每一层的感受野如表1所示.不同感受野的卷积核采样得到的特征信息有所差异.将采样后的结果融合能够解决单个卷积核采样导致的信息缺失问题.

**表1 多感受野下采样模块各分支感受野**

| 分支序号 | 第一层 | 第二层 |
|---|---|---|
| 1 | (1,1) | (3,3) |
| 2 | (5,3) | (7,9) |
| 3 | (3,1) | (3,3) |
| 4 | (2,2) | (2,2) |

### 2.3 空间位置信息补偿特征金字塔

不同大小的目标经过相同的下采样后出现较大差异,最常见的表现就是小目标检测精度比较低.特征金字塔具有在不同尺度下有不同分辨率的特点,不同大小的目标都可以在相应的尺度下用合适的特征表示.融合多尺度信息在不同尺度下对不同大小的目标进行预测,能更好地提升网络的性能.

在深度学习中,特征图随着卷积层的增多呈现分辨率逐渐降低的现象,因此低层特征中仍保留目标的纹理和边缘等信息即为空间位置信息,而高层特征中经过多层卷积后语义信息会越来越明显,但由于卷积层的增多,空间位置信息会逐渐丢失.

特征金字塔[29]将高层特征图通过上采样后与低层特征图进行融合丰富低层特征图的语义信息,但是没有将低层特征图中丰富的空间位置信息融合进高层特征图中.

针对特征金字塔中高层特征图缺失空间位置信息的问题,提出空间位置信息补偿特征金字塔(SL-FPN),丰富高层特征空间细节位置信息.首先,从骨干网中获取$C_2$~$C_5$层特征并通过一个1×1的卷积核将通道压缩至256,压缩通道后的$C_5$层为$P_5$层.$P_5$层应用双线性插值上采样后与$C_4$层压缩特征逐项相加融合得到$P_4$层.同理,$P_4$层应用双线性上采样后与$C_3$层压缩特征逐项相加融合得到$P_3$层.然后,将$C_2$层利用多感受野下采样(MRFS)模块对特征图进行下采样后与$P_3$层经过一次3×3的卷积操作后的特征图进行相加,再经过一次3×3的卷积操作就可以得到$F_3$层.将$F_3$层经过MRFS模块进行下采样后与$P_4$层经过一次3×3的卷积操作后的特征图进行相加得到$F_4$,同理可得$F_5$.最后,将$F_5$依次经过一次卷积得到$F_6$,$F_6$再经过一次卷积得到$F_7$,网络结构如图1中SL-FPN部分.本节提到的多感受野下采样模块已在第1.2节中介绍过,空间位置信息补偿$F_3$~$F_5$表达式如式(10)所示,$F_6$和$F_7$表达式如式(11)所示.

$$F_i^{\mathrm{out}}=\mathrm{Conv}_{3\times 3}\left(\mathrm{MRFS}(F_{i-1})+\mathrm{Conv}_{3\times 3}(F_i)\right) \tag{10}$$

$$F_i^{\mathrm{out}}=\mathrm{Conv}_{3\times 3}(F_{i-1}) \tag{11}$$

其中，$\mathrm{Conv}_{3\times3}(\cdot)$表示卷积核为3×3的卷积操作；MRFS(·)为多感受野下采样模块；$F_{i-1}$为上一层浅层特征图；$F_i$为当前层特征图；$F_i^{\mathrm{out}}$为输出特征图.

通过建立空间位置信息补偿特征金字塔将低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿，增加高层空间细节信息以提高检测精度和分割精度. 将验证得到的特征图在通道维度上取最大值得到的热力特征与原图融合即为热力图，特征金字塔中的$F_3$~$F_7$热力图可视化如图4所示. 从$F_4$层中可看出，SL-FPN热力分布比FPN能突出目标的轮廓；从$F_5$，$F_6$和$F_7$层热力图中可以看出，SL-FPN热力分布相对FPN更集中在目标中，尤其在$F_6$层表现更明显.

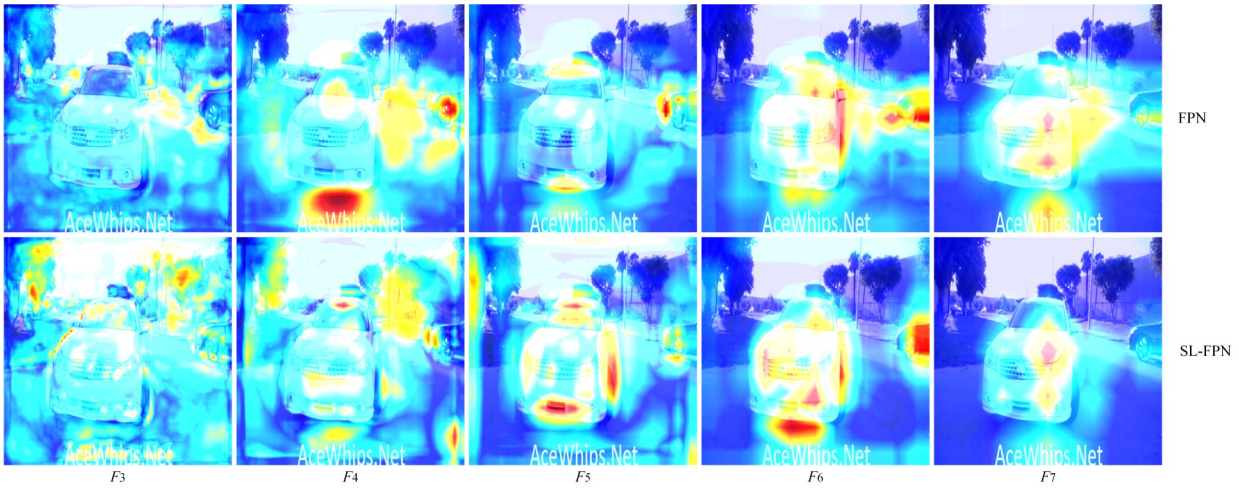



图4　FPN和SL-FPN热力图对比

## 2.4　骨干网络和原型掩膜分支

AS-VIS使用ResNet50作为骨干网络提取街道场景中目标实例的特征，其主要由多个残差块组成. 将$x$输入到残差块中依次经过卷积核大小为1×1，3×3和1×1的卷积操作，每次卷积完成都要经过一次ReLU进行激活得到$F(x)$，然后使用跳跃连接将输入$x$与$F(x)$相加即$H(x)=F(x)+x$. 应用跳跃连接能够解决深度神经网络中带来的梯度消失问题，使网络能够搭建的更深泛化能力更强. 其中，第1个卷积核大小为1×1的卷积操作，其目的是把特征图通道数减少，为卷积核大小为3×3的卷积操作减少计算复杂度. 第2个卷积核大小为1×1的卷积操作，则是为了还原特征图的通道数，这样既能保证信息的完整性，又能简化学习目标和难度.

原型掩膜分支(Protonet)的实现是基于FCN(Full Connected Network)，将整张图像预测成一组$n$个原型掩膜，即$n$个通道. Protonet类似于语义分割网络. 但不同的是，Protonet分支训练不单独设置损失函数，只在整个网络最后输出的掩膜上进行监督. 高分辨率原型掩膜有利于提高小目标检测和分割精度，而空间位置信息补偿特征金字塔的$F_3$层经过特征融合后包含着丰富的语义信息和位置信息，故选择$F_3$层输入Protonet分支生成原型掩膜，如图5所示. 原型掩膜对网络的检测和分割性能是十分重要的，能够使网络对十分确定的原型掩膜(如明显的背景)产生大量的、压倒性的响应. 为了获得更多可解释的原型掩膜，采用ReLU对原型掩膜进行激活.

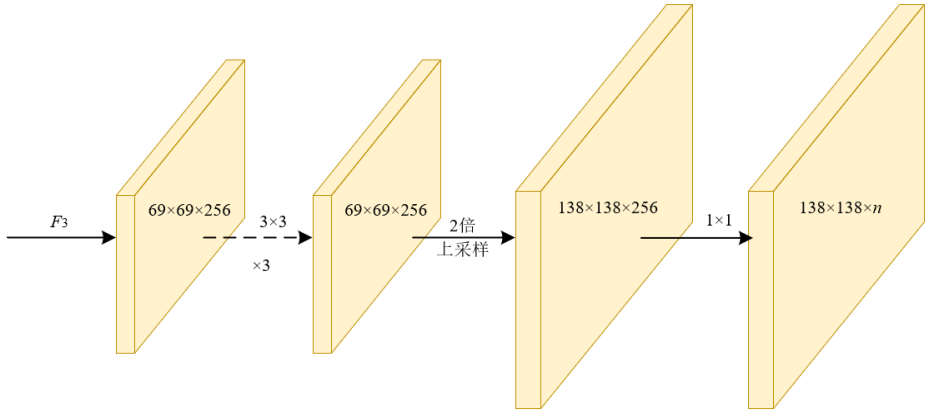



图5　Protonet分支

### 2.5 损失函数和后处理

总体损失函数是分类损失 $L_{cls}$、边界框损失 $L_{box}$ 和掩膜损失 $L_{mask}$ 的加权和，权重分别为 1，1.5 和 6.125，如式(12)所示.

$$L=L_{cls}+1.5L_{box}+6.125L_{mask} \tag{12}$$

其中，$L_{cls}$ 和 $L_{box}$ 采用 SSD[30] 中的定义方式；$L_{mask}$ 使用预测的掩膜 $M$ 和真实值掩膜 $M_{gt}$ 之间的像素级二进制交叉熵，$L_{mask}=\mathrm{BCE}\left(M,M_{gt}\right)$.

预测头中的 Mask 分支生成掩膜系数，通过 Fast NMS 处理后得到矩阵维度为 $n\times k$ 的掩膜系数，记为 $\boldsymbol{M}_c$；与原型掩膜分支得到维度为 $h\times w\times k$ 的掩膜矩阵相乘，记为 $\boldsymbol{P}$；最后，使用 sigmoid 处理获得非线性掩膜.

$$M=\mathrm{sigmoid}\left(\boldsymbol{P}\boldsymbol{M}_c^{\mathrm{T}}\right) \tag{13}$$

Crop：在验证阶段，用预测的边界框剪裁最终的掩膜，而在训练期间改为使用真实边界框进行裁剪掩膜，并将 $L_{mask}$ 除以真实边界框区域，以保留原型中的小目标.

## 3 实验结果与分析

### 3.1 实验准备

#### 3.1.1 数据集建立

本文采用的数据集是在 Youtube-VIS 数据集的基础上，将街道中常见的目标提取出来，包括人、摩托车、滑板、轿车、卡车、火车、狗 7 类. 由于 Youtube-VIS 数据集的验证集中官方没有公开标签信息，因此在 Yolact-Edge 提供的训练集和验证集标签文件中提取出对应的目标，建立街道场景数据集. 其中，训练集含 329 个视频片段，总帧数 7 212 帧，包含 603 个实例，验证集中含 53 个视频片段，总帧数 1 097 帧，包含 88 个实例. 各类实例数量如图 6 所示. 图 6 中，蓝色为训练集各类别实例数量，黄色为验证集各类别实例数量.

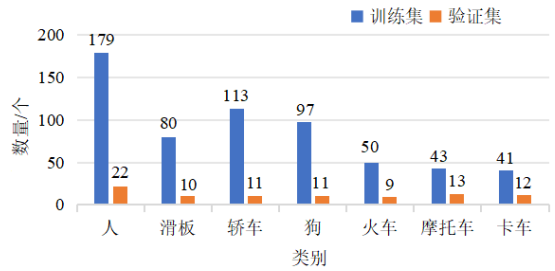



图 6 训练集和验证集各类别实例数量

#### 3.1.2 实验环境

本文实验平台为 Ubuntu18.04 操作系统，内存 16 GB，CPU 为 AMD Ryzen 7 3700X，GPU 为 NVIDIA GTX3070 的台式计算机. 深度学习框架为 Pytorch1.7，采用 CUDA11.0 和 cuDNN8.0.5 加速工具包提高网络训练速度. 实验基于 Pytorch 深度学习框架搭建，训练时共迭代 80 000 次，batch size 大小为 6，初始学习率为 0.000 05 学习率衰减采用 consine.

#### 3.1.3 评价指标

为验证本文所提算法的优越性及有效性，选取 COCO 数据集评价指标中的平均精度(mean Average Precision，mAP)对算法模型进行综合评估. COCO 数据集使用的阈值，即交并比(Intersection over Union，IoU)，以 0.05 为增量，分别计算 0.5~0.95 共 10 个阈值下的 AP，最后将所有阈值的 AP 求取平均值. COCO 评价指标对实验结果的准确性要求远高于单一阈值评价指标，更能体现算法的真实性，mAP 计算如式(14)所示.

$$\mathrm{mAP}=\frac{1}{\mathrm{class}}\left(\sum_c\frac{1}{\mathrm{thresholds}}\sum_t\frac{\mathrm{TP}(t)}{\mathrm{TP}(t)+\mathrm{FP}(t)}\right) \tag{14}$$

其中，class 为类别总数；$c$ 为当前类别；thresholds 为阈值数；$t$ 为当前阈值. 在 COCO 的评价指标中，AP，$\mathrm{AP}_{50}$ 和 $\mathrm{AP}_{75}$ 分别代表 $\mathrm{mAP}^{\mathrm{IoU=0.5:0.05:0.95}}$，$\mathrm{mAP}^{\mathrm{IoU=0.5}}$ 和 $\mathrm{mAP}^{\mathrm{IoU=0.75}}$，本文将用这 3 个指标评价模型精度.

### 3.2 实验结果

#### 3.2.1 锚框校准

通常网络预测时在特征图的每个像素点上生成纵横比不同的锚框，然后采用一个 3×3 的卷积核提取特征. 不同纵横比的锚框使用单一感受野提取特征，导致特征提取不充分影响网络检测和分割性能. 通过应用多类型纵横比的卷积核提取对应纵横比锚框内特征，从而保证多类型纵横比锚框内特征能够充分提取提高检测精度和分割精度. 本文在 Bbox 分支上线性输出 $4a$、Class 分支上输出二进制预测 $c$ 和 Mask 分支上输出掩膜系数 $m$ 时采用锚框校准模块.

在 Bbox，Class 和 Mask 这 3 个分支上单独或两两组合添加锚框校准模块，以及在 3 个分支上同时添加锚框校准模块进行对比实验. 表 2(加粗数据表示最优结果)中可以看出，当在 Bbox，Class 和 Mask 上都添加锚框校准模块时网络的检测和分割的精度最高，Bbox 和 Mask 的 AP 相比基准分别提升 8.63% 和 5.09%，在 $\mathrm{AP}_{50}$ 时则分别提升 7.14% 和 3.33%，而在 $\mathrm{AP}_{75}$ 则分别提升 14.24% 和 7%. 3 个分支中在 Class 分支上添加锚框校准模块对检测和分割精度的影响最大. 仅在 Bbox 和 Mask 分别或者同时加上而 Class 分支不添加锚框校准模块时，检测和分割精度都低于 Class 分支中添加锚框校准模块. 此外，在 Bbox 分支上添加锚框校准模块对检测和分割精度的影响仅次于添加在 Class 分支上，锚框校准模块仅存在于 Bbox 分支时其效果比仅存在于 Mask 分支的分割效果更好. 添加锚框校准使模型参数量增加导致分割速度下降 20% 左右，其中 Bbox 和 Class 分支添加锚框校准模块使模型分割速度下降约 12%，而 Mask 分支添加锚框校准模块速度则下降

表2　锚框校准消融实验

| 添加位置 | | | Bbox | | | Mask | | | 分割速度/FPS |
|---|---|---|---|---|---|---|---|---|---|
| Bbox | Mask | Class | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% | |
| × | × | × | 29.47 | 57.20 | 27.51 | 30.57 | 54.48 | 30.99 | 42~43 |
| √ | × | × | 33.44 | 60.23 | 33.18 | 33.37 | 54.98 | 34.94 | 37~39 |
| × | √ | × | 34.04 | 57.95 | 37.77 | 31.82 | 54.24 | 31.25 | 35~37 |
| × | × | √ | 35.31 | 62.01 | 35.67 | 34.27 | 58.50 | 34.51 | 37~39 |
| √ | √ | × | 29.66 | 57.63 | 27.03 | 28.71 | 50.17 | 27.64 | 35~36 |
| √ | × | √ | 36.01 | 62.15 | 36.99 | 34.90 | **58.68** | 36.04 | 36~37 |
| × | √ | √ | 35.22 | 61.89 | 35.28 | 34.29 | 57.07 | 36.03 | 33~35 |
| √ | √ | √ | **38.10** | **64.34** | **41.75** | **35.66** | 57.81 | **37.99** | 33~35 |

约 17%. 因为 Mask 分支上输出特征图的通道数是 $m$,在网络中 $m$ 取值为 32,所以 Mask 分支的参数要比 Bbox 和 Class 分支多,因此 Mask 分支上添加锚框校准模块对速度的影响最大.

### 3.2.2　多感受野下采样模块

在神经网络中数据的处理和融合需要将维度对齐,对于特征图长和高的维度缩小通常称为下采样. 一般常用卷积或池化作为下采样方法,本文设计 MRFS 模块应用多种感受野对特征图进行下采样操作,然后将多个感受野采样得到的特征图进行融合、激活和归一化. 因为 MRFS 模块是在空间位置信息补偿特征金塔(SL-FPN)中应用,因此本节将在 SL-FPN 中的下采样操作中应用 MRFS、卷积下采样和池化下采样 3 种方式去验证本文设计的 MRFS 的效果. 其中,Conv 表示卷积核为 3×3 步长为 2 的卷积操作,MaxPool 表示最大池化操作.

实验结果如表 3 所示,加粗数据表示最优结果. 当在相同条件下应用卷积和池化操作对特征图进行下采样操作时分割精度 AP 分别为 35.43% 和 33.79%,应用 MRFS 进行下采样分割精度 AP 为 37.03%. 可见 MRFS 对特征图下采样效果要比应用卷积或者池化效果好.

表3　下采样模块实验

| 方法 | Bbox | | | Mask | | | 分割速度/FPS |
|---|---|---|---|---|---|---|---|
| | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% | |
| Baseline+AFC+MRFS | 38.73 | **68.05** | **39.19** | **37.03** | **61.56** | **39.18** | 25~26 |
| Baseline+AFC+Conv | **38.95** | 64.69 | 39.12 | 35.43 | 57.17 | 37.63 | 30~31 |
| Baseline+AFC+MaxPool | 36.81 | 61.92 | 36.78 | 33.79 | 56.39 | 34.31 | 32~33 |

### 3.2.3　空间位置信息补偿特征金字塔

经过多次卷积特征提取后得到高层特征,其包含着丰富的语义特征,但空间位置信息匮乏. 因此,本文提出空间位置信息补偿特征金字塔(SL-FPN),应用多感受野下采样模块将特征金字塔低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿,解决高层特征空间细节位置信息匮乏的问题. 实验结果如表 4 所示,添加 SL-FPN 时检测精度 AP,$AP_{50}$ 和 $AP_{75}$ 分别提升 7.76%,5.27% 和 12.5%,分割精度 AP,$AP_{50}$ 和 $AP_{75}$ 分别提升 4.75%,3.82% 和 6.86%. 实验结果表明,空间位置信息补偿特征金字塔对提升网络分割精度有明显的效果,但参数量的增加使模型分割速度下降 28% 左右.

表4　空间位置信息补偿

| 方法 | Bbox | | | Mask | | | 分割速度/FPS |
|---|---|---|---|---|---|---|---|
| | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% | |
| Baseline | 29.47 | 57.20 | 27.51 | 30.57 | 54.48 | 30.99 | 42~43 |
| Baseline+SL-FPN | 37.23 | 62.47 | 40.01 | 35.32 | 58.30 | 37.85 | 30~31 |

### 3.2.4　AS-VIS 消融实验

通过消融实验验证锚框校准(ACF)与空间位置信息补偿特征金字塔(SL-FPN)单独使用以及两者融合时对网络检测和分割精度的影响. 实验结果如表 5 所示,单独添加 SL-FPN 时检测精度 AP,$AP_{50}$ 和 $AP_{75}$ 分别提升 7.76%,5.27% 和 12.5%,分割精度 AP,$AP_{50}$ 和 $AP_{75}$ 分别提升 4.75%,3.82% 和 6.86%. 同时,添加 SL-FPN 和 AFC 时相比只添加 SL-FPN 时分割精度 AP,$AP_{50}$ 和 $AP_{75}$ 分别提升 1.71%,3.26% 和 1.33%,但由于参数量增多,模型分割速度下降. 从实验结果表 4 中可以看出,SL-FPN 对提高网络精度是有效果的,但 SL-FPN 和 AFC 一起使用对网络精度的提升并没起到叠加的效果. 这两种方法对检测和分割都起到空间位置校准的效果,但是 SL-FPN 和 AFC 分别单独计算并没有建立其联立关系,

表5 AS-VIS消融实验

| 方法 | Bbox | | | Mask | | | 分割速度/FPS |
|---|---|---|---|---|---|---|---|
| | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% | |
| Baseline | 29.47 | 57.2 | 27.51 | 30.57 | 54.48 | 30.99 | 42~43 |
| Baseline+SL-FPN | 37.23 | 62.47 | 40.01 | 35.32 | 58.30 | 37.85 | 30~31 |
| Baseline+AFC | 38.10 | 64.34 | **41.75** | 35.66 | 57.81 | 37.99 | 33~35 |
| Baseline+AFC+SL-FPN | **38.73** | **68.05** | 39.19 | **37.03** | **61.56** | **39.18** | 25~26 |

导致两个模块对网络精度的影响并不能起到叠加的效果.

将模型对数据集各类分割精度进行可视化,如图7所示. 从图中得知,AS-VIS可提高除摩托车类别外所有实例的分割精度. AFC使用同锚框纵横比匹配的卷积采样时,能提高模型对于目标边缘位置的敏感度,从而有效提升网络对目标实例边缘轮廓分割效果. 同时,SL-FPN通过将低层特征中的目标的纹理和边缘等空间细节信息补偿到高层特征中,从而增加高层特征空间细节信息,以提高分割精度. AS-VIS对人、滑板、火车等尺度变化小并且大多数显示齐全的类别分割精度提升显著,而摩托车类别图像大多数只有车头部分,所以模型对摩托车形状和纹理等信息的学习混乱,使模型识别摩托车不准确.

为更好地分析各个模块对于AS-VIS的作用,本文将Baseline, Baseline+AFC 和 Baseline+AFC+SL-FPN 训练过程中总损失值可视化,如图8所示. 从图8中曲线可知,3个模型总损失值随着迭代次数的变化而变化,损失值总体趋势呈逐渐下降,并在60 000次迭代后损失值逐渐平稳. 其中,Baseline的最终损失最大,而带有AFC和SL-FPN的模型最终损失最小. 这说明,本文所提出的AFC和SL-FPN对模型性能有较大提升.

### 3.2.5 AS-VIS交叉验证

为评价本文模型对于不同数据集学习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与分割性能. 在第3.1.1节中所构建的街道场景数据集基础上进行3次随机重复采样,在重复采样过程中保证训练集与验证机划分比例与原文基本接近,即训练集与验证集之间的比例为85%∶15%. 重复采样获得的数据组成3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,重采样的获得的5组数据集具体实例信息分布如图9所示.

在图9所划分的数据基础上,使用本文AS-VIS模型分别对3组数据集训练并验证,最后计算其均值与基准模型作以对比,实验结果分别如表6和表7所示. 可以看出,3组验证模型的平均检测精度均值为40.78,方

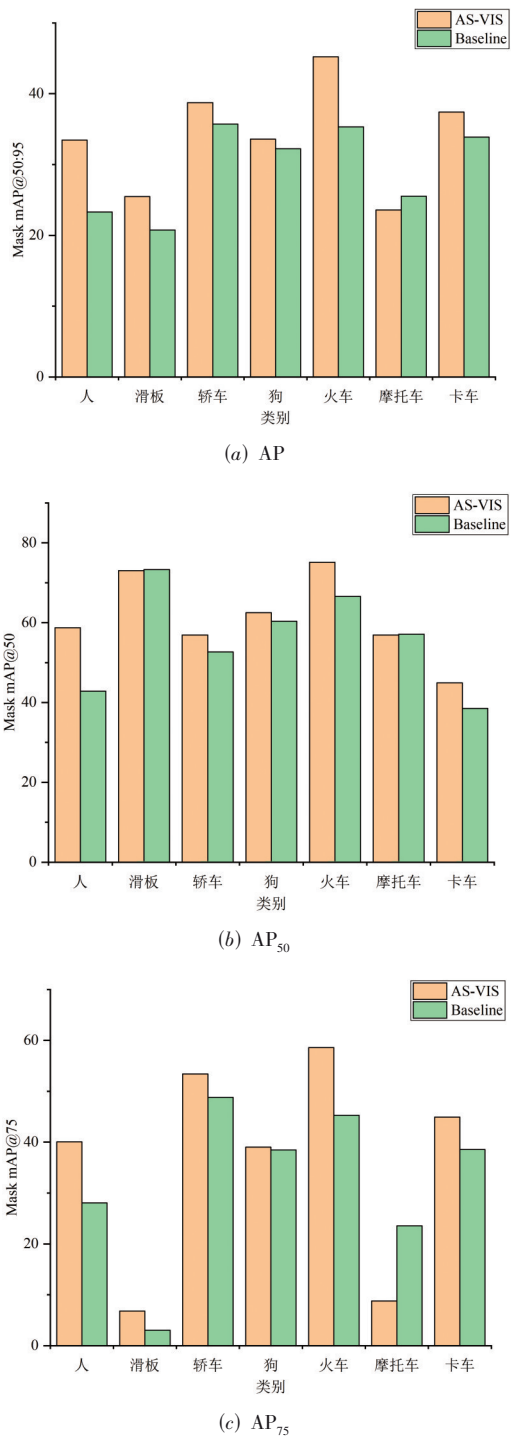



图7 AS-VIS和Baseline类间精度对比

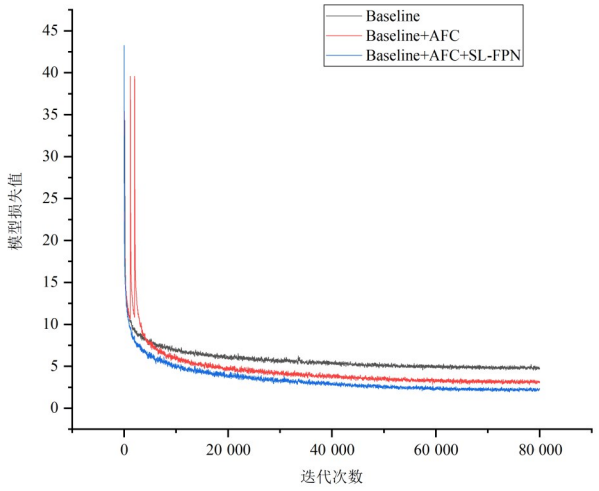



图8　模型损失收敛

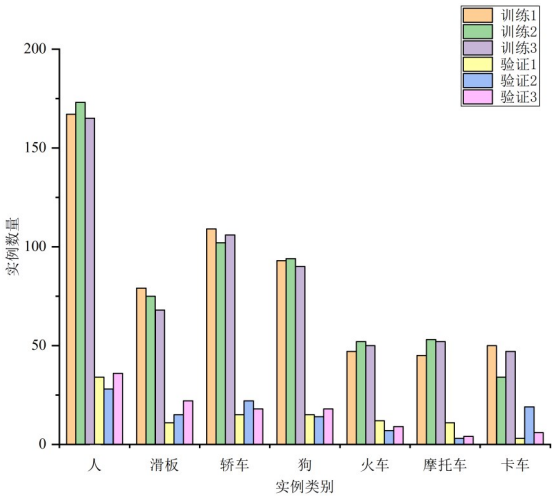



图9　交叉验证组数据集实例样本分布信息

差为0.58,平均分割精度的均值为37.67,方差为0.07,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,本文模型平均检测精度和平均分割精度分别提高了11.31%和7.1%. 这表明本文提出的视频实例分割模型AS-VIS有较好的稳定性与鲁棒性.

**表6**　AS-VIS交叉验证实验

| 实验次数 | Bbox | | | Mask | | |
|---|---|---|---|---|---|---|
| | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% |
| 1 | 39.87 | 66.42 | 40.91 | 37.31 | 61.16 | 35.94 |
| 2 | 40.74 | 69.65 | 42.86 | 37.95 | 65.81 | 37.12 |
| 3 | 41.73 | 66.27 | 46.49 | 37.76 | 62.97 | 38.80 |

**表7**　基础模型与交叉验证实验结果对比　　单位:%

| 模型 | Bbox | Mask |
|---|---|---|
| Baseline | 29.47 | 30.57 |
| AS-VIS | 38.73 | 37.03 |
| AS-VIS(交叉验证) | 40.78(±0.58) | 37.67(±0.07) |

### 3.2.6　AS-VIS与4种先进的单阶VIS模型对比

与YolactEdge相比,AS-VIS的网络分割精度提升6.4%,可见SL-FPN和AFC对网络提升精度是有效的. 两种网络分割可视化效果如图10所示. 第1个和第3个视频帧中AS-VIS分割目标边缘比YolactEdge效果好,主要原因是锚框校准模块通过与锚框纵横比匹配的多类型感受野采样目标特征边缘提取更充分,使AS-VIS对于目标边缘更敏感,从而提高了检测和分割精度. 其中,第3个视频帧中AS-VIS分割目标区域比YolactEdge完整,主要原因是空间位置信息补偿特征金字塔将低层目标区域激活特征映射嵌入到高层中实现空间位置信息补偿,增加高层空间细节信息. 第2个视频帧中,AS-VIS的误检和漏检比YolactEdge低,是因为空间位置信息补偿特征金字塔增加目标空间细节信息,从而使AS-VIS误检和漏检出现概率降低. 总体来看,AS-VIS比YolactEdge能够更全面地将目标分割出来,并且误检漏检出现的概率更低. 将本文方法与4种先进的单阶视频实例分割网络[15,18,23,24]进行比较,为消除骨干网对整体网络精度的影响,AS-VIS与对比网络均使用ResNet50作为骨干网络,实验结果如表8所示,加粗数据表示最优结果. 在分割精度上,AS-VIS比4种先进的单阶视频实例分割网络效果好,但由于参数量增多,模型分割速度只达到25 FPS,低于其他4种单阶视频实例分割模型. 不能实现实时分割,自动驾驶车辆就不能快速做出准确的判断,这对于行驶车辆是极其危险的.

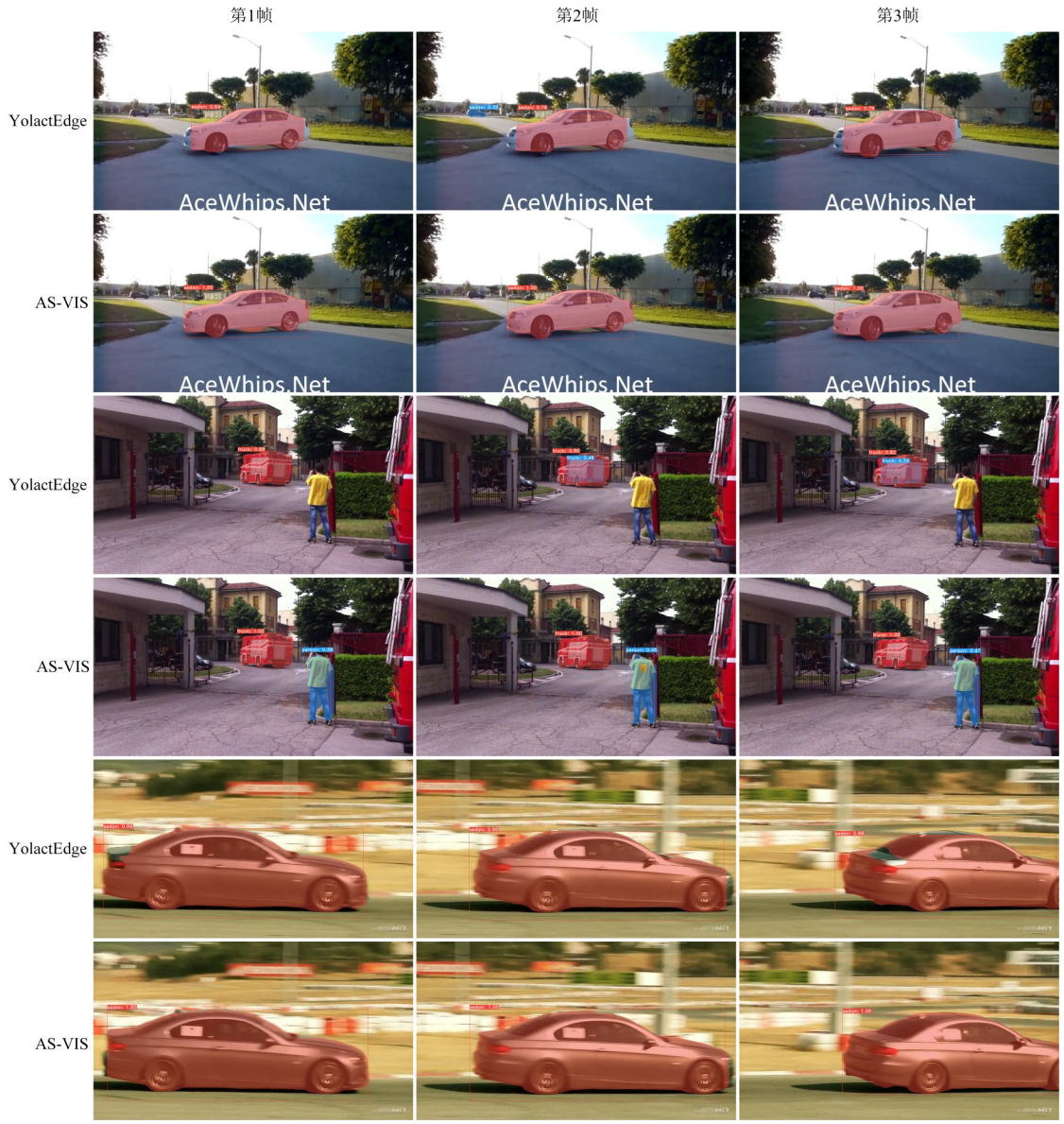



图 10 AS-VIS 与 YolactEdge 分割可视化对比

**表 8** AS-VIS 和 4 种单阶视频实例分割网络性能对比

| 方法 | Bbox | | | Mask | | | 分割速度/FPS |
|---|---|---|---|---|---|---|---|
| | AP/% | $AP_{50}$/% | $AP_{75}$/% | AP/% | $AP_{50}$/% | $AP_{75}$/% | |
| SipMask[15] | — | — | — | 35.6 | 56.6 | 36.4 | 31~33 |
| SG-Net[18] | — | — | — | 33.2 | 58.9 | 36.1 | 25~26 |
| YolactEdge[25] | 29.5 | 57.2 | 27.5 | 30.6 | 54.5 | 31.0 | 42~43 |
| STMask[20] | — | — | — | 34.5 | 56.6 | 36.2 | 29~30 |
| AS-VIS | **38.7** | **68.1** | **39.2** | **37.0** | **61.6** | **39.2** | 25~26 |

## 4 结语

本文提出锚框校准和空间位置信息补偿视频实例分割，实现对街道场景中的目标实例进行检测和分割. 针对特征金字塔高层特征图的空间位置信息匮乏严重的问题，设计 SL-FPN 实现对高层特征图空间位置信息补偿. 多感受野下将包含丰富位置信息的低层特征图进行采样，然后与高层特征图进行融合，对高层特征图空间位置信息进行补偿. 针对锚框校准模块，采用与锚框纵横比匹配的感受野采样，实现锚框校准效果. 与 YolactEdge 基础网络相比，AS-VIS 网络的检测精度和分

割精度都有较大的提升,其中平均检测精度和平均分割精度分别提升9.26%和6.46%. 虽然锚框校准和SLFPN能够有效提升算法整体分割精度,但卷积层的增加导致模型参数量增多使得本文模型推理速度变慢. 未来工作中将探索更简单有效的方法. 另外,视频数据中存在运动模糊、遮挡、形变等问题也是导致模型分割精度低的原因. 未来工作中将探索显式视觉记忆去解决数据中存在的问题.

**参考文献**


[1] 曾凡. 基于激光雷达的低速无人物流车的环境感知算法研究[D]. 重庆: 重庆理工大学, 2020.
ZENG F. Research on Environmental Perception Algorithm of Low Speed Unmanned Logistics Vehicle Based on Lidar[D]. Chongqing: Chongqing University of Technology, 2020. (in Chinese)

[2] 徐国艳, 牛欢, 郭宸阳, 等. 基于三维激光点云的目标识别与跟踪研究[J]. 汽车工程, 2020, 42(1): 38-46.
XU G Y, NIU H, GUO C Y, et al. Research on target recognition and tracking based on 3D laser point cloud[J]. Automotive Engineering, 2020, 42(1): 38-46. (in Chinese)

[3] 王阳阳, 刘之光, 邓航云, 等. 电动小车自动变道环境感知系统[J]. 同济大学学报(自然科学版), 2019, 47(8): 1201-1206.
WANG Y Y, LIU Z G, DENG H Y, et al. Automatic lane change environment perception system of electric vehicle [J]. Journal of Tongji University (Natural Science), 2019, 47(8): 1201-1206. (in Chinese)

[4] 张硕, 叶勤, 史婧, 等. 改进RangeNet++损失函数的车载点云小目标语义分割方法[J]. 计算机辅助设计与图形学学报, 2021, 33(5): 704-711.
ZHANG S, YE Q, SHI J, et al. A semantic segmentation method of in-vehicle small targets point cloud based on improved RangeNet++ loss function[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(5): 704-711. (in Chinese)

[5] 陈治宇. 无人驾驶中多传感器融合环境感知算法研究[D]. 南京: 南京邮电大学, 2020.
CHEN Z Y. Research on Multi-sensor Fusion Environment Perception Algorithm in Autonomous Driving[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[6] WEON I S, LEE S G. Environment recognition based on multi-sensor fusion for autonomous driving vehicles[J]. Journal of Institute of Control, Robotics and Systems, 2019, 25(2): 125-131.

[7] 郑少武, 李巍华, 胡坚耀. 基于激光点云与图像信息融合的交通环境车辆检测[J]. 仪器仪表学报, 2019, 40(12): 143-151.
ZHENG S W, LI W H, HU J Y. Vehicle detection in the traffic environment based on the fusion of laser point cloud and image information[J]. Chinese Journal of Scientific Instrument, 2019, 40(12): 143-151. (in Chinese)

[8] 王新竹, 李骏, 李红建, 等. 基于三维激光雷达和深度图像的自动驾驶汽车障碍物检测方法[J]. 吉林大学学报(工学版), 2016, 46(2): 360-365.
WANG X Z, LI J, LI H J, et al. Obstacle detection based on 3D laser scanner and range image for intelligent vehicle [J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(2): 360-365. (in Chinese)

[9] 王中宇, 倪显扬, 尚振东. 利用卷积神经网络的自动驾驶场景语义分割[J]. 光学精密工程, 2019, 27(11): 2429-2438.
WANG Z Y, NI X Y, SHANG Z D. Autonomous driving semantic segmentation with convolution neural networks [J]. Optics and Precision Engineering, 2019, 27(11): 2429-2438. (in Chinese)

[10] 孟琭, 徐磊, 郭嘉阳. 一种基于改进的MobileNetV2网络语义分割算法[J]. 电子学报, 2020, 48(9): 1769-1776.
MENG L, XU L, GUO J Y. Semantic segmentation algorithm based on improved MobileNetV2[J]. Acta Electronica Sinica, 2020, 48(9): 1769-1776. (in Chinese)

[11] 刘强, 何自芬, 张印辉. 分支空洞卷积神经网络的机加工车间场景语义分割[J]. 计算机辅助设计与图形学学报, 2021, 33(1): 126-141.
LIU Q, HE Z F, ZHANG Y H. Semantic segmentation of mechanical workshop scenes with branch-atrous convolutional neural networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(1): 126-141. (in Chinese)

[12] 邹逸群, 肖志红, 唐夏菲, 等. Anchor-free的尺度自适应行人检测算法[J]. 控制与决策, 2021, 36(2): 295-302.
ZOU Y Q, XIAO Z H, TANG X F, et al. Anchor-free scale adaptive pedestrian detection algorithm[J]. Control and Decision, 2021, 36(2): 295-302. (in Chinese)

[13] YANG L J, FAN Y C, XU N. Video instance segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 5187-5196.

[14] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2980-2988.

[15] CAO J L, ANWER R M, CHOLAKKAL H, et al. SipMask: Spatial information preservation for fast image and video instance segmentation[C]//European Conference on Computer Vision. Cham: Springer, 2020: 1-18.

[16] BOLYA D, ZHOU C, XIAO F Y, et al. YOLACT: Real-time instance segmentation[C]//2019 IEEE/CVF Interna-

tional Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 9156-9165.
[17] BOLYA D, ZHOU C, XIAO F Y, et al. YOLACT++: Better real-time instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 1108-1121.
[18] LIU D F, CUI Y M, TAN W B, et al. SG-net: Spatial granularity network for one-stage video instance segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 9811-9820.
[19] FU Y, YANG L J, LIU D, et al. CompFeat: Comprehensive feature aggregation for video instance segmentation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1361-1369.
[20] LI M H, LI S, LI L D, et al. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 11210-11219.
[21] ZHU X Z, XIONG Y W, DAI J F, et al. Deep feature flow for video recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 4141-4150.
[22] ZHU X Z, WANG Y J, DAI J F, et al. Flow-guided feature aggregation for video object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 408-417.
[23] ZHU X Z, DAI J F, YUAN L, et al. Towards high performance video object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7210-7218.
[24] BERTASIUS G, TORRESANI L. Classifying, segmenting, and tracking object instances in video with mask propagation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9736-9745.
[25] GOEL V, LI J, GARG S, et al. MSN: Efficient online mask selection network for video instance segmentation[EB/OL]. (2021-06-19)[2022-11-25]. https://arxiv.org/abs/2106.10452.
[26] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 9992-10002.
[27] YANG Z X, WEI Y C, YANG Y. Collaborative video object segmentation by foreground-background integration [C]//European Conference on Computer Vision. Cham: Springer, 2020: 332-348.
[28] LIU H T, RIVERA SOTO R A, XIAO F Y, et al. Yolact-Edge: Real-time instance segmentation on the edge[C]// 2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2021: 9579-9585.
[29] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.
[30] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

## 作者简介

**张印辉** 男，1977 年生，河北衡水人．博士．教授．主要研究方向为图像处理、机器视觉和机器智能．
E-mail: yinhui_z@163.com

**赵崇任** 男，1997 年生，广西钦州人．硕士研究生．主要研究方向为视频实例分割．
E-mail: chongren_z@163.com

**何自芬** 女，1976 年生，河北南宫人．博士．副教授．主要研究方向为计算机视觉、图像处理．
E-mail: zyhhzf1998@163.com

**杨宏宽** 男，1995 年生，云南保山人．硕士研究生．主要研究方向为目标检测．
E-mail: kustyhk@163.com

**黄 滢** 女，1997 年生，陕西渭南人．博士研究生．主要研究方向为视频实例分割．
E-mail: erying_h@163.com