doi:10.19562/j.chinasae.qcgc.2025.04.003

基于实例激活图的自动驾驶实时实例分割算法*

秦启瑞[1],王 海[1],蔡英凤[2],陈 龙[2],李祎承[2]
(1. 江苏大学汽车与交通工程学院,镇江 212013; 2. 江苏大学汽车工程研究院,镇江 212013)

[摘要] 基于深度学习的实例分割算法能够帮助智能汽车获取精确的感知信息。但受到制造成本的限制,通常智能汽车上的计算资源有限,为在有限的计算资源下获取高精度的识别与分割,要求算法本身能够充分利用已提取到的特征。同时,一阶段的实例分割算法虽然有较快的推理速度,但其在精度方面有所欠缺。为此,本文对一阶段的实例分割算法SparseInst进行了改进,以提升模型对有效特征的利用率。具体来说,首先在主干网络基础构建块中增加了残差连接。其次,在编码器部分,设计了三尺度特征融合模块克服了原先跨尺度特征不能进行直接交互的问题。本文还设计了解耦的实例激活模块,增强模型对实例特征的学习能力。除此以外,改进的算法充分利用细节特征对掩码特征进行修正,提高了生成掩码的质量。最后,本文用内核去初始化目标物体得分,提高了已提取特征的利用率。改进的算法在多个数据集上的掩码精度超越了同类型算法,且具有较强的实时性。为进一步验证改进算法的有效性,本文利用实车平台收集的数据进行了实验,在输入图片分辨率为640×480时,模型推理速度达到了54 FPS,并精确地分割出了实例掩码。

关键词:自动驾驶;深度学习;实时实例分割;特征利用率

Real-Time Instance Segmentation Algorithm for Autonomous Driving Based on Instance Activation Maps

Qin Qirui[1], Wang Hai[1], Cai Yingfeng[2], Chen Long[2] & Li Yicheng[2]
1. *School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang* 212013;
2. *Institute of Automotive Engineering, Jiangsu University, Zhenjiang* 212013

[**Abstract**] Instance segmentation algorithms based on deep learning are capable of helping intelligent vehicles to obtain accurate perception information. However, due to the limitation of manufacturing cost, the computing resources on intelligent vehicles are usually limited. In order to obtain high-precision recognition and segmentation under limited computing resources, the algorithm itself is required to make full use of the extracted features. Meanwhile, although the one-stage instance segmentation algorithm has a relative fast inference speed, it has poor performance in accuracy. To this end, structural improvement based on the one-stage instance segmentation algorithm SparseInst is conducted to enhance the model's utilization of effective features. Specifically, firstly, residual connection is added inside the basic building block of the backbone. Secondly, a three-scale feature fusion module is designed to overcome the problem of indirect interaction of cross-scale features in the encoder. A decoupled instance activation module is designed to enhance the model′s ability to learn instance features. In addition, the improved algorithm makes full use of detail features to refine the mask features to improve the quality of the generated masks. Finally, the kernel is used to initialize the score of the target object, which improves the utilization rate of the extracted features. The improved algorithm surpasses similar algorithms in mask accuracy on multiple datasets and has strong real-time performance. To further verify the effectiveness of the improved algorithm, experiments using data

*国家自然科学基金(52225212, U20A20333, U20A20331, 52072160)资助。
原稿收到日期为 2024 年 07 月 14 日,修改稿收到日期为 2024 年 10 月 06 日。
通信作者:王海,教授,博士,Email:wanghai1019@163.com。

collected from a real vehicle platform are conducted. When the input image resolution is 640×480, the model inference speed reaches 54 FPS, and the instance mask is segmented accurately.

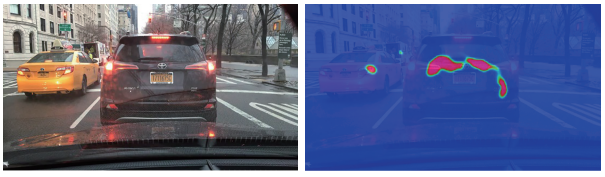
Keywords: autonomous driving; deep learning; real-time instance segmentation; feature utilization

前言

在复杂交通场景下,环境感知系统可以帮助智能汽车获取重要的交通要素信息[1-2],对智能汽车的行驶安全至关重要。视觉相机的成像更贴近于人类驾驶员的视觉,且视觉相机制造成本相对较低,利于在智能汽车上的部署。近年来,深度学习的飞速发展促使基于视觉的感知算法成为了研究热点。在主流的感知算法中,实例分割算法能在区分不同实例物体的基础上逐像素精确预测出对应物体的掩码,有着高效的特征利用方式,可为智能汽车提供精确的感知信息。

主流的基于深度学习的实例分割算法可以分为二阶段、一阶段与基于查询的3种方法[3]。其中,二阶段算法通常是在目标检测算法的基础上添加了一条语义分支。Mask R-CNN[4]是最具代表性的二阶段算法,后续的RefineMask[5]与PatchDCT[6]都致力于提升掩码的质量。二阶段的算法虽然能生成高细粒度的掩码,但因为依赖于目标检测算法而导致整体结构冗余,分割效率较低。为此,许多研究者设计了更高效的一阶段算法[7-10],此类算法通常用特殊方法定位和区分实例物体,摆脱了对目标检测算法的依赖。如SOLO[7-8]利用物体的中心位置来对实例进行定位和区分。CondInst[10]则利用动态卷积[11]区分实例目标。虽然一阶段算法分割效率较高,但是其生成的掩码质量不如二阶段算法。除此之外,基于查询的算法如Mask2Former[12]和OneFormer[13]取得了较好的分割质量,但此类方法大量运用了注意力模块[14],运算量巨大,无法实现实时运算。

由于二阶段与基于查询的算法难以满足自动驾驶要求的实时性,本文选取一阶段算法进行研究。而目前大多数一阶段算法倾向于用物体的中心像素来表示和定位实例。但雨天、黄昏、黑夜等视线不佳的场景会导致物体中心特征的模糊而难以识别。相比之下,SparseInst[15]则采用稀疏实例激活图来表示实例,对于物体的表示与定位方式更加灵活。图1(b)为实例激活图的可视化,其高亮之处代表该处有实例物体存在。且SparseInst无须非极大值抑制操作,推理速度较快。因此,本文选择SparseInst作为研究对象。



(a) 原图 (b) 实例激活图

图1 实例激活图可视化

SparseInst虽然有较快的推理速度,但模型精度有待提升。为在有限算力下实现高效地分割,在设计算法时应充分考虑提升特征的利用率,避免已被提取的特征因未得到充分利用而造成计算资源浪费。除此之外,由于遮挡、光影变化、天气变化与形状变化等导致的物体特征不显著问题会造成重要特征的遗漏。而通过提升模型的特征利用率可以使模型发现微小或不清晰特征,并对其进一步提取,使模型在道路场景下具有更强的鲁棒性。

首先,本文针对原主干网络未能对基础构建块内部特征进行有效保留这一问题,在基础构建块内部增加了残差连接。其次,为使模型能更好学习不同尺寸交通物体的特征,设计了三尺度特征融合模块,并以此解决了原编码器中跨尺度特征图不能进行直接交互的问题。由于道路场景中实例数量众多,本文对实例激活模块进行了解耦,增强实例特征学习能力的同时抑制了额外的噪声干扰。除此之外,为改善掩码质量,本文利用细节特征对掩码进行精修。最后,针对原结构中直接粗暴地用目标物体得分与掩码交并比计算交叉熵损失这一问题,本文用内核去初始化目标物体得分,实现了更好的训练结果。用改进的算法在多个数据集上进行了实验,探究并验证所改进部分的有效性,并利用实车收集的真实道路场景数据进行了实验。

1 实时实例分割算法设计

1.1 框架概览

改进算法的主要结构包括:主干网络、编码器与解码器。其整体结构如图2所示。假设输入图片的尺寸为$H \times W$,H与W分别为图片的高与宽,经过主干网络后分别获取了特征图$F_{1/4}$、$F_{1/8}$、$F_{1/16}$以及$F_{1/32}$,

分别对应的尺寸为 $\frac{H}{4}\times\frac{W}{4}$、$\frac{H}{8}\times\frac{W}{8}$、$\frac{H}{16}\times\frac{W}{16}$ 以及 $\frac{H}{32}\times\frac{W}{32}$。随后,编码器会对多尺度特征进行融合,并由解码器生成掩码特征与实例特征,掩码特征与实例特征相交互生成最终的实例掩码。具体来说,在主干网络部分,本文在基础构建块的内部增加了残差连接。其次,本文设计了三尺度特征融合模块,增强了多尺度特征的融合。还设计了解耦的实例激活模块以增强模型区分和定位实例的能力。为提高掩码的质量,设计了细节特征补充分支以及掩码细节特征修正模块。最后,本文用内核初始化目标物体得分,提升了训练效果。

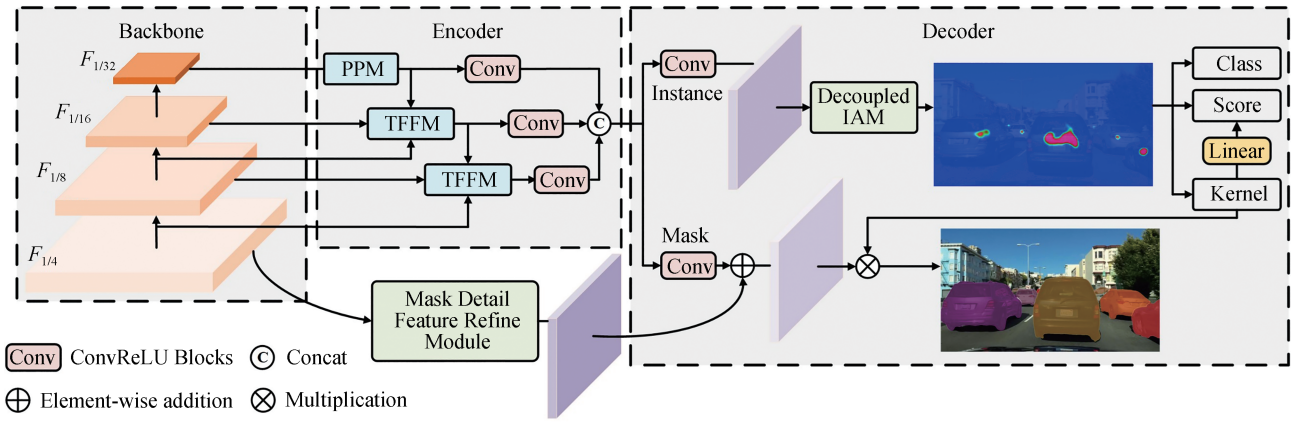


图 2 算法整体结构图

1.2 主干网络

作为模型中参数最大的部分,主干网络的优劣将直接影响后续编码器与解码器对特征的利用效率。为了使改进的算法实现最优的性能,本文以原 SparseInst 算法为基准对主流的主干网络在 BDD100K[16]上进行了筛选实验,包括 ResNet50[17]、CSPDarknet53[18]以及 PVTv2-B1[19]。实验结果如表 1 所示,虽然 PVTv2-B1 实现了最高的精度,但是推理速度较慢。CSPDarknet53 与 ResNet50 相比只提升了 0.2 mAP 的精度,却损失了 2.1 FPS 的速度,未实现速度与精度的均衡。因此,本文选取 ResNet50 作为主干网络。

表 1 不同主干网络在 SparseInst 算法上的实验结果

Backbone	mAP(Mask)	FPS
ResNet50	19.3	34.6
CSPDarknet53	19.5	32.5
PVTv2-B1	20.8	20.1

在 ResNet50 中,基础构建块之间由残差结构进行连接,以避免每个基础构建块中的有效特征因模型层数过深而丢失。但是残差连接只存在于不同基础构建块之间,而在每个基础构建块中还有 3 个卷积层,原结构只有助于不同基础构建块之间的特征保留,忽略了内部特征的保留。因此,为了避免内部有效特征的丢失,本文在基础构建块的内部增加了残差连接,改进的结构如图 3 所示。

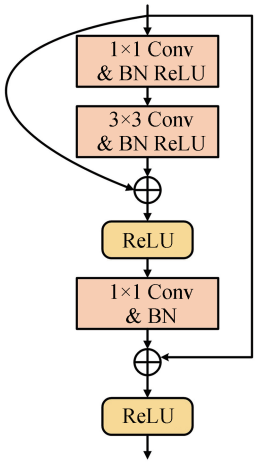


图 3 改进的基础构建块

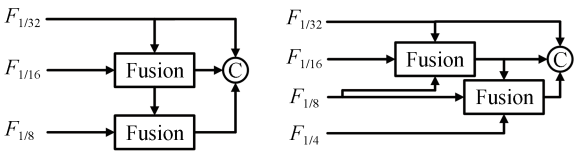
1.3 编码器

编码器对主干网络输出的不同尺寸的特征图进行融合,促使多尺度特征的交互。尤其在物体尺寸各异且与摄像头距离不断变化的道路场景下,充分的多尺度特征融合可以提供丰富的多尺度语义信息。图 4(a)为原编码器的大体结构,该结构在前段融合的过程中忽略了跨尺度特征图之间的直接交互,而仅仅只有相邻尺度特征图之间的交互,其跨尺度的特征图交互只存在于后段的拼接融合中。受到 Gold-YOLO[20]中跨尺度特征融合方式的启发,也为

充分利用主干网络提取出的多尺度特征,本文对编码器进行了改进。改进后的大体结构如图4(b)所示,与图4(a)相比,除了将$F_{1/32}$、$F_{1/16}$以及$F_{1/8}$作为输入之外,改进结构还将$F_{1/4}$作为输入。与其他特征图相比,$F_{1/4}$有更大的分辨率,包含更丰富的细节信息。且为在前段融合过程中实现3种不同尺度特征图的融合,本文专门设计了三尺度特征融合模块(three-scale feature fusion module, TFFM),使得跨尺度的特征图之间能进行直接交互,其结构如图5所示。在输入该模块的3种不同尺寸的特征图中,最小尺寸的特征图有最大的感受野,其对物体位置信息更加敏感,因此将其作为其他特征图的位置引导特征。此外,为进一步发挥最小尺寸特征图的位置信息引导作用,本文在位置注意力机制[21](coordinate attention, CA)的基础上设计了经大核卷积增强的位置注意力机制(large kernel convolution enhanced coordinate attention, LKCA),其结构如图5所示。大核卷积相较于小核卷积有更大的感受野,而更大的感受野则可以捕获更多的全局特征,利于精确位置信息的获取。但大核卷积计算消耗较大,为此本文采用了轻量化的深度可分离卷积(depth-wise separable convolutions)去构造大核卷积。在LKCA模块中,输入特征经过5×5的深度可分离卷积和ReLU后得到了有较大感受野的特征,并分别经过水平方向的平均池化(X AvgPooling)与垂直方向的平均池化(Y AvgPooling)获取了位置编码特征。同时,在平行的另外两条分支上,输入特征直接经过水平和垂直方向的平均池化后与已获取的位置编码特征逐元素相加,以此获取增强的位置编码特征。

1.4　解码器

解码器对编码器输出的特征进行解码以得到最终的预测结果。具体来说,该解码器包含实例分支与掩码分支,其中实例分支用以对实例物体进行定位和分类,而掩码分支用以学习掩码特征,实例分支与掩码分支相交互得到最终的实例掩码。



(a) 原编码器的大体结构　　(b) 改进编码器的大体结构

图4　编码器大体结构对比图

在原实例分支中,实例激活模块(instance activation module,IAM)用以生成实例特征。而实例特征是由实例激活图与传入实例分支的特征矩阵相乘得到的。其中,实例激活图本质上为权重图,是由传入实例分支的特征经过Sigmoid后得到的,所有权重值被映射为0~1之间的概率值。在得到实例特征后,分别由3个线性层输出实例物体的种类、目标物体得分以及内核。内核与掩码特征经矩阵相乘生成最终的实例掩码。如图6(a)所示,原实例激活模块用一个3×3的卷积来提取所有实例激活图的特征,但是在拥堵的交通场景下,实例物体众多,仅仅用一个卷积去提取所有的实例特征效率欠佳。为此,本文提出了解耦的实例激活模块,其结构如图6(b)所示。该模块增加了一条平行的激活分支来学习实例激活图的特征,且为使所提取的特征有更丰富的尺度信息,该分支采用了5×5的大核卷积,两个分支的权重图逐元素相乘得到最终的权重图。在解耦的实例激活模块中,假设输入特征为X_D,输出的权重图为O_D,其产生权重图的过程为

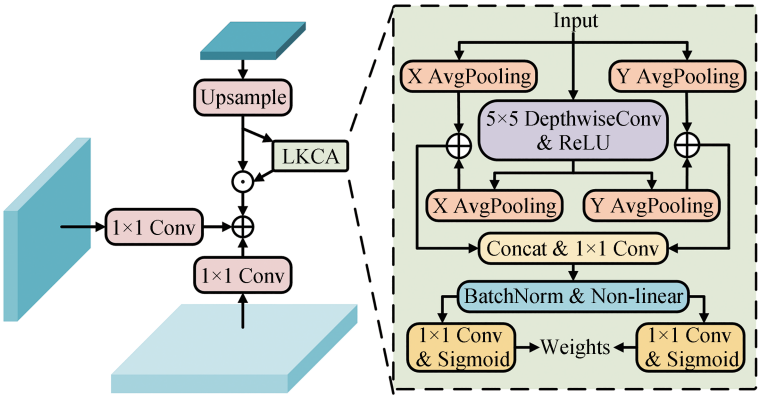
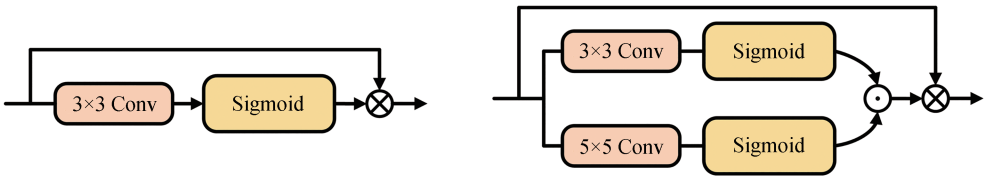


图5　三尺度特征融合模块结构图



(a) 原实例激活模块 (b) 解耦的实例激活模块

图6 实例激活模块结构对比图

$$O_D = Sigmoid(Conv_{3\times3}(X_D))\odot Sigmoid(Conv_{5\times5}(X_D)) \quad (1)$$

O_D中的每个元素$o_d \in (0,1)$。由于改进模块的权重图是由两个元素值为0~1之间的矩阵逐元素相乘得到的,而0~1之间的两个数值相乘会得到相比原数值更小的数值,因此改进模块相比原模块不仅有更精细的特征学习能力,还利于噪声抑制。在训练时,为平衡分类与分割任务,将实例分支输出的目标物体得分与生成掩码的交并比用于计算交叉熵损失,而掩码则是通过内核与掩码特征交互生成的,但目标物体得分与内核分别由两个不同的线性层输出,计算该损失的方式较为粗暴,不利于模型对特征的学习。为此,本文将变换维度的内核与目标物体得分逐元素相加,以此初始化目标物体得分,增强不同特征的内在联系,提高模型对特征的利用率。

掩码分支的输入特征在编码器中经过了多次下采样和上采样,虽然这样有利于学习多尺度特征和增大感受野,但是会造成部分细节信息的丢失,不利于对细粒度要求较高的分割任务。受到Refinemask[5]中用语义分支不断对掩码进行修正的启发,也为了充分利用主干网络提取的特征,本文将主干网络输出的$F_{1/4}$作为掩码分支的细节特征补充分支,为掩码分支提供精细的掩码特征。与$F_{1/32}$、$F_{1/16}$和$F_{1/8}$相比,$F_{1/4}$有着更大的分辨率,从而有更丰富的细节特征。但是,将未经编码的特征与原掩码特征直接融合,可能会造成特征对不齐的问题,从而导致精度下降。为此,本文设计了掩码细节特征修正模块(mask detail feature refine module, MDRM),充分利用主干网络提取出的细节特征的同时,柔和地实现了精修掩码特征与原掩码特征的融合,其结构如图7所示。该模块包含两条分支,其中细节特征增强分支(detail feature enhancement branch)用来提取高细粒度的细节特征,鉴于中心差分卷积[22](central difference convolution, CDConv)能够通过聚合强度和梯度信息来发掘深层次的细节特征,该分支用中心差分卷积来去除噪声,并提取主要的细节特征以进行精修。但中心差分卷积对于细节特征极度敏感,且输入特征未经编码器处理,导致对形状变化、大小变化及位置变化的鲁棒性较差。因此另一条位置引导分支(location-guided branch)用于提取物体的外形特征和位置特征,用以弥补掩码细节特征修正模块对全局信息的缺失。可变性卷积[23](deformable convolutional, DConv)可产生发散效果,能更好捕捉物体形状与位置的变化,由此该分支用可变形卷积提取的特征去引导细节特征增强分支提取恰当位置的细节特征,增强细节特征增强分支对全局信息的敏感性。位置引导分支对细节特征增强分支的引导作用是通过逐元素相乘实现的,且是二阶段的修正过程。

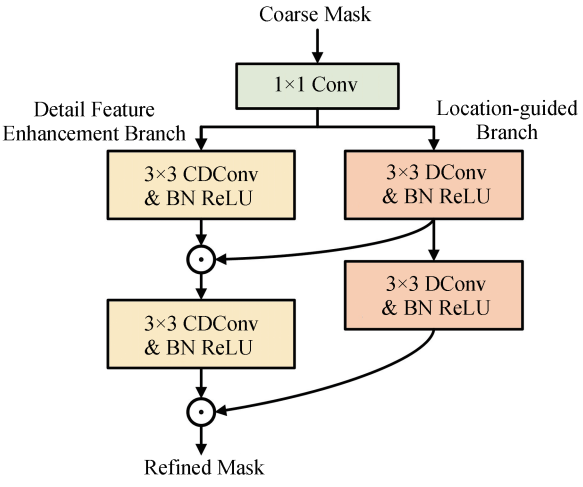


图7 掩码细节特征修正模块结构图

1.5 标签分配策略与损失函数

由于本算法直接输出固定大小的预测结果的集合,因此在分配真值标签时采用了二值匹配[24]的方式。

训练损失包括分类损失L_{class}、掩码损失L_{mask}以及目标物体得分损失L_{score}。L_{class}采用了focal loss[25]来对目标进行分类,L_{mask}用以实现精确的掩码特征

学习,其为 dice loss[26]与交叉熵损失的加权之和,L_{mask}的表达式为

$$L_{mask} = \delta_d L_{dice} + \delta_c L_{cross} \tag{2}$$

式中δ_d与δ_c分别为L_{dice}与L_{cross}的加权系数。L_{score}采用了交叉熵损失来缓解分类与掩码预测任务之间的不平衡性。

总的损失L为L_{class}、L_{mask}与L_{score}三者的加权之和,其公式为

$$L = \lambda_c L_{class} + L_{mask} + \lambda_s L_{score} \tag{3}$$

式中λ_c与λ_s分别为L_{class}与L_{score}的加权系数。

2　实验验证

2.1　数据集介绍

BDD100K[16]数据集包含100k张分辨率为1280×720的道路场景图片。该数据集有10k张图片应用于实例分割任务。为进一步测试模型的鲁棒性,本文用图像生成模型 TPSeNCE[27]对 BDD100K 中的图像进行处理,通过生成黑夜、雨天以及模糊场景来提升数据集的识别难度,新生成的数据集命名为 R-BDD(robust BDD),其样本如图8所示。除此以外,nuImages 是 nuScenes[28]数据集中用于2D实例分割的数据集,其包含93k张分辨率为1600×900的图片。Waymo[29]数据集有1 150个不同的道路场景,其包含分辨率为1920×1280和1920×886的高质量图像。

2.2　实施细节

模型训练及精度分析都是在两张 NVIDIA RTX 3090显卡上完成的,而模型推理速度测试是在一张显卡上完成的。在训练时,用 AdamW[30]作为优化器,初始的学习速率设置为5×10^{-5},权重衰减系数设置为0.05,训练时的 batch size 设置为2。为获取更高的模型精度,采用迁移训练的训练方式,即将在 COCO[31]数据集上训练好的 SparseInst 模型作为预训练权重模型。在式(2)中权重系数δ_d与δ_c都设置为2,而式(3)中权重系数λ_c与λ_s分别设置为2与1。

2.3　实验结果

本文使用 mAP(mean average precision)作为掩码精度的评价指标,并选择 FPS(frames per second)作为模型推理速度的评价指标。

首先,本文将 Yolact[9]、Centermask[32]、CondInst[10]、SOLOv2[8]、FastInst[33]、RTMDet-Ins[34]、Boxsnake[35]、Box2mask[36]、YOLOv8-seg[37]、YOLOv11-seg[38]以及基线 SparseInst[15](baseline)与改进算法进行对比,分别在 BDD100K、R-BDD、nuImages 与 Waymo 上进行了实验。由于 Waymo 缺乏实例分割的标签,无法进行训练与定量验证,因此本文用在 BDD100K 上训练所得的模型在 Waymo 上进行了可视化与速度实验。实验结果如表2所示。根据实验结果分析,在 BDD100K、R-BDD 与 nuImages 上,改进的算法均获得了最高的掩码精度。为验证改进算法的鲁棒性,将 BDD100K 与 R-BDD 上的实验结果进行比较,由于 R-BDD 相较于 BDD100K 中的图像有更多的噪声干扰,模型精度都会有一定程度的降低,改进模型精度降低了2.9 mAP,降低的

(a) 原图

(b) 生成图像

图8　R-BDD 图片样例

表2 改进算法与其他实例分割算法实验结果对比

算法	BDD100K		R-BDD		nuImages		Waymo
	mAP(Mask)	FPS	mAP(Mask)	FPS	mAP(Mask)	FPS	FPS
Yolact	19.5	26.0	14.1	26.0	34.7	23.7	19.2
Centermask	19.4	23.9	13.8	23.9	34.9	21.7	16.6
CondInst	19.9	22.1	14.8	22.1	34.4	18.8	14.1
SOLOv2	20.3	23.8	16.4	23.8	35.2	21.2	15.9
RTMDet-Ins	21.7	24.2	18.0	24.2	35.6	22.3	17.3
Boxsnake	18.3	14.1	13.9	14.1	32.1	11.9	6.1
YOLOv8-seg	17.7	30.0	14.5	30.1	35.3	28.0	22.1
FastInst	21.6	25.7	18.6	25.5	36.7	23.2	19.0
Box2mask	20.2	16.6	14.2	16.6	34.8	13.1	7.2
YOLOv11-seg	19.8	33.0	15.8	33.0	35.7	29.6	23.7
Baseline	19.3	34.6	16.0	34.5	35.1	31.6	25.4
Ours	22.4	31.0	19.5	31.0	37.7	28.8	22.5

精度最少,有较强的鲁棒性。而FastInst精度降低了3 mAP,该模型也具有较强的鲁棒性,但该模型在总体精度与推理速度上不及改进模型。在分辨率为1280×720时,改进模型的推理速度达到了31.0 FPS,且即使在最大分辨率为1920×1280的Waymo上,也能达到22.5 FPS,有较强的实时性。改进模型在BDD100K、R-BDD、nuImages与Waymo上的可视化结果分别如图9~图12所示。在多种场景下,改进模型成功地将不同的实例对象定位并区分出来,且较为精确地预测出了每个实例物体的掩码。

其次,本文对基础构建块中残差连接位置对模型性能的影响进行了探究。如图13所示,Case Ⅰ为改进主干所采取的残差连接方式,Case Ⅱ与Case Ⅰ选用了不同的连接位置,而Case Ⅲ则是仅隔一层卷积进行残差连接。实验结果如表3所示,跨一层卷积进行残差连接不仅不会提升精度,还会造成精度的降低,且残差连接位置的选择至关重要,在基础构建块的第1个卷积前和第2个卷积后添加残差连接会取得较好的效果。

本文还探究了解耦的实例激活模块中解耦分支的数量以及解耦分支中卷积核的大小对模型精度的影响。实验结果如表4所示,与原结构中未解耦的方式相比,解耦的方式提升了模型精度。但3条解耦分支对精度的提升不及2条解耦分支,且在2条解耦分支中,不同的分支分别采用3×3的卷积与5×5的大核卷积会带来更显著的精度提升。这是由于不同卷积核大小的卷积促进了多尺度特征的提取,且大核卷积在一定程度上扩充了感受野,尤其对实例识别与定位会产生显著的提升。

2.4 消融实验

为验证各改进部分的有效性,本文对改进算法进行了消融实验。实验结果如表5所示,改进的主干网络在没有额外堆叠更多参数的情况下,使模型的精度提升了0.3 mAP。而TFFM、解耦的实例激活模块与掩码细节特征修正模块分别使模型精度提升了0.4 mAP、1.0 mAP与0.6 mAP。用内核初始化目标物体得分使模型精度提升了0.8 mAP,而速度仅损失了0.3 FPS,该改进对特征的有效利用率较高。

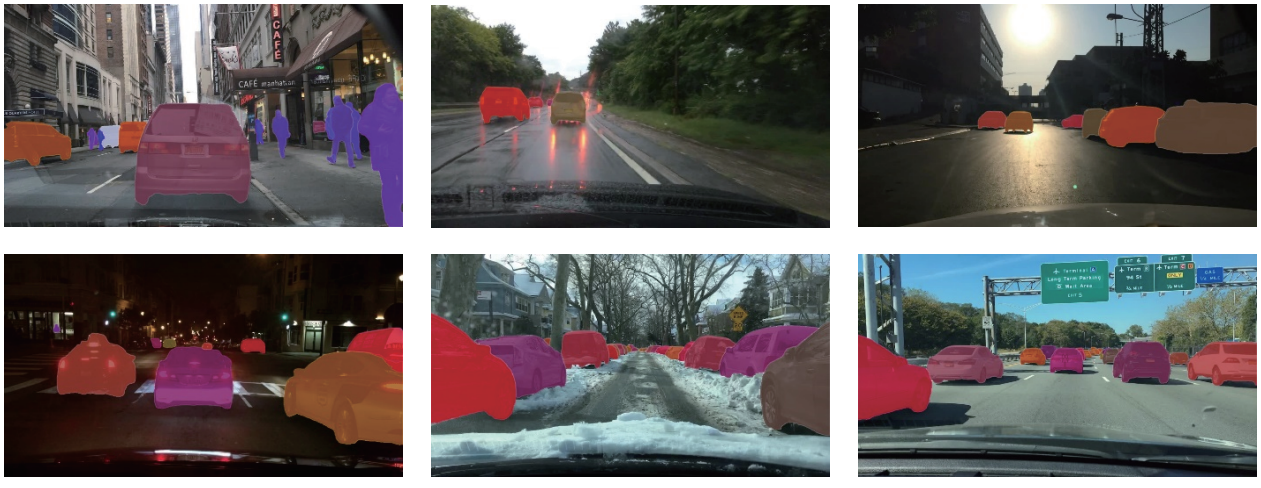


图9 改进算法在BDD100K上的可视化结果

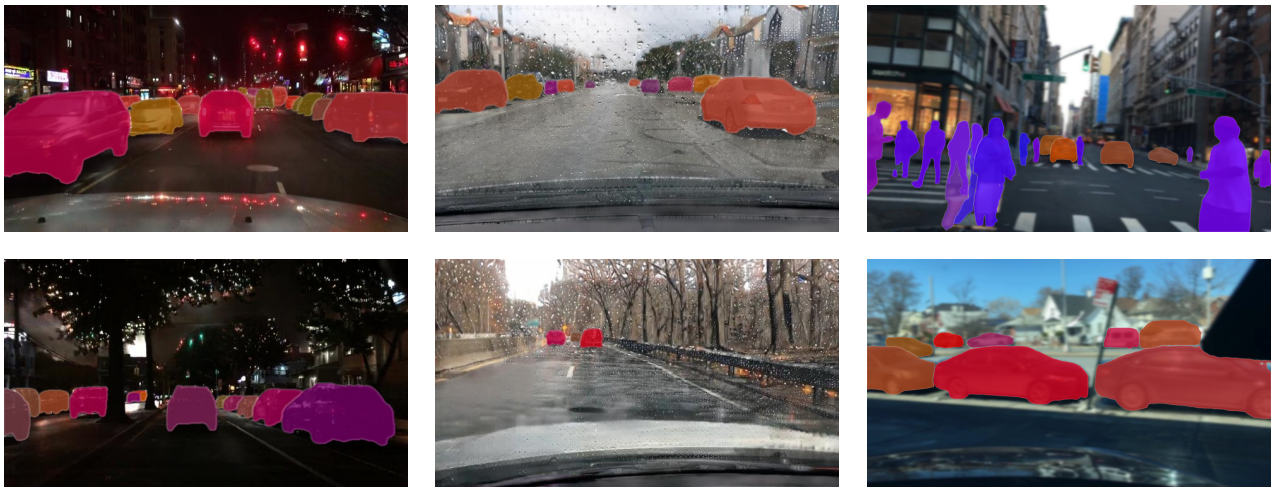


图10　改进算法在R-BDD上的可视化结果

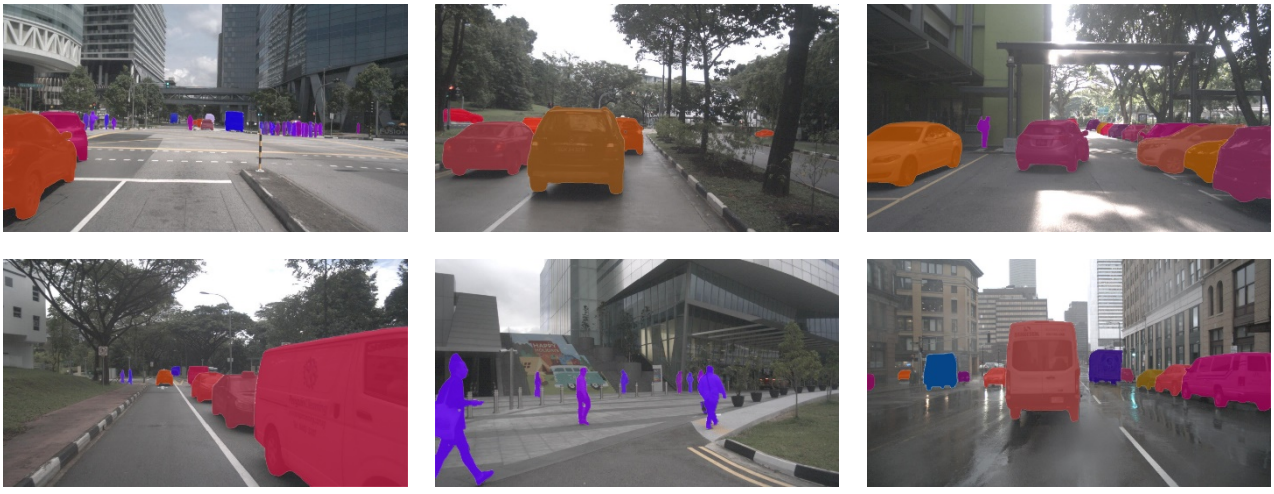


图11　改进算法在nuImages上的可视化结果

图12　改进算法在Waymo上的可视化结果

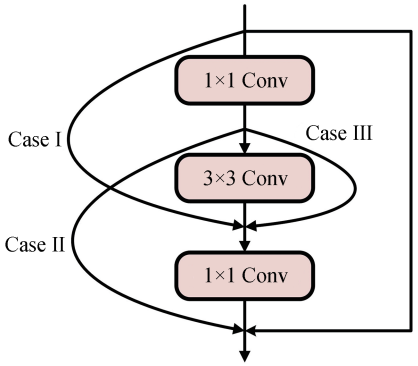


图13 基础构建块内不同的残差连接方式

表3 基础构建块内不同的残差连接方式对模型精度的影响

连接方法	mAP(Mask)
Baseline	19.3
Case Ⅰ	19.6
Case Ⅱ	18.7
Case Ⅲ	18.5

表4 不同的解耦方式对模型精度的影响

Branch Ⅰ	Branch Ⅱ	Branch Ⅲ	mAP(Mask)
3×3 Conv			19.3
3×3 Conv	3×3 Conv		19.9
3×3 Conv	5×5 Conv		20.3
1×1 Conv	3×3 Conv	5×5 Conv	19.9

表5 改进算法在BDD100K验证集上的消融实验结果

Improved ResNet50	TFFM	Decouple IAM	MDRM	Initializing score with kernel	mAP (Mask)	FPS
					19.3	34.6
√					19.6	34.5
√	√				20.0	33.2
√	√	√			21.0	32.7
√	√	√	√		21.6	31.3
√	√	√	√	√	22.4	31.0

2.5 实际场景实验结果

为进一步验证改进算法的性能,本文用改进算法在BDD100K上训练所得的模型对真实道路场景进行了预测。如图14所示的艾瑞泽5E智能汽车采集了7 328张分辨率为640×480的图片,包括城市、高速及高架3种不同的地点。可视化结果如图15所示,改进模型均较好地识别出每一个实例目标,且能较为准确地分割出掩码,测试时的速度达到了54 FPS。

图14 艾瑞泽5E实验车

3 结论

本文在SparseInst基础上进行了改进。首先,为避免模型层数过深导致的特征丢失,在主干网络基础构建块中增加了残差连接。其次,提出了三尺度特征融合模块克服了原先跨尺度特征不能进行直接交互的问题。为增强模型对实例特征的学习能力并抑制噪声干扰,设计了解耦的实例激活模块。除此之外,充分利用主干网络提取出的细节特征对掩码特征进行修正以提高生成掩码的质量。最后,用内核去初始化目标物体得分,使模型能够获取更好的训练效果,提高了特征的利用率。

在多个数据集上,改进模型的精度均超越了其他实例分割算法,同时具有较强的实时性。在自主搭建的实车平台收集的数据上也取得了精细的分割效果。但是,由于缺乏浓雾天、大暴雨这类极端恶劣天气的数据,改进算法对这类场景的分割效果还有待验证,后续会利用先进的图像生成技术,生成相关极端恶劣天气的数据集,进行进一步的验证并提升算法对极端恶劣天气的鲁棒性。

参考文献

[1] 王海, 李洋, 蔡英凤, 等. 基于激光雷达的3D实时车辆跟踪[J]. 汽车工程, 2021, 43(7): 1013-1021.
WANG Hai, LI Yang, CAI Yingfeng, et al. 3D real-time vehicle tracking based on lidar [J]. Automotive Engineering, 2021, 43(7): 1013-1021.

[2] 武志斐, 李守彪. 基于实例分割的车道线检测算法[J]. 汽车工程, 2023, 45(2): 263-272.
WU Zhifei, LI Shoubiao. Lane detection algorithm based on instance segmentation [J]. Automotive Engineering, 2023, 45(2): 263-272.

图15 改进算法对真实道路场景的可视化结果

[3] 陈妍妍, 王海, 蔡英凤, 等. 基于检测的高效自动驾驶实例分割方法[J]. 汽车工程, 2023, 45(4): 541-550.
CHEN Yanyan, WANG Hai, CAI Yingfeng, et al. Efficient automatic driving instance segmentation method based on detection [J]. Automotive Engineering, 2023, 45(4): 541-550.

[4] HE K, GKIOXARI G, DOLLáR P, et al. Mask R-CNN[C]. Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, F, 2017.

[5] ZHANG G, LU X, TAN J, et al. RefineMask: towards high-quality instance segmentation with fine-grained features[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2021.

[6] WEN Q, YANG J, YANG X, et al. PatchDCT: patch refinement for high quality instance segmentation[C]. Proceedings of the The Eleventh International Conference on Learning Representations, F, 2022.

[7] WANG X, KONG T, SHEN C, et al. SOLO: segmenting objects by locations [C]. Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII 16, F, 2020. Springer.

[8] WANG X, ZHANG R, KONG T, et al. SOLOv2: dynamic and fast instance segmentation [J]. Advances in Neural Information Processing Systems, 2020, 33: 17721-17732.

[9] BOLYA D, ZHOU C, XIAO F, et al. Yolact: real-time instance segmentation [C]. Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2019.

[10] TIAN Z, SHEN C, CHEN H. Conditional convolutions for instance segmentation[C]. Proceedings of the Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I 16, F, 2020. Springer.

[11] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: attention over convolution kernels[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[12] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022.

[13] JAIN J, LI J, CHIU M T, et al. OneFormer: one transformer to rule universal image segmentation [C]. Proceedings of the Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2023.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30.

[15] CHENG T, WANG X, CHEN S, et al. Sparse instance activation for real-time instance segmentation[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022.

[16] YU F, CHEN H, WANG X, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, 2016.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection [J]. arXiv preprint arXiv:200410934, 2020.

[19] WANG W, XIE E, LI X, et al. PVT v2: improved baselines with pyramid vision transformer [J]. Computational Visual Media, 2022, 8(3): 415-424.

[20] WANG C, HE W, NIE Y, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism [J]. Advances in Neural Information Processing Systems, 2024, 36.

[21] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2021.

[22] YU Z, ZHAO C, WANG Z, et al. Searching central difference convolutional networks for face anti-spoofing[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[23] ZHU X, HU H, LIN S, et al. Deformable convnets v2: more deformable, better results[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2019.

[24] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. Proceedings of the European Conference on Computer Vision, F, 2020. Springer.

[25] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, F, 2017.

[26] MILLETARI F, NAVAB N, AHMADI S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]. Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), F, 2016. IEEE.

[27] ZHENG S, LU C, NARASIMHAN S G. TPSeNCE: towards artifact-free realistic rain generation for deraining and object detection in rain[C]. Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, F, 2024.

[28] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[29] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[30] KINGMA D P, BA J. Adam: a method for stochastic optimization [J]. arXiv preprint arXiv:14126980, 2014.

[31] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]. Proceedings of the Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, F, 2014. Springer.

[32] LEE Y, PARK J. CenterMask: real-time anchor-free instance segmentation[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020.

[33] HE J, LI P, GENG Y, et al. Fastinst: a simple query-based model for real-time instance segmentation[C]. Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2023.

[34] LYU C, ZHANG W, HUANG H, et al. RTMDet: an empirical study of designing real-time object detectors [J]. arXiv preprint arXiv:221207784, 2022.

[35] YANG R, SONG L, GE Y, et al. BoxSnake: polygonal instance segmentation with box supervision[C]. Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, F, 2023.

[36] LI W, LIU W, ZHU J, et al. Box2Mask: box-supervised instance segmentation via level-set evolution [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[37] JOCHER G, CHAURASIA A, QIU J. Ultralytics YOLOv8 [Z]. 2023.

[38] JOCHER G, QIU J. Ultralytics YOLO11 [Z]. 2024.