

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2023)02-0495-15

论文引用格式: Wan X J, Zhou Y Y, Shen M F, Zhou T and Hu F Y. 2023. Multi-scale context information fusion for instance segmentation. Journal of Image and Graphics, 28(02): 0495-0509 (万新军, 周逸云, 沈鸣飞, 周涛, 胡伏原. 2023. 融合多尺度上下文信息的实例分割. 中国图象图形学报, 28(02): 0495-0509) [DOI: 10. 11834/jig. 211090]

融合多尺度上下文信息的实例分割

万新军^{1, 2}, 周逸云^{1, 2}, 沈鸣飞¹, 周涛³, 胡伏原^{1, 2*}

1. 苏州科技大学电子与信息工程学院, 苏州 215009; 2. 苏州市虚拟现实智能交互及应用技术重点实验室, 苏州 215009; 3. 北方民族大学计算机科学与工程学院, 银川 750021

摘要: 实例分割通过像素级实例掩膜对图像中不同目标进行分类和定位。然而不同目标在图像中往往存在尺度差异, 目标多尺度变化容易错检和漏检, 导致实例分割精度提高受限。现有方法主要通过特征金字塔网络 (feature pyramid network, FPN) 提取多尺度信息, 但是 FPN 采用插值和元素相加进行邻层特征融合的方式未能充分挖掘不同尺度特征的语义信息。因此, 本文在 Mask R-CNN (mask region-based convolutional neural network) 的基础上, 提出注意力引导的特征金字塔网络, 并充分融合多尺度上下文信息进行实例分割。**方法** 首先, 设计邻层特征自适应融合模块优化 FPN 邻层特征融合, 通过内容感知重组对特征上采样, 并在融合相邻特征前引入通道注意力机制对通道加权增强语义一致性, 缓解邻层不同尺度目标间的语义混叠; 其次, 利用多尺度通道注意力设计注意力特征融合模块和全局上下文模块, 对感兴趣区域 (region of interest, RoI) 特征和多尺度上下文信息进行融合, 增强分类回归和掩膜预测分支的多尺度特征表示, 进而提高对不同尺度目标的掩膜预测质量。**结果** 在 MS COCO 2017 (Microsoft common objects in context 2017) 和 Cityscapes 数据集上进行综合实验。在 MS COCO 2017 数据集上, 本文算法相较于 Mask R-CNN 在主干网络为 ResNet50/101 时分别提高了 1.7% 和 2.5%; 在 Cityscapes 数据集上, 以 ResNet50 为主干网络, 在验证集和测试集上进行评估, 比 Mask R-CNN 分别提高了 2.1% 和 2.3%。可视化结果显示, 所提方法对不同尺度目标定位更精准, 在相互遮挡和不同目标分界处的分割效果显著改善。**结论** 本文算法有效提高了网络对不同尺度目标检测和分割的准确率。

关键词: 实例分割; Mask R-CNN; 特征金字塔网络 (FPN); 多尺度上下文信息; 多尺度通道注意力 (MSCA)

Multi-scale context information fusion for instance segmentation

Wan Xinjun^{1, 2}, Zhou Yiyun^{1, 2}, Shen Mingfei¹, Zhou Tao³, Hu Fuyuan^{1, 2*}

1. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China;
2. Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, Suzhou 215009, China;
3. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China

Abstract: **Objective** Case-relevant segmentation is one of the essential tasks for image and video scene recognition. Its precise segmentation is widely used in real scenes like automatic driving, medical image profiling, and video surveillance. To classify and locate multiple targets of image, this kind of segmentation can be used for pixel-level case-related masks. However, different targets interpretation often has featured of multiple scales. For larger-scale targets, the receptive field can be

收稿日期: 2021-11-26; 修回日期: 2022-01-10; 预印本日期: 2022-01-17

* 通信作者: 胡伏原 fuyuanhu@usts.edu.cn

基金项目: 国家自然科学基金项目 (61876121); 江苏省重点研发计划项目 (BE2017663); 江苏省教育厅高等学校自然科学研究面上项目 (19KJB520054)

Supported by: National Natural Science Foundation of China (61876121)

covered its local area only, which is possible to get detection error or insufficient or inaccurate segmentation. For smaller-scale targets, the receptive field is often affected by much more background noise, and it is easy to be misjudged as the background category and lead to detection error. The recognition and segmentation accuracy are lower at the target boundary and occlusion. To enhance the segmentation accuracy effectively, most of case-relevant segmentation methods are improved in consistency without a multi-scale targets-oriented solution. To optimize segmentation accuracy further, we develop a mask region-based convolutional neural network (Mask R-CNN) based case-relevant segmentation network in terms of the improved feature pyramid network (FPN) and multi-scale information. **Method** First, an attention-guided feature pyramid network (AgFPN) is illustrated, which optimizes the fusion method of FPN adjacent layer features through an adaptive adjacent layer feature fusion module (AFAFM). To learn multi-scale features effectively, the AgFPN is based on content-oriented reconstruction for features-upsampled and a channel attention mechanism is used to weight channels before adjacent layer feature fusion. Then, we design an attention feature fusion module (AFFM) and a global context module (GCM) in relation to multi-scale channel attention. We enhance the multi-scale feature representation of the mask prediction branch and the classification and regression branch for region of interest (RoI) features via multi-scale contextual information-adding. Hence, our analysis can improve the quality of mask prediction for multi-scale objects. First, we utilize AgFPN to extract multi-scale features. Next, multi-scale context information extraction and fusion are carried out in the network. The inner region proposal network (RPN) can be used to develop the bounding boxes of target regions and filters. Meanwhile, multi-scale context information is derived from the output of AgFPN in accordance with AFFM and GCM. Then, to obtain a fixed-size feature map, the network-based RoIAlign algorithm can be used to map the RoI to the feature map, which is fused with the following multi-scale context information. Finally, the bounding box regression and mask prediction are performed in terms of features-fused. We use the deep learning framework PyTorch to implement the algorithm proposed. The experimental facility is equipped with the Ubuntu 16.04 operating system, and a sum of 4 NVIDIA 1080Ti graphics processing units (GPUs) are used to accelerate the operation. The ResNet-50/101 network is used as the backbone network and the pre-trained weights on ImageNet are utilized to initialize the network parameters. For the Microsoft common objects in context 2017 (MS COCO 2017) dataset, we use stochastic gradient descent (SGD) for 160 000 iterations of training optimization. The initial learning rate is 0.002 and the batch size is set to 4. When the number of iterations is 130 000 and 150 000, the learning rates can be reached to 10 times lower. For the Cityscapes dataset, we set the batch size to 4 and the initial learning rate to 0.005. The number of iterations is 48 000. When it reaches 36 000, the learning rate can be down to 0.000 5. The weight decay coefficient is set to 0.000 5 and the momentum coefficient is configured to 0.9. The loss function and hyperparameters-related are set and initialized of the strategy-described following. **Result** Our method effectiveness is evaluated through comprehensive experiments on the two datasets of MS COCO 2017 and Cityscapes. For the COCO dataset, the algorithm value can be increased by 1.7% and 2.5% of each compared to the benchmark of Mask R-CNN when the backbone network is based on ResNet50 and ResNet101. For the Cityscapes dataset, ResNet50 is used as the backbone network to evaluate on the validation set and test set, which are 2.1% and 2.3% higher than Mask R-CNN for the two sets. The ablation results show that the AgFPN has its potential performance and is easy to be integrated into multiple detectors. Furthermore, feature-related augmentation is utilized to improve average accuracy of 0.6% and 0.7% each for attention feature fusion module and the global context module. When we combine the two modules, the performance-benched is improved by 1.7%. The visualization results show that our method is more accurate in positioning multi-scale targets. The segmentation effect is improved significantly on the two aspects of mutual occlusion and the boundary of multiple targets. **Conclusion** The experimental results show that our algorithm is based on the overall multi-scale context information of the target and the multiple feature representation of the target can be improved. Therefore, the algorithm effectiveness is demonstrated that it can improve the accuracy of the network for target detection and segmentation at different scales further.

Key words: instance segmentation; mask region-based convolutional neural network (Mask R-CNN); feature pyramid network (FPN); multi-scale context information; multi-scale channel attention (MSCA)

0 引言

实例分割(王子愉等,2019)是图像及视频场景理解的基础任务,精确的实例分割在自动驾驶(Zhou等,2020)、医学影像分割(林成创等,2020)和视频监控(黄泽涛等,2021)等实际场景中具有广泛应用。随着深度卷积神经网络的快速发展,实例分割技术取得显著进展,主要包括单阶段实例分割方法和两阶段实例分割方法。

单阶段实例分割方法形式多样。YOLACT(you only look at coefficients)(Bolya等,2019)为每个实例预测一组原型掩膜和掩膜系数,并通过矩阵乘法组合;PolarMask(Xie等,2020)利用实例中心点分类和密集距离回归,基于极坐标系对实例掩膜进行建模;SOLO(segmenting objects by locations)(Wang等,2020)将实例类别定义为目标的位置和尺寸,将实例分割任务分为类别预测和生成实例掩膜两个子任务。单阶段方法需要同时定位、分类和分割对象,特别是不同尺度目标,难度更大,虽然速度较快,但分割精度提升受限,在解决目标多尺度问题上不具有优势。

两阶段实例分割方法首先使用检测器生成候选区域,然后针对候选区域进行分割,并为每个实例生成像素级掩膜。Mask R-CNN(mask region-based convolutional neural network)(He等,2017)通过扩展Faster R-CNN(Ren等,2017),增加掩膜预测分支来分割候选框中的目标,该算法对于检测和分割两个阶段的高效利用,极大提高了分割精度。以下算法都是在Mask R-CNN框架基础上的改进。PANet(path aggregation network)(Liu等,2018)通过添加一条自底向上的路径,增强了特征金字塔网络(feature pyramid network, FPN)(Lin等,2017)的多层次特征表示;MaskLab(Chen等,2018)产生边界框检测、语义分割和方向预测3个输出,通过组合语义和方向预测来执行前景和背景分割;MS RCNN(mask scoring region-based convolutional neural network)(Huang等,2019a)缓解了掩膜质量和评分之间的偏差;Wen等人(2020)提出联合多任务级联结构,并在全卷积网络分支中引入特征融合,有效联合了高低层特征;BMask R-CNN(boundary-preserving mask region-based convolutional neural network)(Cheng等,2020)利用额外的分支直接估计边界来增强掩膜特

征的边界感知。DCT-Mask(discrete cosine transform mask)(Shen等,2021)使用离散余弦变换将高分辨率二进制掩膜编码成紧凑的向量。两阶段方法经过不断改进,有效提高了分割精度。但上述方法未从多尺度目标变化角度提出解决方案,因此分割精度仍有提升空间。

多尺度上下文信息通过增强特征表示,可以有效提高分割性能。在图像分割领域已有一些工作致力于提取和融合多尺度上下文信息。PSPNet(pyramid scene parsing network)(Zhao等,2017)通过金字塔池模块和金字塔场景解析网络,利用不同尺度的上下文信息聚合来实现高质量的场景分割;CCNet(criss-cross network)(Huang等,2019b)通过循环交叉注意模块获取密集的上下文信息进行语义分割;HTC(hybrid task cascade)(Chen等,2019a)增加了语义分割分支来整合FPN各层特征的上下文信息,以增强目标前景和背景的判别性特征用于实例分割;Zhang等人(2022)设计了语义注意模块和尺度互补掩膜分支,以充分利用多尺度上下文信息解决遥感图像实例分割问题;吉淑滢和肖志勇(2021)使用金字塔卷积和密集连接的集成提取多尺度信息,并充分融合上下文和多尺度特征进行胸部多器官分割;丁宗元等人(2021)提出融合不同尺度交互映射的双路网络结构用于提取目标的多尺度特征,显著提升交互式图像分割性能;RefineMask(Zhang等,2021)使用空洞卷积设计了语义融合模块,将捕获的多尺度上下文信息用于实例分割。

但是,目标多尺度变化导致实例分割精度提升受限。对此,本文在两阶段实例分割模型Mask R-CNN的基础上,提出了融合多尺度上下文信息的实例分割算法。首先,提出注意力引导的特征金字塔网络(attention-guided feature pyramid network, AgFPN),通过邻层特征自适应融合模块(adjacent-layer feature adaptive fusion module, AFAFM)对FPN邻层特征融合方式进行优化,使用内容感知重组(Wang等,2019)对特征上采样,并在邻层特征融合前使用通道注意力机制(Hu等,2020)对通道加权,增强语义一致性。其次,引入多尺度通道注意力(multi-scale channel attention, MSCA)(Dai等,2021)构造了注意力特征融合模块(attentional feature fusion module, AFFM)和全局上下文模块(global context module, GCM)来整合多尺度特征,并将感

兴趣区域(region of interest, RoI)特征与目标多尺度上下文信息(multi-scale contextual information, MSCl)进行融合,增强了分类回归和掩膜预测两个分支的多尺度特征表示。通过在 MS COCO 2017 (Microsoft common objects in context 2017) (Lin 等, 2014) 和 Cityscapes (Cordts 等, 2016) 两个数据集上进行训练和评估,所提方法有效提高了实例分割的精度,显著提升了不同尺度目标在相互遮挡和分界处的定位、识别和分割性能。

1 本文算法

为解决目标多尺度变化问题,本文提出了融合多尺度上下文信息的实例分割算法。如图 1 所示,

所提算法网络结构以 Mask R-CNN 框架为基础,首先,使用注意力引导的特征金字塔网络 AgFPN 提取图像多尺度特征,主干网络的特征层次表示为 $\{f_2, f_3, f_4, f_5\}$,邻层特征自适应融合后得到的自顶向下的特征表示为 $\{p_2, p_3, p_4, p_5\}$ 。接着,进行多尺度上下文信息提取与融合。其中,区域建议网络(region proposal network, RPN)对目标区域建议边界框,进行前景和背景的分类和边界框的回归,并筛选感兴趣区域 RoI;同时,多尺度上下文信息通过注意力特征融合模块和全局上下文模块从 AgFPN 中获得。然后,使用 RoIAlign 算法根据目标检测框的位置,将 RoI 映射到特征图中获得固定尺寸的特征图,进而与多尺度上下文信息进行融合。最后,利用融合特征进行边界框回归和掩膜预测。

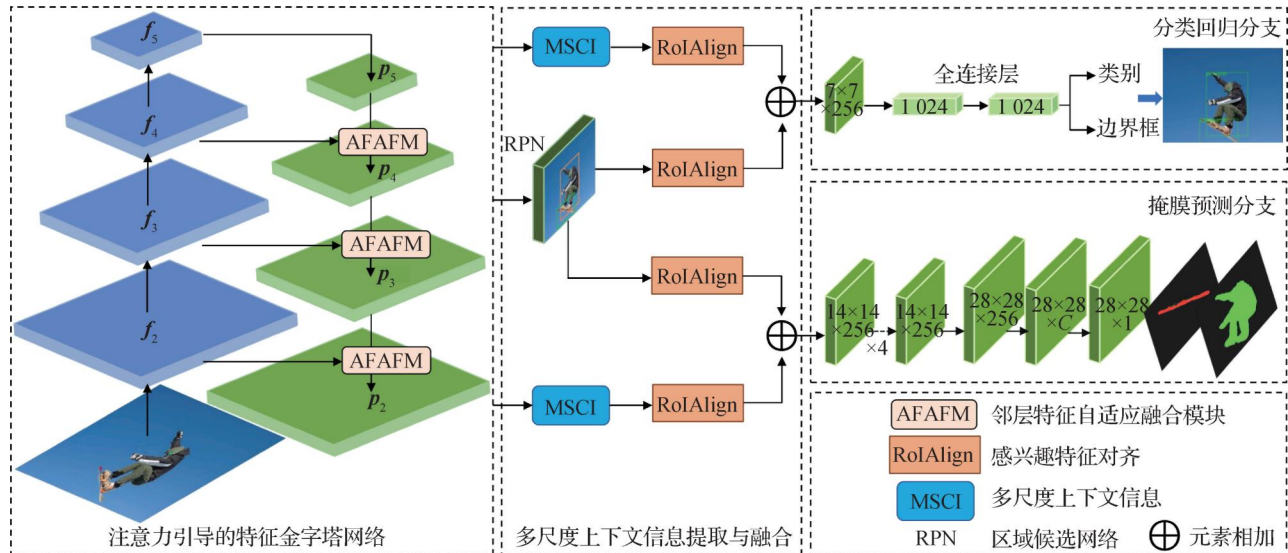


图 1 融合多尺度上下文信息的实例分割模型

Fig. 1 An instance segmentation model incorporating multi-scale context information

通过更高效的 AgFPN 进行特征提取以及多尺度上下文信息聚合,可以有效提高不同尺度目标的实例分割性能。

1.1 注意力引导的特征金字塔网络

多尺度特征表示是检测和分割不同尺度目标的有效方法,为了充分利用高层语义特征和底层细粒度特征,FPN 成为实例分割算法的通用网络。

但是,在 FPN 自顶向下的特征融合路径中,不同层的特征融合采用最近邻插值和元素相加的方法,插值只依赖特征的相对位置而无法利用丰富的语义信息,直接元素相加忽略了相邻特征之间的语义差距而产生混叠效应。FPN 邻层特征的融合方式

不能充分利用不同尺度的特征,因此,本文提出 AgFPN,通过邻层特征自适应融合模块 AFAFM 对 FPN 邻层特征融合方式进行优化。

AFAFM 结构如图 2 所示。 $f_i \in \mathbf{R}^{c \times h \times w}$ 是主干网络提取的基础特征, $p_{i+1} \in \mathbf{R}^{c \times (h/s) \times (w/s)}$ 是自顶向下路径的上层特征,其中 $s=2$ 为尺度因子。首先,采用最大值池化对特征 f_i 进行 2 倍下采样得到特征 $\tilde{f}_i$,并与特征 p_{i+1} 通道拼接,然后将融合特征输入 CARAFE (content-aware reassembly of features) 模块 (Wang 等, 2019),与特征 p_{i+1} 进行内容感知重组上采样得到 $\tilde{p}_{i+1}$ 。同时,对融合特征使用 1×1 卷积 t_1 和 softmax 函数归一化得到预测权重 M_i ,并与融

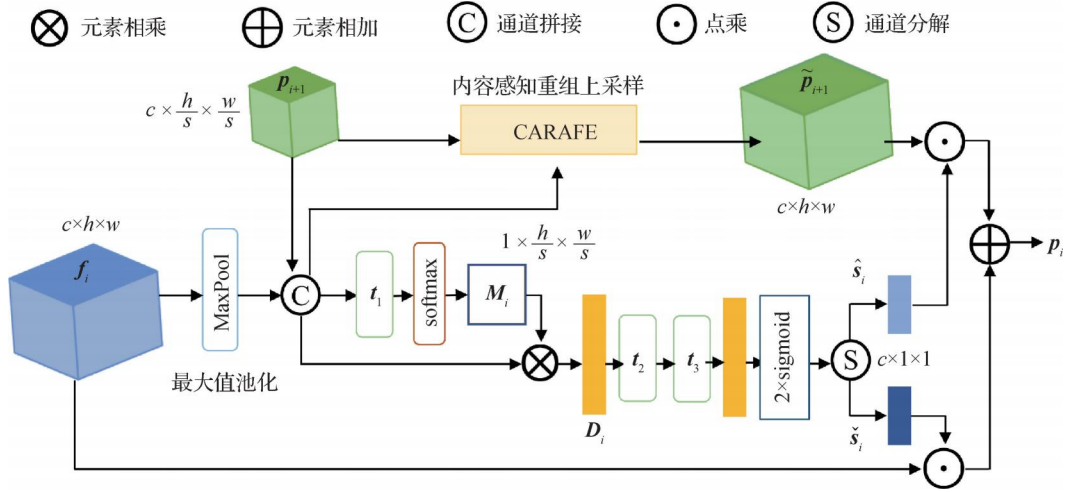


图 2 邻层特征自适应融合模块结构图

Fig. 2 Structure diagram of the adjacent layer feature adaptive fusion module

合特征进行矩阵相乘得到通道统计 D_i ；为了限制 AFAFM 模块的复杂性,在非线性周围加入 1×1 卷积 t_2 用于降维和 1×1 卷积 t_3 用于升维。然后,采用 $2 \times \text{sigmoid}$ 函数激活的简单门控机制来模拟通道间的相互依赖关系。最后,利用通道注意力 (Hu 等, 2020) 学习两个权重向量 $\hat{s}_i$ 和 $\check{s}_i$ 作为融合系数,并将相邻特征 $\tilde{p}_{i+1}$ 和 f_i 进行元素加权融合,重新校正邻层特征的语义信息。

上述过程可具体描述为

$$M_i = \text{softmax}(t_1[p_{i+1}, \tilde{f}_i])^T \quad (1)$$

$$D_i = [p_{i+1}, \tilde{f}_i] M_i \quad (2)$$

$$[\hat{s}_i, \check{s}_i] = 2\sigma(t_3\delta(LN(t_2D_i))) \quad (3)$$

$$p_i = \hat{s}_i \odot \tilde{p}_{i+1} + \check{s}_i \odot f_i \quad (4)$$

式中, $t_1 \in \mathbf{R}^{1 \times 2c \times 1 \times 1}$, $t_2 \in \mathbf{R}^{(c/s) \times 2c \times 1 \times 1}$, $t_3 \in \mathbf{R}^{2c \times (c/s) \times 1 \times 1}$, $s=2$ 为尺度因子; LN 代表层归一化, δ 表示激活函数 ReLU (rectified linear unit); 2σ 表示激活函数 $2 \times \text{sigmoid}$, 该函数可使通道权值连续相乘后的均值为 1, 并可选择性地激发或抑制特征, $\odot$ 表示点乘。

1.2 多尺度上下文信息提取与融合

Mask R-CNN 算法在检测和掩膜分支中仅利用到 RoI 特征, 由于缺乏多尺度上下文信息, 掩膜预测质量提高受限。

因此, 本文通过引入多尺度通道注意力, 设计了 AFAFM 模块来整合多尺度特征, 以及 GCM 模块来挖掘融合特征中的多尺度上下文信息, 并将上下文信息与 RoI 特征融合, 从而使模型能够更好地预测实

例分割结果。

具体来说, 给定 RoI, 使用 RoIAlign 算法从相应层次的 FPN 输出中提取小的特征块 (例如 7×7 或 14×14)。同时, 对多尺度上下文信息特征应用 RoIAlign, 得到相同尺寸的特征块, 然后将两个分支的特征按元素求和进行组合。MSCI 的结构设计如图 3 所示。首先, 利用 AFAFM 聚合相邻层特征; 然后, GCM 提取融合特征的上下文信息为新的特征层, 并与下一层特征进行注意力特征融合, 依次迭代; 最后, 得到来自不同层的多尺度上下文信息。

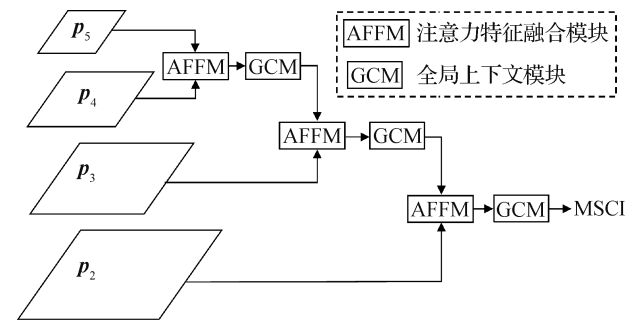


图 3 多尺度上下文信息提取与融合

Fig. 3 Multi-scale context information extraction and fusion

1.2.1 注意力特征融合模块

将不同层或分支的特征进行融合通常利用加法求和或通道拼接等简单操作, 无法有效利用上下文信息。因此, 本文提出注意力特征融合模块 AFAFM, 通过引入多尺度通道注意力 MSCA 有效融合跨层特征, 利用多尺度上下文信息缓解多尺度变化的影响。

MSCA 对不同尺度目标具有较强的适应性, 其

结构如图 4(a) 所示。MSCA 使用双分支并行结构, 其中一个分支利用全局平均池化提取和增强特征图的全局上下文信息, 另一个分支保持原始特征分辨率以获取局部上下文信息, 避免忽略较小尺度目标。MSCA 利用两个分支的逐点卷积沿通道维度压缩和恢复特征, 从而聚合多尺度通道上下文信息, 便于网络识别和检测极端尺度变化下的目标。

AFFM 结构如图 4(b) 所示, 表示为

$$\begin{aligned} F_{XY} = & A(F_X \oplus \tilde{F}_Y) \otimes F_X \oplus (1 - \\ & A(F_X \oplus \tilde{F}_Y)) \otimes \tilde{F}_Y \end{aligned} \quad (5)$$

式中, A 表示 MSCA 操作, F_X 和 F_Y 是两个输入特

征, 特征 F_Y 更高级但分辨率更低。首先, 对特征 F_Y 进行 2 倍上采样得到 $\tilde{F}_Y$ 。然后, 将特征 F_X 与 $\tilde{F}_Y$ 相加得到的初始融合特征作为 MSCA 的输入; $\otimes$ 和 $\oplus$ 分别表示元素级乘法和加法; $1 - A(F_X \oplus \tilde{F}_Y)$ 对应图 4(b) MSCA 右侧中的虚线, 融合权值 $A(F_X \oplus \tilde{F}_Y)$ 和 $1 - A(F_X \oplus \tilde{F}_Y)$ 都属于 $0 \sim 1$ 之间。最后, 将特征 F_{XY} 输入到 3×3 卷积层中, 进行批处理归一化和 ReLU 激活操作, 得到跨层融合特征 F 。

AFFM 引入了多尺度通道注意力, 通过挖掘通道之间的相互依赖关系, 对不同层次的多尺度特征进行融合, 获得了注意力信息引导的融合特征。

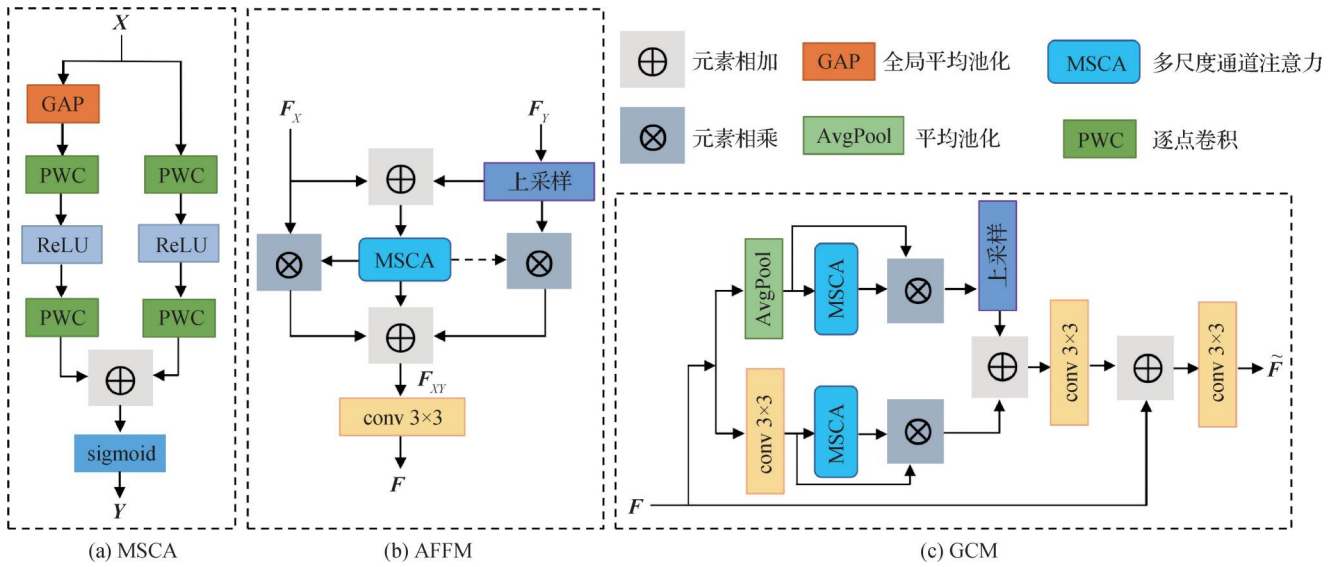


图 4 3 种网络结构图

Fig. 4 Structure diagrams of three kinds of network ((a) MSCA; (b) AFFM; (c) GCM)

1.2.2 全局上下文模块

为了充分利用跨层融合特征中丰富的全局上下文信息, 提出全局上下文模块 GCM, 结构如图 4(c) 所示。具体来说, 设 AFFM 的输出特征 $F \in \mathbf{R}^{C \times H \times W}$, 两个分支分别通过卷积运算和平均池化得到两个子特征 $F_c \in \mathbf{R}^{C \times H \times W}$ 和 $F_p \in \mathbf{R}^{C \times (2/H) \times (2/W)}$ 。为了获得多尺度特征表示, 首先将特征 F_c 和 F_p 输入到 MSCA 模块。接着采用元素级乘法将 MSCA 的输出与对应的特征 F_c 和 F_p 相融合, 得到 $F_{ca} \in \mathbf{R}^{C \times H \times W}$ 和 $F_{pa} \in \mathbf{R}^{C \times (2/H) \times (2/W)}$ 。然后使用加法运算融合两个分支特征得到 F_{cpa} 。最后利用残差结构将 F 和 F_{cpa} 融合得到 $\tilde{F}$ 。上述过程可具体描述为

$$F_c = \text{conv}(F), F_{ca} = F_c \otimes A(F_c) \quad (6)$$

$$F_p = \text{conv}(\text{Pool}(F)), F_{pa} = F_p \otimes A(F_p) \quad (7)$$

$$\tilde{F} = \text{conv}(F \oplus \text{conv}(F_{ca} \oplus \text{upsample}(F_{pa}))) \quad (8)$$

式中, conv , Pool , A 和 upsample 分别代表卷积、平均池化、MSCA 和上采样操作。

GCM 用于在特定级别自适应地提取多尺度上下文信息, 改进不同尺度和特定语义的特征表示, 自适应地整合全局和局部特征, 可以有效提高多尺度目标的分割精度。

2 实验与结果分析

为了验证本文算法的性能, 在 MS COCO 2017 和 Cityscapes 数据集上进行实验, 与相关方法进行视觉效果和定量结果对比, 使用平均精度 (average

precision, AP) 作为评价指标, 并在 MS COCO 2017 数据集上进行消融实验。

2.1 实验数据集与评价指标

MS COCO 2017 数据集包含 80 个实例级标签类别, 模型使用训练集的 115 000 幅图像进行训练, 对 5 000 幅验证集图像进行测试, 最终展示了 20 000 幅测试数据集图像上的定量结果。

Cityscapes 数据集包含大量城市街道场景图像, 提供了语义、实例特定和像素特定的注释, 分别有 2 975、500 和 1 525 幅图像用于训练、验证和测试。对于实例分割任务, 有 8 个实例类别。

评估指标采用标准的掩膜平均精度 AP。AP 表示 IoU(intersection over union) 阈值从 0.5 ~ 0.95 每隔 0.05 取值情况下得到的 AP 的平均值。

2.2 实验环境与参数

利用深度学习框架 PyTorch 实现所提算法, 实验环境为 Ubuntu 16.04 操作系统, 使用 4 块 NVIDIA 1080Ti 图形处理器 (graphics processing unit, GPU) 加速运算。

在 MS COCO 2017 数据集上, 本文方法分别使用 ResNet-50 和 ResNet-101 作为主干网络, 并利用 ImageNet 上预训练的权重来初始化网络参数。实验采用随机梯度下降法 (stochastic gradient descent, SGD) 进行 16 万次迭代训练优化, 初始学习率为

0.002, batch size 设为 4, 当迭代次数为 13 万次和 15 万次时, 学习率分别降低 10 倍。设置权重衰减 (weight decay) 系数为 0.000 5, 动量 (momentum) 系数设为 0.9。损失函数和其他超参数均按照 mmdetection (Chen 等, 2019b) 中描述的策略进行设置和初始化。

在 Cityscapes 数据集上, 使用 ResNet-50 作为主干网络, batch size 设为 4, 迭代次数为 48 000 次, 初始学习率为 0.005, 当迭代到 36 000 次时, 学习速率降至 0.000 5。其他设置与在 MS COCO 2017 数据集上的实验相同。

2.3 定量结果分析

在 MS COCO 2017 测试集上, 将所提方法与经典的两阶段方法和其他单阶段方法进行分割精度的对比, 结果如表 1 所示。其中, AP_{50} 和 AP_{75} 分别表示 IoU 阈值为 0.5 和 0.75 时的平均精度, AP_S 、 AP_M 和 AP_L 分别是小、中、大 3 种不同尺度目标的平均精度。可以看出, 所提算法相较于基线 Mask R-CNN 在主干网络为 ResNet50 和 ResNet101 时分别提高了 1.7% 和 2.5%; 在多尺度目标分割精度上, 以主干网络 ResNet50 为例, AP_S 和 AP_M 分别提高了 1.6% 和 2.6%, 说明利用 AgFPN 进行特征提取, 并在 RoI 特征中引入多尺度上下文信息, 有效提高了中小目标的掩膜预测质量。

表 1 实例分割模型在 MS COCO 2017 测试集上的平均精度对比

Table 1 Comparison of average accuracy of instance segmentation models on MS COCO 2017 test dataset

方法	主干网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN (He 等, 2017)	ResNet-50-FPN	35.6	57.6	38.1	18.7	38.3	49.7
PANet (Liu 等, 2018)	ResNet-50-FPN	36.6	58.0	39.3	16.3	38.1	53.1
PointRend (Kirillov 等, 2020)	ResNet-50-FPN	36.3	—	—	—	—	—
DCT-Mask (Shen 等, 2021)	ResNet-50-FPN	36.5	56.3	39.6	17.7	38.6	51.9
本文	ResNet-50-FPN	37.3	60.4	39.7	20.3	40.9	51.8
FCIS (Li 等, 2017)	ResNet-101-C5	29.2	49.5	—	7.1	31.3	50.0
Mask R-CNN (He 等, 2017)	ResNet-101-FPN	36.2	58.6	38.4	19.4	38.4	52.1
MaskLab (Chen 等, 2018)	ResNet-101-FPN	35.4	57.4	37.4	16.9	38.3	49.2
MS RCNN (Huang 等, 2019a)	ResNet-101-FPN	37.5	58.7	40.2	17.2	39.5	53.0
YOLOACT (Bolya 等, 2019)	ResNet-101-FPN	31.2	50.6	32.8	12.1	33.3	47.1
TensorMask (Chen 等, 2019)	ResNet-101-FPN	37.1	59.3	39.4	17.4	39.1	51.6
MEInst (Zhang 等, 2020)	ResNet-101-FPN	33.9	56.2	35.4	19.8	36.1	42.3
PolarMask (Xie 等, 2020)	ResNet-101-FPN	32.1	53.7	33.1	14.7	33.8	45.3
SOLO (Wang 等, 2020)	ResNet-101-FPN	37.8	59.5	40.4	16.4	40.6	54.2
本文	ResNet-101-FPN	38.7	61.5	41.6	21.1	41.7	53.1

注: 加粗字体表示各列最优结果, “—”表示数据不可用, MEInst: mask encoding based instance segmentation, FCIS: fully convolutional instance-aware semantic segmentation。

本文方法与其他两阶段方法如 PANet 和 MS RCNN 等相比具有一定的竞争优势,且分割精度高于流行的 YOLACT (you only look at coefficients)、PolarMask 和 SOLO 等单阶段方法。但是,本文方法在大尺度目标的分割精度 AP_L 上低于 SOLO 算法,表明所提方法在大型目标的边缘分割精度上还有提升空间。

在 Cityscapes 数据集上,对比了部分实例分割模型的平均精度,主干网络均采用 ResNet-50。实验结果如表 2 所示。其中, $AP_{[val]}$ 表示 Cityscapes 验证

子集的结果, AP 和 AP_{50} 表示 Cityscapes 测试子集的结果。fine 表示只使用精细数据进行训练,coarse 表示粗糙数据, fine + coco 表示使用精细数据并在 MS COCO 2017 数据集上进行预训练。

从表 2 可以看出,所提方法使用 fine + coco 训练策略,在验证子集和测试子集上进行性能评估,比 Mask R-CNN 分别提高了 2.1% 和 2.3%,有效提高了实例分割精度,同时优于 PANet 和 BMask RCNN 等实例分割方法。实验结果表明,所提方法具有较强的模型泛化性和对不同尺度目标的识别鲁棒性。

表 2 Cityscapes 数据集上实例分割模型的平均精度对比
Table 2 Comparison of average accuracy of instance segmentation models on Cityscapes dataset

方法	训练策略	/%										
		$AP_{[val]}$	AP	AP_{50}	person	rider	car	truck	bus	train	mcycle	bicycle
SGN(Liu 等,2017)	fine + coarse	29.2	25.0	44.9	21.8	20.1	39.4	24.8	33.2	30.8	17.7	12.4
InstanceCut(Kirillov 等,2017)	fine	15.8	13.0	27.9	10.0	8.0	23.7	14.0	19.5	15.2	9.3	4.7
Mask R-CNN(He 等,2017)	fine	31.5	26.2	49.9	30.5	23.7	46.9	22.8	32.2	18.6	19.1	16.0
Mask R-CNN(He 等,2017)	fine + coco	36.8	32.6	59.2	36.7	29.2	52.8	30.0	40.3	27.9	25.0	19.0
PANet(Liu 等,2018)	fine	36.5	31.8	57.1	36.8	30.4	54.8	27.0	36.3	25.5	22.6	20.8
Deep snake(Peng 等,2020)	fine	37.4	31.7	58.4	37.2	27.0	56.0	29.5	40.5	28.2	19.0	16.4
BMask RCNN(Cheng 等,2020)	fine	35.0	29.4	54.7	34.3	25.6	52.6	24.2	35.1	24.5	21.4	17.1
本文	fine + coco	38.9	34.9	61.0	38.2	31.6	56.2	31.8	41.7	29.0	26.5	21.4

注:加粗字体表示各列最优结果。

2.4 可视化结果展示与对比分析

2.4.1 在 MS COCO 2017 数据集上多尺度目标下的可视化结果展示

MS COCO 2017 数据集上多尺度目标实例分割的可视化结果如图 5 所示。可视化结果表明,本文方法对多尺度目标具有较好的定位、分类和分割效果。由于远、小目标信息较少,通过在分类回归和掩膜预测分支中弥补目标多尺度上下文信息可以有效提高小目标的识别精度,更有利于分割。同时,AgF-PN 可以有效缓解 FPN 邻层不同尺度目标的语义特征混叠,减少多尺度目标错检和漏检概率,显著提高多尺度目标的分割精度。

此外,本文方法在不同目标边界位置以及存在遮挡的情况下有较好的预测结果。如图 6 第 1、2 行所示,本文方法可以识别到被“足球”遮挡的“手”属于“运动员”,能准确识别出被“轿车”遮挡的目标是“马”,而其他算法或漏检或错检。如图 6 第 3、4 行

所示,在不同实例的边界处,本文方法处理的边界更为精准,分割质量更高。可视化对比表明,所提算法取得了良好的性能。

2.4.2 在 Cityscapes 数据集上的可视化结果展示与分析

为了验证本文方法的有效性和泛化性,在 Cityscapes 数据集上进行多尺度目标实例分割,可视化结果如图 7 所示。Cityscapes 数据集注释质量较高,且城市街景中更容易造成视觉形变,产生多尺度目标,对于实例分割任务具有更多挑战性。从图 7 可以看出,本文方法有效解决了不同尺度和不同类别的实例分割任务,多尺度目标得到了准确的识别、分类和像素级掩膜生成,甚至有效缓解了目标遮挡问题,表现了较好的分割性能,证明了所提方法的有效性和泛化性。

为进一步验证本文方法的性能,在 Cityscapes 数据集上与基线 Mask R-CNN 进行对比,结果如图 8

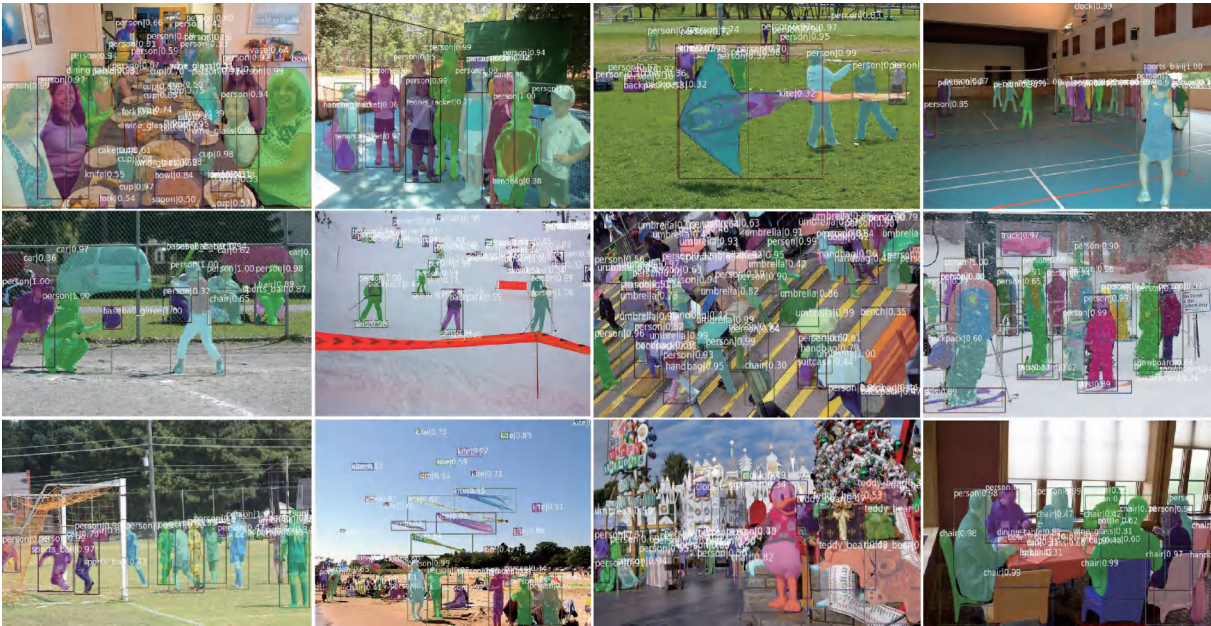


图 5 在 MS COCO 2017 数据集上多尺度目标实例分割的可视化结果

Fig. 5 Visualization results of multi-scale target instance segmentation on MS COCO 2017 dataset

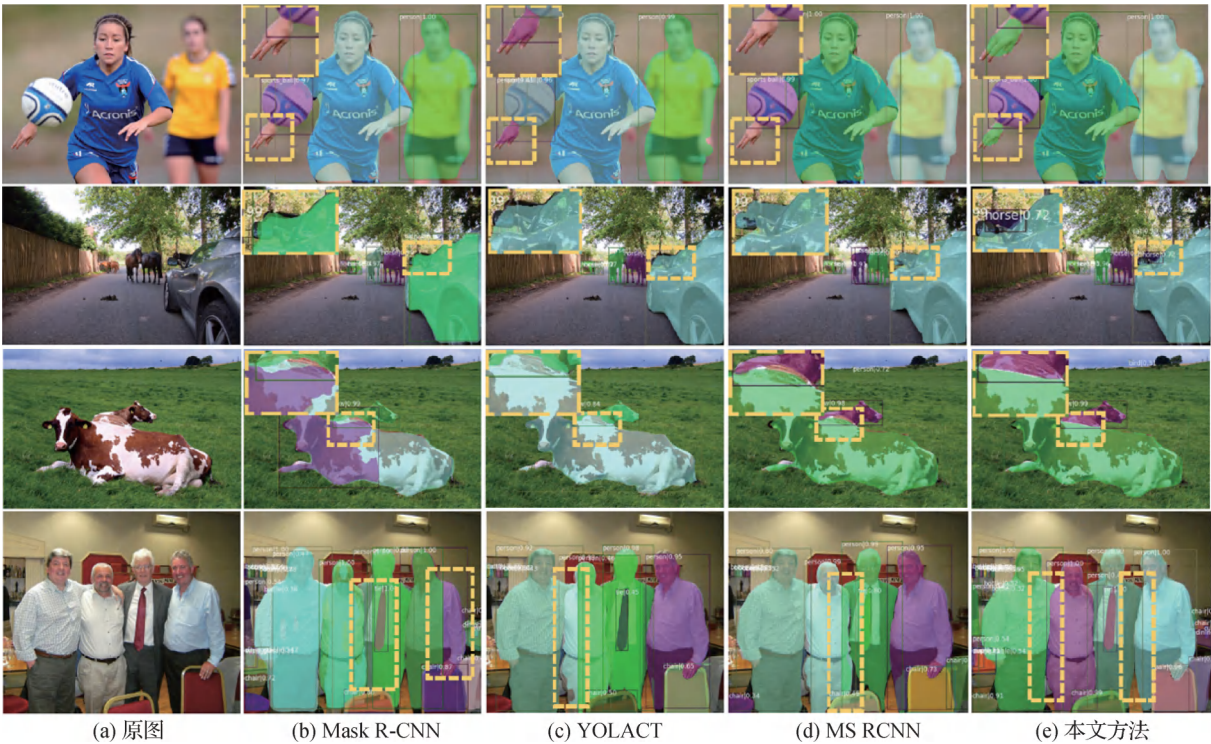


图 6 在 MS COCO 2017 数据集上在目标边界处和遮挡情况下的可视化结果对比

Fig. 6 Comparison of visualization results on MS COCO 2017 dataset at the target boundary and under occlusion

((a)original images;(b)Mask R-CNN;(c)YOLACT;(d)MS RCNN;(e)ours)

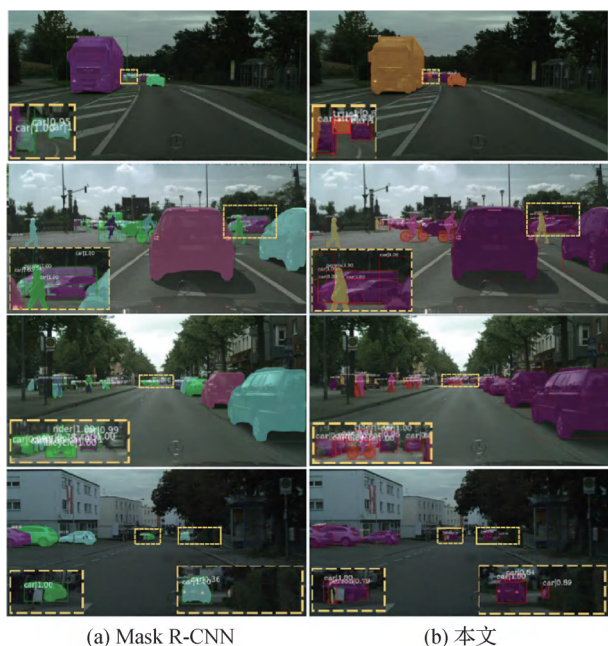
所示。从图 8 第 1、2 行可以看出,本文方法在不同尺度目标存在遮挡和目标边界处分割效果较好;从图 8 第 3、4 行可以看出,所提方法有效改善了小尺

度目标的漏检和错检。可视化结果显示,所提方法在具有挑战性的 Cityscapes 数据集上也有较好的效果。



图7 在 Cityscapes 数据集上多尺度目标实例分割的可视化结果

Fig.7 Visualization results of multi-scale target instance segmentation on Cityscapes dataset



(a) Mask R-CNN

(b) 本文

图8 本文方法与 Mask R-CNN 在 Cityscapes 数据集上的可视化结果对比

Fig.8 Comparison of visualization results between Mask R-CNN and ours on Cityscapes dataset((a)Mask R-CNN;(b)ours)

2.4.3 本文训练模型在不同场景图像上的测试结果

图9为在 MS COCO 2017 和 Cityscapes 数据集上训练的本文方法模型在不同场景图像上的测试结果,测试图像来源网络 and 实地拍摄。在 Cityscapes 数据集上训练的模型,主要测试具有不同尺度目标的城市街景图,如图9(a)所示,不同尺度目标得到了准确的识别和分割。在 MS COCO 2017

数据集上测试了室内场景、城市街景和河景,以及白天、黑夜和雨天等特殊场景,如图9(b)所示。测试结果表明,所提方法具有一定的泛化性和实用价值。



(a) Cityscapes数据集

(b) MS COCO 2017数据集

图9 训练模型在不同场景图像上的测试结果

Fig.9 Test results of training model on images of different scenes ((a) Cityscapes dataset;(b) MS COCO 2017 dataset)

2.5 消融实验

为验证所提模型设计的注意力引导的特征金字塔网络 AgFPN、注意力特征融合模块 AFFM 和全局上下文模块 GCM 的有效性,进行消融实验。

2.5.1 AgFPN 的作用

本文提出的 AgFPN 易于集成到当前流行的两阶段实例分割网络,只需要将 AgFPN 直接替换基线

模型中的 FPN 即可。表 3 为 AgFPN 对实验结果的影响对比。其中, * 表示重新实现的结果, AP^b 表示检测框的精度, R 代表 ResNet。可以看出, 使用不同

的实例分割框架和骨干网络, 与基线模型相比, AgFPN 带来了更好的性能提升, 表明 AgFPN 在提高分割和检测精度上具有一定效果。

表 3 AgFPN 对实验结果的影响对比

Table 3 Comparison of the influence of AgFPN on experimental results

方法	主干网络	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L	AP^b	AP_{50}^b	AP_{75}^b	AP_s^b	AP_m^b	AP_L^b
Mask R-CNN *	R-50-FPN	34.5	56.3	36.7	18.6	37.3	44.7	37.6	59.5	40.6	21.8	40.8	46.4
Mask R-CNN	R-50-AgFPN	36.1	58.0	38.3	19.2	39.6	49.2	38.9	61.5	42.3	22.6	41.5	48.2
Mask R-CNN *	R-101-FPN	36.3	58.5	38.9	19.4	39.3	47.8	39.9	61.6	43.6	23.1	43.2	50.0
Mask R-CNN	R-101-AgFPN	37.2	59.6	39.8	20.9	40.8	50.2	40.8	62.4	45.3	23.9	44.8	51.4
HTC *	R-50-FPN	38.4	60.0	41.4	20.3	40.6	51.2	43.5	62.6	47.3	24.5	45.9	55.9
HTC	R-50-AgFPN	39.2	61.3	42.2	21.1	41.3	52.1	44.2	63.8	48.2	25.3	46.8	57.0

2.5.2 AFFM 和 GCM 的作用

注意力特征融合模块 AFFM 和全局上下文模块 GCM 都是即插即用的特征关系增强模块。为评估 AFFM 和 GCM 模块的作用, 以 ResNet-50 + FPN 为主干网络, 在 MS COCO 2017 验证集上进行消融实验。表 4 为 AFFM 和 GCM 模块对实验结果的影响对比。可以看出, 每个模块都有效提高了基线的性能。具体来说, AFFM 和 GCM 模块将平均精度分别提高了 0.6% 和 0.7%, 当组合两个模块时, 基线的性能进一步提高了 1.7%。实验结果表明, 两个模块有助于整合多尺度特征并充分挖掘多尺度上下文信息, 提高了实例分割的准确度。

表 4 AFFM 和 GCM 模块对实验结果的影响对比

Table 4 Comparison of the influence of AFFM and GCM modules on experimental results

方法	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L
Mask R-CNN	34.5	56.3	36.7	18.6	37.3	44.7
Mask R-CNN + AFFM	35.1	56.8	37.4	19.2	38.0	45.3
Mask R-CNN + GCM	35.2	57.0	37.2	19.3	37.9	45.5
Mask R-CNN + AFFM + GCM	36.2	57.5	37.8	20.1	38.4	45.9

注: 加粗字体表示各列最优结果。

2.5.3 MSCI 网络结构有效性分析

为了验证多尺度上下文信息 MSCI 结构的有效性, 对利用不同层、不同融合顺序的结构进行测试,

消融实验结果如表 5 所示。其中, 原始结构记为“P5、P4、P3 和 P2”, 则“P2、P3、P4 和 P5”表示从 P2 层和 P3 层开始进行融合迭代, “P2 和 P3”表示只利用 P2 层和 P3 层特征进行融合, “P4 和 P5”表示只利用 P4 层和 P5 层特征进行融合。可以看出, 从高层开始进行特征融合比从低层开始更加有效, 高层特征具有较强的语义信息, 将高级的语义特征从顶至下传播到底层, 有助于多尺度特征融合与表达。此外, 只融合 P2 层和 P3 层特征比只融合 P4 层和 P5 层特征更有利于精度提高, 由于低层特征包含的是颜色、边缘、轮廓和纹理等信息, 能使分割预测结果更加细致、精准。

表 5 多尺度融合策略对实验结果的影响对比

Table 5 Comparison of the effects of multi-scale fusion strategies on experimental results

MSCI 结构	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_L
P5、P4、P3 和 P2	36.2	57.5	37.8	20.1	38.4	45.9
P2 和 P3	35.4	56.9	37.3	19.4	37.8	45.3
P4 和 P5	35.2	57.0	37.2	18.9	37.5	45.0
P2、P3、P4 和 P5	35.8	57.2	37.5	19.6	38.0	45.6

注: 加粗字体表示各列最优结果。

为了验证 MSCI 的网络结构对目标定位和多尺度目标识别的作用, 使用 Grad-CAM (gradient-weighted class activation mapping) (Selvaraju 等, 2017) 对 MS COCO 2017 数据集图像进行热力图可视化。

图 10 为ResNet-50 和 ResNet-50 + MSCI 网络热力图可视化结果对比。

可以看出,更强的可视化类激活映射(CAM)区域被更亮的颜色覆盖。与 ResNet-50 相比,ResNet-



图 10 在 MS COCO 2017 数据集上 ResNet-50 和 ResNet-50 + MSCI 网络热力图可视化结果对比
Fig. 10 Comparison of visualization results of ResNet-50 and ResNet-50 + MSCI network heat map on MS COCO 2017 dataset((a) original images; (b) ResNet-50; (c) ResNet-50 + MSCI)

50 + MSCI 网络的激活区域更集中,与目标重叠度更高,如图 10 第 1 行中的飞机和第 2 行中的人等,表明它能更好地定位目标、利用目标区域特征。而 ResNet-50 的定位能力相对较差,只覆盖部分对象或受背景干扰。此外,ResNet-50 + MSCI 也可以准确预测小尺度的目标,如图 10 第 3 行和第 5 行中远处的人,图 10 第 4 行中远处的动物等,这体现了 MSCI 网络具有充分表达多尺度特征的能力。

2.5.4 本文算法推理速度的讨论分析

为了测试所提模型在推理速度上的性能,以 ResNet-50 + FPN 为主干网络(本文方法替换为 Ag-FPN),使用单个 V100 GPU,利用预先训练的模型在同一台本地机器上测试每个模型的推理时间。

表 6 为本文方法与其他方法推理速度的对比。可以看出,本文方法在分割精度上略低于 RefineMask 和 HTC,但推理速度明显高于这两种方法。RefineMask 和 HTC 均利用多尺度上下文信息,并且 RefineMask 采用掩膜多阶段细化策略,HTC 使用级联架构,二者都显著增加了计算量,分割精度得到大幅提升的同时,推理速度受限。SOLO 算法是单阶段实例分割方法,采用轻量化模型,有效提高了推理速度,但分割精度不足。本文方法在 Mask R-CNN 的基础上增加了一定的计算复杂度,提高了分割精度,也影响了一定的推理速度,但在精度与速度的权衡上具有一定优势。

表 6 本文方法与其他方法推理速度的对比		
Table 6 Comparison of inference speed between our method and other methods		
方法	AP/%	推理速率/(帧/s)
Mask R-CNN	34.7	15.7
PointRend	35.6	11.4
RefineMask	37.3	11.4
HTC	37.4	4.4
SOLO	34.2	22.5
本文	37.1	13.6

注:加粗字体表示各列最优结果。

3 结 论

为解决目标多尺度变化问题,本文充分考虑

FPN 在邻层特征融合时信息损失和语义特征混叠, 以及 RoI 特征多尺度上下文信息不足问题, 提出一种融合多尺度上下文信息的实例分割方法。通过邻层特征自适应融合模块优化 FPN 邻层特征的融合方式, 减少了信息衰减并增加了语义一致性, 有利于多尺度特征的表达; 同时, 通过引入多尺度通道注意力设计了注意特征融合模块和全局上下文模块, 增强了 RoI 特征的目标多尺度上下文信息。实验结果表明, 所提方法有效提高了多尺度目标的实例分割精度。

但是, 由于分割网络中存在多次卷积和下采样操作, 且边界像素比例较低, 本文方法在较大尺度目标边界分割精度上提升有限。此外, 本文方法在 Mask R-CNN 的基础上增加了一定的计算开销, 影响了推理速度, 使得将本文算法应用于实时应用程序或部署在边缘设备上具有一定挑战性。因此, 改善较大尺度实例边界分割准确度和模型轻量化设计是今后需要继续研究的问题。

参考文献 (References)

- Bolya D, Zhou C, Xiao F Y and Lee Y J. 2019. YOLACT: real-time instance segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9156-9165 [DOI: 10.1109/ICCV.2019.00925]
- Chen K, Pang J M, Wang J Q, Xiong Y, Li X X, Sun S Y, Feng W S, Liu Z W, Shi J P, Ouyang W L, Loy C C and Lin D H. 2019a. Hybrid task cascade for instance segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4969-4978 [DOI: 10.1109/CVPR.2019.00511]
- Chen K, Wang J Q, Pang J M, Cao Y H, Xiong Y, Li X X, Sun S Y, Feng W S, Liu Z W, Xu J R, Zhang Z, Cheng D Z, Zhu C C, Cheng T H, Zhao Q J, Li B Y, Lu X, Zhu R, Wu Y, Dai J F, Wang J D, Shi J P, Ouyang W L, Loy C C and Lin D H. 2019b. MMDetection: open MMLab detection toolbox and benchmark [EB/OL]. [2021-10-10]. <https://arxiv.org/pdf/1906.07155.pdf>
- Chen L C, Hermans A, Papandreou G, Schroff F, Wang P and Adam H. 2018. MaskLab: instance segmentation by refining object detection with semantic and direction features//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4013-4022 [DOI: 10.1109/CVPR.2018.00422]
- Chen X L, Girshick R, He K M and Dollar P. 2019. TensorMask: a foundation for dense object segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2061-2069 [DOI: 10.1109/ICCV.2019.00215]
- Cheng T H, Wang X G, Huang L C and Liu W Y. 2020. Boundary-preserving mask R-CNN//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 660-676 [DOI: 10.1007/978-3-030-58568-6_39]
- Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The Cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]
- Dai Y M, Gieseke F, Oehmcke S, Wu Y Q and Barnard K. 2021. Attentional feature fusion//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3559-3568 [DOI: 10.1109/WACV48630.2021.00360]
- Ding Z Y, Sun Q S, Wang T and Wang H Y. 2021. Deep interactive image segmentation based on fusion multi-scale annotation information. Journal of Computer Research and Development, 58(8): 1705-1717 (丁宗元, 孙权森, 王涛, 王洪元. 2021. 基于融合多尺度标记信息的深度交互式图像分割. 计算机研究与发展, 58(8): 1705-1717) [DOI: 10.7544/issn1000-1239.20210195]
- He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]
- Hu J, Shen L, Albanie S, Sun G and Wu E H. 2020. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8): 2011-2023 [DOI: 10.1109/tpami.2019.2913372]
- Huang Z J, Huang L C, Gong Y C, Huang C and Wang X G. 2019a. Mask scoring R-CNN//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6402-6411 [DOI: 10.1109/CVPR.2019.00657]
- Huang Z L, Wang X G, Huang L C, Huang C, Wei Y C and Liu W Y. 2019b. CCNet: Criss-cross attention for semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 603-612 [DOI: 10.1109/ICCV.2019.00069]
- Huang Z T, Liu Y, Yu C L, Zhang J J, Wang X and Qi S H. 2021. Video instance segmentation based on temporal feature fusion. Journal of Image and Graphics, 26(7): 1692-1703 (黄泽涛, 刘洋, 于成龙, 张加佳, 王轩, 漆舒汉. 2021. 时序特征融合的视频实例分割. 中国图象图形学报, 26(7): 1692-1703) [DOI: 10.11834/jig.200521]
- Ji S Y and Xiao Z Y. 2021. Integrated context and multi-scale features in thoracic organs segmentation. Journal of Image and Graphics,

- 26(9): 2135-2145 (吉淑滢, 肖志勇. 2021. 融合上下文和多尺度特征的胸部多器官分割. 中国图象图形学报, 26(9): 2135-2145) [DOI: 10.11834/jig.200558]
- Kirillov A, Levinkov E, Andres B, Savchynskyy B and Rother C. 2017. InstanceCut: from edges to instances with MultiCut//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 7322-7331 [DOI: 10.1109/CVPR.2017.774]
- Kirillov A, Wu Y X, He K M and Girshick R. 2020. PointRend: image segmentation as rendering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 9796-9805 [DOI: 10.1109/CVPR42600.2020.00982]
- Li Y, Qi H Z, Dai J F, Ji X Y and Wei Y C. 2017. Fully convolutional instance-aware semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 4438-4446 [DOI: 10.1109/CVPR.2017.472]
- Lin C C, Zhao G S, Yin A H, Ding B C, Guo L and Chen H B. 2020. AS-PANet: a chromosome instance segmentation method based on improved path aggregation network architecture. Journal of Image and Graphics, 25(10): 2271-2280 (林成创, 赵淦森, 尹爱华, 丁笔超, 郭莉, 陈汉彪. 2020. AS-PANet: 改进路径增强网络的重叠染色体实例分割. 中国图象图形学报, 25(10): 2271-2280) [DOI: 10.11834/jig.200236]
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]
- Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland; Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]
- Liu S, Jia J Y, Fidler S and Urtasun R. 2017. SGN: sequential grouping networks for instance segmentation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 3516-3524 [DOI: 10.1109/ICCV.2017.378]
- Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]
- Peng S D, Jiang W, Pi H J, Li X L, Bao H J and Zhou X W. 2020. Deep snake for real-time instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 8530-8539 [DOI: 10.1109/CVPR42600.2020.00856]
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]
- Shen X, Yang J R, Wei C B, Deng B, Huang J Q, Hua X S, Cheng X L and Liang K W. 2021. DCT-Mask: discrete cosine transform mask representation for instance segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 8716-8725 [DOI: 10.1109/CVPR46437.2021.00861]
- Wang J Q, Chen K, Xu R, Liu Z W, Loy C C and Lin D H. 2019. CARAFE: content-aware reassembly of features//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 3007-3016 [DOI: 10.1109/ICCV.2019.00310]
- Wang X L, Kong T, Shen C H, Jiang Y N and Li L. 2020. SOLO: segmenting objects by locations//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 649-665 [DOI: 10.1007/978-3-030-58523-5_38]
- Wang Z Y, Yuan C and Li J C. 2019. Instance segmentation with separable convolutions and multi-level features. Journal of Software, 30(4): 954-961 (王子愉, 袁春, 黎健成. 2019. 利用可分离卷积和多级特征的实例分割. 软件学报, 30(4): 954-961) [DOI: 10.13328/j.cnki.jos.005667]
- Wen Y L, Hu F Y, Ren J C, Shang X R, Li L Y and Xi X F. 2020. Joint multi-task cascade for instance segmentation. Journal of Real-Time Image Processing, 17(6): 1983-1989 [DOI: 10.1007/s11554-020-01007-5]
- Xie E Z, Sun P Z, Song X G, Wang W H, Liu X B, Liang D, Shen C H and Luo P. 2020. PolarMask: single shot instance segmentation with polar representation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 12190-12199 [DOI: 10.1109/CVPR42600.2020.01221]
- Zhang G, Lu X, Tan J R, Li J M, Zhang Z X, Li Q Q and Hu X L. 2021. RefineMask: towards high-quality instance segmentation with fine-grained features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Nashville, USA; IEEE: 6857-6865 [DOI: 10.1109/CVPR46437.2021.00679]
- Zhang R F, Tian Z, Shen C H, You M Y and Yan Y L. 2020. Mask encoding for single shot instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10223-10232 [DOI: 10.1109/CVPR42600.2020.01024]
- Zhang T Y, Zhang X R, Zhu P, Tang X, Li C, Jiao L C and Zhou H Y. 2022. Semantic attention and scale complementary network for instance segmentation in remote sensing images. IEEE Transactions on Cybernetics, 52(10): 10999-11013 [DOI: 10.1109/TCYB.2021.3096185]

Zhao H S, Shi J P, Qi X J, Qi X G, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhou D F, Fang J, Song X B, Liu L, Yin J B, Dai Y C, Li H D and Yang R G. 2020. Joint 3D instance segmentation and object detection for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1836-1846 [DOI: 10.1109/CVPR42600.2020.00191]

作者简介

万新军,男,硕士研究生,主要研究方向为计算机视觉与图像

实例分割。E-mail: wanxinjun1030@163.com

胡伏原,通信作者,男,教授,主要研究方向为机器学习与计算机视觉。E-mail: fuyuanhu@usts.edu.cn

周逸云,女,硕士研究生,主要研究方向为计算机视觉与图像实例分割。E-mail: yyy_zhou@yeah.net

沈鸣飞,男,高级工程师,主要研究方向为计算机视觉和智慧城市。E-mail: shenmf@xgmsz.com

周涛,男,教授,主要研究方向为基于医学影像的计算机辅助诊断、医学大数据分析、智能计算。

E-mail: zhoutaonxmu@126.com