DOI:10.19651/j.cnki.emt.2212108

基于改进 Mask R-CNN 的建筑工地实例分割算法*

宋艳飞[1] 王恒友[1,2] 何 强[1,2] 陈琳琳[1]
(1.北京建筑大学理学院 北京 100044;2.北京建筑大学大数据建模理论与技术研究所 北京 100044)

摘 要: 实例分割对排除建筑工地不规则机械设备带来的安全隐患以及监测工人具有重要意义。然而当前主流的实例分割模型存在着边界检测精度不高的问题。结合实例分割的特点,提出了一种基于全局上下文通道注意力(GCCA)机制多阶段细化掩码的改进 Mask R-CNN 模型。首先,在 Mask 头部以多阶段的方式逐步融合细粒度特征,细化高质量掩码。其次,为了更好的融合细粒度特征,构建了 GCCA 注意力机制,其通过简化的全局上下文模块聚合全局特征,并利用一维卷积实现无降维的局部通道交互。实验结果表明,在 COCO 和 MOCS 数据集上均取得了较好的效果。其中,相较于 Mask R-CNN 模型,此算法在检测和分割的平均精度分别提高了 2.4%和 7.6%。

关键词: 实例分割;注意力机制;细粒度特征;掩码质量
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 510.4050

Building site instance segmentation algorithm based on improved Mask R-CNN

Song Yanfei[1] Wang Hengyou[1,2] He Qiang[1,2] Chen Linlin[1]
(1. School of Science, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;
2. Institute of Big Data Modeling and Technology, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: Instance segmentation is of great significance for eliminating safety hazards brought by irregular machinery and equipment on construction sites and for monitoring workers. However, the current mainstream instance segmentation models have the problem of low boundary detection accuracy. Combining the characteristics of instance segmentation, this paper proposes an improved Mask R-CNN model of multi-stage refining mask based on the global context channel attention (GCCA) mechanism. First, this paper gradually fuses fine-grained features in the mask head in a multi-stage manner to refine high quality masks. Second, in order to better fuse fine-grained features, a GCCA attention mechanism is constructed, which aggregates global features through a simplified global context module, and utilizes one-dimensional convolution to achieve local channel interactions without dimensionality reduction. The experimental results show that this paper has achieved great results on both COCO and MOCS datasets. Among them, compared with the Mask R-CNN model, the average accuracy of the algorithm in this paper in detection and segmentation is improved by 2.4% and 7.6% respectively.

Keywords: instance segmentation;attention mechanism;fine-grained features;mask quality

0 引 言

近年来,深度学习的快速发展推动了计算机视觉和图像分析的广泛应用[1-2]。同时以数据为驱动的深度学习方法在建筑工地上也备受关注,在建筑工地大场景施工范围的工程建设中,不规则物体比如吊车上的吊头、起重机的前臂等存在一定的安全隐患,因此在建筑工地安全检测中需要对其采用像素级识别,即实例分割。

实例分割算法主要分为一阶段和两阶段的实例分割算法。其中一阶段的实例分割算法统一到 FCN[3]框架下,如 Instance FCN[4]将原有 FCN 单一输出通道变为多个对实例位置敏感的通道,通过聚合位置敏感图得到每个实例掩码;YOLACT[5]通过融合语义分割原型图和检测框,获取最终的掩码;SOLOv1-v2[6-7]在全卷积特征图上输出相应类别概率直接输出实例掩码;BlendMask[8]则通过 blender 模

收稿日期:2022-11-16
*基金项目:国家自然科学基金(62072024, 61971290)、建大杰青(JDJQ20220805)、北京市教委科技计划面上项目(KM202110016001, KM202210016002)资助

块融合高层和底层的语义信息来提取更准确的实例分割特征。两阶段的实例分割算法代表算法如 Mask R-CNN[9]，它是在 Faster R-CNN[10] 基础上先生成候选框，然后在检测分支添加一个并行的语义分割分支用于掩码预测；PANet[11] 基于 Mask R-CNN 引入了自上而下的路径并自适应融合了不同层次的特征信息。MS R-CNN[12] 通过添加 Mask IOU 分支来预测掩码并给其打分以提升模型实例分割性能。

然而，这些算法普遍存在一个问题，即边界分割精度不高，尤其针对缺少细节信息的小目标来说更是如此。针对这一问题。2020 年 BMask R-CNN[13] 则将目标边缘信息加入 Mask R-CNN 中用于监督网络以增强掩码预测。RefineMask[14] 则利用边缘信息和语义分割信息逐阶段细化 Mask R-CNN 生成的粗糙掩码边缘，其在 ROI Align[9]（感兴趣区域对齐）操作之后逐渐扩大预测大小，并逐步融合低层次高分辨率的特征，尽管该方法在一定程度上改善了边界细节损失的问题，但仍有进一步改进的空间。

因此，本文提出了一种基于改进 Mask R-CNN 模型的建筑工地实例分割算法，该算法首先在 Mask 分支采用 RefineMask 中的多阶段逐步融合细粒度特征的方式以改善掩码质量，其次，结合空间及通道注意力机制提出了全局上下文通道注意力（global context channel attention，GCCA）机制以更好的融合细粒度特征。最后，将本文模型与常用实例分割算法以及不同注意力机制在 MOCS[15] 和 MS COCO[16] 数据集上进行大量对比实验，结果表明，该算法可以有效提高在 2 个数据集上的检测及分割精度。

1　方法和原理

本节将主要介绍本文提出的改进 Mask R-CNN 模型，图 1 为模型的整体架构示意图。其中 GCCA 模块，SFM 模块以及 BAR 模块分别如图 2～4。其基于强大的特征金字塔网络[17]（feature pyramid network，FPN），利用多阶段融合细粒度特征进行高质量的实例分割，其中语义头以特征金字塔中分辨率最高的特征图作为输入，然后经过 GCCA 模块以实现更细致的语义分割，该注意力模块结合空间和通道注意力，建模全局上下文特征以及通道依赖关系以自适应细粒度特征细化。掩码头部分逐步融合语义特征，并从细粒度特征中提取语义掩码，增大特征空间的大小，以实现更准确的实例掩码预测。为了预测出更清晰的边界，该方法引入基于边界感知的优化策略，更加关注边界区域。

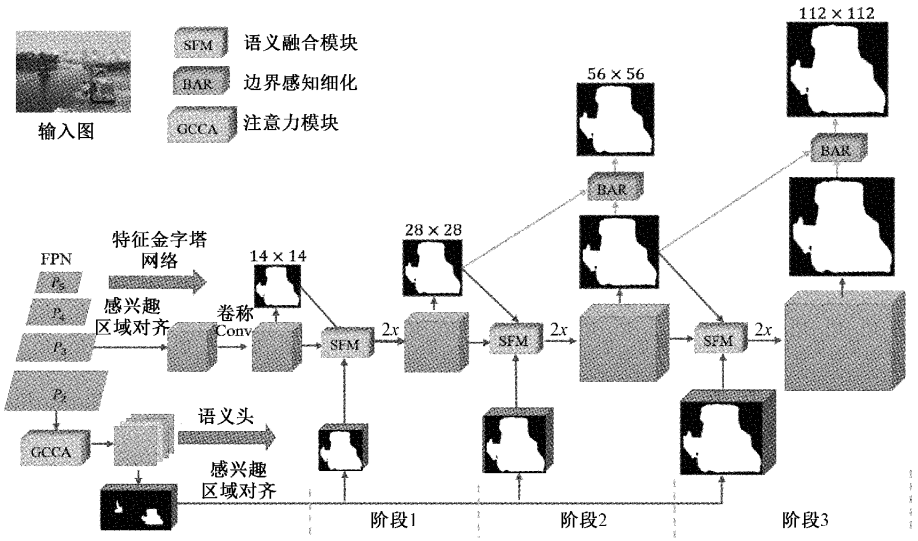


图 1　改进的 Mask R-CNN 模型架构

1.1　语义头

语义头（semantic head）是一个连接到 P_2（FPN 的最高分辨率特征图）上的全卷积神经网络。它通过 4 个卷积层来提取整个图像的语义特征，并使用一个二值分类器来预测各像素属于前景的概率。在二值交叉熵损失的监督下，预测整个图像的高分辨率语义掩码。我们将细粒度特征定义为语义特征和语义掩码的结合，这些细粒度特征用于补充掩码头中丢失的细节，从而实现高质量的掩码预测。

1.2　全局上下文通道注意力（GCCA）模块

为了更好的融合细粒度特征，本文提出了 GCCA 模块，添加在 P_2 和语义头之间，该模块受 GCnet[18] 以及 ECANet[19] 的启发，结合空间和通道注意力机制，首先通过全局上下文模块聚合全局特征，再利用一维卷积实现无降维的局部通道交互，既充分利用了空间上下文信息，又避免了降维带来的通道信息损失，GCCA 结构图如图 2 所示。其中 GAP 表示全局平均池化，1Dconv 表示一维卷积，$k=\psi(C)$ 如式(8)，σ 表示 sigmoid 函数。

由于对不同查询位置的注意图几乎相同[18]，全局上下文模块仅通过计算一个的全局注意图来聚合全局上下文特征，并对所有查询位置共享这个全局注意图，该模块通过一个 1 * 1 卷积和一个 softmax 函数实现，其公式如式(1)所示。

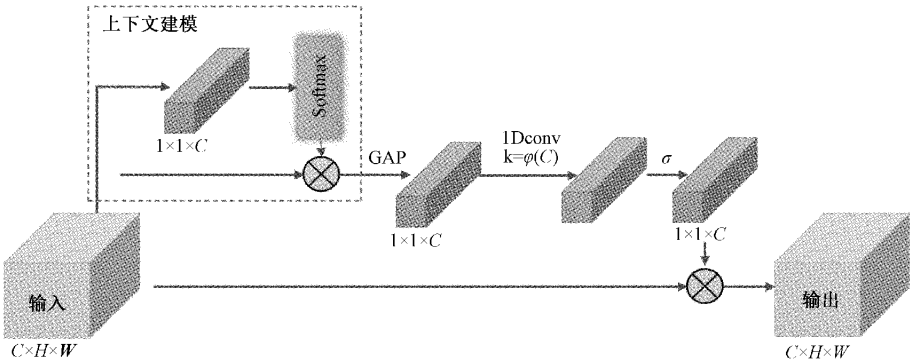


图 2 GCCA 结构示意图

$$y=\sum_{i=1}^{N_p}\frac{e^{\boldsymbol{W}_q x_i}}{\sum_{m=1}^{N_p}e^{\boldsymbol{W}_q x_m}}x_i \tag{1}$$

其中,$x_i(i=1,\cdots,N_p)$ 为输入特征图的每一个像素点,N_p 为特征图的像素点总数,通常为 $H\times W$,$\boldsymbol{W}_q$ 代表线性变换矩阵,y 为全局上下文模块输出的特征变量。

ECANet 提出通道降维操作会使得通道与其权重不能直接对应,从而不能有效学习通道注意。因此为了建模通道注意力,本文避免 SENet[20] 中的降维操作,并且使用卷积核为 k 的一维卷积自适应的实现通道的局部交互,具体来说,给定不降维的聚合特征,通道注意可以通过下式学习:

$$w=\sigma(\boldsymbol{W}y) \tag{2}$$

其中,$\boldsymbol{W}$ 为 $C\times C$ 的参数矩阵,σ 表示 sigmoid 函数。

而为了有效实现局部跨通道互动本文用一个波段矩阵 $\boldsymbol{W}_k$ 来学习信道注意力,$\boldsymbol{W}_k$ 公式如下:

$$\boldsymbol{W}_k=\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k-1} & \cdots & w^{C,C} \end{bmatrix} \tag{3}$$

可以看出 $\boldsymbol{W}_k$ 仅包含 $k\times C$ 个参数,将其代入式(2)中,得到通道注意如下式:

$$w_i=\sigma\left(\sum_{j=1}^{k}w_i^j y_i^j\right),y_i^j\in\Omega_i^k \tag{4}$$

其中,Ω_i^k 表示 y_i 的 k 个相邻通道的集合。可以看出 y_i 的权值只考虑 y_i 与其 k 个近邻之间的相互作用。

一种更有效的方法是让所有通道共享相同的学习参数,则变为下式:

$$w_i=\sigma\left(\sum_{j=1}^{k}w^j y_i^j\right),y_i^j\in\Omega_i^k \tag{5}$$

所以,式(5)可以通过卷积核大小为 k 的快速一维卷积实现,即:

$$w=\sigma(C1D_k(y)) \tag{6}$$

其中,$C1D$ 表示一维卷积,可以看出,该模块仅涉及 k 个参数。

鉴于组卷积[21]中高维(低维)信道在固定分组数量的情况下涉及长(短)卷积,即通道维度 C 和卷积核大小 k 成比例,但以线性函数为特征的关系过于有限,又由于通道维度通常设置为 2 的幂,所以通道 C 可以用下面的非线性函数表示:

$$C=\phi(k)=2^{(\gamma\cdot k-b)} \tag{7}$$

因此给定通道数 C,可以自适应的确定内核大小 k:

$$k=\psi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{odd} \tag{8}$$

其中,$|t|_{odd}$ 表示距离 t 最近的奇数,参考 ECANet,我们在所有实验中设置 γ 和 b 分别为 2 和 1。

因此,整个 GCCA 模块可以由下式表示:

$$z=x\otimes\sigma\left(C1D_k\left(\Delta\left(\sum_{i-1}^{N_p}\frac{e^{\boldsymbol{W}_q x_i}}{\sum_{m=1}^{N_p}e^{\boldsymbol{W}_q x_m}}x_i\right)\right)\right) \tag{9}$$

其中,z 为输出特征图,Δ 为全局平均池化函数,$\otimes$ 为逐元素相乘。

1.3 掩码头

在掩码头部,通过 14×14 大小的 ROI Align 操作提取的特征首先被送入 2 个 3 * 3 卷积层以生成实例特征,然后和 Mask R-CNN 一样,使用 1 * 1 卷积层来预测实例掩码,此时掩码图大小仅为 14×14,这个粗糙的掩码作为后期细化阶段的初始掩码。然后通过 1 个多阶段的细化过程,以迭代的方式细化掩模,每个阶段的输入由 4 个部分组成,即由上一阶段获得的实例特征和实例掩码,由语义头输出汇聚的语义特征和语义掩码。再使用语义融合模块(semantic fusion module, SFM)来集成这些输入,融合的特征被放大到更高的空间尺寸。最后通过迭代运行这个细化过程,输出分辨率高达 112×112 的高质量实例掩码。其中,SFM 模块如图 3 所示,其中 d 代表膨胀因子,它将上面提到的每个阶段的 4 个输入部分连接起来,并且使用 1 * 1 卷积层来融合这些特征并降低通道维数,然后使用 3 个并行的以不同膨胀因子的 3 * 3 膨胀卷积以融合单个神经元周围的信息,同时保持局部细节,最后,实例掩码和语义掩码再次与融合特征进行拼接,为后续预测提供指导。

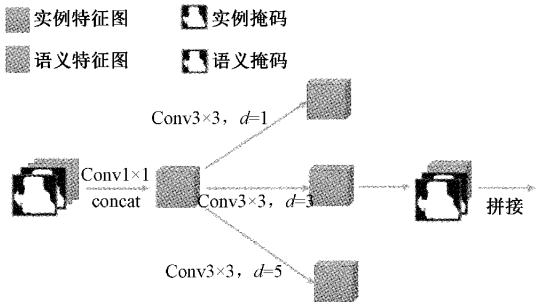


图 3　SFM 模块

同时，为了准确预测边界，本文参考 RefineMask 算法进行边界细化操作。首先，利用拉普拉斯边缘检测算子计算边界区域，若第 T 阶段的二进制实例掩码为 $\boldsymbol{M}^T$，则边界区域 $\boldsymbol{B}^T$ 为：

$$\boldsymbol{B}^T = \boldsymbol{M}^T * \boldsymbol{RC}_d \tag{10}$$

其中，$\boldsymbol{RC}_d$ 表示拉普拉斯边缘检测算子，$*$ 代表卷积操作，$d = 1,2$。当 $d = 1$ 时，边界宽度为 1，则：

$$\boldsymbol{RC}_1 = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{11}$$

当 $d = 2$ 时，

$$\boldsymbol{RC}_1 = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 24 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix} \tag{12}$$

在训练阶段，边界细化策略只在后两个输出大小为 56×56 和 112×112 的阶段中进行，只在特定的边界区域使用监督信号进行训练。这些区域由真实掩模和前一阶段的预测掩模决定：

$$\boldsymbol{R}^T = f_{up}(\boldsymbol{B}_G^{T-1} \cup \boldsymbol{B}_P^{T-1}) \tag{13}$$

其中，f_{up} 表示尺度因子为 2 的双线性插值上采样，$\boldsymbol{B}_G^{T-1}$ 表示第 $T-1$ 阶段真实掩码的边界区域，$\boldsymbol{B}_P^{T-1}$ 表示第 $T-1$ 阶段的预测掩码的边界区域，$\cup$ 表示两个边界区域的并集。则训练的第 T 阶段输出大小为 $S_T \times S_T$ 的损失为：

$$L^T = \frac{1}{\delta_n} \sum_{n=0}^{N-1} \sum_{i=0}^{S_T} \sum_{j=0}^{1S_T\ 1} R_{nij}^T l_{nij} \tag{14}$$

其中，N 为实例数目，l_{nij} 为实例 n 的 (i,j) 处的交叉熵损失，且

$$\delta_n = \sum_{n=0}^{N\ 1} \sum_{i=0}^{S_T} \sum_{j=0}^{1S_T\ 1} R_{nij} \tag{15}$$

在推理阶段，对于每个实例，第 1 阶段输出大小为 28×28 的粗糙掩码 $\boldsymbol{M}^1$ 和边界掩码 $\boldsymbol{B}_P^1$，则第 T 阶段的输出可以按照下式推理：

$$\boldsymbol{M}'^1 = \boldsymbol{M}^1 \tag{16}$$

$$\boldsymbol{M}'^T = f_{up}(\boldsymbol{B}_P^{T-1}) \otimes \boldsymbol{M}^T + (1 - f_{up}(\boldsymbol{B}_P^{T-1})) \otimes f_{up}(\boldsymbol{M}'^{T\ 1}) \tag{17}$$

其中第 2 阶段的推理如图 4 所示。其中 f_{up} 表示尺度因子为 2 的双线性插值上采样，$\boldsymbol{B}_P^1$ 表示第 1 阶段预测掩码的边界区域，$\boldsymbol{M}^2$ 表示第 2 阶段的真实掩码。

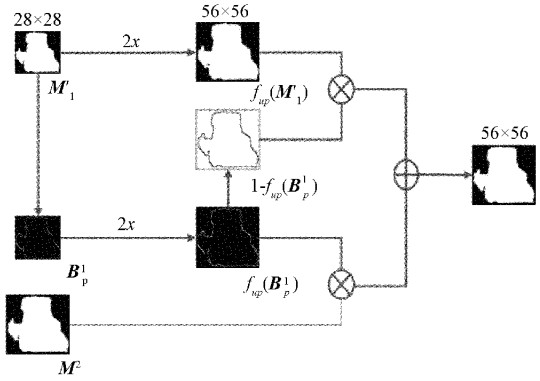


图 4　BAR 模块网络架构

2　实验分析

2.1　实验环境及参数配置

本文实验硬件环境皆为 NVIDIA Tesla V100-SXM2 32 GB 显存的 GPU，且所有实验皆基于 1 块 GPU。本文训练时采用 ResNet-50 预训练模型，除了 BlendMask 模型，所有实验皆基于 mmdetection 目标检测工具箱，采用 1.7.0Pytorch 框架，Python3.7 版本，并采用随机梯度下降(stochastic gradient descent，SGD)算法来优化浅层特征增强网络的权重，其动量设置为 0.9，权重衰减因子为 0.000 1，基础学习率为 0.002 5，最大训练 epoch 为 12，并且采用学习率热身算法来保证模型训练的稳定性，在初始的 500 次迭代中学习率逐渐增加，在第 8 和 11 个 epoch 时降低学习率。BlendMask 实验基于 detectron2 工具箱，采用 Pytorch-1.9.0 框架，Python3.8 版本，其动量和权重衰减因子和上述实验一样，基础学习率为 0.001 25，最大迭代次数为 720 000，在迭代第 480 000 和 640 000 次的时候降低学习率。

2.2　数据集及评价指标

MOCS 数据集是清华大学提出的一个大规模的、公开的用于检测建筑工地物体的图像数据集，该数据集包含从 174 个不同建筑工地收集的 41 668 幅图像，包含 13 个类别，其中用于训练的有 19 404 幅图像，用于验证的有 4 000 幅图像，用于测试的有 18 264 幅图像。MS COCO 数据集包含 80 个类别，由 115 000 张用于训练的图像(train-2017)和 5 000 张用于验证的图像(val-2017)组成以及 20 000 张测试图像组成，而由于这两个数据集的测试图像都没有标注，所以本文仅使用它们的训练集培训模型，用验证集报告消融实验结果。本文采用的评价指标是不同交并比(intersection over union，IOU)和不同目标尺寸的均值平均精度(mean average precision，mAP)，根据 IOU 阈值的不同，划分为 AP (IOU 阈值为 0.5～0.95 的 AP)，AP_{50}

(IOU 阈值为 0.5 的 AP),AP_{75}(IOU 阈值为 0.75 的 AP);根据目标尺寸的不同,划分为 AP_s(检测小目标的 AP),AP_m(检测中等目标的 AP),AP_l(检测大目标的 AP)。此外,算法的检测分割时间也是评估算法性能的重要指标,本文采用评估算法速度的指标是 FPS(每秒传输帧数),具体来说,随机挑选测试集中的 1 000 张图片,利用上述几种算法分别进行推理测试,得到每个算法测试这 1 000 张图片的总时间 t_{sum},则算法的运行速度可以用以下公式计算:

$$Runtime = t_{sum}/1\ 000\ \text{fps} \tag{18}$$

2.3 实验分析

本文先在 MOCS 数据集上进行实验,然后在经典的 MS COCO 数据集验证评估以证明本文算法的有效性,并且与其他算法进行了对比实验,主要包括一阶段实例分割算法 Yolact 以及 BlendMask,两阶段实例分割算法 Mask R-CNN 以及 RefineMask,为了公平比较,所用骨干网络均为 ResNet-50,实验结果如表 1、2 所示。表 1 为不同算法在建筑工地数据集 MOCS 上的检测和分割的性能对比,经典高性能实例分割 Mask R-CNN 在 11.5 FPS 的运行速度下,检测和分割的平均精度为 48.1%和 40.1%,而本文算法在增加很少的时间成本情况下,检测和分割的平均精度就达到了 50.5%和 47.7%,分别提高了 2.4%和 7.6%;虽然 BlendMask 在检测的平均精度超越了本文算法 0.2%,但是在分割的精度确远不如本文算法,尤其是小目标的分割精度仅达到 12.8%,与本文算法的 21.6%相差 8.8%,运行速度也仅达到 8.4 FPS。而相较于 RefineMask,不管是检测还是分割方面,加入注意力机制的本文算法在精度和速度上都带来了有效的提升。综上所述,本文的实例分割算法具有最好的性能。

表 1 不同算法在 MOCS 数据集上检测和分割的性能对比

模型	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_s	AP^{bb}_m	AP^{bb}_l	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_s	AP^{seg}_m	AP^{seg}_l	Runtime/FPS
Yolact	38.4	64.8	39.8	8.3	25.6	50.4	32.7	55.9	33.0	3.5	18.3	48.9	14.4
BlendMask	**50.7**	73.3	**56.2**	17.2	**39.9**	61.3	46.2	70.8	49.8	12.8	32.9	58.7	7.3
Mask R-CNN	48.1	72.7	52.8	17.4	37.9	59.4	40.1	68.1	42.0	15.7	32.6	50.5	11.5
RefineMask	49.6	73.7	54.3	20.4	38.5	61.2	47.1	71.4	51.0	19.9	35.7	58.9	9.2
本文算法	50.5	**74.0**	55.1	**22.1**	39.3	**61.4**	**47.7**	**71.9**	**51.6**	**21.6**	**36.5**	**58.9**	9.2

注:bb:检测,seg:分割

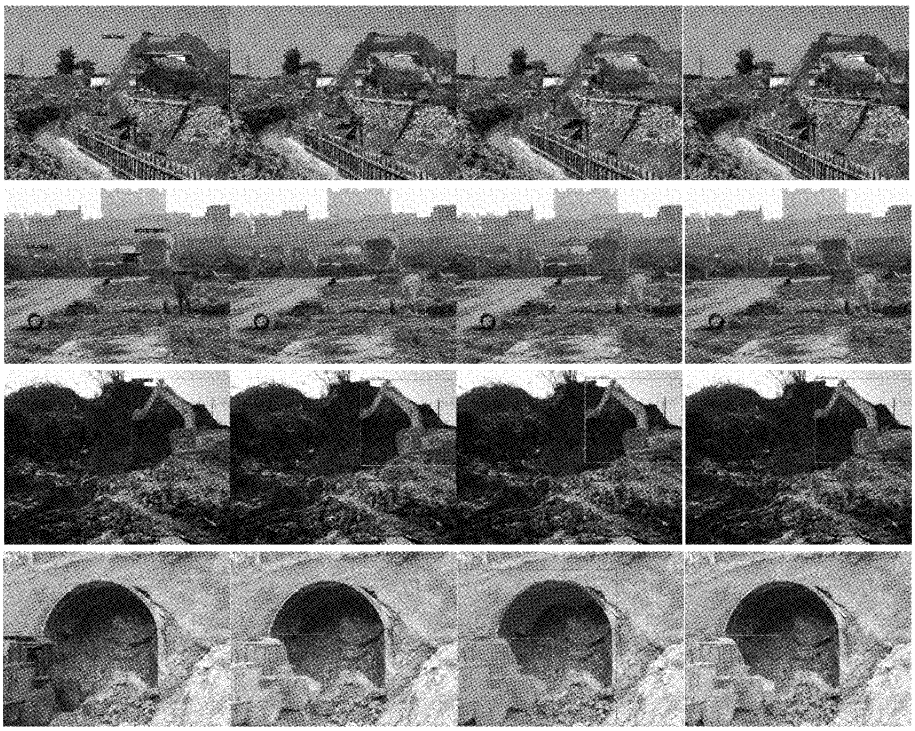
表 2 不同算法在 MS COCO 数据集上检测和分割的性能对比

模型	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_s	AP^{bb}_m	AP^{bb}_l	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_s	AP^{seg}_m	AP^{seg}_l	Runtime/FPS
Yolact	27.4	47.1	28.1	11.8	29.5	38.8	25.4	43.1	25.8	7.7	26.9	40.7	17.5
BlendMask	**39.3**	58.6	**43.2**	19.2	**43.2**	51.6	35.4	55.6	37.9	16.6	38.7	51.5	8.4
Mask R-CNN	38.1	58.6	41.4	21.8	41.7	49.4	34.6	55.5	36.8	16.0	37.2	50.8	11.4
RefineMask	38.9	59.7	43.2	21.8	42.1	51.5	37.4	57.5	40.1	19.6	40.1	51.9	8.7
本文算法	39.1	**59.8**	42.7	**22.2**	42.4	**51.7**	**37.4**	**57.6**	**40.5**	**19.7**	**40.2**	**52.0**	8.9

注:bb:检测,seg:分割

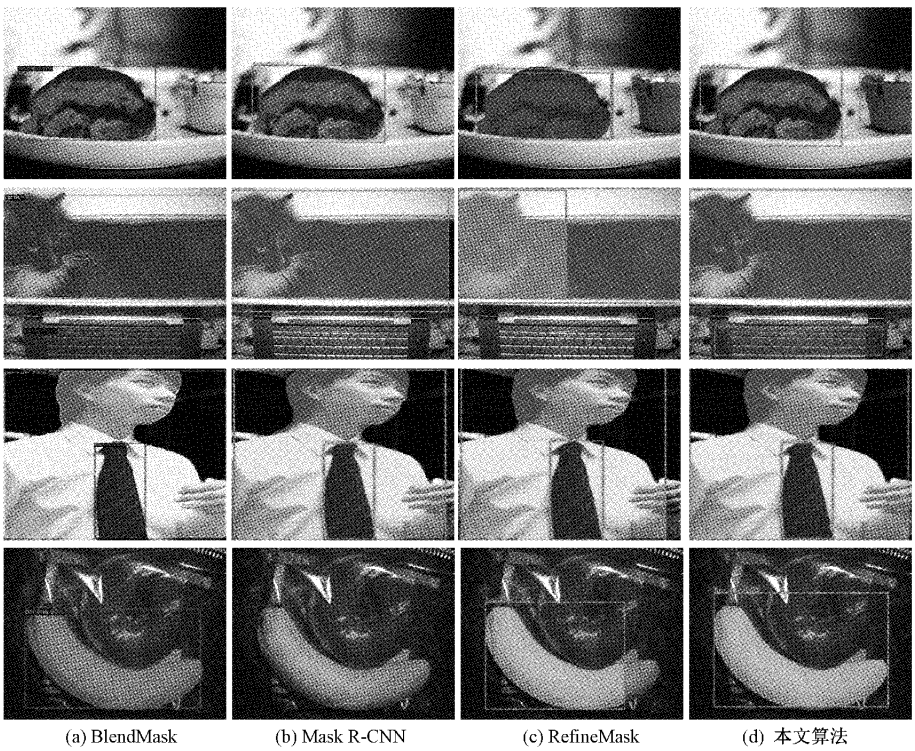
为了进一步验证本文算法在建筑工地数据集 MOCS 实例分割精度的有效提升,将本文算法与 BlendMask、Mask R-CNN 以及 RefineMask 算法在建筑工地数据集 MOCS 的可视化检测效果进行对比,结果如图 5 所示。其中从上到下为检测的不同图片,从左到右依次是 4 种算法的可视化结果图,可以看出,本文算法的检测分割效果最佳,其他算法皆出现漏检、错检、以及分割边缘质量不高的问题。例如,对比第 1 张检测图,只有本文算法能够较为正确的检测并分割图,BlendMask 检测分割混凝土搅拌机不够准确,边缘部分还包含了其他目标;Mask R-CNN 则错误检测到了挖掘机下面的两个人;而 RefineMask 出现了重叠框。而对于第 2、3、4 张图片来说,只有本文算法能够较为准确并分割图像中的人,且边缘质量最好。

此外,为了证明本文算法的泛化性能,将本文算法在常用目标检测 MS COCO 数据集上进行实验,并与其他算法进行比较,实验结果如表 2 所示。从表 2 的实验结果可以看到,本文算法在 MS COCO 数据集上也能带来有效的提升。例如,相较于实例分割高性能算法 Mask R-CNN,本文算法在检测和分割平均精度方面分别提升了 1.0%和 2.8%,而只增加了一点时间成本,证明本文算法在 mask 头部的优化是非常有效的。尽管 BlendMask 模型在检测的平均精度稍高于本文算法,但是其在小目标的检测精度以及所有分割精度都远低于本文算法。同样,本文也在 MS COCO 数据集上进行了可视化效果对比,实验结果如图 6 所示,从上到下依次是 COCO 验证集里 4 张不同的图片,从左到右依次是 BlendMask、Mask R-CNN、RefineMask 以及本文改进 Mask R-CNN 的可视化检测图。从图中可以看出,本文算法具有最好的检测分割效



(a) BlendMask　(b) Mask R-CNN　(c) RefineMask　(d) 本文算法

图 5　不同算法在 MOCS 数据集上的可视化检测图对比



(a) BlendMask　(b) Mask R-CNN　(c) RefineMask　(d) 本文算法

图 6　不同算法在 MS COCO 数据集上的可视化检测图对比

果。例如，只有本文算法正确检测到了第 1 张图片的食物和碗，其他算法均出现漏检、错检和重复检测的问题。而针对第 3 张图片，虽然 4 种算法都正确检测到了图片中的目标，但是从边缘质量和分割的角度来看，本文算法最为精准。

最后，为了验证本文提出的 GCCA 注意力机制的有效性，本文将文中的 GCCA 模块依次替换成 SENet、

ECANet、GCnet 注意力机制在 MOCS 数据集上进行实验,实验结果如表 3 所示。其中 RefineMask 不加注意力机制的基本模型,本文算法为在 RefineMask 的 Mask 分支上添加了 GCCA 注意力机制。可以看出,本文算法相较于原始 RefineMask 模型在检测和分割精度上都有所提升,而且与其他注意力机制相比,本文的注意力机制表现最佳。具体来说,相较于 RefineMask 模型,加入 GCCA 注意力机制的 RefineMask 模型即本文算法检测精度和分割精度分别提升了 0.9%和 0.6%,其中小目标的检测和分割精度分别提升了 1.7%和 1.7%。

表 3 基于不同注意力机制的 RefineMask 在 MOCS 数据集上的平均精度对比

模型	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{s}	AP^{bb}_{m}	AP^{bb}_{l}	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_{s}	AP^{seg}_{m}	AP^{seg}_{l}
RefineMask	49.6	73.7	54.3	20.4	38.5	61.2	47.1	71.4	51.0	19.9	35.7	58.9
RefineMask(+SE)	49.8	74.0	54.9	19.6	38.5	61.2	47.2	72.0	51.2	19.4	36.4	58.6
RefineMask(+ECA)	49.8	73.8	53.3	20.4	39.2	61.0	47.3	71.8	51.2	19.8	37.0	58.5
RefineMask(+GC)	50.1	74.0	55.1	20.8	39.0	61.1	47.2	71.8	51.3	21.1	36.5	58.7
本文算法	**50.5**	**74.0**	**55.1**	**22.1**	**39.3**	**61.4**	**47.7**	**71.9**	**51.6**	**21.6**	**36.5**	**58.9**

3 结 论

为了实现建筑工地的安全监控,解决当前实例分割算法分割质量差的问题,本文提出了改进的 Mask R-CNN 模型:首先本文通过多阶段融合高分辨率细粒度特征并扩大掩码大小以细化掩码质量,其次提出了结合空间和通道的全局上下文通道注意力机制即 GCCA 以获取更丰富有效的特征信息,最后通过与边缘检测算子生成的边界部分进行融合进一步细化掩码质量,提升了检测和分割精度。并且通过在 MOCS 数据集以及常用 MS COCO 数据集上与其他算法以及不同注意力机制的大量对比实验证明了本文算法的一致有效性和泛化性,有助于深度学习方法在建筑工地安全监控中的应用。由于本文方法建立在两阶段算法 Mask R-CNN 模型上,在检测速度方面不如一些一阶段算法,因此在后续工作中需要考虑本文算法的实用性,可以把本文的思想迁移到一阶段算法中,深入研究在建筑施工场景下的实例分割应用。

参考文献

[1] 李衍照,于镭,田金文. 基于改进 YOLOv5 的金属焊缝缺陷检测[J]. 电子测量技术,2022,45(19):70-75.

[2] 黄聪,杨珺,刘毅,等. 基于改进 DeeplabV3+的遥感图像分割算法[J]. 电子测量技术,2022,45(21):148-155.

[3] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pat-tern Recognition, 2015: 3431-3440.

[4] DAI J, HE K, LI Y, et al. Instance-sensitive fully convolutional networks[C]. European Conference on Computer Vision. Springer, Cham, 2016: 534-549.

[5] BOLYA D, ZHOU C, XIAO F, et al. Yolact: Real-time instance segmentation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9157-9166.

[6] WANG X, KONG T, SHEN C, et al. Solo: Segmenting objects by locations[C]. European Conference on Computer Vision. Springer, Cham, 2020: 649-665.

[7] 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.

[8] CHEN H, SUN K, TIAN Z, et al. Blendmask: Top-down meets bottom-up for instance segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8573-8581.

[9] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[10] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[11] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[12] HUANG Z, HUANG L, GONG Y, et al. Mask scoring r-cnn[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 6409-6418.

[13] CHENG T, WANG X, HUANG L, et al. Boundary preserving mask r-cnn[C]. European Conference on Computer Vision. Springer, Cham, 2020: 660-676.

[14] ZHANG G, LU X, TAN J, et al. Refinemask: Towards high-quality instance segmentation with fine-

grained features[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6861-6869.

[15] AN X H, ZHOU L, LIU Z G, et al. Dataset and benchmark for detecting moving objects in construction sites[J]. Automation in Construction, 2021, 122: 103482.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]. European Conference on Computer Vision. Springer, Cham, 2014: 740-755.

[17] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[18] CAO Y, XU J, LIN S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 1971-1980.

[19] WANG Q, WU B, ZHU P, et al. Eca-net: Efficient channel attention for deep convolutional neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[21] ZHANG T, QI G J, XIAO B, et al. Interleaved group convolutions[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 4373-4382.

作者简介

宋艳飞，硕士研究生，主要研究方向为计算机视觉、深度学习目标检测和实例分割等。

E-mail: yfssdau@163.com

王恒友(通信作者)，博士，副教授，硕士生导师，主要研究方向为计算机视觉图像分析、稀疏表示、低秩矩阵理论及其图像重构等。

E-mail: wanghengyou@bucea.edu.cn