智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于反馈注意力机制和上下文融合的非模式实例分割

董俊杰， 刘华平， 谢珺， 续欣莹， 孙富春

引用本文:
董俊杰, 刘华平, 谢, 等. 基于反馈注意力机制和上下文融合的非模式实例分割[J]. 智能系统学报, 2021, 16(4): 801–810.
DONG Junjie, LIU Huaping, XIE Jun, et al. Feedback attention mechanism and context fusion based amodal instance segmentation[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 801–810.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202007042

DOI: 10.11992/tis.202007042
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210507.1642.002.html

基于反馈注意力机制和上下文融合的非模式实例分割

董俊杰[1], 刘华平[2], 谢珺[1], 续欣莹[3], 孙富春[2]
（1. 太原理工大学 信息与计算机学院, 山西 晋中 030600; 2. 清华大学 智能技术与系统国家重点实验室, 北京 100084; 3. 太原理工大学 电气与动力工程学院, 山西 太原 030024）

摘　要: 非模式实例分割是最近提出的对实例分割的扩展, 其任务是对每个对象实例的可见区域和被遮挡区域都进行预测, 感知完整的物理结构和语义概念。在预测对象被遮挡部分的形状和语义时, 往往由于特征表示的识别能力不够和对上下文信息缺乏而导致对遮挡区域预测欠拟合甚至错误。针对这个问题, 提出一个上下文注意模块和反馈注意力机制的特征金字塔结构, 引入反馈连接进行再学习。该方法能够有效捕获全局语义信息和精细的空间细节, 通过在 COCO-amodal 数据集训练和验证, 非模式实例分割掩码平均精确率从 8.4% 提高到 14.3%, 平均召回率从 16.6% 提高到 20.8%。实验结果表明, 该方法能够显著提高对物体被遮挡部分预测的准确率, 有效解决欠拟合问题。

关键词: 非模式实例分割; 遮挡预测; 反馈连接; 注意力机制; 上下文信息; 深度学习; 神经网络; 计算机视觉
中图分类号: TP183　　**文献标志码**: A　　**文章编号**: 1673−4785(2021)04−0801−10

中文引用格式: 董俊杰, 刘华平, 谢珺, 等. 基于反馈注意力机制和上下文融合的非模式实例分割 [J]. 智能系统学报, 2021, 16(4): 801–810.
英文引用格式: DONG Junjie, LIU Huaping, XIE Jun, et al. Feedback attention mechanism and context fusion based amodal instance segmentation[J]. CAAI transactions on intelligent systems, 2021, 16(4): 801–810.

Feedback attention mechanism and context fusion based amodal instance segmentation

DONG Junjie[1], LIU Huaping[2], XIE Jun[1], XU Xinying[3], SUN Fuchun[2]
(1. College of Information and Computer, Taiyuan University of Technology, Jinzhong 030600, China; 2. State Key Lab. of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China; 3. College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: Recently, model instance segmentation has been proposed as an extension of instance segmentation to predict the visible and occluded areas of each object instance and perceive the complete physical structure and semantic concepts. When the shapes and meanings of occluded objects are being predicted, underfitting or even wrong results are obtained in the occlusion prediction due to the insufficient recognition capability of feature representation and the lack of contextual information. To solve this problem, this paper proposes a contextual attention module and feature pyramid structure of feedback attention mechanism and introduces feedback connections for relearning. The proposed method can effectively capture global semantic information and fine spatial details. Through training and verification in the COCO-amodal dataset, the average precision of the amodal instance segmentation mask increases from 8.4% to 14.3%, and the average recall rate increases from 16.6% to 20.8%. Experimental results show that this method can significantly improve the accuracy of occlusion prediction and effectively end underfitting.

Keywords: amodal instance segmentation; occlusion prediction; feedback connection; attention mechanism; context information; deep learning; neural network; computer vision

收稿日期: 2020−07−24.　**网络出版日期**: 2021−05−07.
基金项目: 山西省自然科学基金项目 (201801D121144, 201801D221190); 辽宁省科技厅机器人技术国家重点实验室联合基金项目 (2020-KF-22-06).
通信作者: 刘华平. E-mail: hpliu@tsinghua.edu.cn.

近年来, 图像分类[1-2]、目标检测[3-6]、语义分割[7-8]、实例分割[9-10]等视觉识别任务取得了巨大的进展。计算机视觉系统的性能在精度上越来越

接近于甚至超过人类水平。尽管如此，人类的视觉系统具有感知物体完整物理结构的能力，即使物体在部分遮挡甚至重度遮挡的情况下也能准确预测物体的形状，这种能力被称为非模式感知[11](即 amodal perception)，使得人类对物体的不可见的、被遮挡的部分进行推理，针对遮挡有一定的鲁棒性，从而仅在部分可见的情况下感知物体完整形状和语义概念。

在非模式实例分割[11]的任务中，amodal masks 定义为物体 visible masks 和 occlusion masks 的并集。预测 amodal mask 和 visible mask 可以加深对场景的理解；例如，非模式感知可以使自动驾驶汽车能够在视觉范围内推断出车辆和行人的整体形状，即使其中的一部分是看不见的，这能够显著降低碰撞的风险；还有机器人在拾取或放置对象时，需要知道对象是否被一个或多个其他对象所遮挡，通过对被遮挡部分的感知计算遮挡区域，然后指导机器人朝哪个方向移动或者将某些对象移除，帮助机器人获取感兴趣对象的完整结构和语义。

预测物体的不可见部分是非常具有挑战性的。一个计算机视觉系统如果要感知物体被遮挡、不可见部分的形状和语义概念，首先需要识别和定位这个物体，这涉及到了目标检测的技术；第二，需要从可见部分提供的线索去推断出物体被遮挡部分的最可能外观特征，并且为物体遮挡部分的像素进行标记，这涉及到语义分割的技术，与此同时在此过程中往往会遭受来自附近其他对象实例信息的误导；第三，实际上物体往往被多个不同的实例遮挡，这将导致被遮挡的对象会处于不同的深度顺序中，如何探索对象之间的相对深度顺序关系也给非模式感知带来很大的困难。

为了有效地完成非模式实例分割的任务，需要区分一些容易混淆的类别，并考虑不同外观的对象。我们可以粗略地将所有对象分为两类："Things"和"Stuff"，其中"Things"是感兴趣的对象且具有相对规范的形状，例如行人、汽车等，"Stuff"可以具有相对任意的范围，例如树木、墙壁等。因此，在像素级识别中，有必要提高特征表示的识别能力，充分融合上下文信息，捕获全局特征。

研究人员基于深度神经网络对目标分割提出了一系列方法，提升了分割算法的精度和灵活性。Jonathan 等[7]提出全卷积网络 (fully convolutional networks, FCN)，将 CNN 网络中的全连接层转化为卷积层，采用反卷积的方法进行上采样，并引入跳跃连接来改善上采样效果，实现对图像中的每个像素预测和分类的任务，但是 FCN 没有考虑像素之间的关系，导致分割精度不够；He 等[9]提出了一种高效的实例分割框架——Mask-RCNN，该算法通过特征金字塔 (feature pyramid network, FPN) 和 ROI Align 层，极大地提升了算法的精度，但仍然没有考虑到卷积特征的通道和空间的上下文依赖关系；Zhang 等[12]提出 SLN 模型，该方法完成了非模式实例分割的任务，引入一种新的表示方法即语义感知距离映射，根据物体的可见性将物体的不同区域放置在不同的层次上，进而对物体被遮挡部分预测，尽管实现了对不可见部分的感知，但是忽略了空间相关性，使算法缺乏全局上下文语义信息，很容易导致预测的欠拟合。

在识别物体的时候，人类的视觉感知通过反馈连接和注意力机制传递高级语义信息，选择性地增强和抑制神经元的激活[13]。然而非模式实例分割任务的特点是仅仅从对象的可见部分提供的线索来合理预测该对象的被遮挡部分，并且被遮挡部分的空间形状具有不确定性。通过反馈过程和注意力机制的学习建立丰富的全局上下文关系是非常有帮助的，从而根据已有的提示推断出对象最可能的外观特征。针对以上问题，受人类视觉系统的启发，为了提高非模式实例分割中特征表示的鉴别能力和充分聚合上下文信息，本文精心设计了一个反馈注意力机制的特征金字塔结构，以及在实例分割分支引入 Context Attention Module。

1　相关工作

1.1　实例分割

两阶段实例分割通常将此任务描述为"先检测后分割"。它们通常先检测到边界框，然后在每个边界框的区域中执行像素分割。FCIS[14]的主要思想是利用全卷积层预测一组位置敏感的输出通道，这些通道能够同时预测对象的类别、边界框和掩码；Mask R-CNN[9]建立在 Faster-RCNN 基础上，只需添加一个额外的 Mask 分支，并使用 ROI Align 层代替 ROI 池化操作，以提高精确度；在 Mask R-CNN 之后，PANet[15]引入了自底向上的路径扩充、自适应特征池化和全连接融合，以提高实例分割的性能；单阶段实例分割的研究受到了单阶段目标检测器的启发，如 YOLACT[10]将图像分割分成两个子任务：原型掩码 (prototype

masks) 的生成和预测每个实例掩码的掩码系数 (mask coefficients); 然后, 通过原型掩码和掩码系数的线性组合来生成实例掩码; TensorMask[16] 研究了在密集滑动窗口上的实例分割, 使用结构化的 4D 张量来表示空间域上的掩码; PolarMask[17] 提出使用极坐标表示对掩码进行编码, 并将逐像素掩码预测转换为距离回归。但这些方法都仅仅预测对象可见部分的掩码, 而没有对被遮挡、不可见部分进行预测。

1.2　非模式实例分割

图像非模式实例分割的研究才刚刚起步。Zhu 等[11] 基于 COCO 原始数据集为非模式实例分割提供了一个新的开创性数据集 COCO-amodal, 并提出了 AmodalMask 模型, 该模型对于提议的对象候选具有较高的召回率, 实现对物体不可见部分的推理; Zhang 等[12] 提出的 SLN 模型, 引入语义感知距离映射, 根据物体的可见性程度将物体的不同区域放置在不同的层次上来对物体不可见部分进行预测; Li 等[18] 提出的 amodal 实例分割方法, 主要通过迭代地将对象的边界框扩大到具有高热力图值的方向, 并且重新计算热力图来实现对物体被遮挡部分的预测; Follmann 等[19] 提出 ARCNN 模型, 该模型基于 Mask RCNN, 通过扩展其预测分支, 分为 amodal mask 预测分支和 visible mask 预测分支, 实现非模式实例分割; Ehsani 等[20] 试图通过生成对抗网络 (GAN) 来生成物体的不可见部分。

1.3　注意力机制

注意力机制可以建立长期依赖关系, 提高特征表达能力, 并且成为了许多具有挑战性任务的有效方法, 包括图像分类、语义和实例分割等。目前视觉识别领域主要包括 3 种注意力机制的方式: 通道注意力机制、空间注意力机制和混合注意力机制。Hu 等[21] 提出了 SENet 模型, 该模型通过在卷积网络的不同通道间探索各个通道的重要程度, 从而显式地建模通道之间的相互依赖关系, 自适应地重新校准通道的特征响应; Wang 等[22] 提出的空间注意力机制利用特征图中所有位置的加权和计算出一个特征的响应, 建立起像素之间的上下文依赖关系; Sanghyun 等[23] 提出的混合注意力机制通过融合通道和空间两种注意力机制, 充分挖掘全局语义信息, 极大地提升了图像识别的性能。

1.4　基于注意力机制的实例分割

实例分割试图为输入图像中的每个像素获取类和实例标签, 然而组成每个实例的不同像素点之间具有紧密联系, 同时有必要区分相同类别的不同实例对象, 由于注意力机制可以获取全局信息, 建立上下文长期依赖关系, 因此一些研究引入注意力机制来提高实例分割的性能。Cao 等[24] 提出了 GCNet, 引入 Context Modeling 和 Transform 模块从而建立一个轻量级的注意力机制模型, 进行全局上下文建模和捕获通道间的相互依赖, 并采用逐像素加法进行特征融合, 极大提高了实例分割的效率; FGNet[25] 是将一般实例分割和 Few-shot 学习范式结合起来, 在 Mask-RCNN 的各个关键组件中引入了不同的指导机制, 包括注意力指导 RPN, 关系指导检测器和注意力指导 FCN 用于指导基本实例分割的预测, 能够更好地适应类间泛化; Liu 等[26] 提出的 Cell r-cnn v3 模型属于生物医学图像领域的实例分割, 包含残差注意力特征融合机制、掩码质量预测分支, 前者促进实例分支中语义上下文信息的学习, 后者使每个目标的置信度得分与掩码预测的质量对齐, 提高了实例分割的性能。

以上方法, 模仿人类视觉系统引入注意力机制, 通过对全局上下文信息的建模, 捕获远程长期依赖关系, 使得实例分割性能显著提升; 然而对于非模式的实例分割任务而言, 对被遮挡、不可见区域的像素点进行语义预测则具有更大的挑战, 并且由于需要对物体被遮挡部分进行补全, 这将导致同一个像素点可能会分配多个标签, 对提取的特征表达能力和空间细节具有更高的要求。因此, 本文工作将注意力机制引入非模式实例分割的任务中, 引入反馈连接进行再学习, 建立丰富的上下文融合关系, 有选择地聚合全局信息, 显著提高了预测的精度, 极大地解决了分割欠拟合问题。

2　本文方法

2.1　反馈注意力机制的 FPN

鉴别特征表示是图像理解的关键, 它可以通过捕获远程上下文信息来获得。然而, 许多研究表明, 由传统 FCN(全卷积网络) 生成的局部特征可能导致“Things”和“Stuff”的错误分类; 与此同时, 基于特征金字塔 (feature pyramid network, FPN) 的分层检测方法虽然取得了很好的效果, 但是 FPN 仍然主要对局部特征进行建模, 并没有充分考虑全局上下文依赖关系。

在利用卷积神经网络对图像进行特征提取的

过程中，高层次特征的每个通道图都可以看作是一个特定于类的响应，不同的语义响应相互关联；通过挖掘通道图之间的相互依赖关系，强调特征图之间的相关性，提高特定语义的特征表示；与此同时，人脑是具有层级结构的，不仅执行从下层到上层的前馈过程，而且执行从上层到下层的反馈过程。因此，本文引入反馈过程和注意力机制来学习语义信息，首先构建一个通道注意模块[27]，它可以捕获通道维度远程上下文依赖关系，然后将通道注意模块嵌入到 FPN 中；第一轮获取的特征经过通道注意模块建立全局依赖关系后引入反馈过程进行再学习提取第二轮的特征，构成一个循环特征金字塔结构，并将两次提取的特征进行自适应加权，整体的结构如图 1 所示。

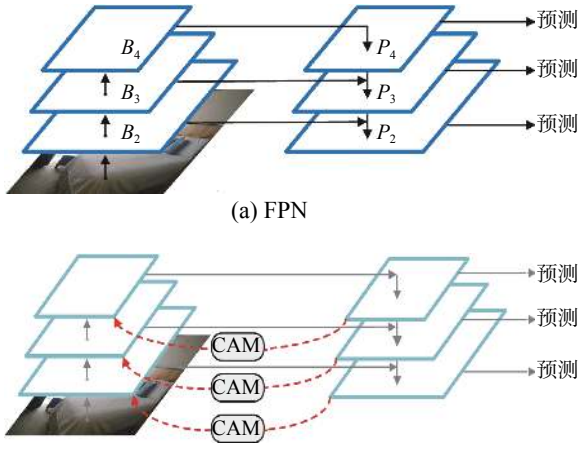

图 1　反馈注意力机制的 FPN 结构

Fig. 1　FPN with feedback attention mechanism

本文基于 ResNet101 引入反馈连接的 FPN。首先如图 1(a) 所示为特征金字塔结构 (FPN)，左侧为 ResNet101 主干网络，右侧为金字塔网络，主干网络提取图像特征，然后经过自顶向下和横向连接将每一阶段的特征图进行融合。例如，P_3 层经过 B_3 层和 P_4 层融合得到，P_4 层是 B_4 层经过 1×1 卷积和上采样得到，具有高级语义信息，而 B_3 层是位于主干网络的较浅层，具有底层的细节信息。

为了更充分合理地模拟人脑捕获高级语义信息，本文在原来 FPN 结构的基础上，添加了反馈连接和通道注意模块，构成反馈注意力机制 FPN 结构，简记为 FCAM-FPN，如图 1(b) 所示。将第一轮 FPN 提取的特征，经过通道注意模块建立远程依赖关系后得到的输出特征，采用反馈连接输入到主干网络得到第二轮 FPN 提取的特征，引入了注意力机制和反馈过程的二次学习，这样将捕获富含注意力的前后两次特征。

通道注意模块 (channel attention module, CAM) 的结构如图 2 所示。我们直接从原始特征图 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$ 计算通道注意图 $\boldsymbol{D}\in\mathbf{R}^{C\times C}$，通道注意图 $\boldsymbol{D}$ 相当于一个相关矩阵，它代表了 C 个通道之间的相关性程度。

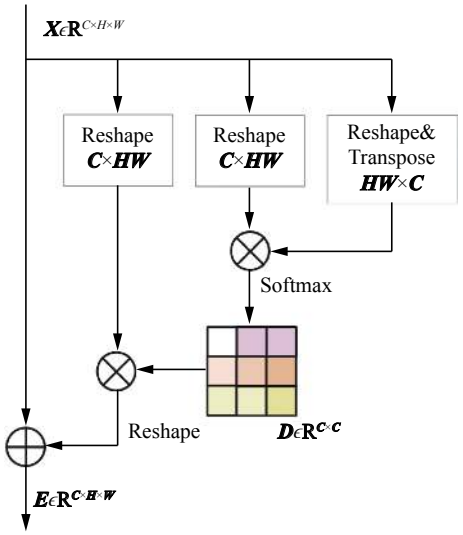

图 2　通道注意模块

Fig. 2　Channel attention module

具体来说，首先将原始特征图 $\boldsymbol{X}$ 变换为 $\boldsymbol{X}\in\mathbf{R}^{C\times N}$，这里 $N=H\times W$，然后在 $\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{T}}$ 之间执行矩阵乘法，最后应用 Softmax 函数来获得通道注意图 $\boldsymbol{D}\in\mathbf{R}^{C\times C}$：

$$\boldsymbol{d}_{ji}=\frac{\exp(\boldsymbol{X}_i\cdot\boldsymbol{X}_j)}{\sum_{i=1}^{C}\exp(\boldsymbol{X}_i\cdot\boldsymbol{X}_j)}$$

式中：$\boldsymbol{d}_{ji}$ 表示第 i 个通道和第 j 个通道之间的相关程度。与此同时，对通道注意图 $\boldsymbol{D}^{\mathrm{T}}$ 与原始特征图 $\boldsymbol{X}$ 变换后的 $\boldsymbol{X}\in\mathbf{R}^{C\times N}$ 执行矩阵乘法并重新变换为 $\mathbf{R}^{C\times H\times W}$，将此结果与原始特征图 $\boldsymbol{X}$ 执行逐元素求和运算，以获得最终输出特征 $\boldsymbol{E}\in\mathbf{R}^{C\times H\times W}$：

$$\boldsymbol{E}_j=\sum_{i=1}^{C}(\boldsymbol{d}_{ji}\boldsymbol{X}_i)+\boldsymbol{X}_j \tag{1}$$

式 (1) 表明，每个通道的最终特征与所有通道建立了紧密的相互依赖关系，通过跳跃连接使得输出特征表示为所有通道特征和原始特征的和，建立了通道特征图之间的上下文依赖关系模型。

接下来将从通道注意模块 (CAM) 得到的输出特征，采用反馈连接，重新输入到 ResNet101 主干网络中，进行第二次特征提取。假设 $\boldsymbol{B}_i$ 表示自底向上主干网络的第 i 级，$\boldsymbol{F}_i$ 表示自顶而下 FPN 操作的第 i 级，$\boldsymbol{R}_i$ 表示经过通道注意模块 (CAM) 后的输出特征，那么具有反馈过程的输出特征 $\boldsymbol{f}_i$ 定义为

$$\boldsymbol{f}_i=\boldsymbol{F}_i(\boldsymbol{f}_{i+1},\boldsymbol{x}_i),\ \ \boldsymbol{x}_i=\boldsymbol{B}_i(\boldsymbol{x}_{i-1},\mathrm{R}_i(\boldsymbol{f}_i))$$

式中：$\boldsymbol{x}_0$ 表示输入图像；$\boldsymbol{x}_i$ 表示主干网络经过多

个阶段生成的输出特征, f_i 表示经过自顶而下 FPN 的输出特征, $i=1,2,\cdots,S$, S 是主干残差网络的阶段数, 令 $f_{S+1}=0$。这样就使得 FPN 形成一个循环网络; 可将其展开成具有序列的网络结构, 那么此时输出特征 f_i 表示为

$$f_i^t = F_i^t(f_{i+1}^t, x_i^t),\quad x_i^t = B_i^t(x_{i-1}^t, R_i^t(f_i^{t-1}))$$

其中 $\forall i=1,2,\cdots,S$, $t=1,2,\cdots,T$, 令 $f_i^0=0$, 在本文的实验中令 $T=2$。最后基于反馈注意力机制的 FPN 结构的输出特征 q_i 表示为

$$q_i = f_i^{t-1} + \alpha f_i^t \tag{2}$$

其中 α 被初始化为 0, 并逐渐学习分配更多的权重。式 (2) 表明输出特征为经过通道注意模块前后两次 FPN 提取到特征 f_i^t 和 f_i^{t-1} 的自适应加权和; 这样既可以保留前一次 FPN 的信息, 又可以充分利用反馈注意力机制再学习到的特征表达, 建立起通道间的上下文关系, 提取更丰富的语义信息。

2.2　上下文注意模块

为了主动捕获像素之间的语义依赖关系, 引入了基于自注意机制的上下文注意模块[28](context attention module, CxAM)。对于非模式实例分割任务而言, 物体之间的位置关系复杂, 并且被遮挡部分的外观具有不确定性。基于这些特征, CxAM 编码了一个像素级别的远程上下文依赖关系, 能够自适应地关注更相关的区域。因此, CxAM 的输出特征将具有全局的语义信息, 并包含周围对象中的上下文关系。

CxAM 的结构如图 3 所示, 本文将 CxAM 模块仅用于 Mask head, 在语义分割时用来捕获像素之间的语义和位置依赖关系。图 3 中, 在给定的特征图 $F\in\mathbf{R}^{C\times H\times W}$ 的情况下, 分别使用 1×1 卷积层 W_m 和 W_n, 按式 (3) 计算得到转换后的特征图为

$$M = W_m^{\mathrm{T}}F \text{ 和 } N = W_n^{\mathrm{T}}F \tag{3}$$

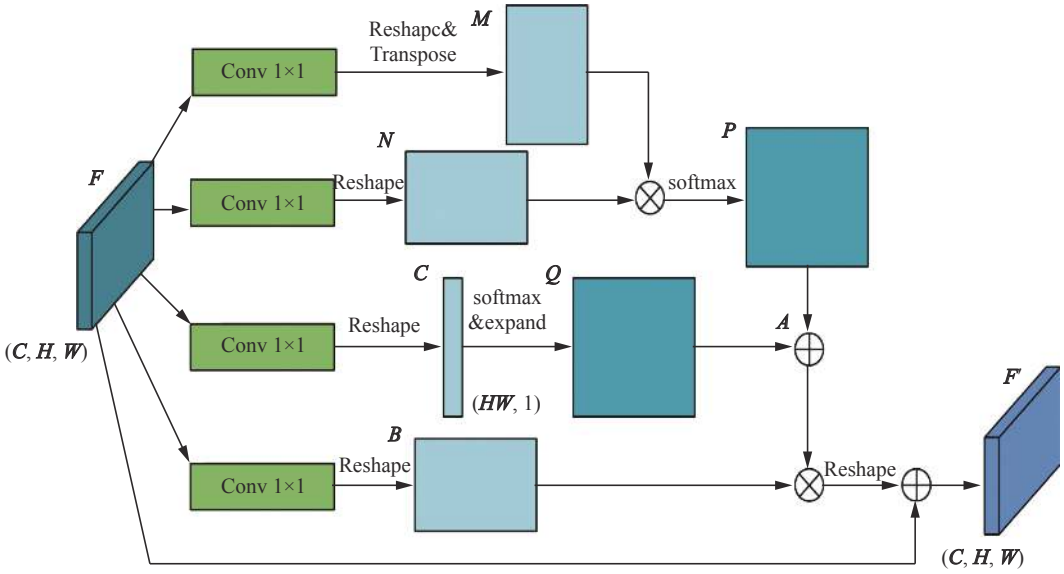

图 3　上下文注意模块

Fig. 3　Context attention module

这里, $\{M,N\}\in\mathbf{R}^{C'\times H\times W}$, 接着将 M 和 N 变换为 $\mathbf{R}^{C'\times K}$, 其中 $K=H\times W$。为了建立每个像素之间的依赖关系, 将 M^{T} 和 N 执行矩阵乘法, 然后应用 Softmax 函数来计算它的上下文注意力特征映射得到一个相关性矩阵 $P\in\mathbf{R}^{K\times K}$:

$$p_{ji} = \frac{\exp(M_i\cdot N_j)}{\sum_{i=1}^{K}\exp(M_i\cdot N_j)}$$

式中: p_{ji} 表示第 i 个像素与第 j 个像素之间的相关程度。与此同时, 另外一条路径将原始特征图 $F\in\mathbf{R}^{C\times H\times W}$ 经过一个 $1\times1\times1$ 卷积层后得到一个大小为 $1\times H\times W$ 的特征融合图, 将此特征融合图变换为 $C\in\mathbf{R}^{HW\times1}$, 将 C 再经过一个 Softmax 函数得到大小为 $HW\times1$ 的相关性矩阵, 然后复制自身大小变为 $Q\in\mathbf{R}^{K\times K}$, 将 Q 和 P 执行逐元素求和得到 A。

将原始特征图 $F\in\mathbf{R}^{C\times H\times W}$ 使用另外一个 1×1 卷积层 W_b 变换为 $B=W_b^{\mathrm{T}}F$, 这里 $B\in\mathbf{R}^{C\times K}$,将 B 和 A 执行矩阵相乘操作并将其结果变换为 $\mathbf{R}^{C\times H\times W}$, 原始特征图 $F\in\mathbf{R}^{C\times H\times W}$ 经过跳跃连接与此结果执行逐元素求和, 得到最后的输出特征图 F'。

经过 CxAM 模块后, 每个位置产生的特征 F' 是跨越所有位置的特征和原始特征的加权和。因此, 它可以有选择地聚合全局信息, 建立上下文依赖关系, 相似的语义特征相互促进, 从而提高了语义一致性。

3　实验验证

为了验证本文提出的算法，本节对改进的 SLN[12] 算法进行实验。采用 COCO-amodal 数据集对该模型进行训练，实验运行环境为深度学习框架 Pytorch 0.4.0，操作系统为 Ubantu 16.04，Python 3.6，GPU 显卡型号为 NVIDIA GeForce RTX 2060。

3.1　实验数据集

本文采用文献 [11] 中发布的 COCO-amodal 数据集。COCO-amodal 数据集是由 5 072 幅图像组成的非模式实例分割数据集，其中 2 500、1 250 和 1 322 幅图像分别用于训练、验证和测试。COCO-amodal 数据集的注释包括每个对象的可见/不可见区域以及每张图像中所有对象的相对深度顺序，作者没有将注释限制为通常的 COCO 类，可以为对象指定任意名称；此外，作者还提供了背景区域的注释，这些区域有时扩展到整个图像域，标记为“Stuff”。因此 COCO-amodal 数据集中的所有对象可以分为两类：“Things”和“Stuff”，其中“Things”是具有规范形状的对象，“Stuff”具有一致的视觉外观，但可以具有任意范围。

3.2　实验细节

实验首先利用在 COCO2014 数据集上预训练的 Mask RCNN 模型来初始化网络参数，算法的主干网络是 ResNet101。在训练区域提议网络 (RPN) 时，本文对 RPN 的网络参数进行了适当的调整，设置非最大抑制的阈值为 0.6，以便生成更多的区域提议。模型使用的损失函数和其他超参数均按照文献 [12] 中描述的策略进行设置和初始化。具体的训练过程中，借鉴离散下降学习率设置方法，以初始学习率 $l_r = 0.001$ 训练网络的 head 部分，训练 12 个 epochs，然后将学习率降低为 $l_r = 0.000\,1$ 来微调整个网络，训练 8 个 epochs，总计训练 20 个 epochs。所有目标均采用随机梯度下降法 (SGD) 进行优化，并设置 weight_decay = 0.000 1，momentum = 0.9。

3.3　评价指标

为了验证本文所提出的基于反馈注意力机制和上下文注意模块算法在非模式实例分割任务中的有效性，采用平均精确率 (average precision，AP) 和平均召回率 (average recall，AR) 作为对该算法性能的评价指标。平均精确率是指在图像分割时，将 IoU 阈值在 0.5~0.95 进行十等分，计算这 10 个不同 IoU 阈值下交并比的平均值；同样地，平均召回率指的是在 0.5~0.95 内 10 个不同 IoU 阈值下召回率的平均值。本文分别计算每幅图像在 All regions、Things only 和 Stuff only 情况下 AP 和 AR 值来评估算法性能，其中 AR^{10} 和 AR^{100} 分别表示每张图片中每个类别分类置信度最高的前 10 和前 100 个预测框的平均召回率。

由于本文研究的是非模式实例分割的任务，重点关注的是物体在被遮挡情况下对不可见部分的预测，所以有必要关注物体在不同遮挡强度下预测的准确性。因此，本文还统计了每幅图像中“Things”和“Stuff”在部分遮挡 (partial occlusion) 或重度遮挡 (heavy occlusion) 状态下的 AR 值，将其表示为 AR^P 和 AR^H。

3.4　实验结果分析

在 COCO-amodal 数据集上，将本文所提出的方法与 AmodalMask[11]、ARCNN[19]、ARCNN ++[19]、SLN[12] 在平均精确率和平均召回率进行对比，其中 ARCNN ++表示 ARCNN with visible mask，实验结果如表 1 所示。从表 1 可知，在 COCO-amodal 数据集上比较结果，本文所提出的方法，在 AP 和 AR 两个方面都有显著的提升，在 All regions 情况下的 AP 从 8.4% 提高到 14.3%，AR^{10} 从 16.6% 提高到 20.8%，AR^{100} 从 36.5% 提高到 40.3%，分别具有 5.9%、4.2% 和 3.8% 的收益。实验数据表明，本文的方法通过反馈注意力机制再学习和上下文注意模块，有效建立远程上下文依赖关系，捕获丰富的全局语义信息，增强了非模式实例分割的性能。

表 1　COCO-amodal 测试集上的分割结果对比

Table 1　Comparison of segmentation results on COCO-amodal test set

算法	All regions					Things only					Stuff only				
	AP	AR^{10}	AR^{100}	AR^P	AR^H	AP	AR^{10}	AR^{100}	AR^P	AR^H	AP	AR^{10}	AR^{100}	AR^P	AR^H
AmodalMask	5.74	13.5	29.23	31.0	21.3	6.12	16.5	33.1	37.0	23.6	0.78	5.4	18.1	16.1	18.0
ARCNN	4.1	10.2	21.3	22.0	13.3	4.4	12.0	23.9	34.7	15.2	0.3	4.8	13.8	15.1	10.1
ARCNN++	6.6	15.3	32.4	34.8	17.1	7.8	19.5	37.6	40.8	19.9	0.5	3.3	17.1	19.9	12.5
SLN	8.4	16.6	36.5	40.1	22.5	9.6	20.5	40.5	43.6	24.9	0.8	5.3	25.0	31.3	18.6
OURS	14.3	20.8	40.3	44.3	25.5	16.3	24.8	44.3	48.0	28.1	1.4	9.3	28.6	35.3	21.4

观察表 1 在 All regions 下 AR^P 和 AR^H 的结果可知, 当物体处于部分遮挡或重度遮挡状态下, 本文提出的方法对于像素的平均召回率仍然具有很大的提高, AR^P 从 40.1% 提高到 44.3%, AR^H 从 22.5% 提高到 25.5%, 分别具有 4.2% 和 3% 的收益, 这表明, 通过反馈注意力机制的再学习和上下文注意模块, 使网络能够学习到全局的语义依赖关系, 充分挖掘像素的空间相关性, 在非模式实例分割的任务中, 该方法能够有效帮助检测器仅仅从物体可见部分提供的线索去准确推断出物体被遮挡部分的最可能外观特征, 定性的可视化结果见图 4。

(a) 原图　(b) Ground truth　(c) SLN　(d) 本文模型

图 4　在 COCO-amodal 数据集上非模式实例分割的定性结果

Fig. 4　Qualitative results of amodal instance segmentation on coco-amodal dataset

与此同时, 在“Things only”和“Stuff only”的情况下, 本文算法无论是在 AP 还是 AR, 即使在部分遮挡或重度遮挡的状态下, 都表现出一致的优势: 在“Things only”时 AP 从 9.6% 提高到 16.3%, 具有 6.7% 的绝对收益, 69.8% 的相对收益; 同样地, 在“Stuff only”时 AP 从 0.8% 提高到 1.4%, 具有 0.6% 的绝对收益, 75% 的相对收益。

为了进一步证明本文提出方法的有效性, 本文对 COCO-amodal 测试集的一些图片进行定性分析, 非模式实例分割的定性实验结果如图 4 所示, 观察第 1 行的对比图可以看出, 图中的“冰箱”存在部分遮挡, SLN 算法在预测被遮挡、不可见

部分时存在一定的欠拟合问题，本文提出的方法通过建立丰富的上下文依赖关系，获取全局语义信息，实现了更准确的预测；从第 3 行对比图可知，在复杂的场景情况下，SLN 算法对小目标分割存在一定程度的漏分割现象，如图中有的“人”没有检测出来，并且这些样本属于小目标，本文的方法由于捕获了像素级的全局语义信息，加强了上下文信息的融合，对于小目标对象的漏分割、分割不准确的情况有了显著改善，不仅检测到图像中的小目标，同时对小目标对象的遮挡部分也能合理预测，使得分割质量得到大幅提升。

3.5　统计检验分析

为了对比不同的算法在数据集上的性能差异，本文采用 Friedman 检验来分析本文提出的算法是否具有显著性。本文在 COCO-amodal 数据集上对该方法进行了充分的实验，表 1 从 All regions、Things only 和 Stuff only 3 个维度分析了不同算法之间的性能差异。本文把表 1 转换成按 AP 从高到低排序的排序表，最后获得不同方法在 COCO-amodal 数据集不同维度上的排序情况，结果如表 2 所示。

表 2　不同算法在 COCO-amodal 测试集不同维度的 AP 排序表

Table 2　AP ranking tables of different algorithms in different dimensions on COCO-amodal test set

算法	All regions	Things only	Stuff only	平均序值 r_i
AmodalMask	4	4	3	3.67
ARCNN	5	5	5	5
ARCNN++	3	3	4	3.33
SLN	2	2	2	2
OURS	1	1	1	1

在获得不同算法的 AP 排序表之后，采用 Frie-dman 检验来判断这些算法是否性能都相同，同时做出假设“所有的算法性能相同”。变量 τ_F 服从自由度为 $(k-1)$ 和 $(k-1)(N-1)$ 的 F 分布，计算方法为

$$\tau_{\chi^2}=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}r_i^2-\frac{k(k+1)^2}{4}\right)$$
$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \tag{4}$$

通过式 (4) 计算得到的变量值 τ_F 与临界值 $F_{\alpha=0.05}$ 进行比较，假设“所有的算法性能相同”被拒绝，说明了不同算法之间的性能显著不同，其中 $F_{\alpha=0.05}=3.837\,9$，$k=5$，$N=3$。为了进一步区分各算法，采用 Nemenyi 检验作为“后续检验”。Nemenyi 检验临界值域 CD 的计算公式为

$$\mathrm{CD}=q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{5}$$

由式 (5) 计算出临界值域 CD 后，画出 Friedman 检验结果图，如图 5 所示。其中，中心圆点表示每个算法的平均序值，以圆点为中心的横线段表示临界值域的大小。Friedman 检验结果表示，如果两种方法的横线段有较多重叠，则表明两种算法的差异性较小，否则，说明两种算法具有显著的差异性。

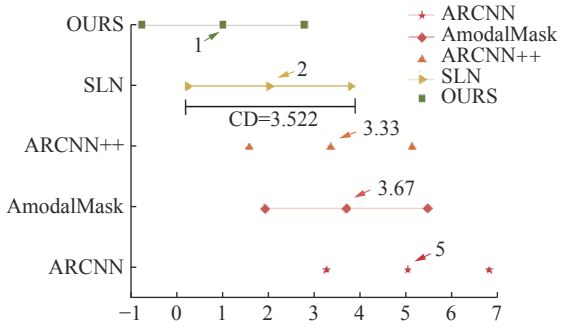

图 5　Friedman 检验结果

Fig. 5　Graph of Friedman test result

观察图 5 可知，直线 AmodalMask 与 ARCNN++重叠的部分比例最高，说明了算法 AmodalMask 和 ARCNN++没有显著差别；直线 OURS 与直线 SLN 有较多重叠部分，直线 OURS 与直线 AmodalMask、直线 ARCNN++具有较少的重叠部分，直线 OURS 与直线 ARCNN 基本无重叠部分。也就是说本文所提出的算法仍然优于其他 4 种算法，显著优于算法 ARCNN，这也验证了表 1 的实验结果。

4　结束语

本文提出一个反馈注意力机制的特征金字塔结构和上下文注意模块的方法并将其应用到非模式实例分割任务中。该方法在特征金字塔结构基础上引入反馈连接进行再学习，有效建立起通道之间的远程上下文依赖关系，并结合像素上下文注意力模块学习特征的空间相关性，捕获精细的空间细节，充分利用全局信息。在 SLN 网络的基础上，加入本文提出的方法构成新的网络结构，通过在 COCO-amodal 数据集上训练和测试，实验结果表明，本文方法能对物体被遮挡、不可见部分的最可能外观做出合理预测，并改善了其他方法中存在的漏分割、分割不准确的情况，但离实时处理仍有较大差距，后续将对此进行优化。

参考文献：

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (20

15-04-10)[2020-07-21] https://arxiv.org/abs/1409.1556.

[2] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016: 770−778.

[3] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21−37.

[4] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137−1149.

[5] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2999−3007.

[6] GIRSHICK R. Fast R-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1440−1448.

[7] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Pr- oceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, America, 2015: 3431−3440.

[8] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segme- ntation[EB/OL]. (2017-10-05)[2020-07-21] https://arxiv.org/abs/1706.05587.

[9] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980−2988.

[10] BOLYA D, ZHOU Chong, XIAO Fanyi, et al. YOLACT: real-time instance segmentation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019: 9156−9165.

[11] ZHU Yan, TIAN Yuandong, METAXAS D, et al. Semantic amodal segmentation[C]//Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 3001−3009.

[12] ZHANG Ziheng, CHEN Aapei, XIE Ling, et al. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation[C]//Pro- ceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 2124−2132.

[13] QIAO Siyuan, CHEN L C, YUILLE A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution[EB/OL]. (2020-06-03)[2020-07- 21] https://arxiv.org/abs/2006.02334.

[14] LI Yi, QI Haozhi, DAI Jifeng, et al. Fully convolutional instance-aware semantic segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 4438−4446.

[15] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 8759−8768.

[16] CHEN Xinlei, GIRSHICK R, HE Kaiming, et al. Tensormask: a foundation for dense object segmentation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019: 2061−2069.

[17] XIE Enze, SUN Peize, SONG Xiaoge, et al. PolarMask: single shot instance segmentation with polar representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, United States, 2020: 12190−12199.

[18] LI Ke, MALIK J. Amodal instance segmentation[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 677−693.

[19] FOLLMANN P, KÖNIG R, HÄRTINGER P, et al. Learning to see the invisible: end-to-end trainable amodal instance segmentation[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, United States, 2019: 1328−1336.

[20] EHSANI K, MOTTAGHI R, FARHADI A. SeGAN: segmenting and generating the invisible[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 6144−6153.

[21] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 7132−7141.

[22] WANG Xiaolong, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 7794−7803.

[23] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 3−19.

[24] CAO Yun, XU Jiarui, LIN S, et al. GCNet: non-local net-

works meet squeeze-excitation networks and beyond[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, South Korea, 2019: 1971−1980.

[25] FAN Zhibo, YU Jingang, LIANG Zhihao, et al. FGN: fully guided network for few-shot instance segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, United States, 2020: 9169−9178.

[26] LIU Dongnan, ZHANG Donghao, SONG Yang, et al. Cell R-CNN v3: a novel panoptic paradigm for instance segmentation in biomedical images[EB/OL]. (2020-02-15) [2020-07-21] https://arxiv.org/abs/2002.06345.

[27] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, United States, 2019: 3141−3149.

[28] CAO Junxu, CHEN Qi, GUO Jun, et al. Attention-guided context feature pyramid network for object detection[EB/OL]. (2020-05-23)[2020-07-21] https://arxiv.org/abs/2005.11475.

作者简介：

董俊杰，硕士研究生，主要研究方向为智能信息处理、计算机视觉和图像识别。

刘华平，副教授，博士生导师，IEEE Senior Member、中国人工智能学会理事、中国人工智能学会认知系统与信息处理专业委员会秘书长。主要研究方向为机器人感知、学习与控制、多模态信息融合。发表学术论文 340 余篇。

谢珺，副教授，主要研究方向为粗糙集、粒计算、数据挖掘和智能信息处理。

大数据智能高峰论坛

举办时间： 2021 年 10 月−11 月

举办地点： 重庆

会议官网： http://bdaiid.cqupt.edu.cn/#/introduction

会议简介：

论坛始办于 2015 年，是全国第一个关于大数据智能主题的专业学术会议，已连续成功举办 6 届。本次论坛将围绕智能科技前沿和智能产业创新发展的焦点问题展开研讨，为大数据与人工智能领域的专家学者、企业技术领军者提供一个高端交流、合作的平台。