计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):261−268 [doi: 10.15888/j.cnki.csa.009617]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于对比学习及背景挖掘的少样本语义分割①

王善杰

(南京信息工程大学 软件学院, 南京 210044)
通信作者: 王善杰, E-mail: 20211221038@nuist.edu.cn

摘　要: 少样本语义分割是在具有少量标注样本的查询图像的条件下, 对潜在对象类别进行分割的计算机视觉任务. 然而, 现有方法仍然存在两个问题, 这对它们构成了挑战. 首先是原型偏差问题, 这导致原型具有较少的前景目标信息, 难以模拟真实的类别统计信息. 另一个是特征破坏问题, 这意味着模型只关注当前类别而不关注潜在类别. 本文提出了一个基于对比原型以及背景挖掘的新网络. 该网络主要思想是使模型学习更具代表性的原型, 并从背景中识别潜在类别. 具体而言, 特定类学习分支构建了一个大且一致的原型字典, 然后使用 InfoNCE 损失使原型更具区分性. 另一方面, 背景挖掘分支初始化背景原型, 并使用构建的背景原型与字典之间的注意力机制来挖掘潜在类别. 在 PASCAL-5^i 和 COCO-20^i 数据集上的实验证明模型有优秀的性能. 在使用 ResNet-50 网络的 1-shot 设置下, 达到了 64.9% 和 44.2%, 相较于基准模型分别提升了 4.0% 和 1.9%.

关键词: 图像分割; 少样本语义分割; 对比学习; 背景挖掘

引用格式: 王善杰.基于对比学习及背景挖掘的少样本语义分割.计算机系统应用,2024,33(9):261–268. http://www.c-s-a.org.cn/1003-3254/9617.html

Few-shot Semantic Segmentation Based on Contrastive Learning and Background Mining

WANG Shan-Jie

(School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Few-shot semantic segmentation is a computer vision task that involves segmenting potential object categories in query images with a small number of annotated samples. However, existing methods still face two challenges. Firstly, there is a prototype bias problem, resulting in prototypes having less foreground object information and making it difficult to simulate real category statistics. The other issue is feature degradation, which means that the model only focuses on the current category rather than potential categories. This study proposes a new network based on contrastive prototypes and background mining. The main idea of the network is to enable the model to learn more representative prototypes and identify potential categories from the background. Specifically, a specific class learning branch constructs a large and consistent prototype dictionary and then uses InfoNCE loss to make the prototypes more discriminative. On the other hand, the background mining branch initializes background prototypes and uses an attention mechanism between the constructed background prototypes and the dictionary to mine potential categories. Experimental results on the PASCAL-5^i and COCO-20^i datasets demonstrate excellent performance of the model. Under the 1-shot setting using the ResNet-50 network, 64.9% and 44.2% are achieved, an improvement of 4.0% and 1.9%, respectively, compared to the baseline model.

Key words: image segmentation; few-shot semantic segmentation; contrastive learning; background mining

① 收稿时间: 2024-03-22; 修改时间: 2024-04-16; 采用时间: 2024-04-23; csa 在线出版时间: 2024-07-26
CNKI 网络首发时间: 2024-07-29

卷积神经网络[1−3]已经成为解决计算机视觉问题的强大平台, 就速度和精度而言, 它们在计算机视觉领域取得了巨大成功. 然而, 大多数计算机语义分割任务[4−6]需要大量的像素级注释, 这需要进行大量的手工标记. 如果要对训练集中不存在的新类别进行分割, 那么需要大量的图像和像素级标签来表示该类别.

为了缓解这个问题, 一些学者提出了少样本分割方法[7,8], 该方法是在给定少量像素级注释支持图像的情况下分割查询图像中的目标对象. 但是少样本语义分割领域存在两个固有问题, 分别是原型偏差问题和特征破坏问题. 原型偏差问题意味着支持原型的质量不佳, 因为原型只具有少量的前景信息, 特别是当支持图像的外观和尺度与查询图像中的分割目标之间存在很大差异时. 这是由于少量的样本无法模拟真实的类别统计信息, 使得仅利用当前支持来估计原型是次优的. 特征破坏问题源于以前的分割方法仅关注当前基础类别, 而在每个训练阶段将其他潜在类别视为背景. 因此, 在测试阶段难以识别和分割潜在的新类别. 潜在的类别与真实的背景本质上不同, 并且应该更好地利用.

本文主要通过使用特定类学习分支和背景挖掘分支来解决上述问题. 首先, 特定类学习分支用于生成更多的类别特定表示, 以解决原型偏差问题. 具体来说, 本文使用对比学习方法使类别嵌入空间更加均匀分布, 并保留更多的特征信息, 从而产生了一个强大的原型表示来引导查询分支的分割. 其次, 为了解决特征破坏问题, 本文提出了背景挖掘分支, 给分支主要通过挖掘潜在的类别, 从而使模型能够区分前景和背景类别, 使它们在嵌入空间的距离最大化, 从而能提升模型识别物体的鲁棒性和精确性.

1 相关工作

1.1 少样本语义分割

近年来, 少样本语义分割受到了广泛关注. 现有的方法可以分为以下 3 类: 1) 基于原型的方法: 这些方法[9−11]从支持分支中提取单个或多个原型, 以指导查询分支的分割. 具体而言, Shaban 等人[9]和 Li 等人[10]提出了一个双分支网络, 采用参数匹配方法. 另一项工作 PANet[12]使用非参数度量学习来双向对齐支持和查询, 其中每个都可以成为另一个的参考. SGC[13]提出一种简单有效的自我引导学习方法, 通过构造和聚合主支持向量和辅助支持向量来挖掘原型丢失的关键信息, 并且提出一个用于多镜头的交叉引导模块, 使用多个注释样本的预测融合最终的掩码. 2) 基于注意力的方法[14,15]: 这些方法倾向于采用查询-键-值交互的非局部自注意机制, 探索支持和查询特征图之间的关系. 例如, Igbal 等人[14]提出了一种任务感知自适应注意力, 从支持图像中提取任务特定信息, 并将其整合到通道维度和空间维度的特征表示中以进行自适应增强. 3) 基于半监督/自监督学习的方法[16,17]: 这些方法执行多尺度、像素级和区域级监督, 生成伪标签以区分可能的非目标类别. 具体而言, Yang 等人[18]提出了一种新颖的联合训练框架, 通过引入额外的基类掩码, 在训练过程中生成不同类别的伪掩码表示.

虽然之前的工作都专注于建立支持图像和查询图像之间的高维对应关系, 但却忽略了在支持图像的基础上寻找良好的特征表示. 另一个问题是, 他们只关注当前支持图像中的目标类别, 而忽略了背景中的潜在类别. 本文的工作遵循非参数度量学习范式, 不同之处在于本文使用特定类别学习分支来生成优秀的代表性特征, 并从背景挖掘分支中生成背景原型, 然后逐像素地将它们与目标的前景进行匹配, 从而挖掘出潜在类别.

1.2 对比学习

近年来, 随着自监督学习的兴起, 对比学习进入了新的阶段. 自监督对比学习通过构建代理任务, 让模型学会利用数据内在的结构和关系进行学习, 而无需人工标注的标签. 自监督对比学习使得模型在大规模无标签数据上进行预训练成为可能, 为各种下游任务提供了有力的特征表示. 鉴于对比学习需要大量的负样本, 文献[19,20]分别提出了 MoCo v1 和 MoCo v2, MoCo v1[19]通过一个动量编码器和一个内存库来增加负样本数量. MoCo v2[20]在 MoCo v1 的基础上进行了优化, 通过改进损失函数和训练策略, 以及引入更大的模型规模, 取得了更好的自监督学习效果. Chen 等人[21,22]提出了 SimCLR, 利用当前批次中的各种数据增强组合以采样足够的负样本. 对比学习为无监督和弱监督学习提供了强大的工具, 对计算机视觉和自然语言处理的发展有着积极的影响.

2 网络架构

2.1 整体框架

本文提出模型的整体框架如图 1 所示. 模型采用了元学习中情景训练的框架, 使得模型能够在每个情

景训练阶段进行训练和推断.

2.2 特定类学习分支

尽管非参数化的少样本分割模型能够生成高质量的原型, 但现有的研究不可避免地会面对原型偏差的问题. 在先前的研究中, 为了解决这一问题, 一些方法尝试通过生成多个原型或者增强原型中的语义信息. 然而, 这些方法未能充分最大化不同类别之间的特征距离. 为了解决这个挑战, 本文采用对比学习方法, 通过增加不同类别之间的距离来避免由于训练样本不足而导致的过拟合问题.

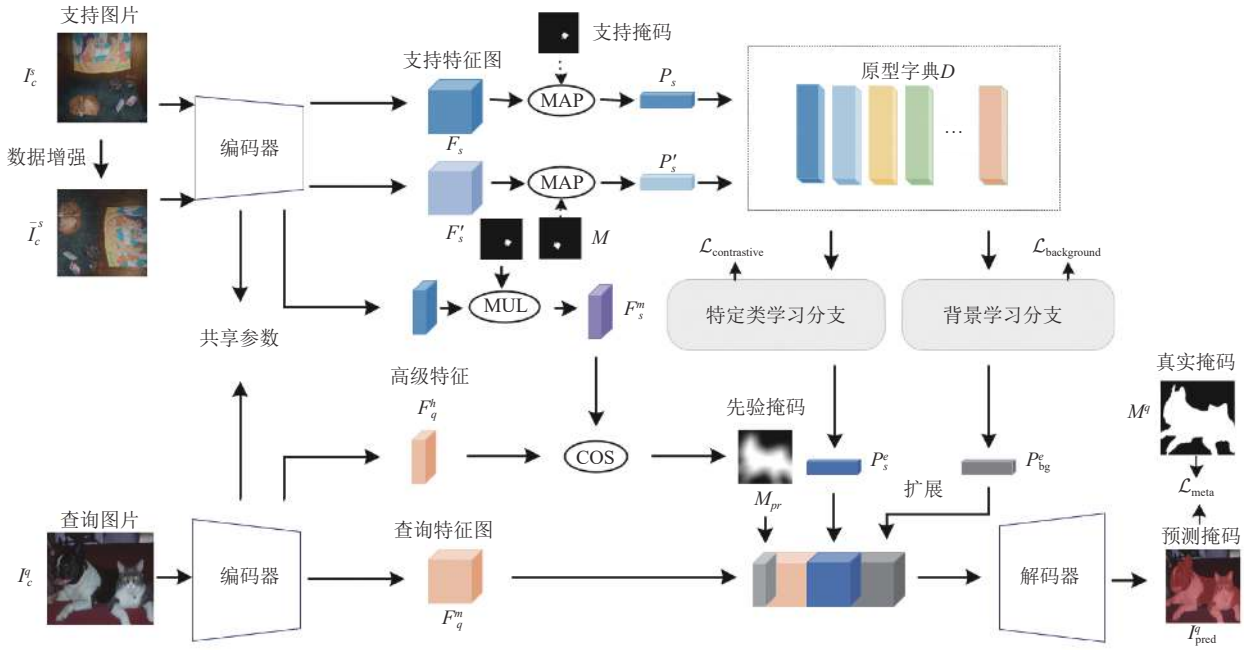

图 1 本文模型框架图

获得代表性原型的关键是确保生成的原型具有良好的特征表达能力, 并能够与其他类别的原型区分开. 受 MoCo 系列工作的启发, 本文根据每个情景训练中获取的原型及其类别标签构建了一个大型且一致的字典. 字典可以表示为$D=\{L_c,P_c\}_{c=1}^{N}$, 其中L_c和P_c表示第c类标签和该类别的原型, N是原型存储在字典D中的总类别数. 为了提取支持图片的原型, 本文采用掩码平均池化策略. 关于第c类的第i张图片的原型$P_i^c\in\mathbb{R}$可以计算为:

$$P_i^c=\frac{\sum_{x,y}F_i^{x,y}1[M_i^{x,y}=c]}{\sum_{x,y}1[M_i^{x,y}=c]} \tag{1}$$

其中, $F_i\in\mathbb{R}^{C\times H\times W}$表示从第$i$张支持图片提取的中层特征. $1(\cdot)$表示指示函数, (x,y)表示图片中的索引空间位置, M_i是第i张支持图片的掩码.

对于每个元学习的训练情景, 给定一个属于第c类的支持图像I_c^s及其对应的数据增强图像$\overline{I}_c^s$. 然后, I_c^s和$\overline{I}_c^s$被送入编码器以产生原型P_c^s和$\overline{P}_c^s$. 本文方法从存储的字典中设置正样本组成$(P_c^s,\overline{P}_c^s)$, 相应的负样本对组成$(P_c^s,\overline{P}^s)$, 其中$\overline{P}^s$是不属于第c类的原型集合. 因此, 本文使用信息噪声对比估计 (information noise-contrastive estimation, InfoNCE) 损失作为损失函数, 计算过程如下:

$$\mathcal{L}_{\text{contrastive}}=-\log\frac{\exp(P_c^s\cdot\overline{P}_c^s/\tau)}{\sum_{i=0}^{K}\exp(P_c^s\cdot\overline{P}^s/\tau)} \tag{2}$$

其中, τ是一个温度, K表示负样本对的个数. 当计算特定类学习分支的对比损失后, 该类别的原型P_c^s及其类别标签c被放入字典D中, 字典D采用 FIFO 策略, 最旧的键和值被出队.

2.3 背景挖掘分支

背景挖掘分支来辅助模型训练, 并实现对潜在类别的泛化. 通常情况下, 训练图像的背景中包含少量对象, 这些对象往往容易被忽略. 先前的方法对所有背景进行简单平均的做法无法学习背景中的类别. 受到显著性检测方法的启发, 本文利用大量图像学习背景原

型是可行的. 背景挖掘分支由背景原型生成模块和背景挖掘模块组成. 两个模块的具体内容如下阐述.

(1) 背景原型生成模块

根据图 2 所示, 首先, 随机初始化背景的原型P_B. 由于残差网络的中层特征比高层特征更具有泛化性, 因此将P_B扩展为与支持和查询的中层特征图F_s^m、F_q^m相同的维度, 并将P_B与中层特征图F_s^m、F_q^m进行拼接. 最后, 利用卷积、下采样和分类头来生成最终的背景预测:

$$y_b^s = F_{\text{cls}}\{Conv_{3\times3}\{F_{\text{down}}(P_B \oplus F_s^m)\}\} \tag{3}$$

$$y_b^q = F_{\text{cls}}\{Conv_{3\times3}\{F_{\text{down}}(P_B \oplus F_q^m)\}\} \tag{4}$$

其中, $Conv_{3\times3}$和F_{down}分别代表两个共享权重的卷积层, $\oplus$表示连接操作. 特征图经过最终的预测头F_{cls}之后, 生成了背景预测概率图y_b^s和y_b^q.

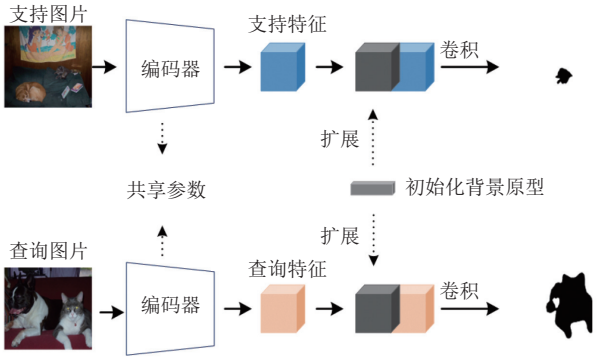

图 2 背景原型生成模块示意图

前景和背景的任务相关预测是具有挑战性的, 因为背景的真实标签在训练的过程中不存在. 因此, 引入背景挖掘损失来预测背景区域, 计算过程如下:

$$\mathcal{L}_{\text{background}} = -\frac{1}{N}\sum \log(1-y_b^{q/s}(i))M^{q/s}(i) - \partial\frac{1}{Z}\sum \log(y_b^{q/s}(j)) \tag{5}$$

其中, N表示特征图中的像素个数, Z是特征图高度H和宽度W的乘积, $M^{q/s}$是支持和查询特征图的真实标签, i和j是空间位置的索引, 超参数∂用于约束模块的训练, 以防止预测的背景区域的掩码值全部为 0. 通过背景挖掘损失优化了背景原型在训练阶段的学习过程.

(2) 背景挖掘模块

本文方法在背景原型生成模块之后使用注意力机制来匹配字典中的原型和背景原型, 以便模型可以挖掘更多的潜在类别. 整个过程如图 3 所示.

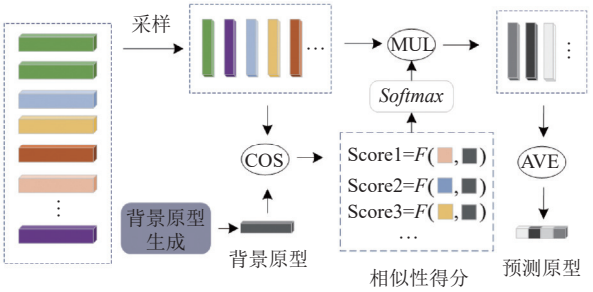

图 3 背景挖掘模块的框架图

具体来说, 背景原型P_{bg}和字典中的原型P_{dict}^i通过W_v和W_k进行线性变换, 通过P_{dict}^i生成一组键, 键的集合表示为$K_i = \{K_1, K_2, \cdots, K_N\}$, 其中$N$表示从字典中随机抽样的原型数量. 在背景原型$P_{\text{bg}}$和$K_i$之间进行余弦相似度比较, 此过程可以表示为:

$$S(P_{\text{bg}}, K_i) = \frac{P_{\text{bg}} \cdot K_i}{\|P_{\text{bg}}\| \cdot \|K_i\|} \tag{6}$$

其中, S表示P_{bg}和K_i之间的相似度分数, $\|\cdot\|$表示 Frobenius 范数. 然后, 通过$Softmax$函数计算注意力权重, 可以表示为:

$$\alpha(P_{\text{bg}}, K_i) = Softmax(S(P_{\text{bg}}, K_i)) = \frac{\exp(S(P_{\text{bg}}, K_i))}{\sum_{i=1}^{N}\exp(S(P_{\text{bg}}, K_i))},\ i = 1,2,\cdots,N \tag{7}$$

其中, N是键的总数. 通过$Softmax$操作后得到的注意力权重表示不同类别的原型与背景原型之间的相关性. 然后, 使用注意力权重更新背景原型$P_{\text{bg}}^{\text{update}}$:

$$P_{\text{bg}}^{\text{update}} = \sum_{i=1}^{N}\alpha(P_{\text{bg}}, K_i) \cdot P_{\text{dict}}^i \tag{8}$$

这里更新的背景原型是属于不同类别的语义信息的混合. 与通过聚类预测背景原型相比, 生成的背景原型可以更好地进行预测. 最后, 将更新的背景原型和查询的中层特征图与支持原型拼接传递给解码器FEM进行分割预测I_{pred}^q:

$$I_{\text{pred}}^q = FEM(concat(M_{\text{pr}}, P_s^e, P_q^m, P_{\text{bg}}^e)) \tag{9}$$

其中, P_s^e、P_{bg}^e分别表示支持和背景的扩展原型. M_{pr}和P_q^m分别是查询图像的先验掩码和中层特征图.

2.4 损失函数

本文模型的分割损失$\mathcal{L}_{\text{seg}}$的计算公式如下:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{meta}} + \mu\mathcal{L}_{\text{contrastive}} + \lambda\mathcal{L}_{\text{background}} \tag{10}$$

其中, $\mathcal{L}_{\text{meta}}$是最终预测图像$I_{\text{pred}}^q$与真实二值掩码图像

M^q之间的交叉熵损失, $\mathcal{L}_{\text{contrastive}}$表示来自特定类学习分支的 InfoNCE 损失, $\mathcal{L}_{\text{background}}$是来自背景挖掘分支的交叉熵损失. 这里$\mu$和$\lambda$是超参数, 分别设置为 0.15 和 0.25.

3 实验分析

3.1 实验环境和参数

为了与先前的工作进行实验比较, 本文方法采用了 ResNet-50 作为主干网络. 这个主干网络在 ImageNet 上进行了预训练, 以改善模型的性能. 训练的过程中采用随机梯度下降优化器, 学习率为 0.002 5, 迭代次数为 200. 每次并行输入网络的样本个数为 8, 对比学习温度超参数τ设置为 1, 本文方法使用了 MoCo 中的数据增强, 这些数据增强包括随机裁剪、随机颜色抖动、随机水平翻转、随机灰度转换和随机模糊增强. 字典中的原型初始化为 1 000 个. 最后, 在测试阶段使用 5 个随机种子进行测试, 以消除实验误差. 本文模型采用 PyTorch 框架, 并在一台 NVIDIA RTX 3090 24 GB 服务器上运行所有实验.

3.2 对比实验

表 1 和表 2 为实验结果, 表中加粗为最优值, 下划线为次优值. 如表 1 所示, 当使用 PASCAL-5^i 数据集时, 本文方法在 ResNet-50 网络结构的 1 类别 1 个样本设置下的 mIoU 得分相较于得分第 2 的 NTRNet[23]提高了 0.7%, 并且在第 2 个子集上的得分为 68.35%, 相较于次优的 MLC[18]提升 3.2%. 在 ResNet-50 网络结构的 1 类别 5 个样本设置下表现一般, 在子集 2 和子集 3 上的 mIoU 得分相较于 DCP[24]和 SSP[25]分别下降了 2.0% 和 1.8%. 本文方法在 COCO-20^i 数据集上进行了实验, 结果如表 2 所示. 从表 2 中可以看出本文方法在 ResNet-50 网络结构的 1 类别 1 个样本设置下的 mIoU 得分相较于得分第 2 的 BAM[26]提高了 1.9%, 并且在子集 1 和子集 2 上的得分相较于次优的 BAM[26]分别提升 3.1% 和 2.4%. 在 ResNet-50 网络结构的 1 类别 5 个样本设置下表现优秀, 在子集 1、子集 2 和子集 3 上的 mIoU 得分相较于 DCP[24]和 BAM[26]分别提高了 2.6%、3.5% 和 0.2%.

表 1 在 PASCAL-5^i 数据集上 1-shot 和 5-shot 设置下的实验结果 (%)

方法	1-shot					5-shot				
	子集0	子集1	子集2	子集3	平均	子集0	子集1	子集2	子集3	平均
PGNet[27]	56.0	66.9	50.6	50.4	56.0	54.9	67.4	51.8	53.0	56.8
PANet[12]	44.0	57.5	50.8	44.0	49.1	55.3	67.2	61.3	53.2	59.3
PPNet[28]	48.6	60.6	55.7	46.5	52.8	58.9	68.3	66.8	58.0	63.0
PMMs[29]	55.2	66.9	52.6	50.7	56.3	60.8	67.3	54.5	51.0	57.3
IPMT[30]	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
MLC[18]	59.2	71.2	65.6	52.5	62.1	63.5	71.6	71.2	58.1	66.1
NTRNet[23]	**65.4**	**72.3**	59.4	**59.8**	64.2	66.2	72.8	61.7	62.2	65.7
DCP[24]	63.8	70.5	61.2	55.7	62.8	67.2	**73.2**	66.4	**64.5**	67.8
SSP[25]	61.4	67.2	65.4	49.7	60.9	**68.0**	72.0	**74.8**	60.2	**68.8**
本文方法	62.3	70.5	**68.3**	58.8	**64.9**	64.3	70.8	72.8	62.7	67.6

表 2 在 COCO-20^i 数据集上 1-shot 和 5-shot 设置下的实验结果 (%)

方法	1-shot					5-shot				
	子集0	子集1	子集2	子集3	平均	子集0	子集1	子集2	子集3	平均
PANet[12]	31.5	22.6	21.5	16.2	22.9	33.0	45.9	29.2	30.6	33.8
PPNet[28]	34.5	25.4	24.3	18.6	25.7	48.3	30.1	36.7	30.2	33.8
PMMs[29]	29.5	36.8	28.9	27.0	30.6	33.8	42.0	33.0	33.3	35.5
IPMT[30]	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	37.4
MLC[18]	**46.8**	35.3	26.2	27.1	33.9	**54.1**	41.2	34.1	33.1	40.6
SSP[25]	46.4	35.3	27.3	25.4	33.6	53.8	41.5	36.0	33.7	41.3
DCP[24]	40.9	43.7	42.6	38.2	41.3	45.8	49.6	43.6	46.5	46.8
BAM[26]	38.2	45.5	43.1	**41.6**	42.1	46.8	45.6	42.1	46.6	45.2
本文方法	42.4	**48.6**	**45.5**	41.4	**44.2**	46.2	**48.2**	**45.6**	**46.8**	**47.1**

3.3 可视化实验

为了更深入地分析和理解本文提出的模型的效果, 实验中将模型与查询图像的真实二值掩码 (第 2 行)、NETRNet[23] (第 3 行)、IPMT[30] (第 4 行) 以及本文的

方法 (第 5 行) 进行了比较. 模型分割的可视化结果展示在图 4 中. 可视化结果证明模型成功地克服了不同图像中前景目标尺度不一的难题, 并且显著提高了前景目标识别的准确性. 这进一步证明了本文模型在解决图像语义分割任务的有效性.

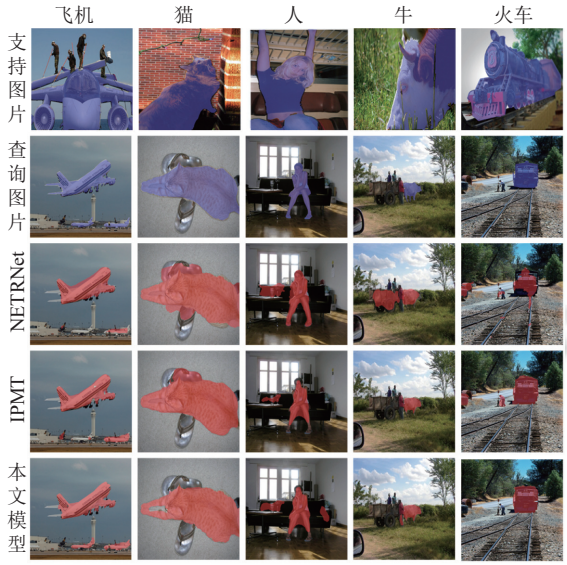

图 4 模型在 PASCAL-5^i 和 COCO-20^i 数据集上的分割结果

图 5 展示了模型在不同 PASCAL-5^i 数据集子集上的 IoU 测试得分, 可以看出从子集 0、子集 1、子集 2 和子集 3 的最高 IoU 得分分别为 62.3%、70.5%、68.3% 和 58.8%, 证明模型有良好的分割精度.

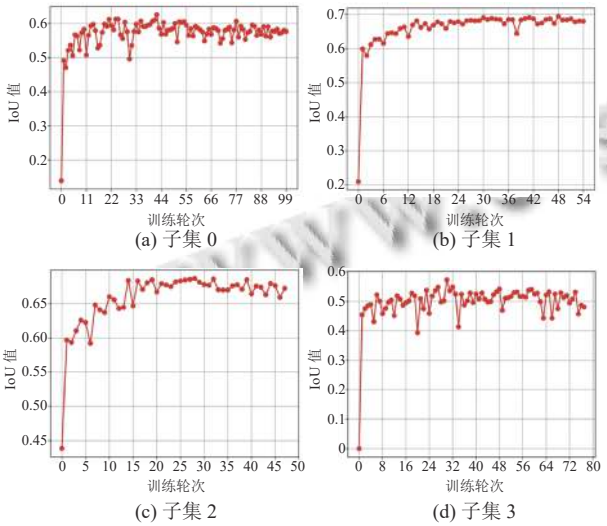

图 5 模型在不同子集上的 IoU 测试得分

3.4 消融实验

对于实验中有关损失函数的超参数μ和λ, 在 ResNet-50 网络结构上使用 3 组超参数进行测试, 3 组超参数的实验结果如图 6 所示. 实验选择了 3 组常用的超参数, 分别是$\mu=0.01,\lambda=0.02$和$\mu=0.15,\lambda=0.25$以及$\mu=0.4,\lambda=0.6$. 其中$\mu=0.15,\lambda=0.25$这一组参数在不同的数据集子集上取得了出色的实验结果, 因此选择这组参数作为实验超参数. 值得注意的是, 虽然$\mu=0.4,\lambda=0.6$这一组超参数在许多工作的实验中被证明是有效的, 但它们在本次消融实验中未能取得良好的结果. 这是由于对比学习中有关损失的超参数不应设置得太大, 因为对比学习的过程会产生许多原型, 原型之间的相似度计算会带来实验误差.

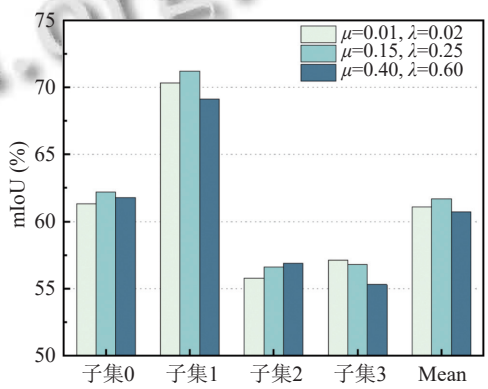

图 6 损失超参数的消融实验

为了直观分析和比较模型中不同分支的效果, 对分支的消融实验有 3 个对象, 分别是数据增强、特定类学习分支和背景挖掘分支, 实验结果如表 3 所示. 从实验结果可以看出, 特定类学习分支比第 1 行基准模型的实验结果高出了 2.2%. 这表明该特定类学习分支可以有效区分不同的类别.

表 3 本文模型每个分支的消融结果 (%)

DA	CL	BM	子集0	子集1	子集2	子集3	平均
—	—	—	56.5	68.1	53.3	52.6	57.6
—	√	—	58.2	69.3	54.5	53.4	58.4
√	√	—	58.4	69.5	55.1	54.2	59.3
—	—	√	59.6	70.8	56.1	55.7	60.5
—	√	√	60.3	71.1	56.4	55.4	60.8
√	√	√	**62.2**	**71.2**	**56.6**	**56.8**	**61.7**

注: DA: 数据增强; CL: 特定类学习分支; BM: 背景挖掘分支

背景原型的通道维度是非常重要的, 因为它决定了背景原型能够编码通用的背景信息的数量. 本文进行了消融实验来探索适当的通道维度值, 实验结果如表 4 所示. 结果表明, 在 4 个子集上通道维度为 256 的性能表现较好. 因此, 在本文实验设置中选择使用通道数为 256 的支持和查询特征图来生成背景原型.

表 4 初始化阶段对背景原型 D_{bg} 的通道维度的消融研究 (%)

D_{bg}	子集0	子集1	子集2	子集3	平均
128	57.5	69.1	54.3	53.6	58.6
256	62.2	71.2	56.6	56.8	**61.7**
512	58.4	70.8	55.1	55.2	59.9
1 024	60.3	71.1	56.4	55.4	60.8

在图 7 中展示了不同子集的 t-分布随机邻域嵌入技术的降维实验, 在嵌入空间中, 不同原型类别之间的距离是衡量特定类学习分支性能的关键. 为了更清楚地展示该分支对分割性能的影响, 实验中使用查询特征图来生成新类的原型, 并通过 t-分布随机邻域嵌入技术将这些特征投影到 2D 空间. 在图 7 中, (a) 是通过初始化 ImageNet 的预训练权重得到的嵌入空间中新类的特征分布. (b) 是通过基线模型的训练得到的, (c) 是通过本文模型得到的特征降维图像. 实验表明, 本文中特定类学习分支可以有效增加类内相似性, 属于同一类对象的特征能够有效地聚类.

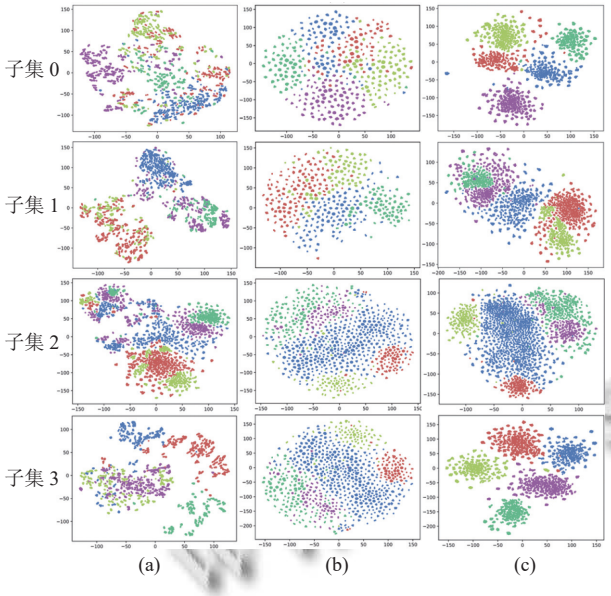

图 7 模型的 t-分布随机邻域嵌入技术的降维可视化实验

为了验证模型是否能够识别背景中的潜在类别, 实验中使用热力图的方式展示分类器的最终分割特征图处理. 处理的结果如图 8 所示, 其中 (a) 为查询图片, (b) 为特定类查询分支, (c) 为背景挖掘分支. 在这些结果中, 红色和蓝色区域分别表示分割的查询集中高相关性和低相关性的区域. 从图 8(c) 可以看出, 模型不仅识别了支持集中的前景对象 (猫、自行车、玻璃、飞机、汽车), 还识别了背景中的对象 (人、公共汽车).

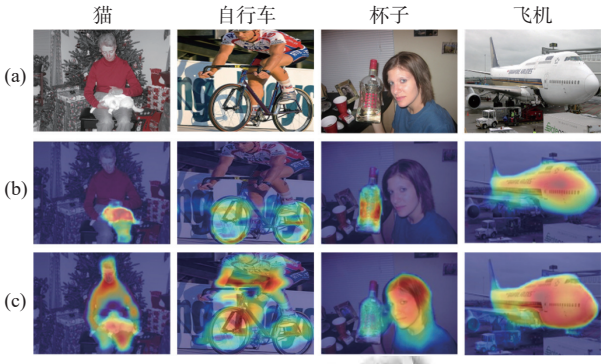

图 8 热力图实验

4 结论与展望

本文提出了一个有效的框架来解决少样本分割问题. 具体而言, 本文设计了一个特定类的原型学习范式和一个挖掘潜在类别的策略, 使用对比学习增加了相同类别原型之间的相似度, 并在背景中识别了潜在类别. 在 PASCAL-5^i 和 COCO-20^i 数据集上的定性和定量结果表明, 所提出的方法取得了出色的性能. 但是仍然存在着一些不足, 一方面, 由于支持图像中原型提供的特征信息不足, 模型无法准确识别查询图片中的目标区域; 另一方面, 由于查询图像中的目标对象和背景像素非常相似, 导致分割失败. 因此, 在少样本语义分割领域, 可靠的原型特征表示和查询图像的复杂性是一个难题, 这是未来可以继续研究的方向.

参考文献

1 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229–1251. [doi: 10.11897/SP.J.1016.2017.01229]
2 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述. 计算机学报, 2020, 43(5): 755–780. [doi: 10.11897/SP.J.1016.2020.00755]
3 常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络. 自动化学报, 2016, 42(9): 1300–1312.
4 徐辉, 祝玉华, 甄彤, 等. 深度神经网络图像语义分割方法综述. 计算机科学与探索, 2021, 15(1): 47–59. [doi: 10.3778/j.issn.1673-9418.2004039]
5 何淼楹, 崔宇超. 面向自动驾驶的交通场景语义分割. 计算机应用, 2021, 41(S1): 25–30.
6 张祥甫, 刘健, 石章松, 等. 基于深度学习的语义分割问题研究综述. 激光与光电子学进展, 2019, 56(15): 150003.
7 Chen S, Meng FM, Zhang RT, *et al*. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. IEEE Transactions on Multimedia, 2024, 26:

7197–7209. [doi: 10.1109/TMM.2024.3361181]
8 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090.
9 Shaban A, Bansal S, Liu Z, *et al*. One-shot learning for semantic segmentation. Proceedings of the 2017 British Machine Vision Conference. London: BMVC, 2017.
10 Li G, Jampani V, Sevilla-Lara L, *et al*. Adaptive prototype learning and allocation for few-shot segmentation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8330–8339.
11 Lang CB, Tu BF, Cheng G, *et al*. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation. Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022. 1024–1030.
12 Wang KX, Liew JH, Zou YT, *et al*. PANet: Few-shot image semantic segmentation with prototype alignment. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9196–9205. [doi: 10.1109/ICCV.2019.00929]
13 Zhang BF, Xiao JM, Qin T. Self-guided and cross-guided learning for few-shot segmentation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8312–8321.
14 Iqbal E, Safarov S, Bang S. MSANet: Multi-similarity and attention guidance for boosting few-shot segmentation. arXiv:2206.09667, 2022.
15 曾伟辉, 唐欣, 胡根生, 等. 基于卷积块注意力胶囊网络的小样本水稻害虫识别. 中国农业大学学报, 2022, 27(3): 63–74. [doi: 10.11841/j.issn.1007-4333.2022.03.08]
16 Li YW, Data GWP, Fu YG, *et al*. Few-shot semantic segmentation with self-supervision from pseudo-classes. Proceedings of the 32nd British Machine Vision Conference. BMVC, 2021. 164.
17 Alfaro-Contreras M, Ríos-Vila A, Valero-Mas JJ, *et al*. Few-shot symbol classification via self-supervised learning and nearest neighbor. Pattern Recognition Letters, 2023, 167: 1–8. [doi: 10.1016/j.patrec.2023.01.014]
18 Yang LH, Zhuo W, Qi L, *et al*. Mining latent classes for few-shot segmentation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 8701–8710. [doi: 10.1109/ICCV48922.2021.00860]
19 He KM, Fan HQ, Wu YX, *et al*. Momentum contrast for unsupervised visual representation learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/CVPR42600.2020.00975]
20 Chen XL, Fan HQ, Girshick R, *et al*. Improved baselines with momentum contrastive learning. arXiv:2003.04297, 2020.
21 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. 149.
22 Chen T, Kornblith S, Swersky K, *et al*. Big self-supervised models are strong semi-supervised learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1865.
23 Liu YW, Liu N, Cao QL, *et al*. Learning non-target knowledge for few-shot semantic segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11563–11572. [doi: 10.1109/CVPR52688.2022.01128]
24 Zhang JW, Sun YF, Yang Y, *et al*. Feature-proxy Transformer for few-shot segmentation. Proceedings of the 36th Advances in Neural Information Processing Systems. New Orleans, 2022. 6575–6588.
25 Fan Q, Pei WJ, Tai YW, *et al*. Self-support few-shot semantic segmentation. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 701–719.
26 Lang CB, Cheng G, Tu BF, *et al*. Learning what not to segment: A new perspective on few-shot segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8047–8057. [doi: 10.1109/CVPR52688.2022.00789]
27 Zhang C, Lin GS, Liu FY, *et al*. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9586–9594. [doi: 10.1109/ICCV.2019.00968]
28 Liu YF, Zhang XY, Zhang SY, *et al*. Part-aware prototype network for few-shot semantic segmentation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 142–158.
29 Yang BY, Liu C, Li BH, *et al*. Prototype mixture models for few-shot semantic segmentation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 763–778.
30 Tian ZT, Zhao HS, Shu M, *et al*. Prior guided feature enrichment network for few-shot segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 1050–1065. [doi: 10.1109/TPAMI.2020.3013717]

(校对责编: 张重毅)