计算机科学
COMPUTER SCIENCE

基于多尺度原型分层匹配的小样本分割方法

孙开伟, 刘虎, 冉雪, 郭豪

引用本文

孙开伟, 刘虎, 冉雪, 郭豪.基于多尺度原型分层匹配的小样本分割方法[J]. 计算机科学, 2023, 50(6A): 220300275-7.

SUN Kaiwei, LIU Hu, RAN Xue, GUO Hao. Few-shot Segmentation Based on Multi-scale Prototype Hierarchical Matching [J]. Computer Science, 2023, 50(6A): 220300275-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300275

基于多尺度原型分层匹配的小样本分割方法

孙开伟　刘　虎　冉　雪　郭　豪
重庆邮电大学数据工程与可视计算重点实验室　重庆 400065
(sunkw@cqupt.edu.cn)

摘　要　传统语义分割任务通常需要大量带标注的数据来进行训练,并且难以泛化至新的类别。小样本分割,旨在使用少量带标注的支持图像从查询图像中分割新类别目标对象。由于支持图像数据较少,从有限的支持图像中提取具有代表性的指导信息是小样本分割任务的重要挑战。为了解决这个问题,提出一种基于多尺度原型分层匹配的小样本分割方法。首先通过残差网络 ResNet 得到查询图像和支持图像的中层特征和高层特征;为进一步提取目标对象丰富的上下文特征信息,将提取的中层特征输入金字塔池化模块进行多尺度特征提取;最后基于原型学习的思想,对中层特征和高层特征分层生成原型并匹配修正,得到最终预测分割掩码。在 PASCAL-5^i 数据集上进行实验研究,实验结果表明,在 1-way 5-shot 的设定下,提出的方法在 mIoU 指标上达到了 66.7%,比当前主流模型 PANet 和 PFENet 分别提高了 11.0%和 4.8%,表明了该方法的有效性和先进性。

关键词:小样本分割;多尺度;语义分割;原型学习;残差网络
中图法分类号　TP391

Few-shot Segmentation Based on Multi-scale Prototype Hierarchical Matching

SUN Kaiwei,LIU Hu,RAN Xue and GUO Hao
Key Laboratory of Data Engineering and Visual Computing,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　Traditional semantic segmentation tasks usually need a lot of labeled data for training,and it is difficult to generalize to new categories. Few-shot segmentation aims to segment new categories of target objects from query images using a small number of annotated supporting images. Due to the limited supporting image data,how to extract representative guidance information from limited support images is an important challenge for few-shot segmentation task. In order to solve this problem,a few-shot segmentation method based on multi-scale prototype hierarchical matching is proposed in this paper. Firstly,the middle-level and high-level features of the query image and the support image are obtained through the residual network ResNet. In order to further extract the rich context feature information of the target object,the extracted middle-level features are fed into the pyramid pooling module for multi-scale feature extraction. Based on the idea of prototype learning,middle-level features and high-level features are layered to generate prototypes and matched to obtain the final predicted segmentation mask. Experiments are carried out on the PASCAL-5^i dataset and experimental results show that the mIoU of the proposed method achieves 66.7% in 1-way 5-shot setting,which is 11% and 4.8% higher than the current mainstream PANet and PFENet models,respectively,demonstrating the effectiveness and advanced nature of the method.

Keywords　Few-shot segmentation,Muti-scale,Semantic segmentation,Prototype learning,ResNet

1　引言

近年来,随着深度卷积神经网络和大数据技术的飞速发展[1-2],众多计算机视觉领域,例如图像分类[3]、图像分割[4-7]、目标检测[12]、目标跟踪[13]等,都在大规模数据集[8-9]的支撑下表现出了优异的性能。与此同时,卷积神经网络的缺点也暴露出来。为了充分发挥卷积神经网络的优越性,深度卷积神经网络往往需要使用大量的带标注的训练数据来训练复杂的模型,这将会耗费大量人工来进行手工标注,特别是在密集预测任务图像分割中,像素级别的图像标注获取成本昂贵、耗时长。此外,在传统语义分割任务中,模型只能准确分割出训练时包含的类别,当面临新的类别时,性能急剧下降,因而又需要再次获取该类别的标注数据。为了解决这个问题,有研究者提出了一些半监督方法和弱监督方法[12-13],一定程度上缓解了数据不足的困境。但是这类方法由于使用的带标注的数据不够多,训练出的模型往往泛化能力较差,并且也不能将模型推广至新的类别,仍然无法解决标注数据不足带来的问题。近年来,小样本分割引起了越来越多的关注,它参考人类视觉系统,在极其有限的监督条件下,能够轻易识别目标对象。举个例子,假设你从未见过狗这个类别,当给定你一张或几张包含狗的图片时,得益于人类强大的学习能力,你能轻松识别出任意一张其他图片中狗的部分。小样本分割模仿人类

基金项目:重庆市自然科学基金面上项目(cstc2019jcyj-msxmX0021);重庆市教委项目(KJCXZD2020027);国家自然科学基金(61806033)
This work was supported by the Natural Science Foundation of Chongqing,China(cstc2019jcyj-msxmX0021),Science and Technology Research Program of Chongqing Municipal Education Commission(KJCXZD2020027) and National Natural Science Foundation of China(61806033).
通信作者:刘虎(845904963@qq.com)

的视觉认知，通过少量标注的数据学习泛化能力强的图像分割模型。它将数据集划分为查询集和支持集两个集合，支持集负责提供目标信息，查询集在支持集的指导下完成对目标物体的分割。这种方法能够从带标注的数据中训练出一个具备一定快速学习能力的分割模型。利用该模型，在分割新的目标类别时，只需要使用少量的支持图像就能对查询图像进行标注。现有的方法[14-16]大多从支持图像中学习得到一组原型，通过支持图像的原型来指导查询图像上的像素级分割。然而这类方法大多将注意力集中在原型度量方法之上，对支持图像的信息提取缺少足够的关注，没有深入地探索支持图像的潜在语义信息，特别是图像中的上下文信息和图像的高层特征信息，从而使得学习到的原型对查询特征提供的语义指导有限，限制了分割模型的通用性。针对上述问题，本文基于原型学习的思想，提出了一种基于多尺度原型分层匹配的小样本分割方法。首先，将查询图像和支持图像输入残差网络 ResNet[3] 特征值提取器中，得到各自的中层特征和高层特征。然后，对于图像的中层特征，该方法引入了一个金字塔池化特征提取模块来进一步细化查询图像特征和支持图像特征，充分挖掘图像的上下文信息。最后，通过掩码平均池化来生成支持图像的目标类原型以及背景原型，原型与查询图像特征进行相似度匹配，生成相似度矩阵。为了准确使用高层特征，本文方法使用高层特征得到的原型对中层特征生成的原型进行相似度修正，得到最终的预测结果。在 PASCAL-5^i 数据集上进行的实验表明，本文所提方法在 1-way 5-shot 的设定下，mIoU 达到了 66.7%，比当前主流模型 PANet[15] 和 PFENet[20] 分别提高了 11.0%和 4.8%。

2 相关工作

语义分割是计算机视觉最基本的任务之一，其主要任务是对图像中的每一个像素按照预先设定的类别进行分类。相比传统的图像分类而言，语义分割的难度更大。全卷积神经网络 FCN[7] 是语义分割领域最早提出的深度学习架构，它使用 1×1 的卷积层取代了传统分类网络的全连接层，并且融合了不同尺度的特征，开辟了深度学习语义分割的研究道路。目前主流的语义分割方法[4-7]通常采用编码器-解码器结构。编码器负责提取图像中各个维度的特征，分步细化分割，最后对特征进行聚合。解码器负责对编码器提取的特征进行解码还原，并预测最后的分割掩码。尽管这些方法在语义分割领域中取得了重大突破，并不断刷新图像分割的精度，但它们仍然存在着“数据驱动性能”的明显缺点，即在训练数据不足的情况下缺乏通用性。

为了解决语义分割泛化能力不足以及数据匮乏问题，研究者提出了小样本分割任务。小样本分割任务的目的是从有限的带标注的支持图像中学习可转移知识，从而完成查询图像中新的类别目标的分割。Shaban 等[17]引入了第一个基于双分支结构的小样本分割网络 OSLSM，奠定了小样本分割双分支架构的基础。它包括一个支持分支以及一个查询分支。支持分支从支持图像标注中提取具有代表性的高层特征，查询分支整合从支持分支学习到的知识，最后在查询图像上生成分割掩码。PL[14]首次将原型学习引入小样本分割，生成支持图像中的类别原型，并使用非参数化模块计算原型与查询图像像素的余弦相似度来预测分割掩码。PANet[15]在此基础上引入了一种正则化原型对齐方法，对原型进行一个双向学习，优化了预测结果。这种基于原型学习的方法大多忽略了特征提取的过程，导致图像中的一些空间信息没有被充分利用，甚至出现了误用的情况，生成的原型代表性较差。因此，如何有效提取图像中丰富的上下文信息和全局信息，成为语义分割领域的研究热点。越来越多的研究发现，多尺度特征处理对模型理解上下文信息至关重要。DeepLab[5]设计了一种空洞空间金字塔池化模块 ASPP，采用空洞卷积在多个尺度上扩大感受野来解决多尺度分割目标的问题，效果非常明显。HRNet[18]提出了一种全新的特征增强方法，将特征提取划分为多个阶段，同时对各阶段的特征信息进行相互交汇，实现了特征多尺度融合效果。本文采用了 PSPNet[4]提出的金字塔池模块 PPM 来实现多尺度特征处理，该模块可以在不同区域进行上下文信息聚合，同时缓解了图像规模不一致的问题。

3 本文方法

3.1 问题描述

小样本分割的目标是使用一些基本的类别来生成一个通用的分割模型，生成的模型在不需要重新训练的情况下，仅仅通过少量新类别的标注数据就能够对任意一张图像中与标注数据类别相同的目标类进行分割。设训练数据为 D_{train}，测试数据为 D_{test}，其中，对任意属于 D_{train} 的类别 C_{train} 和属于 D_{test} 的类别 C_{test}，有 $C_{\text{train}} \neq C_{\text{test}}$。在训练阶段，将数据集 D_{train} 分为多个片段，每个片段包含一个查询集 $Q=\{I_q, M_q\}$ 和一个支持集 $S=\{I_s^i, M_s^i\}(i=1,\cdots,k)$，$I_q$ 和 I_s 分别表示查询图像和支持图像，M_q 和 M_s 分别表示查询图像和支持图像各自的真实掩码，i 表示使用多个支持图像时的支持图像和真实掩码的编号。在每一个片段中，查询集中的类别与支持集的类别保持一致，支持集中通常包含 K 个训练样本和 C 个类别，将其定义为 C-way K-shot 设定，本文主要采用 1-way 1-shot 和1-way 5-shot 两种设定进行实验。在训练阶段，每个片段最终得到一个预测掩码 M_q^p，通过计算其与真实掩码 M_q 的交叉熵损失 $L(M_q, M_q^p)$ 来更新模型权重，直至模型收敛。在测试阶段，按照训练阶段相同的数据获取方式获得测试的片段，将测试集 D_{test} 中每个片段的查询集和支持集输入到训练阶段生成的分割模型中，完成对查询集查询图像的分割。整个测试阶段分割目标的类别都是没有出现在训练过程的、全新的类别。

3.2 网络结构

本文方法基于原型学习思想，设计了一种多尺度原型分层匹配的小样本分割网络，它的整体结构如图 1 所示。与大部分主流小样本分割框架相似，该方法采用查询分支来提取查询图像特征，支持分支来提取支持图像特征。整个网络流程如下：1）将查询图像和支持图像输入各自对应的分支，首先通过预训练特征提取器 ResNet[1]，生成各自的中层特征 f_q，f_s 和高层特征 F_q^h，F_s^h；2）为了充分利用图像中丰富的上下文信息，将中层特征 f_q 和 f_s 输入一个金字塔池化模块 PPM 中对查询特征和支持特征进行多尺度细化，得到最终的中层查询特征 F_q 和支持特征 F_s；3）使用支持图像的真实掩码与支持特征进行掩码平均池化，得到支持图像中层目标类原型 P_c、高层目标类原型 P_c^h、中层背景原型 P_b、高层背景原型 P_b^h。对目标类原型 P_c 和 P_c^h 与查询特征 F_q 和 F_q^h 计算余弦相似度，

进行原型匹配，得到目标相似度矩阵 M_c 和 M_c^h，使用同样的方式得到背景相似度矩阵 M_b 和 M_b^h；4）使用高层特征生成的相似度矩阵对中层特征生成的相似度矩阵进行修正，并取二者最优作为最终预测结果。

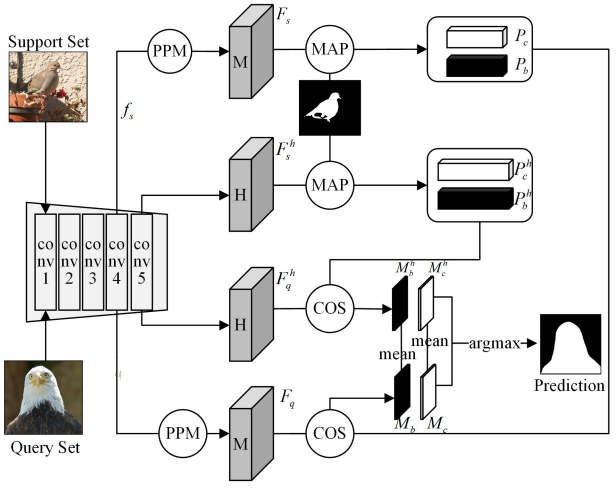


图 1　本文方法总体结构

Fig. 1　Framework of the proposed method

3.3　特征提取

残差网络 ResNet[1] 是常用的特征提取器之一，采用该结构提取特征能够有效防止由网络深度过大而带来的梯度消失现象。残差网络根据层数的不同，可以划分为 ResNet18，ResNet34，ResNet50，ResNet101 4 个版本，本文选择了 ResNet50 和 ResNet101 两个版本作为特征提取器进行实验。ResNet50 以上版本从第二个大层开始，每个大层由多个相同的 bottleneck 堆叠而成。bottleneck 结构如图 2 所示，它由两个 1×1 卷积和一个 3×3 卷积组成，在最后阶段对输入和输出进行残差连接。不同 bottleneck 的区别在于输入通道数 a 和输出通道数 b 不同。

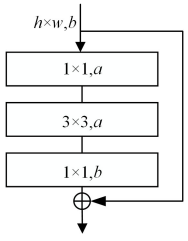


图 2　ResNet 中 bottleneck 的结构

Fig. 2　Structure of bottleneck in ResNet

表 1 列出了 ResNet50 和 ResNet101 的 4 个大层中 bottleneck 输入、输出通道数以及结构组成明细。可以看出 ResNet50 和 ResNet101 的不同之处在于，ResNet50 在第四个大层 conv4 上比 ResNet101 少 17 个 bottleneck。本文方法通过查询分支和支持分支提取出支持图像和查询图像的中层特征和高层特征，查询分支和支持分支使用的特征提取器相同并且共享权重。查询图像 I_q 和支持图像 I_s 经过 ResNet 前 4 个大层后得到中层特征 f_q 和 f_s，其表达式如下：

$$f_q = conv4(conv3(conv2(conv1(I_q)))) \tag{1}$$

$$f_s = conv4(conv3(conv2(conv1(I_s)))) \tag{2}$$

其中，$conv1$，$conv2$，$conv3$，$conv4$ 分别表示 ResNet 的前 4 个大层。将查询图像和支持图像中层特征进一步输入 ResNet 第五个大层后，得到高层特征 F_q^h 和 F_s^h，其表达式如下：

$$F_q^h = conv5(f_q) \tag{3}$$

$$F_s^h = conv5(f_s) \tag{4}$$

其中，$conv5$ 表示 ResNet 的第五个大层，f_q 表示查询图像的中层特征，f_s 表示支持图像的中层特征。

表 1　ResNet50 和 ResNet101 中各个大层明细

Table 1　Details of each large layer in ResNet50 and ResNet101

layer	ResNet50	ResNet101
conv2	(64,256)×3	(64,256)×3
conv3	(128,512)×4	(128,512)×4
conv4	(256,1024)×6	(256,1024)×23
conv5	(512,2048)×3	(512,2048)×3

3.4　金字塔池化模块

由于图像中的目标对象大小及位置不一，并且支持图像与查询图像的规模存在较大的差异，仅仅使用特征提取器生成的特征图来产生原型，无疑会丢失很多子区域的上下文连续信息。为了提高图像的解析性能，充分挖掘探索支持图像和查询图像的全局信息，本文使用种金字塔池化模块[4] PPM 来对中层特征进行多尺度提取。金字塔池化模块的结构如图 3 所示，从左往右看，金字塔池模块包括 4 个步骤：金字塔池化、卷积、上采样和拼接操作。通过金字塔池化，可以得到多种不同空间尺度上的特征，同时捕捉各个子区域间的连续特征，本文最终使用了 4 个池化核，它们的大小分别是 1×1，2×2，3×3 和 6×6，在 4.5 节证实了该组合的有效性。为了提高多尺度特征的非线性学习能力，进一步采用 1×1 卷积来保持特征图的大小，同时将每个特征的通道数减少 $1/N$，N 表示金字塔池化核的个数，在本文中 N 为 4。

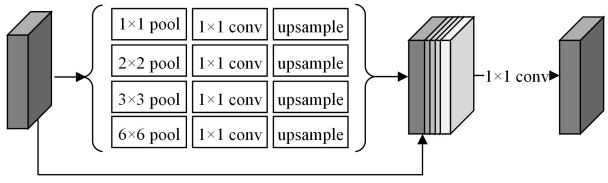


图 3　金字塔池模块结构

Fig. 3　Structure of pyramid pool module

在经过卷积层之后，使用双线性插值方法进行上采样操作，将 4 个尺度的特征图恢复至输入特征图相同的尺寸大小，将输入的特征图与 4 个经过上采样后的特征拼接起来，从而保持多个尺度特征的全局上下文信息，提高模型的学习能力。最后，再通过一个 1×1 卷积将拼接后的特征图的通道数恢复至输入特征图相同的通道数，中层特征 f_q 和 f_s 通过金字塔池化模块后得到最终的查询特征 F_q 和支持特征 F_s，其表达式如下：

$$F_q = conv(concat(Up(conv(P^i(f_q))))) (i=1,2,3,6) \tag{5}$$

$$F_s = conv(concat(Up(conv(P^i(f_s))))) (i=1,2,3,6) \tag{6}$$

其中，$concat()$ 表示向量拼接操作，$Up()$ 表示双线性插值上采样，$conv()$ 表示 1×1 卷积，P_i 表示 $i\times i$ 池化。

3.5　原型匹配及结果修正

基于原型学习的思想，本文使用掩码平均池化[21]（Masked average Pooling，MAP）将支持图像中目标类和背景映射成相同特征空间中的两个原型，这种方法在提取目标原型的同时忽略了背景对象的干扰，能够更好地生产图像原型。具体地，对于支持图像中层特征 F_s 中目标类 c 的原型 P_c 生成有：

$$P_c=\frac{\sum_{i}^{h\times w}F_s(i)\cdot[M_s(i)=c]}{\sum_{i}^{h\times w}[M_s(i)=c]} \tag{7}$$

对于背景原型 P_b 生成有：

$$P_b=\frac{\sum_{i}^{h\times w}F_s(i)\cdot[M_s(i)\neq c]}{\sum_{i}^{h\times w}[M_s(i)\neq c]} \tag{8}$$

其中，i 表示每个像素点的索引；h 和 w 为特征图的高度和宽度；[·]为艾弗森方括号，如果括号内的条件满足则值为 1，不满足则值为 0；M_s 为支持图像的真实掩码，$M_s(i)=c$ 表示第 i 个像素属于 c 类。以上是支持集中只含有一个支持图像的情况下目标类和背景原型的生成及结果预测，对于含 K 个支持图像的支持集，只需要对各自的原型直接向量相加求和并取平均值即可。相关研究[22-23]中的可视化结果展示，高层特征包含的信息大多为抽象级别的信息，比如对象的类别，而这并不适合使用在小样本分割领域，因为它预测的类别是模型从未见过的类别，直接使用经过所有大层生成的高层特征，将导致模型泛化能力下降。因此，本文方法对支持图像的高层特征进行了分层处理，利用高层特征生成的原型对中层特征的原型进行修正。具体地，根据以上式(7)、式(8)，使用支持图像高层特征 F_s^h 生成支持图像高层特征的目标类原型 P_c^h 和背景 P_b^h。接着引入无参化学习方法来对原型进行匹配，分割过程可以看作是对空间位置的分类，如图 4 所示，黑色圆圈和白色圆圈以及灰色三角形分别代表着支持图像目标类、支持图像背景和查询图像在同一空间的映射，五角星代表支持图像学习生成的原型，虚线表示彼此之间的距离。通过计算在同一特征空间上支持图像原型与查询特征之间的距离得到相似度矩阵，将查询图像特征与映射空间中最近、最相似的原型进行匹配，从而对查询图像进行分割。

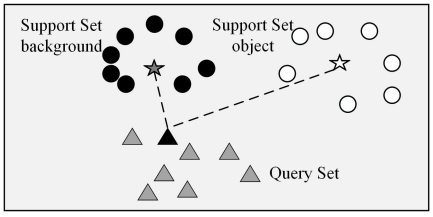


图 4 无参化原型匹配过程

Fig. 4 Nonparametric prototype matching

中层、高层特征生成的目标类相似度矩阵 M_c，M_c^h 和背景相似度矩阵 M_b，M_b^h 的表达式如下：

$$M_c=\cos(F_q,P_c) \tag{9}$$

$$M_b=\cos(F_q,P_b) \tag{10}$$

$$M_c^h=\cos(F_q^h,P_c^h) \tag{11}$$

$$M_b^h=\cos(F_q^h,P_b^h) \tag{12}$$

其中，F_q 表示查询图像的中层特征，F_q^h 表示查询图像的高层特征，cos()表示两个向量的余弦相似度。进一步，使用高层特征对中层特征进行相似度修正，得到两个层融合目标类相似度矩阵 M^c 以及背景相似度矩阵 M^b，其表达式如下：

$$M^c=\frac{1}{2}(M_c+M_c^h) \tag{13}$$

$$M^b=\frac{1}{2}(M_b+M_b^h) \tag{14}$$

其中，M_c 和 M_c^h 分别表示中层特征目标类相似度矩阵和高层特征目标类相似度矩阵，M_b 和 M_b^h 分别表示中层特征背景相似度矩阵和高层特征背景相似度矩阵。最终的预测掩码 M_q^p 计算表达式如下：

$$M_q^p=\text{argmax}(concat(M^c,M^b)) \tag{15}$$

其中，argmax()表示取大值时对应的坐标。

4 实验结果及分析

4.1 数据集

为了保证实验的科学性以及有效性，与已有方法一样，使用 PASCAL-5^i 数据集进行模型的训练和测试。该数据集由 PASCAL VOC 2012[9]以及 SBD[30]组成，其中训练图像样本个数为 10 582，测试图像样本个数为 1 449。它由 Shaban 等[17]首次创建，之后广泛应用于小样本分割任务当中。PASCAL-5^i 数据集总共包含 20 个类别，按照类别将其平均分为 4 份，每份包含 5 个类别，具体的类别划分如表 2 所列。

表 2 PASCAL-5^i 4 份子集的类别划分

Table 2 Four subsets categoriespartition of PASCAL-5^i

Fold	Contained categories
split0	aeroplane, bicycle, bird, boat, bottle
split1	bus, car, cat, chair, cow
split2	dining table, dog, horse, motorbike, person
split3	potted plant, sheep, sofa, train, tv/monitor

本文使用交叉验证的方式来训练模型，将其中 3 份中的类别作为训练集训练模型，剩余 1 份中的类别作为测试集，来评估模型。测试阶段，从测试数据集中随机抽取 1000 个片段进行模型测试，每个片段的类别相同，且属于对应份数的划分类别。主要测试了 1-way 1-shot 和 1-way 5-shot 的设定下的实验结果，为了使实验结果更具有说服力，在测试阶段，本文使用 5 个不同随机种子进行 5 次测试，最后取平均值来作为最后的实验结果。

4.2 评价指标

交并比(Intersection of Union，IoU)是图像分割任务中常用的评价指标之一，其计算式如下：

$$IoU=\frac{TP}{FP+TP+FN} \tag{16}$$

其中，TP，FP 和 FN 分别是分割掩码集合中的真正例、假正例和假负例像素点个数。本文方法采用平均交并比 mIoU 和前景背景交并比 FB-IoU 作为评估指标。mIoU 是所有类别的 IoU 值的平均值，即 $mIoU=(IoU_1+IoU_2+\cdots+IoU_C)/C$，其中 C 是测试集中的类别数，IoU_i 是 i 类别的交并比。FB-IoU 则忽略对象类别，只计算前景和背景 IoU 值的平均值，即 $FB\text{-}IoU=(IoU_F+IoU_B)/2$，其中 IoU_F 和 IoU_B 分别为目标类别和背景的 IoU 值。由于 mIoU 综合了各个目标类别的实验差异，相比 FB-IoU 能更好地反映模型的泛化能力和预测效果，因此实验过程中更多关注的是 mIoU 值，FB-IoU 作为附加指标与其他方法进行对比。

4.3 实验设置

本文使用深度学习框架 Pytorch[24]构建模型并完成模型训练测试，所有实验均在一张 NVIDIA Tesla V100 显卡上进行，骨干网络 ResNet50，ResNet101 预训练权重来自在 ImageNet[8]上训练的公开权重，采用 SGD 优化器进行模型参数优化，初始学习率设置为 0.001，动量为 0.9，权重衰减为 0.0005。训练数据和真实掩码宽高调整至 473×473，模型训练过程中只使用了随机水平翻转进行图像增强。

4.4 实验结果

本文将实验结果同目前的主流方法实验结果进行对比。从表 3 中可以看出,在 1-way 5-shot 的设定下,本文提出的方法取得了所有方法中的最佳实验结果,加粗数据表示所列方法中最好结果,Δ 表示 1-shot 和 5-shot 差值,差值越大说明增加支持图像后模型更为精准。本文使用 ResNet101 作为主干网络的网络结构在 1-way 1-shot 的设定下 mIoU 达到了 57.9%,相比 PANet 提高 11.5%,相比 CANet 提高 2.5%,比表中最好的方法 PFENet 仅低 2.2%;在第三份测试集中,mIoU 达到 58.5%,优于其他所有方法。使用 ResNet101 作为主干网络的网络结构在 1-way 5-shot 的设定下,本文方法的 mIoU 达到了 66.7%,超过了表中所有的方法,超过最接近本文方法的 PFENet 4.8%,并且在每一份测试集中都取得了当前方法中的最佳效果。除此之外,本文方法在以 ResNet50 和 ResNet101 为主干网络下,从 1-shot 增加至 5-shot 时,mIoU 提升了 8.2%和 8.8%,明显优于其他方法,进一步证明了本文提出的方法能够充分使用仅有的几个支持集样本的有效信息,在增加支持图像后能够大幅提升性能。在表 4 中,本文还将 FB-IoU 实验结果同目前的主流模型进行了对比,在 5-shot 设定下 FB-IoU 达到了 76.6%,同样取得了最好的分数,并且从 1-shot 增加至 5-shot 时,FB-IoU 提升了 7.0%,提升效果同样明显。

表 3 不同分割方法在 1-way 1-shot 和 1-way 5-shot 设定下 mIoU 的比较

Table 3 Comparison of mIoU of different segmentation methods under 1-way 1-shot and 1-way 5-shot settings

Methods	1-shot					5-shot					Δ
	0	1	2	3	mean	0	1	2	3	mean	
OSLSM[17]	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9	3.1
co-FCN[25]	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4	0.3
SG-One[21]	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1	0.8
PANet[15]	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7	7.6
PGNet[26]	56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5	2.5
PFENet[20]	**61.7**	**69.5**	55.4	**56.3**	**60.8**	63.1	70.7	55.8	57.9	61.9	1.1
CANet[19]	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1	1.7
PMMs[29]	55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3	1.0
FWB[27]	51.3	64.5	56.7	52.2	56.2	54.8	67.4	62.2	55.3	59.9	3.7
DAN[28]	54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5	2.3
Ours(ResNet50)	53.8	66.9	56.9	50.4	57.0	63.5	70.5	67.7	59.2	65.2	8.2
Ours(ResNet101)	57.8	65.4	**58.5**	50.0	57.9	**63.9**	**73.6**	**68.6**	**60.7**	**66.7**	**8.8**

表 4 不同分割方法在 1-way 1-shot 和 1-way 5-shot 设定下 FB-IoU 的比较

Table 4 Comparison of FB-IoU of different segmentation methods under 1-way 1-shot and 1-way 5-shot settings

Methods	1-shot	5-shot	Δ
OSLM	61.3	61.5	0.2
co-FCN	60.9	60.2	0.5
SG-One	63.1	65.9	2.8
CANet	66.2	69.6	3.4
PANet	66.5	70.7	4.2
PFENet	73.3	73.9	0.6
Ours	69.6	**76.6**	**7.0**

4.5 对比实验

为了有效地对特征进行多尺度提取,本文对金字塔结构中的池化核大小设定进行了对比实验。由于设备有限,本文主要对 6 种组合进行了实验。如表 5 所列,在使用 3 个池化核时,池化核组合 1×1,2×2,3×3 将 3×3 池化核变为 6×6 后 5-shot 下 mIoU 下降了 1.2%,将 2×2 替换为 6×6 后 5-shot 下 mIoU 下降了 2.6%,将 1×1 替换为 6×6 后 5-shot 下 mIoU 下降了 3.4%,可以初步得出较小的池化核起到的作用更大,能够汇聚更多的全局信息。进一步探索池化核个数对实验的影响,可以看出,当池化核达到 4 个时,效果明显好于 3 个池化核,同时 2×2,3×3,6×6,12×12 组合相较于 1×1,2×2,3×3,6×6 组合在 5-shot 设定下 mIoU 下降了 1.9%,这也和上面得出的较小的池化核作用更大保持一致。最终本文选择 1×1,2×2,3×3,6×6 作为金字塔模块中的池化核组合。

表 5 不同池化核组合在 1-way 1-shot 和 1-way 5-shot 设定下 mIoU 的比较

Table 5 Comparison of mIoU with different pooling kernel combinations under 1-way 1-shot and 1-way 5-shot settings

size	1-shot	5-shot
1×1,2×2,3×3	56.4	62.8
1×1,2×2,6×6	56.7	61.6
1×1,3×3,6×6	56.9	60.2
2×2,3×3,6×6	56.9	59.4
2×2,3×3,6×6,12×12	56.8	63.3
1×1,2×2,3×3,6×6	**57.0**	**65.2**

主干网络中不同阶层代表的信息不同,改变主干网络特征的抽取方式能够使网络性能大幅度提升。本文以 ResNet50 为主干网络进行实验,对原型修正前后的实验进行了对比。如表 6 所列,如果对高层特征不做处理,直接使用其生成原型,在 1-shot 和 5-shot 设定下,mIoU 分别仅为 52.3%和 58.5%。在对高层特征进行修正后 mIoU 提升明显,1-shot和 5-shot 设定下分别提高了 4.9%和 6.7%,证明了对特征进行分层匹配并加以修正能够取得更优异的性能。

表 6 高层特征修正消去实验结果

Table 6 Ablation experimental results(mIoU) of fixed high-level feature

conv5	fixed	1-shot	5-shot
√		52.3	58.5
	√	**57.0**	**65.2**

4.6 可视化结果

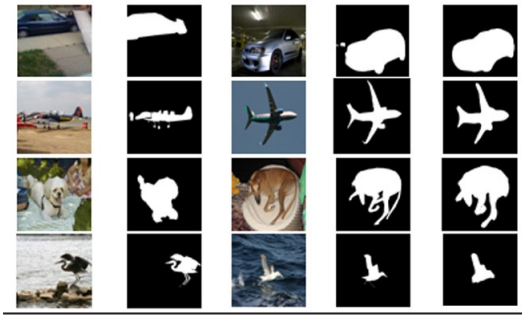
表 7 列出了本文在 PASCAL-5^i集上 1-way 1-shot 设置下部分测试类别的可视化结果,测试类别都是模型训练过程中

从未见过的全新类别。第一列是支持图像，用于指导查询集分割，第二列是支持图像的真实掩码，第三列是查询图像，第四列是查询图像的真实掩码，最后一列是模型最终的分割结果。可以看出，查询图像仅在一张支持图像的指导下，比较精准地区分了背景和类别对象，并生成了目标分割掩码。在给出的展示中，支持图像中的目标很小，并且与查询图像空间位置差异较大，但最终查询图像的分割效果仍然保持较高的准确度，进一步证明了本文方法的有效性。

表 7 1-way 1-shot 设定下部分测试类别的可视化结果

Table 7 Visual results of some test categories under 1-way 1-shot setting

Support	Support Mask	Query	Query Mask	Prediction



尽管当前模型分割精度已经达到较高水平，但仍然存在少部分分割效果较差的结果，本文也对其失败原因进行了分析。如表 8 所列，由于查询图像中目标类别和背景区分度较低，并且支持图像的标注过于精细，导致出现了一定的过拟合现象，最终预测效果较差。这也是小样本分割任务今后研究的重点，即在对精度要求较高的场景下保持分割性能。

表 8 部分分割效果较差测试类别的可视化结果

Table 8 Visual results of some test categories with poor segmentation effects

Support	Support Mask	Query	Query Mask	Prediction

结束语 针对支持集和查询集特征提取方式单一导致图像特征提取不完全和高阶特征导致模型分割精度下降问题，本文提出了一种多尺度原型分层匹配小样本分割网络。通过对主干网络高层特征的融合修正以及对中层特征的多尺度处理，有效获取了支持集和查询集丰富的上下文信息，分层生成原型并匹配的方式显著提高了模型的通用性。在 PASCAL-5^i 数据集上的实验结果以及可视化展示证明了本文方法的有效性。未来工作将在原型度量方式上结合多层特征进一步提升模型效果。

参 考 文 献

[1] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[2] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv: 1409.1556, 2014.

[3] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[4] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.

[5] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[6] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 801-818.

[7] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks for Semantic Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[8] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database [C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.

[9] EVERINGHAM M, ESLAMIS M A, VAN GOOL L, et al. The Pascal Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[10] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[11] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[12] KOH J Y, NGUYEN D T, TRUONG Q T, et al. SideInfNet: A Deep Neural Network for Semi-Automatic Semantic Segmentation with Side Information[C]// European Conference on Computer Vision. Cham: Springer, 2020: 103-118.

[13] LUO W, YANG M. Semi-supervised Semantic Segmentation via Strong-Weak Dual-Bran-ch Network[C]// European Conference on Computer Vision. Cham: Springer, 2020: 784-800.

[14] DONG N, XING E P. Few-Shot Semantic Segmentation with Prototype Learning [J]. British Machine Vision Conference, 2018, 3(4): 79.

[15] WANG K, LIEW J H, ZOU Y, et al. PANet: Few-Shot Image Semantic Segmentation with Prototype Alignment [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9197-9206.

[16] LIU J, QIN Y. Prototype Refinement Network for Few-Shot Segmentation[J]. arXiv: 2002.03579, 2020.

[17] SHABAN A, BANSAL S, LIU Z, et al. One-Shot Learning for Semantic Segmentation [C]// British Machine Vision Confe-

rence. 2017：167. 1-167. 13.

[18] HUANG J，ZHU Z，HUANG G. Multi-Stage HRNet：Multiple Stage High-ResolutionNetwork for Human Pose Estimation[J]. arXiv：1910. 05901，2019.

[19] ZHANG C，LIN G，LIU F，et al. CANet：Class-Agnostic Segmentation Networks with Iterative Refinement and Attentive Few-Shot Learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：5217-5226.

[20] TIAN Z，ZHAO H，SHU M，et al. Prior Guided Feature Enrichment Network for Few-Shot Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2020(1)：1-1.

[21] ZHANG X，WEI Y，YANG Y，et al. SG-One：Similarity Guidance Network for One-Shot Semantic Segmentation[J]. IEEE Transactions on Cybernetics，2020，50(9)：3855-3865.

[22] YOSINSKI J，CLUNE J，NGUYEN A，et al. Understanding Neural Networks Through Deep Visualization[J]. arXiv：1506. 06579，2015.

[23] ZEILER M D，FERGUS R. Visualizing and Understanding Convolutional Networks [C]// European Conference on Computer Vision. Cham：Springer，2014：818-833.

[24] PASZKE A，GROSS S，MASSA F，et al. PyTorch：An Imperative Style，High-Performance Deep Learning Library[J]. Advances in Neural Information Processing Systems，2019，32：8026-8037.

[25] RAKELLY K，SHELHAMER E，DARRELL T，et al. Conditional Networks for Few-Shot Semantic Segmentation[J/OL]. (2018-04-04)[2021-12-11]. https：//openreview. net/pdf? id=SkMjFKJwG.

[26] ZHANG C，LIN G，LIU F，et al. Pyramid Graph Networks with Connection Attentions for Region-Based One-Shot Semantic Segmentation [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9587-9595.

[27] NGUYEN K，TODOROVIC S. Feature Weighting and Boosting for Few-Shot Segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：622-631.

[28] WANG H，ZHANG X，HU Y，et al. Few-Shot Semantic Segmentation with Democratic Attention Networks[C]// Computer Vision-ECCV 2020：16th European Conference. Springer International Publishing，2020：730-746.

[29] YANG B，LIU C，LI B，et al. Prototype Mixture Models for Few-Shot Semantic Segmentation [C]// European Conference on Computer Vision. Cham：Springer，2020：763-778.

[30] HARIHARAN B，ARBELÁEZ P，BOURDEV L，et al. Semantic Contours from Inverse Detectors[C]// 2011 International Conference on Computer Vision. IEEE，2011：991-998.

SUN Kaiwei，born in 1987，Ph. D，associate professor. His main research interests include machine learning，data mining and big data analysis.

LIU Hu，born in 1997，postgraduate. His main research interests include computer tervision，semantic segmentation and so on.