小样本语义分割研究现状与分析

陈善娟 于云龙 李英明
(浙江大学信息与电子工程学院 杭州 310058)

摘 要 传统语义分割任务通常是数据驱动的,需要大规模密集标注样本训练,并且不能泛化到新类,因此在实际应用中受到很大限制.为了缓解传统语义分割中数据匮乏和泛化能力差的问题,人们提出小样本语义分割任务,在未见类别仅提供少量密集标注样本的情况下实现新类分割,在医疗图像分割、自动驾驶等应用领域扮演着重要的角色,已成为计算机视觉领域的重要研究方向之一.本文从基础知识、模型算法和拓展应用等方面对自然图像领域的小样本语义分割研究展开调查,具体包含以下内容:(1)介绍了小样本语义分割的背景知识,包括它的由来、核心思想、概念知识、存在挑战、数据集和性能评价指标;(2)详细分析和比较当前小样本语义分割算法,根据推理过程中是否存在梯度回传将其分为基于优化和基于度量学习的方法,并归纳了其发展现状和不同算法的优缺点;(3)介绍了小样本语义分割与其他技术融合的任务,包括小样本实例分割、广义小样本分割、增量小样本分割、弱监督小样本分割及跨域小样本分割;(4)讨论了小样本语义分割任务仍存在的问题和未来展望.

关键词 小样本学习;图像语义分割;元学习;迁移学习;深度学习
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.02417

Research Status and Analysis of Few-Shot Semantic Segmentation

CHEN Shan-Juan YU Yun-Long LI Ying-Ming
(*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou* 310058)

Abstract The traditional semantic segmentation task is usually data-driven, which requires large-scale intensive annotation data for training. And it cannot be generalized to novel class segmentation which means that when novel class samples emerge, it is necessary to collect a large amount of data to retrain the model. So these issues severely limit its practical application. In order to alleviate the issues of data scarcity and poor generalization ability in traditional semantic segmentation task, the few-shot semantic segmentation (FSS) task has been proposed, which can utilize the past accumulated knowledge to achieve the segmentation of novel classes in images that are not included in the training process with only a limited number of densely labeled samples. FSS can significantly reduce the need for data annotation and demonstrates good generalization performance. It has attracted widespread attention from the academic community. This research filed is crucial for practical applications where data availability is limited and difficult, such as medical field, remote sensing filed and so on. In recent years, there has been rapid development and progress in the field of few-shot semantic segmentation. A large number of novel and high-performing methods have emerged, urgently requiring a comprehensive review to summarize and sort out these algorithms. This paper investigates the research on few-shot semantic segmentation in the field of natural images from the aspects of basic knowledge, algorithms summarization, extended applications and future development. We firstly introduce the basic knowledge of few-shot semantic segmentation

收稿日期:2023-09-10;在线发布日期:2024-07-05.本课题得到浙江省自然科学基金项目(LD24F020016)、国家自然科学基金(62002320,U19B2043)、浙江省省重点研发计划项目(2023C01043,2021C01119)、宁波市重点研发计划项目(2023Z236)资助.**陈善娟**,博士研究生,主要研究领域为机器学习、小样本分割.E-mail:chensj2021@zju.edu.cn.**于云龙**(通信作者),博士,特聘研究员,中国计算机学会(CCF)会员,主要研究领域为机器学习、计算机视觉.E-mail:yuyunlong@zju.edu.cn.**李英明**,博士,副教授,主要研究领域为机器学习、多任务学习.

task, including its origin, core ideas, conceptual knowledge, challenges, dataset benchmarks, and evaluation metrics. Then we have analyzed, compared the current few-shot semantic segmentation methods in detail, and summarized them accordingly. Specifically, we divide those methods into optimization-based and metric-learning-based methods according to whether there is gradient backpropagation during inference. When it comes to optimization-based methods, there is a need for gradient backpropagation to optimize the model during inference, whereas metric-learning-based approaches do not require any optimization of the model during inference. Instead, they utilize densely labeled images to obtain information about the categories to be segmented and use this class-specific information to guide the segmentation process. This allows for a more efficient and direct utilization of the labeled data without the need for further model training or optimization. We have conducted a comparison of the introduced few-shot semantic segmentation algorithms, listing the best performance achieved by each algorithm and providing quantitative analysis on three datasets. The comparison was made from three perspectives: the performance comparison on each dataset under different evaluation metrics, the performance comparison of a single algorithm across different dataset benchmarks, and the results comparison of different algorithms on the PASCAL-5i dataset. In addition, we introduce some tasks of integrating few-shot semantic segmentation and other technologies, such as few-shot instance segmentation, generalized few-shot semantic segmentation, incremental few-shot semantic segmentation, weakly supervised few-shot semantic segmentation, and cross-domain few-shot semantic segmentation. Finally, we discuss the existing problems and future development trend of few-shot semantic segmentation. We believe that this paper will help researchers better understand the current research status of few-shot semantic segmentation, provide a broader perspective for future studies, and promote further development and innovation in this field.

Keywords few-shot learning; image semantic segmentation; meta learning; transfer learning; deep learning

1 引 言

图像语义分割[1-3]作为计算机视觉中的基本任务之一,在图像分析和理解方面具有重要的地位,并广泛应用于自动驾驶、缺陷检测、医学成像和人机交互等领域.虽然深度学习算法的快速发展显著提升了分割性能,但仍存在一些挑战.首先,训练语义分割模型需要大量密集标记的训练数据,这导致了高昂的数据收集和标记成本.其次,传统的语义分割模型无法分割训练阶段未见类别的样本,限制了其在实际应用中的适用性.

传统语义分割模型的局限性促使研究人员寻找新的方法.近年来,受到小样本学习启发,小样本语义分割任务[4]被提出,旨在降低对大规模标记数据的需求,该任务能够在有限标记样本的情况下实现图像中新对象类的分割.在推理时仅利用少量(一个或者几个)具有密集标注的样本就能实现新类泛化,这对于解决实际应用中的数据匮乏等问题具有重要意义.此外,小样本语义分割任务还具有较高的理论研究价值,通过研究可以更好地理解迁移学习[5]和元学习[6]等机制,进一步推动计算机视觉领域的发展.

小样本语义分割任务自提出以来受到了学术界的广泛关注,并且在算法上取得了显著进展.近年来,在国际会议和学术期刊上发表了许多关于小样本语义分割的论文,这些论文主要集中在计算机视觉、机器学习和人工智能等领域.图 1 统计了 2017 年到 2023 年在 NeurIPS、CVPR、ECCV、ICCV、IJCAI 和 AAAI 这六个国际会议上发表的论文数量,从 2017 年到 2022 年论文数量显示该领域的研究呈逐年增长趋势,2023 年在这六大会议上发表论文数有所下降,但仍涌现出了许多新的方法.针对小样本语义分割任务面临的问题和挑战,研究学者们提出了许多创新的模型架构、数据增强技术和训练策略等解决小样本语义分割任务.这些算法的研究内容包括但不限于元学习、自监督学习[7]、迁移学习等方法.

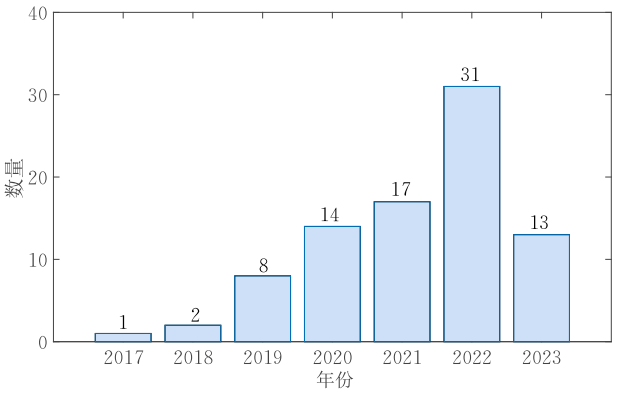


图 1　2017 年～2023 年小样本语义分割论文统计

国内外有一些综述文章对小样本语义分割方法进行了分类和总结，为该领域的研究提供了有益的参考. 陈琼等人[8]根据度量工具是否可学习，将小样本语义分割算法分为基于参数结构和基于原型结构两类. 韦婷等人[9]则基于网络结构对小样本语义分割方法进行了分类，将其分为基于孪生神经网络、基于原型网络和基于注意力机制三大类. 这两篇中文综述主要是对 2022 年之前的工作进行总结归纳，并从不同角度出发对小样本图像语义分割算法进行分类. 2022 年后该研究领域涌现出了大量新颖且性能出色的方法，亟须一篇新的综述对这些新颖算法进行归纳和梳理. Ren 等人[10]对零样本和小样本设置下的图像语义分割、视频分割和三维空间中的点云分割进行分类介绍，该论文基于这几项任务的共性，从概率角度出发对这些任务进行分类，将小样本语义分割问题约束为预测查询图像的后验概率，将其分为基于区分性和基于生成的方法. 其中，基于区分性的方法尝试构建模型来最大化查询图像的后验概率 $P(m|x)$，而基于生成的方法旨在建模 $P(x|m)P(m)$，构建生成器拟合样本的分布或类别特征，其中 $P(m)$ 通常假设为均匀分布. 由于基于生成的方法训练难度大、推理耗时，当前小样本图像语义分割算法大多属于基于区分性的方法，进一步可以分为基于度量学习的、基于参数预测的、基于微调的和基于存储的方法这四类. 该综述旨在介绍在零样本和小样本设置下不同视觉分割任务间的共性和差异，缺少单独对小样本图像语义分割任务的全面介绍.

虽然近年来已有关于小样本语义分割方法的研究综述，但对于该领域的全面综述仍比较少. 此外，随着该领域的飞速发展，又涌现出了大量算法，需要重新对其进行分类. 与之前的综述相比，本文从模型优化的角度出发，根据推理过程中是否存在梯度回传微调模型对当前小样本语义分割方法进行分类，将当前算法进行梳理，并根据其设计思想以及解决小样本分割任务挑战的归属归类到不同的分类中，该分类能够全面地覆盖当前小样本语义分割算法，简单直观. 除提出新的分类方案外，本文深入地对已有方法和存在挑战进行分析，并总结其发展趋势，这将帮助研究者更好地了解其研究现状，为未来的研究提供更广阔的视野，促进该领域的进一步发展和创新.

本文第 2 节介绍小样本语义分割任务的提出背景、问题设置和存在的挑战；第 3 节介绍小样本语义分割任务中常用数据集和性能评价指标；第 4 节详细介绍小样本语义分割方法，并对其进行归纳分类；第 5 节总结当前小样本语义分割算法在三个数据集上的性能比较；第 6 节介绍小样本语义分割任务的一些延伸任务；第 7 节给出小样本语义分割任务的未来发展趋势；最后第 8 节对本文内容进行总结.

2　小样本语义分割技术

2.1　小样本语义分割的引入

图像语义分割是一项具有挑战性的任务. 它对图像中的每个像素进行分类，实现图像理解、分析和应用. 语义分割在计算机视觉领域中扮演着重要的角色，是其他一些任务的基础，如图像编辑[11-12]、图像描述[13-14]和视觉问答任务[15-16]等.

图像语义分割技术在实际生活中有广泛的应用，主要的应用领域包括但不限于：(1) 遥感领域[17-18]. 遥感图像分割可以用于监测土地使用、灾害检测、城市规划等. 例如通过分割卫星或无人机图像，识别出不同类型的目标，如城市、森林和河流等，进而对土地资源进行管理和监测；(2) 自动驾驶[19-20]. 语义分割可以帮助汽车感知和理解周围环境. 通过精确地识别和分割道路、行人、其他车辆和交通标志等物体，可以更准确地规划行驶路径和做出决策；(3) 医疗领域[21-22]. 语义分割在医学影像诊断中起着重要的作用，可以进行疾病诊断和治疗计划. 例如，在肿瘤检测中，精确的语义分割可以帮助确定肿瘤的位置和大小，从而进行更准确的治疗；(4) 农业领域[23-24]. 语义分割在农业领域中有助于提高农业生产效率，减少人工检测和降低生产成本. 例如，通过分割农田照片，可以识别出农作物和杂草，从而判断是否需要除草或施肥.

尽管图像语义分割任务获得了快速发展，然而在实际应用中仍存在较大的挑战，主要包括两点：(1)数据依赖. 对于图像语义分割任务而言，大规模像素级别的标注数据是必需的，但是获取和标注这些数据是一项耗时耗力的工作. 据统计，单张 1280×720 像素的图像标注时间约 1.5 h[25]，对于大规模的数据集，标注时间会更长. 除此之外，在一些特殊应用领域，如医疗图像，某些疾病数据收集困难，并且需要专业人员标注；(2)类别泛化. 传统的图像语义分割模型在训练过程中在已见类分割中表现得好，当分割新类时性能急剧下降，无法实现未见类别的有效分割.

与之相比，人类基于已积累的知识，只要少量标注样本就能够快速识别新类. 为了应对传统的图像语义分割存在的挑战，Shaban 等人[4]受到小样本学习的启发，第一次提出小样本图像语义分割任务(简称为小样本分割). 小样本图像语义分割是指在标注数据有限的情况下对图像执行新类的语义分割任务. 它利用元学习的思想，在已有的知识基础上，利用少量的标记数据实现新类泛化，从而减少对大规模标注数据的需求. 当前小样本分割任务中使用的训练数据中每类仍包含大量标注训练样本，小样本设置体现在推理阶段为每个类别仅提供少量密集标注图像，模型利用训练阶段学习到的知识实现图像中新对象类的语义分割. Shaban 等人提出一个两分支的分割网络(OSLSM)，第一个分支利用密集标注的图像学习分类器参数，第二个分支利用权重哈希算法将其转换为待分割图像的分类器参数进行分割. 小样本分割任务吸引了很多学者的关注，OSLSM 为后续小样本分割任务的发展奠定了基础，其提出的两分支结构也受到了很多方法的青睐.

小样本分割任务是传统图像语义分割任务和小样本学习任务的结合，因此语义分割和小样本学习为小样本分割任务的发展奠定了基础，提供了一些基本思想、模型架构和学习技巧，为小样本图像语义分割任务的研究提供了借鉴和扩展的空间. 在图像语义分割任务中，全卷积网络(FCN)[26]是具有开创性意义的模型架构，它将全连接层替换为卷积层，可以接受任意大小输入，同时处理速度得到了很大的提升. 全卷积网络是专门针对像素级的分类任务提出的网络结构，能够很好保留空间信息，同时减少模型参数，已经成为小样本语义分割任务常用的架构之一. 除此之外，小样本分割还继承了语义分割中的数据集，并且为了使数据集设置满足新类分割的要求，重新对其类别进行划分，为小样本分割任务提供了实验基础，促进了算法的发展和比较. 小样本学习中的一些思想和技巧也被应用于小样本分割任务中，例如元学习的训练方式，模型通过少量标注样本学习如何快速适应新任务从而提高模型的泛化能力. 此外，有些方法借鉴小样本学习任务中原型学习[27]的思想，通过标注图像计算得到原型向量，从而获得类别信息来指导图像分割.

自小样本分割任务提出以来，其主要经历了以下几个发展阶段，具体如图 2 所示. 第 1 个阶段(以 OSLSM[4]为代表)：在这个阶段许多方法将小样本分割任务看作一个像素分类问题，通过标注图像训练优化分类器来实现分割；第 2 个阶段(以 PLNet[28]为代表)：在这个阶段主要利用原型学习思想，利用标注图像(支持集)得到单个原型来表示具有代表性的类别信息，通过选择不同的度量方式来计算样本和原型间的相似度来进行新类分割；第 3 个阶段(以 PGNet[29]和 PMMs[30]为代表)：在这个阶段，一些方法针对单个原型无法捕捉到空间结构和细节信息的问题，引入超像素聚类或 EM 算法等方式生成多个原型来建模支持(有标注图像)和查询(待分割图像)图像间

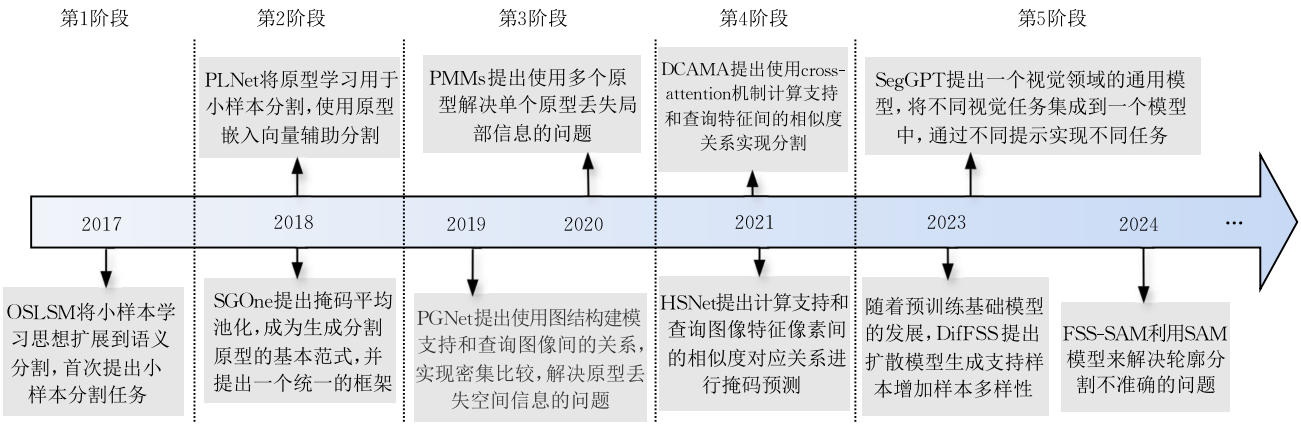


图 2 小样本分割任务发展时间线

的相似关系；第 4 个阶段（以 HSNet[31] 和 DCAMA[32] 为代表）：在这个阶段，随着计算资源和算力的提高，一些方法开始直接计算支持和查询图像特征像素间的相似度关系，获得高维关系矩阵实现密集比较，能够很好地保留图像细节和结构信息；第 5 个阶段（以 SegGPT[33] 和 DifFSS[34] 为代表）：随着通用基础模型和预训练基础模型的发展，人们开始探索视觉通用基础模型的应用，在该阶段，一些方法旨在设计通用的分割模型或借助预训练基础模型提升分割性能，小样本分割任务进入了一个新的发展阶段.

2.2　小样本分割任务设定

具体而言，小样本分割任务旨在利用训练集 D_{train} 训练一个模型 $\mathcal{M}$，推理时测试集 D_{test} 在仅提供少量（一个或几个）标记样本的情况下，实现新对象类的语义分割. 训练类又被称为基类（Base classes）、测试类称为新类（Novel classes），训练类 C_{train} 和测试类 C_{test} 是完全不同的，即 $C_{\text{train}} \cap C_{\text{test}} = \varnothing$. 模型训练通常采用元学习的训练方式，其目的是与推理时的设置保持一致，从而提高模型的泛化能力.

在元学习过程中，每个任务（episode）都由一个支持集和查询集构成 (S, Q)，其中有密集标注的集合称为支持集（support set），可以提供指导新类分割的知识，没有标注的集合称为查询集（query set），是待分割的图像集. 支持集 $S=\{\boldsymbol{I}_j^S, \boldsymbol{M}_j^S\}_{j=1}^{NK}$ 包含 N 个待分割类的标注样本，每个类提供 K 个支持图像 $\boldsymbol{I}^S$ 和相应掩码标签 $\boldsymbol{M}^S$，即 N-way K-shot 分割任务，其中 N 和 K 是指测试时的设置，当前小样本分割任务在训练和测试时一般会保持相同设置. 在查询集 $Q=\{\boldsymbol{I}^Q, \boldsymbol{M}^Q\}$ 中，$\boldsymbol{I}^Q$ 是待分割的查询图像，模型会将 $\boldsymbol{I}^Q$ 中与支持集相同的目标类全部识别分割出来；对于 $\boldsymbol{M}^Q$，它在训练时作为监督信号优化模型，而在推理时被用来评价模型性能. 其中 $\boldsymbol{I}^S$ 和 $\boldsymbol{I}^Q \in \mathbb{R}^{H\times W\times 3}$ 是来自同一个类别的 RGB 图像；$\boldsymbol{M}^S$ 和 $\boldsymbol{M}^Q \in \mathbb{R}^{H\times W}$ 是二分类掩码. 因此小样本分割任务可被约束为

$$\hat{\boldsymbol{M}}^Q = \mathcal{M}(\boldsymbol{I}^S, \boldsymbol{M}^S, \boldsymbol{I}^Q) \tag{1}$$

在训练时从 D_{train} 采样得到 $\{S_j, Q_j\}_{j=1}^{N_{\text{train}}}$ 优化模型，测试时从 D_{test} 中采样得到 $\{S_j, Q_j\}_{j=1}^{N_{\text{test}}}$ 测试模型的泛化性能，其中 N_{train}、N_{test} 是从训练集和测试集采样得到的任务数量.

接下来，我们对小样本分割算法进行归纳，可以得到一个通用框架. 该框架主要有三部分：特征提取器、编码器和解码器，如图 3 所示. (1) 特征提取器. 特征提取器用来从支持和查询图像中提取特征，获得富含结构和语义信息的特征表示. 常用的特征提取器包括预训练的主干网络，如 ResNet[35]、VGGNet[36]、Vision Transformer[37] 等. 在训练过程中，特征提取器参数可以被冻结，不参与优化；(2) 编码器. 编码器用来提供类别信息并实现特征交互. 具体而言，首先利用支持掩码过滤出待分割目标的特征，从而得到类别相关的信息指导分割，然后编码器对提取到的特征和类别信息进行处理，以实现支持和查询图像间的类别信息交互，便于后续分割. 有些方法在交互前会进行特征增强提高分割性能；(3) 解码器. 解码器负责对交互后的特征图进行像素分类，使用上采样等操作逐渐恢复特征图的分辨率，从而得到预测的分割掩码.

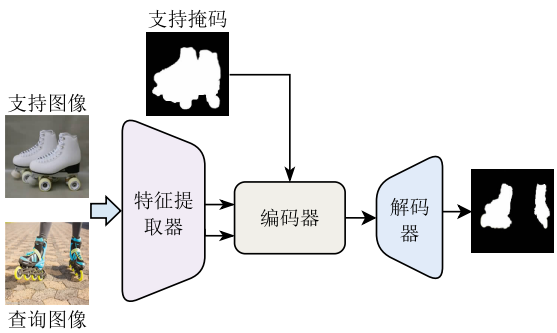


图 3　小样本分割任务框架

为了避免过拟合，除 OSLSM[4]、Co-FCN[38]、PANet[39] 和 MetaSegNet[40] 等方法同时对整个模型架构进行优化外，其他方法通常会使用预训练好的主干网络作为特征提取器，并在训练过程中冻结其参数，从而减少模型对训练数据的依赖，提高模型的泛化能力. 训练时通常利用交叉熵损失优化模型，并采用 episode 训练方式，从支持图像中学习到普遍有效的特征和分割规律，更好地适应小样本分割任务.

2.3　小样本分割任务面临的挑战

与传统语义分割任务不同，小样本分割需要在少量（一个或几个）标注样本的条件下实现新对象类的分割，是一项更具挑战性的任务. 小样本分割任务的最大挑战就是数据稀缺，如何充分利用少量标注样本获得丰富的类别信息指导分割从而实现新类泛化，这是小样本分割的关键.

小样本分割任务中目标类别图像和标注数据稀缺会导致以下问题：

(1) 模型偏向训练类别. 当仅使用少量目标类别的样本优化模型时，极易出现过拟合现象. 在传统的语义分割设置下，使用这些新类样本与基类样本一起训练模型，由于新类别的样本数量与训练过程中出现过的类别样本数量差距很大，模型可能无法

学习到新类别的知识，这会导致模型在基类样本上分割性能表现出色，但在新类分割时性能会急剧下降. 如果只使用新类样本微调模型，由于标注样本有限，不能很好地学习到该类别的泛化知识，导致模型过分拟合训练样本，不具备泛化能力.

(2) 类内差距大. 由于存在许多不同的场景和目标形态，同类数据的外观变化较大，因此当仅提供少量样本时不能够很好地覆盖该类样本所有可能性，这会导致支持特征和查询特征不匹配，很难实现精确的度量，进而影响分割性能. 例如，在图 4(a)、(b)、(c)、(g)，当图像中存在遮挡、干扰物体、多个实例或目标较小时，模型需要能够准确地定位目标位置并排除干扰信息才能获得良好的分割性能. 此外，由于视角、光照等方面的差异，如图 4(d)、(e)所示，同一物体的图像也存在较大差异. 从图中展示的这些情况可以看出在小样本设置下，图像间存在较大的类内差异，这使得在提供少量标记样本的情况下，知识迁移变得十分困难. 此外，还存在一些场景，如图 4(h)所示，当目标包含细节信息或位于复杂背景中时，模型需要具备能够精确识别目标边缘并将目标与背景区分开的能力，然而这在仅提供少量数据的前提下是很难实现的.

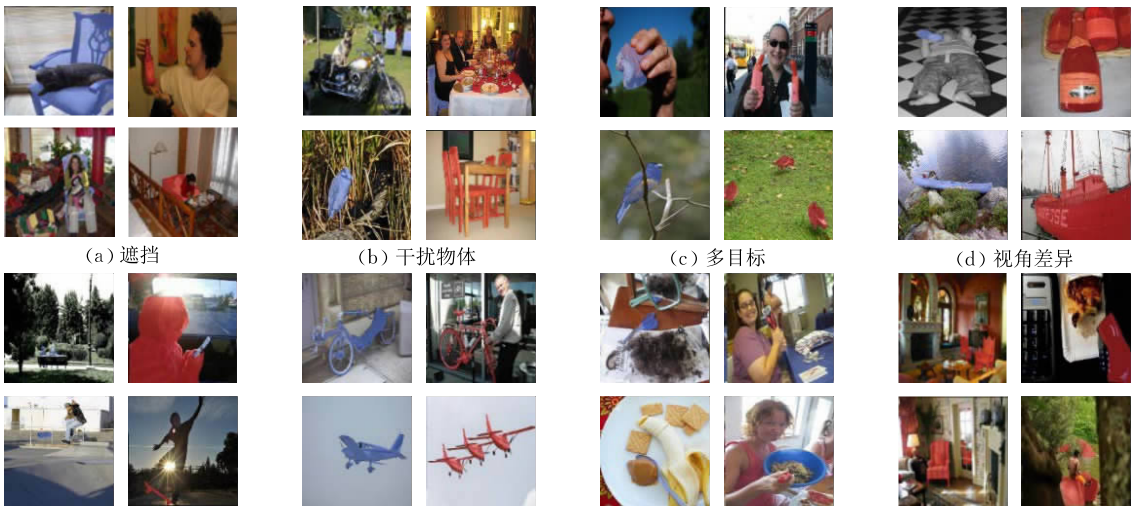


图 4 小样本分割任务的挑战

此外，对于基于优化的小样本分割算法而言，由于样本特征空间稀疏，因此很难找到合适的分类界限将目标类与其他类别完全区分开，特别是对于处于边界线上的目标类别. 总之，无论是基于优化的方法还是基于度量学习的方法，仅依靠少量样本很难学习到可泛化的类别知识，因此导致小样本分割性能受到影响.

在小样本分割任务中，除数据稀缺问题外，训练设置也存在一些问题. 首先，许多算法都使用在 ImageNet[41] 分类任务上得到的预训练参数作为特征提取器权重，并在训练过程中冻结这部分参数. 然而这些权重是在分类任务上训练得到的. 分类任务只关注图像中最具区分性的部分，只要能够找到实现类别辨认的部位即可，而分割任务则是希望关注到图像中的每个像素，对每个位置的像素进行分类. 由于分类和分割任务间的差异，因此当在分割任务中直接冻结住这部分参数时，可能会导致特征提取器得到的特征不能够很好地捕捉上下文信息，影响图像像素级别的理解，进而影响模型分割性能. 其次，当前小样本分割任务均采用二值分割，即分割目标类只有前景和背景两类，图像中待分割类别的目标属于前景区域，非目标类别区域都视为背景. 在训练过程中这种设置会导致模型无法从非目标类中的背景下区域学习到更多有意义的特征表示，模型无法充分利用图像中的多样性信息，在推理时可能会将待分割目标类当做背景从而影响分割性能.

3 常用数据集和性能评价指标

3.1 数据集介绍

目前小样本分割任务中常用基准数据集有三个，分别是 FSS-1000[42]、PSACAL-5i[4] 和 COCO-20i[43]，其图像示例如图 5 所示. 接下来，我们将分别介绍这三个数据集，并对其进行比较和总结.

(a) FSS-1000数据集示例

(b) PASCAL-5i数据集示例

(c) COCO-20i数据集示例

图 5　小样本分割数据集示例

3.1.1　FSS-1000 数据集

FSS-1000[42]数据集是针对小样本分割任务专门提出的一个数据集，每张图像中只包含一个分割类别，共有 1000 个分割目标类别，其中每个类包含 10 张具有二值分类掩码的图像，总计 10 000 张图像. FSS-1000 数据集强调的是类的数量而不是图像的数量，因此每个类别仅提供了少量标注样本. 在这 1000 个类中，其中 584 个类与 ILSVRC[44]数据集中类别有重叠，其余 486 个类是现有数据集中从未出现过的新类，例如微小的日常物品、商品、卡通人物、徽标等. 该数据集具有一些特性，首先是层次性，所有类别可以分为 3 个层级分别为底层、中层和顶层类别，顶层共有 12 个超类，底层共有 1000 个目标类；其次可扩展性，只需要 10 个带有二值掩码标注的样本就能实现新类扩展. 此外，FSS-1000 数据集也支持实例分割任务.

3.1.2　PASCAL-5i 数据集

PSACAL-5i[4]包含 20 个分割类，由 PASCAL VOC[45]和来自 SBD[46]数据集中的额外标记数据构成. 在小样本分割任务中，该数据集被划分为 4 个集合(fold)，每个集合中有 5 个类别，具体目标类别及集合划分如表 1 所示.

表 1　PASCAL-5i 类别及集合划分

集合	类别
0	aeroplane, bicycle, bird, boat, bottle
1	bus, car, cat, chair, cow
2	dining table, dog, horse, motorbike, person
3	potted plant, sheep, sofa, train, TV/monitor

3.1.3　COCO-20i 数据集

COCO-20i[43]是从更具有挑战性的 MS COCO (Microsoft Common Objects in Context)[47]数据集得到的，这些图像主要从复杂的日常场景中截取. 该数据集共包含 80 个分割类，有超过 50 万个目标标注，而且每个类别中包含的图像数目比较多. 在小样本分割任务中，它被划分为 4 个集合(fold)，每个集合包含 20 个类别，具体目标类别及集合划分如表 2 所示.

表 2　COCO-20i 类别及集合划分

集合	类别
0	person, airplane, boat, park meter, dog, elephant, backpack, suitcase, sports ball, skateboard, wine glass, spoon, sandwich, hot dog, chair, dining table, mouse, microwave, fridge, scissors
1	bicycle, bus, traffic light, bench, horse, bear, umbrella, frisbee, kite, surfboard, cup, bowl, orange, pizza, couch, toilet, remote, oven, book, teddy
2	car, train, fire hydrant, bird, sheep, zebra, handbag, skis, baseball bat, tennis racket, fork, banana, broccoli, donut, potted plant, TV, keyboard, toaster, clock, hairdrier
3	motorcycle, truck, stop, cat, cow, giraffe, tie, snowboard, baseball glove, bottle, knife, apple, carrot, cake, bed, laptop, cellphone, sink, vase, toothbrush

3.1.4　小　结

小样本分割将图像分割看作前景和背景的二值分类任务. 不同于 FSS-1000 数据集中的单类目标图像，PASCAL-5i 和 COCO-20i 数据集中的图像同时包含多个分割类目标. 因此，在使用 PASCAL-5i 和 COCO-20i 数据集时会对这两个数据集的原始掩码进行处理，将待分割目标类别的掩码像素值设置为 1，其余不属于待分割类别的区域均看作背景，掩码像素值设置为 0. 在推理时，PASCAL-5i 和 COCO-20i 均采用交叉验证的方式评价模型性能，即每次使用三个集合训练，剩下的一个集合进行测试，如 fold 0 的结果是指模型使用 fold 1～3 中的训练样本训练模型，在 fold 0 的测试图像中得到的结果，最后分别在四个集合上进行测试后得到模型在该数据集上的平均性能.

这三个数据集各有特点，FSS-1000 数据集中每张图像中仅包含 1 个待分割目标类别，图像背景相对纯净，分割简单，但分割类别数最多，在 1000 个类中共使用 520 个类训练，240 个类验证，240 个类测试；PASCAL-5i 数据集中平均每张图像包含来自 2 个不同类别的 3 个实例，分割难度升级，分割类别数最少，总共 20 个类中使用 15 个类训练，5 个类测试；COCO-20i 数据集平均每张图像包含来自 3.5 个类的 7.7 个实例，图像中实例数和类别数增多，背景中包含较多干扰目标，分割难度较大，分割类别数适中，总共 80 个类中使用 60 个类训练，20 个类测试.

3.2 性能评价指标

语义分割是对图像中的每个像素按照语义进行分类,得到的预测像素可以分为四类:真实值是前景并预测为前景的像素 TP(True Positive)、真实值为前景但预测为背景的像素 FN(False Negative)、真实值为背景并预测为背景的像素 TN(True Negative)和真实值为背景但预测为前景的像素 FP(False Positive).

基于上述概念,我们介绍小样本分割任务中常用的性能评价指标,平均交并比($mIoU$)和前景-背景交并比($FBIoU$).关于交并比(IoU),它是模型对某一个类别的预测值和真实值的交集和并集之比,具体如图 6 所示.在小样本分割任务中,我们所提到的交并比主要是计算前景类别的交并比,也就是真实值是前景并预测为前景的像素数与真实值是前景并预测为前景的像素数、真实值为前景但预测为背景数以及真实值为背景但是预测为前景的像素数之和的比值.其计算公式如下:

$$IoU=\frac{Interaction}{Union}=\frac{TP}{TP+FP+FN} \tag{2}$$

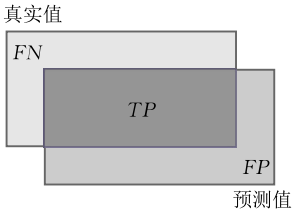


图 6 交并比说明图

均交并比($mIoU$)是模型对每一个前景类别的交并比结果求和再进行平均后得到的结果.均交并比数值越接近 1 说明模型分割性能越好,即预测掩码越接近真实掩码.其计算公式如下:

$$mIoU=\frac{1}{n}\sum_{i=1}^{n}IoU_i \tag{3}$$

其中 n 是分割的目标类别数,IoU_i 是类别 i 的交并比结果.

前景-背景交并比($FBIoU$)是指计算前景交并比和背景交并比进行平均得到的结果.同样,前景-背景交并比数值越接近 1 说明模型前背景区分能力越强,预测掩码越接近真实掩码.其计算公式如下:

$$FBIoU=\frac{1}{2}(IoU_{\text{fg}}+IoU_{\text{bg}}) \tag{4}$$

其中 IoU_{fg} 和 IoU_{bg} 分别指的是前景的交并比结果和背景的交并比结果.

$mIoU$ 考虑了不同类别前景的分割结果,可以得到基于全局的评价,对类别不平衡具有很好的鲁棒性.而 $FBIoU$ 计算时只有前景和背景两个类别,并没有考虑不同目标类别间差异,而且背景部分像素占大多数,因此即使全部像素均分类为背景,也可以获得不错的 $FBIoU$.因此人们一般采用 $mIoU$ 作为小样本分割任务的性能评价指标来比较不同算法.

4 小样本分割方法介绍

本文对当前小样本图像语义分割算法进行总结分类,从模型优化角度出发,根据推理过程中是否存在梯度回传可以将当前方法分为两类:基于优化和基于度量学习的小样本分割方法.

基于优化的方法在推理过程中会利用支持集信息对模型进行微调,即通过计算支持图像预测掩码和真实支持掩码间的损失,将得到的梯度回传到模型实现参数更新,从而使模型学习到当前类别的知识指导分割.具体而言,首先将支持图像 $\boldsymbol{I}^S$ 输入到使用基类数据训练好的模型中得到预测支持掩码 $\hat{\boldsymbol{M}}^S$,即

$$\hat{\boldsymbol{M}}^S=\mathcal{M}(\boldsymbol{I}^S) \tag{5}$$

然后通过损失函数计算预测支持掩码与真实支持掩码间的损失:

$$L_{\text{supp}}=L(\hat{\boldsymbol{M}}^S,\boldsymbol{M}^S) \tag{6}$$

其中 $L(\hat{\boldsymbol{M}}^S,\boldsymbol{M}^S)$ 是指具体论文中使用的损失函数.当前大部分论文中均使用二值交叉熵损失作为模型优化损失,即

$$L(\hat{\boldsymbol{M}}^S,\boldsymbol{M}^S)=-\boldsymbol{M}^S\log\hat{\boldsymbol{M}}^S-(\boldsymbol{1}-\boldsymbol{M}^S)\log(\boldsymbol{1}-\hat{\boldsymbol{M}}^S) \tag{7}$$

最后计算梯度 $\frac{\partial L_{\text{supp}}}{\partial\boldsymbol{\omega}}$,并进行反向传播来更新模型参数 $\boldsymbol{\omega}$,得到优化后的模型来分割查询图像 $\boldsymbol{I}^Q$,即

$$\boldsymbol{M}^Q=\mathcal{M}(\boldsymbol{I}^Q) \tag{8}$$

基于度量学习的方法则利用支持掩码过滤掉与分割类别无关的信息,得到图像前景区域来获得与待分割类别有关的知识.通过选择不同的度量方式 $\mathcal{F}_{\text{metric}}$(如欧式距离、余弦相似度等)对由支持集获得的类别知识与查询图像特征进行相似度度量,实现类别信息和查询图像信息的交互,从而获得最后的预测分割掩码.该类方法将在训练集上训练好的模型直接应用于推理阶段,不再更新模型参数.本质上,基于度量学习的方法旨在学习一个与类无关的模型,该模型只要在提供类别信息的情况下就能直接实现新类图像的语义分割,无须重新优化模型.

我们对当前小样本分割任务中提出的各种算法进行归纳总结,并对其进行分类,具体分类如图 7 所示.接下来,我们将对这些算法的具体分类和该领域的发展现状进行详细介绍.

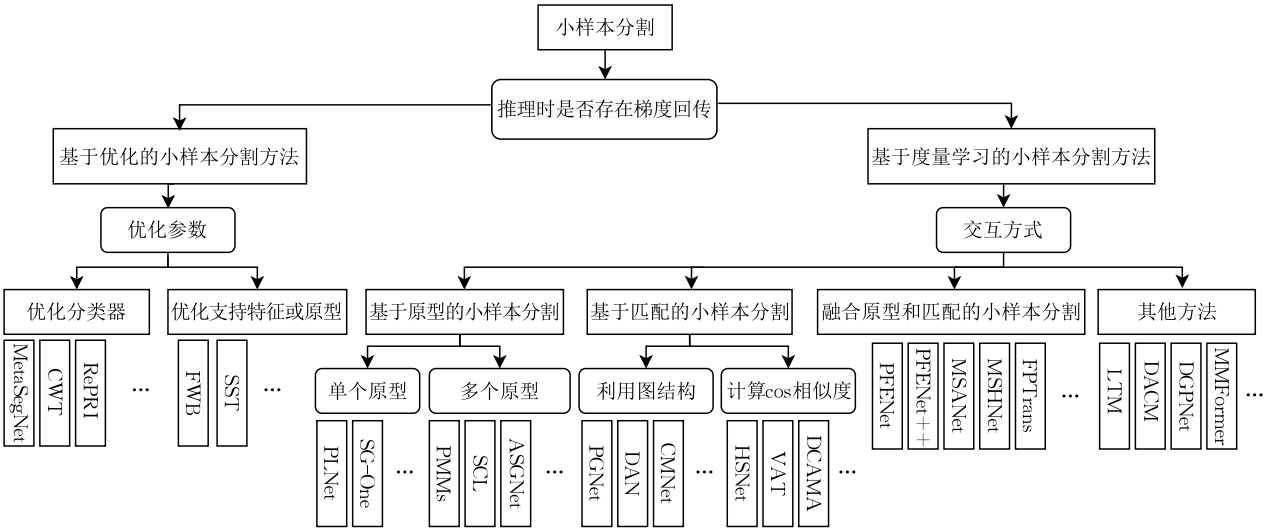
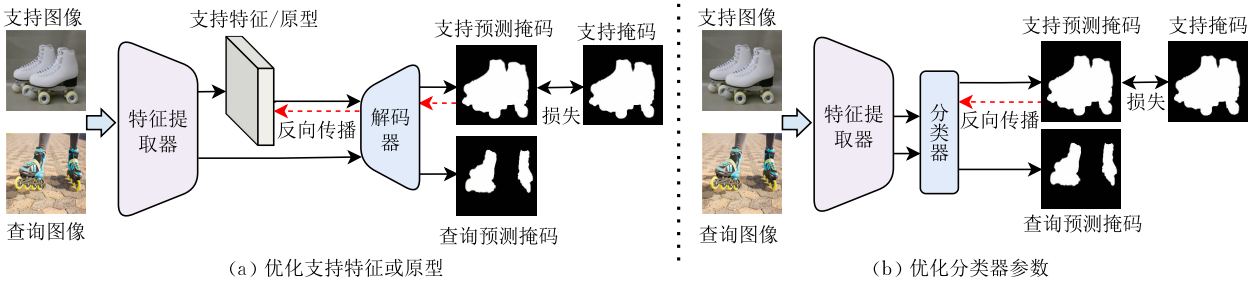


图 7　小样本分割方法分类介绍

4.1　基于优化的小样本分割方法

基于优化的小样本分割方法在推理时旨在利用少量标注样本微调部分模型参数使其适应于新类，其关键思想是在微调时如何避免模型过拟合. 我们对基于优化的小样本分割方法进行概括，可以得到进一步的分类结果. 根据其优化参数，我们将基于优化的方法可以分为两种：一种是使用支持集预测损失优化支持特征或原型，在特征层面进行数据增强从而获得更多与类别相关的知识，以 FWB[43] 和 SST[48] 为代表，具体框架如图 8(a)所示. 另一种则是利用支持集预测损失优化分类器，该类方法将分割视为像素分类任务，通过支持集样本来微调分类器参数，使其适应于新类别，以MetaSegNet[40]、RePRI[49] 和CWT[50] 为代表，具体框架如图 8(b)所示. 其中，属于同一种分类的不同算法间的区别是通过不同优化方式或模块设计实现的.



(a) 优化支持特征或原型　　(b) 优化分类器参数

图 8　基于优化的小样本分割

(1) 基于支持集预测损失优化支持特征或原型的方法.

Nguyen 等人[43] 通过实验发现，卷积神经网络倾向于学习不同类别的非判别性特征，即提取的特征存在类间距离小、不能很好区分不同类别的问题. 对此，FWB[43] 方法引入正则化项 r 来最大化前景与背景区域激活值的差距，从而增加类别间的区分度，即

$$\boldsymbol{\phi}_s=\sum_{i=1}^{wh}\boldsymbol{F}_{s,i}\left[\frac{\widetilde{\boldsymbol{m}}_{s,i}}{|\widetilde{\boldsymbol{m}}_{s,i}|}-\frac{1-\widetilde{\boldsymbol{m}}_{s,i}}{wh-|\widetilde{\boldsymbol{m}}_{s,i}|}\right] \tag{9}$$

$$\max_{\boldsymbol{r}}\boldsymbol{\phi}_s^{\mathrm{T}}\boldsymbol{r},\ \text{s.t.}\ \|\boldsymbol{r}\|_2=1 \tag{10}$$

其中 $\boldsymbol{F}_{s,i}$ 指的是支持图像特征位置 i 处激活值，$\widetilde{\boldsymbol{m}}_{s,i}$ 是支持图像掩码位置 i 处的值. 针对标注样本稀缺导致模型过拟合的问题，该方法在推理时提出一种 boosting 机制，即通过交叉熵损失计算支持集预测掩码和真实掩码间的差距 $L(\hat{\boldsymbol{M}}_s^n,\boldsymbol{M}_s)$，迭代地更新支持原型，最后得到 N 个支持原型来提高模型鲁棒性. 其中第 n 次迭代生成的支持原型为 $\boldsymbol{f}_s^n$：

$$\boldsymbol{f}_s^n=\boldsymbol{f}_s^{n-1}-\nu\frac{\partial L(\hat{\boldsymbol{M}}_s^{n-1},\boldsymbol{M}^{\mathrm{S}})}{\partial \boldsymbol{f}_s^{n-1}} \tag{11}$$

其中 n 是指第 n 次迭代，$\hat{\boldsymbol{M}}_s^{n-1}$ 是利用支持原型 $\boldsymbol{f}_s^{n-1}$ 预测得到的查询掩码，ν 是学习率. 通过生成多个支持掩码增加数据的多样性，从而减轻了过拟合问题. 此外，该方法第一次在 COCO 数据集上验证小样本分割算法，提出 COCO-20i 数据集作为一个新的数

据基准用于后续研究比较. Zhu 等人[48]则提出一个自适应调整框架(SST)来解决样本稀的缺问题,针对每个 episode 中的支持集优化类别信息指导分割. 该方法首先利用交叉熵损失计算支持图像的预测掩码和真实掩码间的损失 L_{supp},通过梯度提供特定类别的语义限制,获得修正后的特征:

$$\boldsymbol{P}_s'=\boldsymbol{P}_s-\frac{\partial L_{\text{supp}}}{\partial \boldsymbol{P}_s} \tag{12}$$

这样一来,模型可以根据反馈信号自适应地调整支持特征,从而提高分割性能.

(2)基于支持集预测损失优化分类器的方法.

这些方法将图像分割看作像素分类任务,继承了小样本学习任务的基本思想,根据支持集信息微调模型分类器参数使其获得新类知识. Tian 等人[40]提出 MetaSegNet,基于元学习的思想训练线性分类器实现像素分类,推理时利用支持集损失仅微调分类器权重. 该方法不使用预先训练的主干网络作为特征提取器,而是使用一个小的网络结构作为特征提取器进行端到端的训练,来解决当前方法中普遍存在的预训练模型任务和分割任务分布不匹配的问题. 此外,MetaSegNet 是针对 K-way N-shot 设置下的小样本分割任务而提出的模型,在不需要任何先验知识的情况下获得了不错的性能. Lu 等人[50]提出分类器加权结构(CWT). 针对预训练模型分布不匹配的问题,采用两阶段的训练方式. 第 1 阶段利用传统语义分割任务训练模型,得到适用于分割任务的特征提取器权重. 第 2 阶段冻结第 1 阶段得到的特征提取器,采用元学习的方式训练执行小样本分割任务的模块. 针对查询和支持图像间类内差异大的问题,首先利用支持集得到的分类器权重 $\boldsymbol{\omega}$ 和查询特征 $\boldsymbol{F}$,通过 Transformer 结构得到查询图像的预测权重:

$$\boldsymbol{\omega}^*=\boldsymbol{\omega}+\phi\left(\text{Softmax}\left(\frac{\boldsymbol{\omega}\boldsymbol{W}_q(\boldsymbol{F}\boldsymbol{W}_k)^{\mathrm{T}}}{\sqrt{d_a}}\right)(\boldsymbol{F}\boldsymbol{W}_v)\right) \tag{13}$$

其中 ϕ 是一个线性函数,$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$是可学习参数,d_a是正则化项. 该方法不是直接将支持集得到分类器参数用来分割查询图像,而是通过元学习训练的 Transformer 结构动态调整分类器参数使其可以适应测试样本. Boudiaf 等人[49]放弃对模型结构进行复杂设计,而是提出区域比例正则化推理模型(RePRI),利用查询图像中未标记像素的统计信息和直推式推理进行分割,推理时通过优化一个简单的线性分类器实现新类分割. 该方法放弃了元学习的思想,使用传统语义分割任务的训练方式,在训练过程中重新考虑标准的交叉熵对模型进行监督. 在推理时利用特征提取器将支持和查询图像投影到低秩的特征空间 ϕ,同时优化三个互补的损失函数进行像素类别预测. 第一个损失函数是支持图像的交叉熵损失:

$$L_{\text{CE}}=-\frac{1}{K|\phi|}\sum_{k=1}^{K}\sum_{j\in\phi}\tilde{y}_k(j)^{\mathrm{T}}\log(p_k(j)) \tag{14}$$

其中 $\tilde{y}_k(j)$、$p_k(j)$分别是类 k 下采样位置 j 处的真实标签值和预测标签值. 第二个损失函数是查询图像像素的后验信息熵,将线性分类器的决策边界推向了查询图像特征空间的低密度区域:

$$\mathcal{H}=-\frac{1}{\phi}\sum_{j\in\phi}p_Q(j)^{\mathrm{T}}\log(p_Q(j)) \tag{15}$$

其中 $p_Q(j)$是查询图像像素的预测值. 第三个损失函数是基于预测前景像素在查询图像中的比例的全局 Kullback-Leibler(KL)散度正则化,有助于避免由前两个损失最小化引起的退化:

$$\mathcal{D}_{\text{KL}}=\hat{p}_Q^{\mathrm{T}}\log\left(\frac{\hat{p}_Q}{\pi}\right) \tag{16}$$

$$\hat{p}_Q=\frac{1}{|\phi|}\sum_{j\in\phi}p_Q(j) \tag{17}$$

其中 π 是模型预测的前景/背景比例匹配参数,$\pi\in[0,1]^2$. 该方法在推理时优化分类器并预测前景和背景占比参数,可以适配于任何特征提取器.

基于优化的小样本分割方法在推理时通过优化模型能够学习到新类知识,从而提高分割性能. 但由于推理时只能提供一个或几个有标记的支持样本指导分割,因此这类方法极易出现过拟合现象,需要设计复杂精巧的结构来避免该现象发生. 同时该方法推理时要调整模型参数,不能够进行实时推理. 其基本思想主要是借鉴小样本学习的思想和研究方案,由于在分割任务中存在一定的问题和性能提升困难,该类方法的研究主要集中在小样本分割任务的发展初期.

4.2 基于度量学习的小样本分割方法

基于度量学习的方法是解决小样本分割任务的一种有效途径,在该领域得到了广泛关注和研究. 它们将查询和支持图像投影到一个低维特征空间,并使用参数或非参数学习的方式来度量支持和查询特征间的相似性,从而得到查询图像的预测掩码. 基于度量学习的方法在分割过程中会使用一个度量函数 $\mathcal{F}_{\text{metric}}(f(\boldsymbol{I}^S),f(\boldsymbol{I}^Q),\boldsymbol{M}^S)$实现支持和查询特征间的交互,其中 f 表示特征处理函数,推理时无需优化调整模型参数使得该过程更加高效. 属于该类方法中的不同算法是通过设计不同的特征处理方式 f 或度量函数$\mathcal{F}_{\text{metric}}$实现的.

同样地，我们对基于度量学习的小样本分割方法进行概括，可以得到进一步的分类结果，根据支持和查询图像间的交互方式可以将其分为四种：基于原型的方法、基于匹配的方法、融合原型和匹配的方法以及其他方法. 第一种是基于原型的方法，即首先利用特征处理操作 f 处理输入支持图像，然后通过池化操作获得一个或多个原型向量作为类别指导信息：

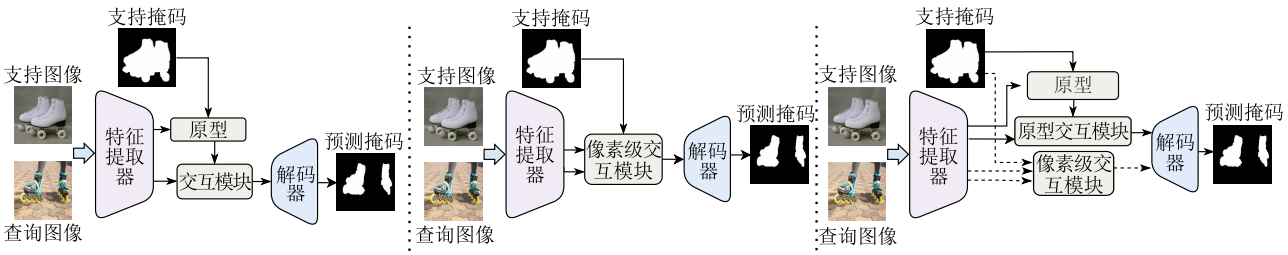
$$\boldsymbol{P}_i=\text{Pooling}(f(\boldsymbol{I}^S),\boldsymbol{M}^S) \tag{18}$$

通过选择不同的度量方法 $\mathcal{F}_{\text{metric}}$（如欧式距离、余弦相似度等）将查询特征与原型向进行交互，

$$\boldsymbol{M}^Q=\mathcal{F}_{\text{metric}}(\boldsymbol{P}_i,f(\boldsymbol{I}^Q)) \tag{19}$$

交互后的特征经过处理后得到最终的查询预测掩码，以 PLNet[28] 和 PMMs[30] 为代表，其基本框架如图 9(a). 第二种是基于匹配的方法，即对经过特征处理操作 f 得到的特征，通过逐像素地计算支持和查询特征间的对应关系或相似度 $\boldsymbol{R}$ 实现支持和查询图像间的信息交互：

$$\boldsymbol{R}=\text{Pixel-Interaction}(f(\boldsymbol{I}^S),f(\boldsymbol{I}^Q),\boldsymbol{M}^S) \tag{20}$$



(a) 基于原型的小样本分割　(b) 基于匹配的小样本分割　(c) 融合原型和匹配的小样本分割

图 9　基于度量学习的小样本分割

然后选择合适的度量方法 $\mathcal{F}_{\text{metric}}$（如 4D 卷积等）对高维相似度矩阵进行处理，并用于指导查询图像的分割. 它们使用图结构或直接计算余弦相似度来建模像素间的对应关系实现信息交互，以 PGNet[29] 和 HSNet[31] 为代表，其基本框架如图 9(b). 第三种是融合原型和匹配的方法，它们同时利用原型和像素级交互指导新类分割，既使用原型向量 $\boldsymbol{P}_i$ 获得类别信息，又建模支持和查询特征像素间的相似度关系 $\boldsymbol{R}$ 来更好地利用支持图像信息. 它们通常包含两个分支，一个分支处理原型和查询特征间的交互，另一个分支计算支持和查询特征像素间相似度，最后融合两个分支信息生成查询图像的预测掩码，以 PFENet[51] 和 FPTrans[52] 为代表，其基本框架如图 9(c). 从使用一个确定性的向量原型到计算支持和查询特征像素间的相似度，再到融合原型和相似度计算，基于度量学习的小样本分割方法朝着更好更精确地实现支持和查询特征交互的方向发展. 除了前述方法外，还有一些方法从其他角度出发解决小样本分割问题. 例如，LTM[53] 利用线性代数知识提出了一种新的转换模块来直接生成查询掩码；DACM[54] 改变了支持和查询特征间的度量方式，使用可学习的协方差矩阵来代替传统的余弦相似度矩阵；DPGNet[55] 利用高斯过程知识，从高斯过程回归的角度解决小样本分割问题；MMFormer[56] 则将分割和匹配过程进行解耦，类似于实例分割的思想，首先得到所有目标的预测掩码，然后对预测掩码分类和融合实现语义分割. 下面将详细介绍这四类方法.

4.2.1　基于原型的小样本分割方法

Dong 等人[28] 受到小样本分类中原型学习的启发，将其应用于小样本分割任务. 这类方法通过支持集生成一个或多个具有代表性的特征向量作为类原型，这种原型学习的思想为小样本分割发展提供了新的思路，为该领域的发展奠定了坚实的基础.

基于原型的方法具有较好的泛化能力和抗噪声能力，计算方法相对简单方便且易于理解. 通过全局平均池化或掩码平均池化操作来获得类别信息，这种池化操作能够过滤掉特征中存在的一些噪声信息，提高类别特征的鲁棒性，从而提高分割的准确性. 当前许多研究都在原型方法的基础上进行改进，并取得了显著进展.

(1) 基于单个原型的小样本分割方法.

基于单个原型的方法利用支持集图像和掩码生成一个具有代表性的类原型向量用来指导分割. PLNet[28] 和 Co-FCN[38] 这两种方法是直接将支持图像和掩码相乘后得到的掩码图像输入到特征提取器，然后将提取到的特征进行全局平均池化来生成类原型指导分割. 在 PLNet 方法中通过直接计算查询特征和类原型间的余弦相似度获得查询预测掩码；Co-FCN 方法首次将全卷积网络应用于小样本分割任务，通过卷积操作实现图像分割. Co-FCN 采取了两分支结构，其中条件分支用来获得原型提供类别信息，分割分支则用来实现信息交互并进行语义分割. 其中条件分割除接收具有掩码信息的支持

集数据外,还可以接受正负标记点的稀疏注释支持集作为输入,即使只有一个正负点标记,也可以获得具有竞争力的结果.

Zhang等人[57]提出一个相似性引导网络(SG-One),首次提出掩码平均池化(Mask Average Pooling,MAP)操作,该操作成为分割任务中获取原型的范例.与前两个方法不同,SG-One在获取类别信息时是将支持特征 $\boldsymbol{F}$ 与支持掩码 $\boldsymbol{M}$ 相乘,然后进行掩码平均池化得到类原型,即

$$\boldsymbol{P}=\frac{\sum_{x=1,y=1}^{w,h}\boldsymbol{M}_{x,y}\times\boldsymbol{F}_{x,y}}{\sum_{x=1,y=1}^{w,h}\boldsymbol{M}_{x,y}} \tag{21}$$

其中 w,h 是特征图的宽和高,x,y 是特征图和掩码的位置坐标.掩码平均池化直接利用利用支持掩码消除特征中背景噪声的影响,使得原型更具代表性.自此之后,掩码平均池化成为小样本分割任务中获取类原型的常用方式.此外,该方法使用同一个网络提取支持和查询图像特征,并在特征层面融合掩码信息,确保网络输入的一致性.SG-One方法摒弃了并行的网络结构,建立了一个统一的框架,为小样本分割任务发展奠定了结构基础.

Zhang等人[58]提出一个与类无关的分割网络(CANet),并指出由卷积神经网络获得的高层特征(最后一层特征)与类别语义相关,而底层特征才是有可能与新类别共享的.因此,CANet提出使用底层特征进行分割,在后续方法得到了广泛的应用.受图像分类任务中度量学习的启发,CANet提出密集比较模块,通过距离函数评估支持原型与查询特征中每个像素点的相似性来指导分割,其中支持原型是通过全局池化掩码支持图像的特征得到的.此外,该方法还提出了一个迭代优化模块,对查询图像预测进行迭代优化预测结果.

以上提到的方法仅利用支持图像和掩码生成的原型来提供类别信息指导分割,忽略了支持和查询图像间的关系.为提高分割性能,一些方法提出挖掘查询图像中的类别知识来缓解类内差异大的问题.Wang等人[39]和Liu等人[59]均通过预测支持掩码来挖掘更一致的类别知识,Wang等人[39]提出的原型对齐分割网络(PANet)旨在使支持和查询特征生成更一致的类原型,而Liu等人[59]提出的交叉参考网络(CRNet)旨在获得更一致的支持和查询特征.PANet提出一种新的原型对齐正则化方法,先利用支持图像和支持掩码预测查询掩码,然后利用查询图像及其预测掩码分割支持图像,从而优化网络结构.具体而言,PANet先利用支持图像特征和掩码通过掩码平均池化生成支持原型,然后利用余弦相似度对原型与查询图像特征进行度量得到预测查询掩码.得到预测查询掩码后,利用预测掩码与查询图像特征获得查询原型,用来预测支持掩码.通过交替预测支持和查询掩码,使得模型学习一个一致的嵌入空间,从而获得更加一致的支持和查询原型.CRNet是同时预测支持和查询掩码,通过交叉参考机制对支持和查询特征进行加权,挖掘支持和查询图像中共同出现的特征来增强特征表示.

针对查询图像中包含多个干扰类别会导致分割性能下降的问题,Li等人[60]提出了一种自监督任务来关注查询图像中的非目标类信息.通过超像素分割为查询图像背景中出现的目标生成伪掩码,利用该伪掩码通过掩码平均池化获得支持原型来训练网络,其中输入网络的支持和查询图像是相同的.通过添加一个额外的自监督损失优化,使模型能够学习到一些新类知识,以学习更具区分性的特征空间.

除了上述提到的方法外,Lang等人[61]和Kayabaşı等人[62]提出利用模型学习到的基类知识辅助新类分割,可以消除训练模型中的基类偏好.Lang等人[61]提出的BAM方法采用两阶段的训练方式:第1阶段,采用传统的分割任务训练一个特征提取器和基类学习器,经过分割训练的特征提取器可以降低预训练分类任务和当前分割任务的分布差异,基类学习器能够帮助分割出新类图像中出现的基类目标消除部分干扰;第2阶段,冻结第1阶段训练好的网络,通过元学习的方式训练执行小样本分割任务的其他模块,并使用Frobenius范数来指导基类 $\boldsymbol{m}_b$ 和新类掩码 $\boldsymbol{m}_n$ 的融合:

$$\boldsymbol{m}_n^0=\mathcal{F}_{\text{ensemble}}(\mathcal{F}_\phi(\boldsymbol{m}_n^0),\boldsymbol{m}_b^f) \tag{22}$$

$$\boldsymbol{m}_n=\boldsymbol{m}_n^0\oplus\mathcal{F}_\phi(\boldsymbol{m}_n^1) \tag{23}$$

以消除特征提取器对已见类的偏好,其中上标'0'和'1'表示背景和前景掩码,$\boldsymbol{m}_b^f$ 是基类前景掩码,$\mathcal{F}_{\text{ensemble}}$ 和 $\mathcal{F}_\phi$ 分别是具有特定初始化参数的 1×1 卷积.Kayabaşı等人[62]提出的BAM++在BAM的基础上提出了一种多尺度融合机制,旨在消除空间不一致的问题.BAM++利用多尺度的支持和查询特征进行交互,通过融合多尺度的基类掩码和新类掩码生成更准确的分割掩码,进一步提升分割性能.

上述提到的方法均使用单个原型获得具有代表性的类别知识,利用掩码与原始图像或特征相乘过滤掉背景噪声,将类别信息凝练为一个向量有效地指导新类分割.使用池化操作得到的全局类原型能够过滤掉图像前景中出现的干扰信息,可以更好地应对数据中的噪声和异常值,增强模型的鲁棒性.

(2) 基于多个原型的小样本分割方法.

尽管单个原型可以过滤掉前景区域中存在的一些噪声信息,但它忽略了图像的局部信息,不能够捕捉图像的细节和结构信息,从而影响分割性能. 为了更好地捕捉具有局部结构的类别信息并提高分割性能,人们提出使用多个原型来建模图像的局部特征,将不同的图像区域关联起来解决语义混淆和局部信息丢失的问题,从而让模型能够更好地适应不同的数据分布,提高模型的泛化能力.

一些方法旨在从支持图像的前景区域来获得不同的原型提升分割性能. Liu 等人[63]提出将目标类整体的类原型分解为多个局部原型,能够捕获多样化和细粒度的对象特征,在语义对象区域产生更好的空间覆盖. 首先利用 K-means 聚类对支持前景特征进行划分,通过平均池化得到 N 个初始原型 $\tilde{\boldsymbol{p}}_i$;然后,将语义类的全局上下文信息合并到部分感知原型中,即

$$\boldsymbol{p}_i=\tilde{\boldsymbol{p}}_i+\lambda_p\sum_{j=1\&\&j\neq i}^{N}\mu_{i,j}\tilde{\boldsymbol{p}}_j \tag{24}$$

$$\mu_{i,j}=\frac{d(\tilde{\boldsymbol{p}}_i,\tilde{\boldsymbol{p}}_j)}{\sum_{j\neq i}d(\tilde{\boldsymbol{p}}_i,\tilde{\boldsymbol{p}}_j)} \tag{25}$$

其中 λ_p 是缩放因子,d 是相似性度量方式. 最后通过超像素聚类利用未标记的同类图像数据来丰富部分感知原型,使用图注意力网络来平滑未标注样本特征并将其结合到局部原型中,从而更好地建模语义对象的类内变量,即

$$\boldsymbol{p}_i^r=\boldsymbol{p}_i+\lambda_r\sum_{j=1}^{|\tilde{R}_k^u|}\phi_{i,j}\bar{\boldsymbol{r}}_j \tag{26}$$

$$\phi_{i,j}=\frac{d(\boldsymbol{p}_i,\bar{\boldsymbol{r}}_j)}{\sum_{j\neq i}d(\boldsymbol{p}_i,\bar{\boldsymbol{r}}_j)} \tag{27}$$

其中 λ_r 是缩放因子,$\bar{\boldsymbol{r}}_j$ 是图注意力网络节点 j 处的特征表示,$|\tilde{R}_k^u|$ 是指未标记数据中与类 k 中相关特征的个数. 此外,该方法还利用一个额外的传统语义分割分支来提升分割性能. Li 等人[64]仅利用超像素聚类处理支持前景特征来生成多个具有代表性的支持原型,并按语义信息与查询特征进行匹配实现分割;而 Wang 等人[65]是通过超像素和 K-means 聚类来处理支持前景特征,生成互相补充并且能够适应支持和查询图像尺度差异的多个原型指导分割,并设计了一种 round-way 反馈机制将多尺度的具有区分性的信息添加到解码器中. Zhang 等人[66]引入自指导和交叉指导机制(SCL),利用支持原型预测支持掩码 $\hat{\boldsymbol{M}}^S$,将 $\hat{\boldsymbol{M}}^S$ 和真实掩码 $\boldsymbol{M}^S$ 相减,利用缺失的前景掩码补充支持原型所丢失的关键信息,为分割提供边缘和细节信息. Zhang 等人[67]提出生成三个具有丰富语义的原型辅助分割:类原型是利用全局掩码池化获得通用性的特征,峰值原型通过选取前景区域内的最大值向量来捕捉最具区分性的特征,自适应原型通过学习的方式得到一个平均原型来捕捉特征内部长期依赖性. 而 Rao 等人[68]则提出对目标的频率差异进行建模,在频域中将支持信息分解为多个频率的原型指导目标的语义对齐. 它将目标的语义信息分为低频、中频和高频部分. 低频信息包含更一致的信息,中频信息包含必要的目标语义信息,而高频部分包含形状信息等. Lang 等人[69]利用分而治之思想提出一个代理网络(DCP),利用支持原型分割支持图像,并将支持图像的预测掩码 $\hat{\boldsymbol{M}}^S$ 划分为具有不同属性的多个区域:

$$\begin{cases}\boldsymbol{M}_\alpha^{(x,y)}=\mathbf{1}[\hat{M}_s^{(x,y)}=M_s^{(x,y)}=1],\\ \boldsymbol{M}_\beta=\boldsymbol{M}_s-\boldsymbol{M}_\alpha,\\ \boldsymbol{M}_Y^{(x,y)}=\mathbf{1}[\hat{M}_s^{(x,y)}=M_s^{(x,y)}=0],\\ \boldsymbol{M}_\delta=\boldsymbol{G}-\boldsymbol{M}_s-\boldsymbol{M}_Y\end{cases} \tag{28}$$

利用这些掩码生成多个具有不同信息的原型,其中 $\mathbf{1}$ 是指示函数,当输入为 True 的时候,输出为 1,输入为 False 的时候,输出为 0;$\boldsymbol{G}$ 是一个全 1 的矩阵. Adaptive FSS[70]基于 adapter 机制提出了一种新的框架,利用新类样本微调少量模型参数,提出的 prototype adaptive 模块(PAM)可以嵌入到任何 FSS 方法中. PAM 在微调阶段利用支持集提供的掩码获得类原型,通过不断更新类原型库中的原型得到更通用的类别表示,然后根据类原型库中的原型对支持和查询特征进行增强.

鉴于图像背景中可能会包含有助于分割的信息,一些方法提出利用支持图像背景区域提取指导信息辅助分割. Yang 等人[30]提出原型混合模型(PMMs),使用 EM 算法生成多个支持原型,激活查询图像中的目标区域并抑制背景信息. PMMs 是一个概率混合模型,定义为

$$p(\boldsymbol{s}_i|\boldsymbol{\theta})=\sum_{k=1}K\omega_k p_k(\boldsymbol{s}_i|\boldsymbol{\theta}) \tag{29}$$

其中 K 是原型个数,ω_k 是混合权重,满足 $0\leq\omega_k\leq1$ 并且 $\sum_{k=1}K\omega_k=1$,$\boldsymbol{\theta}$ 是模型参数,$p_k(\boldsymbol{s}_i|\boldsymbol{\theta})$ 表示第 k 个基于核距离的概率模型. 对于 EM 算法,它包含两个步骤:E-steps 和 M-steps. E-steps 计算样本 $\boldsymbol{s}_i$ 的均值 $\boldsymbol{E}_{ik}$:

$$\boldsymbol{E}_{ik}=\frac{p_k(\boldsymbol{s}_i|\boldsymbol{\theta})}{\sum_{k=1}K\omega_k p_k(\boldsymbol{s}_i|\boldsymbol{\theta})} \tag{30}$$

在 M-steps 计算得到的均值用来更新 PMMs 的均

值向量 $\boldsymbol{\mu}_k$：

$$\boldsymbol{\mu}_k=\frac{\sum_{i=1}^{N}\boldsymbol{E}_{ik}\boldsymbol{s}_i}{\sum_{i=1}^{N}\boldsymbol{E}_{ik}} \tag{31}$$

其中 N 是样本个数，对前景和背景部分进行计算得到前景和背景的多个原型辅助分割. Gairola 等人[71]提出了一种前景-背景注意力融合分割网络(SimPropNet)，利用支持图像掩码得到前景和背景原型，通过计算这些原型与查询特征间的相似度关系获得查询特征的前景和背景激活，然后使用注意力机制获得特征的前景和背景注意力图，实现支持和查询图像的信息交互. 此外该方法还同时预测支持和查询掩码，通过共享模型架构获得更一致的特征. Pambala 等人[72]提出语义元学习分割网络(SML)，利用视觉特征得到的前景和背景原型 $\boldsymbol{\phi}^s_{i,c}$，$\boldsymbol{\phi}_{i,\mathrm{bg}}$来校正语义特征前景和背景原型 $\boldsymbol{a}^s_{i,c}$，$\boldsymbol{a}_{i,\mathrm{bg}}$，通过标准的脊回归：

$$\boldsymbol{\Phi}=\left[\left\{\boldsymbol{\phi}^s_{i,c}\,|\,\boldsymbol{\phi}_{i,\mathrm{bg}}\right\}_{i=1}^{K}\right]_{c=1}^{C}\in\mathbb{R}^{d\times 2|S|} \tag{32}$$

$$\boldsymbol{A}=\left[\left\{\boldsymbol{a}^s_{i,c}\,|\,\boldsymbol{a}_{i,\mathrm{bg}}\right\}_{i=1}^{K}\right]_{c=1}^{C}\in\mathbb{R}^{d_a\times 2|S|} \tag{33}$$

$$L_W=\left\|\boldsymbol{\Phi}-\boldsymbol{W}\boldsymbol{A}\right\|_2^2+\lambda\left\|\boldsymbol{W}\right\|_2^2 \tag{34}$$

学习权重矩阵 $\boldsymbol{W}$，增强语义嵌入 $\boldsymbol{a}_c$和 $\boldsymbol{a}_{\mathrm{bg}}$来生成的前景和背景原型指导分割：

$$\boldsymbol{h}_W(\boldsymbol{a}_c)=\boldsymbol{W}\boldsymbol{a}_c \tag{35}$$

$$\boldsymbol{h}_W(\boldsymbol{a}_{\mathrm{bg}})=\boldsymbol{W}\boldsymbol{a}_{\mathrm{bg}} \tag{36}$$

Xie 等人[73]提出一个循环存储网络(CMN)，在支持前景和背景原型的指导下，与不同尺度下的查询特征进行交互获得多对多分辨率的增强查询特征，将其存储起来，分割时通过循环读取这些特征更好地捕捉不同分辨率下的目标变化. Liu 等人[74]提出学习非目标区域分割网络(NERTNet)，通过在大量图像上学习得到背景掩码，并用其去除背景区域干扰来提高分割性能. Yang 等人[75]则引入一个新类挖掘分支，通过从基类图像中获得的可迁移的语义子簇，在离线状态下可以直接标记图像中潜在的新类. Wu 等人[76]提出元类存储网络(MMNet)，元类是在所有类中共享的信息. 它在训练时引入可学习的嵌入来存储基类的元类信息，并将其迁移到新类中，避免将图像中非目标类都看作背景导致分割性能下降. 这些方法通过使用支持图像的背景区域提取指导信息，来改善分割的性能. SimPropNet、SML、CMN、NERTNet 和 DCP 均利用支持背景掩码获得支持背景原型提升查询图像前景和背景的可区分性.

由于存在较大类内差异，无论是前景目标还是背景信息，来自同一张图像的目标和背景要比来自其他图像的目标和背景更相似，从这个角度出发，有许多方法直接从查询图像中挖掘分割指导信息. Liu 等人[77]提出一个查询引导的分割网络(QGNet)，第一次利用无监督的查询图像进行自监督特征学习. QGNet 提出使用 Felzenszwalb 算法的基于图结构分割方式和线性迭代聚类方式生成局部图像区块，并进行局部对比学习；全局对比学习是在整张图像层面上进行对比学习. 通过全局-局部对比学习一个先验特征提取器：

$$L_{\mathrm{global/local}}=-\log\frac{\exp\left(\boldsymbol{q}\times\frac{\boldsymbol{k}_+}{\tau}\right)}{\sum_{i=0}^{K}\exp\left(\boldsymbol{q}\times\frac{\boldsymbol{k}_i}{\tau}\right)} \tag{37}$$

其中 $\boldsymbol{k}_+$是来自同一张图像的正样本或同一个图像块的不同视角，$\boldsymbol{q}$ 是编码的查询图像向量或查询图像块向量，τ 是温度超参数. 该方法能够从未标记的图像中提取查询先验信息，并根据生成的先验信息定位查询图像中的目标. 然后利用特征提取器得到的特征和支持掩码计算原型和查询先验掩码用于分割. IPMT[78]引入中间原型 $\boldsymbol{G}$，从支持特征中 $\boldsymbol{F}^s$ 挖掘决定性的类别信息，并从查询特征中 $\boldsymbol{F}^q$ 挖掘自适应的类别知识来缓解类别信息差距，利用支持掩码 $\boldsymbol{M}^s$ 和预测查询掩码 $\boldsymbol{P}^q$ 以迭代的方式学习学习中间原型：

$$\boldsymbol{IPM}=\mathrm{MLP}(MA(\boldsymbol{G},\boldsymbol{F}^s,\boldsymbol{M}^s)+MA(\boldsymbol{G},\boldsymbol{F}^q,\boldsymbol{P}^q)+\boldsymbol{G}) \tag{38}$$

其中 MA 是一种使用掩码增强后的交叉注意力机制. QPENet[79]利用查询特征来辅助前景和背景原型的生成，遵循 support-query-support 过程(即利用支持掩码预测查询图像，然后使用预测的查询掩码得到支持预测掩码)来生成前景原型. 此外，该方法设计了一个全局背景清除模块来消除不利于前景分割的特征部分. 还有一些方法利用查询特征生成查询原型辅助分割. PST[80]在训练时使用支持和查询掩码通过 EM 算法分别生成支持和查询原型，并使用最小费用最大流算法：

$$\min\sum_{k,k'}p_{k,k'}(1.0-\boldsymbol{\mu}^+_{sk}\times\boldsymbol{\mu}^+_{qk'}) \tag{39}$$

其中 $p_{k,k'}$是节点 k 和 k'间的流，将支持前景原型 $\boldsymbol{\mu}^+_{sk}$ 和查询前景原型 $\boldsymbol{\mu}^+_{qk'}$进行语义匹配，通过语义分解匹配将支持和查询图像中的目标语义对齐. Zhao 等人[81]提出使用潜在原型进行对比增强的小样本分割方法(CELP)，通过计算高层查询特征的不同位置间的相似度 $\boldsymbol{D}_q$，得到属于同一类别的区域作为初始查询掩码，并使用掩码平均池化操作得到查询原型. 为了提高初始查询掩码的可靠性，CELP 提出一

种采样机制，并利用对比增强进一步关注相似区域并增强未见类的激活. 不同于 CELP，Yang 等人[82]提出一个先验语义协调网络(PSHNet)，利用支持和查询高层特征得到的相似度矩阵作为查询的初始掩码，并通过 EM 算法得到多个查询原型与支持原型一起分割查询图像. Mao 等人[83]提出双原型网络(DPNet)，基于查询前景特征构建伪原型，选择与支持像素最相似的查询像素，并将其映射回支持像素，保留支持像素均属于前景的查询像素生成伪原型. Fan 等人[84]和 Tang 等人[85]均使用支持原型得到的预测查询掩码生成查询原型. Fan 等人提出的自支持原型网络(SSP)选择具有较高置信度的查询图像预测掩码生成前景和背景原型；而后者是利用查询损失对预测查询掩码进行优化，并利用其前景掩码计算得到查询原型. 然后利用支持和查询原型预测支持掩码得到融合权重，加权得到最终用于分割的原型. QSR[86]和 CobNet[87]方法都从查询图像的背景挖掘知识分割背景区域. 前者可根据已见类和潜在类向量与查询原型间的相似度对查询特征进行处理，并利用已知类和查询图像背景类中的未知类标签消除其前景信息，即将待分目标类标签设置为 0，其余类别设置为 1 来生成查询背景原型. 后者对查询特征中未交叠的、指定大小的特征进行平均池化，将得到的向量作为查询背景原型并与查询特征进行交互，利用支持原型得到的初始查询预测掩码并与多尺度查询背景特征融合用于后续分割.

除了前面提到的方法，还有一些方法通过对生成多个原型增加样本多样性或增强类别知识来辅助分割. Siam 等人[88]提出自适应掩码代理，根据当前支持图像在不同分辨率下的原型和历史代理进行加权，得到当前任务的正代理. 为了获得更具多样性的原型信息，Wang 等人[89]将原型 $\boldsymbol{z}$ 建模为概率分布：

$$\boldsymbol{z}\sim p_\theta(\boldsymbol{z}|S)=\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{\text{prior}},\boldsymbol{\sigma}^2_{\text{prior}}) \tag{40}$$

经过两层全连接层将特征向量映射到 $\boldsymbol{\mu}_{\text{prior}}$ 和 $\boldsymbol{\sigma}^2_{\text{prior}}$，类原型不再是一个确定性的向量，可以消除有限数据和类内差异大带来的不确定性. 该方法将问题约束为变分推理问题，通过优化找到变分后验分布以逼近原型的真实后验概率分布. 分割时从建模的分布中通过蒙特卡洛采样得到多个的原型. Okazawa[90]提出一个类间原型关系网络(IPRNet)，通过计算出该批次中所有类别的类原型，利用关系损失训练网络，

$$L_r=\frac{\sum\limits_{c_s}^{n}\sum\limits_{c_t}^{n}\text{Sim}(\boldsymbol{P}^{C_s},\boldsymbol{P}^{C_t})\mathbf{1}[c_s\neq c_t]}{\sum\limits_{c_s}^{n}\sum\limits_{c_t}^{n}\mathbf{1}[c_s\neq c_t]} \tag{41}$$

其中 c_s 和 c_t 指的是类别编号，$\boldsymbol{P}^{C_s}$ 和 $\boldsymbol{P}^{C_t}$ 指的是不同类别的原型向量，n 是所有类别的原型数，Sim 是指余弦相似度计算. 通过优化该损失降低目标类和其他类间的相似性来提高可区分性.

对于过分拟合基类导致分割边界模糊的问题，Cheng 等人[91]提出整体原型激活的方法(HPA)，利用无需训练的机制获得基类原型，通过基类和新类原型构成整体原型，使用这些基类原型过滤掉与目标不相关的高置信度区域来获得与查询特征更好匹配的原型. 为了解决小样本分割任务中的类分布偏移，MENUA[92]提出一个类共享的存储模块来保存基类特征，然后利用高斯函数和基类特征对支持和查询特征进行增强，与支持原型一起输入解码器进行分割.

上述方法使用多个原型建模图像的局部信息，捕捉到图像中的不同特征和上下文信息，每个原型都可以代表一个特定的概念或物体，具有更灵活的适应性，在一定程度上能够缓解单个原型建模存在的问题. 后来，随着基于多个原型的小样本分割算法的发展，这类方法已经不局限于仅使用多个原型建模局部信息，还有一些方法利用不同方式生成多个原型来增加特征的多样性，或者通过获得不同功能的原型提升分割性能等.

在小样本语义分割任务中，使用单个或几个向量原型来表示类别信息是十分粗糙的，即使使用多个原型也无法捕捉不同目标的细节信息，并丢失了大量有用的结构和空间信息，这会妨碍目标区域进行细粒度的匹配，在解决目标遮掩和对于存在较大类内差异的图像对时仍存在一些问题，进而影响分割性能.

4.2.2　基于匹配的小样本分割方法

为了充分利用支持集信息，更好地捕捉图像中目标的细节信息和空间结构，人们提出了基于匹配的方法. 开始时这类方法使用图结构建模支持和查询图像间的关系，后来开始直接计算支持和查询特征像素间相似度，并使用 4D 卷积或注意力机制等方法处理得到的高维关系矩阵. 基于匹配的方法在近年来取得了飞速发展，并且大大提升了小样本分割性能. 该类方法能够最大程度地保留特征的空间结构信息，更好地捕捉支持和查询特征间的上下文信息，实现视觉密集特征匹配.

(1) 利用图结构建模支持和查询图像间关系.

针对原型的方法存在数据结构损失，基于图结构的方法被提出来，直接建模图像像素间的对应关系，将像素点看作图的节点，可以充分利用支持图像

的空间和结构等细节信息,以减少信息损失.

Zhang 等人[29]提出金字塔图网络(PGNet),首次建模支持和查询特征像素级的密集对应关系,并利用图结构来处理结构化的分割数据.它使用图注意力单元来传递支持图像的标签信息,并使用金字塔式的图推理结构来处理不同分辨率的特征,传递不同语义层面的标签信息.Wang 等人[93]则引入图注意力机制建立像素级别的联系,提出图注意力网络(DAN),通过抑制高权重连接和增强低权重连接来激活更多的区域,使前景的所有像素点都能够参与到连接当中.通过处理更大的区域,而不是之前较小的特定激活区域,使得网络能够建立更加稳固的连接.此外,DAN 还引入了类似 U-Net[94]的结构来融合多尺度信息提升分割性能.Liu 等人[95]和 Xie 等人[96]均提出使用图卷积网络建模并挖掘支持和查询图像间的关系.Liu 等人[95]提出关系匹配网络(CMNet),在同一张图像中通过内部流动传递上下文信息,在支持和查询图像间通过跨图像流动传递类别信息;Xie 等人[96]提出尺度感知图神经网络(SAGNN),引入 self-node 协作机制,通过边 $\boldsymbol{e}_{ij}^{t-1}$ 的其他节点 $\hat{\boldsymbol{h}}_j^{t-1}$ 丰富当前节点 $\hat{\boldsymbol{h}}_i^{t-1}$ 的特征,增强后的节点为 $\boldsymbol{g}_{ji}^{t}$:

$$\boldsymbol{g}_{ji}^{t}=\mathrm{Softmax}(\boldsymbol{e}_{ij}^{t-1})(\hat{\boldsymbol{h}}_i^{t-1}+\hat{\boldsymbol{h}}_j^{t-1}) \tag{42}$$

对与点 $\hat{\boldsymbol{h}}_i^{t-1}$ 作为顶点的所有边进行相同的增强处理操作,并进行求和获得最后的节点.

这些方法利用图神经网络的知识来解决小样本语义分割中的问题,通过直接建模像素间的关系,能够充分利用图像上下文和空间结构信息,在提高分割精度、捕捉上下文信息和建立密集对应关系方面都有较为显著的贡献.

(2)计算相似度建模支持和查询图像间关系.

使用图结构或图卷积网络建模支持和查询图像间相似度关系时需要对图的节点和边进行学习,不可避免会增加模型参数量,同时增大过拟合的可能性,导致模型学习到知识的泛化性差,影响模型分割性能.此外,随着计算设备的进步,高复杂度计算变得现实.因此直接人们提出直接计算支持和查询图像特征像素间的高维相似度关系矩阵来保留图像结构关系,从而提高分割性能.其中支持和查询图像间的相似度关系度量通常是通过计算余弦相似度实现的,这种方式简单易于理解,且不会增加其他参数量,同时能够很好地保留图像细节和结构信息.

有些方法首先将支持特征与支持掩码相乘获得支持前景特征后,通过计算支持前景特征与查询特征间的相似度实现信息交互.Min 等人[31]提出超关系挤压网络(HSNet),计算支持前景和查询特征多层次的余弦相似度实现信息交互.它设计了一种高效轻量化的 center-pivot 4D 卷积,将 4D 卷积核参数的主元中心化,并分离成两个 2D 空间中的卷积之和来高效处理高维相似度关系矩阵,

$$(c*K_{cp})(x,x')=\sum_{p'\in P(x')}c(x,p')k_c^{2\mathrm{D}}(p'-x')+\sum_{p\in P(x)}c(p,x')k_{c'}^{2\mathrm{D}}(p-x) \tag{43}$$

其中 $k_c^{2\mathrm{D}}$ 和 $k_{c'}^{2\mathrm{D}}$ 分别是两个 2D 空间的卷积核,$P(x)$ 和 $P(x')$ 分别指点 x 和 x' 的邻域点集合.同时采用金字塔式的设计来捕捉高层语义和底层几何线索,从粗到细进行掩码预测.这种方法为小样本分割提供了一种新的思路,后续许多方法也采用了直接计算支持和查询特征像素之间相似度的方式来进行分割.Hong 等人[97]提出 4D 卷积 Swin Transformer 处理多尺度的支持前景特征和查询特征间的相似度矩阵.它通过一系列具有较小感受野的卷积操作,将局部上下文传递给所有像素.同时,它引入了卷积归纳偏差并扩展了补丁嵌入模块,使其能够处理高维输入.这些方法通过不同的方式处理支持前景特征和查询特征像素间的高维相似度关系矩阵,通过自注意力机制、交叉注意力机制、多尺度特征和高维卷积等技术手段来提高分割的性能和效率.

前面提到的方法只利用了支持图像的前景信息,考虑到其背景部分可能也包含一些利于分割的信息,一些方法在计算相似度时考虑其背景信息,即先计算支持和查询特征的相似度矩阵,然后通过掩码过滤出前景部分,获得待分割类别知识.Zhang 等人[98]提出循环一致的 Transformer 解决小样本分割任务(CyCTR),并利用自注意力机制和交叉注意力机制增强查询特征.CyCTR 提出循环一致的注意力机制,通过计算支持和查询特征间的余弦相似度,找到与支持像素最相似的查询像素,然后利用该查询像素找到支持特征中最相似的支持像素,利用支持掩码保留两个支持像素标签一致且正确的查询像素,以过滤掉有害的查询特征像素,实现查询特征与信息丰富的支持像素之间的交互.Shi 等人[32]提出密集交叉查询和支持注意力加权掩码聚合网络(DCAMA),使用交叉注意力机制,通过计算支持和查询特征间的相似度来对支持掩码加权得到查询掩码,即

$$\mathrm{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})/\sqrt{d})\boldsymbol{V} \tag{44}$$

其中 DCAMA 将查询特征当作 Transformer 结构的 Query($\boldsymbol{Q}$),支持特征当作 Key($\boldsymbol{K}$),支持掩码当做 Value($\boldsymbol{V}$)进行分割.Zhang 等人[99]提出基于上下文和相似性的 Transformer(CATrans),使用一个

层次化的结构来结合上下文和相似性信息，生成更具区分性的表示. 除此之外，它通过自注意力和交叉注意力机制分别计算支持特征和查询特征自身的上下文信息以及它们之间的上下文信息实现特征增强和交互，有助于产生更精确、更鲁棒的对应关系. 同时利用一个小的卷积网络编码掩码信息帮助交互后的特征获得类别信息，然后通过 U-Net 结构逐渐上采样得到分割掩码. Kang 等人[100]提出注意力挤压网络(ASNet)，通过自注意力机制和支持掩码将支持和查询特征间的相似度矩阵 $\boldsymbol{A}$ 以及经过投影得到的矩阵 $\boldsymbol{V}$，转换为关系张量

$$\hat{\boldsymbol{A}}(p_t,p_k)=\frac{\exp(\boldsymbol{A}(p_t,p_k)Y_s(p_k))}{\sum_{p_k'}\exp(\boldsymbol{A}(p_t,p_k')Y_s(p_k'))} \tag{45}$$

$$\boldsymbol{C}^{s'}=\mathcal{F}_{\text{process}}(\hat{\boldsymbol{A}}\boldsymbol{V}) \tag{46}$$

其中 $Y_s(p_k)$是掩码操作，属于前景部分赋值为 1，否则赋值为 $-\infty$，$\mathcal{F}_{\text{process}}$ 是一系列卷积或线性投影操作. 该方法将小样本学习和小样本分割任务整合到一个统一的框架中，这个前景图可以用于多标签的分类和像素的分割任务. Cao 等人[101]提出分别计算支持图像的前景特征和背景特征与查询特征间的相似度来帮助优化前景信息，并通过层次化的迭代消除具有高置信度的前景区域，激活剩余区域的细节信息来捕捉隐藏的细节信息，抑制背景的错误激活，不断挖掘目标的语义和边缘信息. Liu 等人[102]在 HSNet 基础上，针对原始相似性矩阵中存在噪声的问题，提出特征增强的上下文网络(FECANet)，通过一种新的交叉注意力和通道机制来增强支持和查询特征表示，抑制不同类间的局部相似性，增强同类间的全局相似性，从而减少由类内差异性和类间相似性带来的噪声. 针对缺乏上下文信息的问题，该方法计算每个像素与特定邻近区域间的相似度来得到局部上下文关系度量，与支持和查询特征得到的全局关系矩阵融合得到最后的关系矩阵指导分割. TBTNet[103]设计了目标感知 Transformer 层(TTL)来处理相似度度量矩阵，通过计算支持和查询特征间的 cross-similarity 和不同层次查询特征融合后的 self-similarity 迫使模型更加关注前景信息. 它将超相关视为一个特征，从而显著减少了特征通道的数量. TBSNet[104]强调了背景在分割中的重要性，基于 query-relevant 和 target-relevant 得分来抑制具有破坏性支持背景特征，利用处理后的特征计算相似度关系. SCCAN[105]设计了一个自校准交叉注意力(SCCA)模块来解决背景不匹配和前景-背景纠缠的问题，并缓解无效的支持 patch 和不匹配的支持 patch 给分割带来的干扰. SCCA 模块首先计算支持和查询特征 patch 间的相似度进行对齐，然后计算支持和查询特征间的 cross-attention 得分和查询特征自身的 self-attention 得分对查询特征进行增强校正. DAM[106]使用双向 3D 卷积捕捉每个支持-查询对中的像素到像素和像素到 patch 的关系，并提出滞后空间滤波模块(HSFM)利用支持背景过滤掉查询特征的背景特征，从而增强查询特性前景特征. 这些方法利用支持图像的背景信息来结合上下文信息和增强特征相似度，从而提高分割准确性和鲁棒性. 利用背景信息辅助分割任务可以提供更全面的图像理解，帮助模型更好地区分前景和背景，并产生更准确的分割结果.

由于大多数特征提取器是卷积神经网络，因此经过一系列卷积处理后的支持特征与掩码相乘后会丢失一些细节信息，然而这些细节信息对于小物体等目标分割来说尤为重要. 为此，一些方法通过将输入图像与掩码相乘来保留细节信息. 例如，HM[107]方法不仅考虑了特征与掩码相乘(Feature Masking，FM)，还引入了输入图像与掩码相乘(Input Masking，IM). HM 优先考虑 FM 特征，在 FM 未激活的区域中，使用 IM 特征来替代，可以更好地保留细节信息. 而 MSI[108]仅使用掩码与支持图像乘积和原始支持特征分别与查询特征计算余弦相似度，通过这种方式可以消除掩码标记不充分或目标很小的情况下导致的分割不准确问题，并利用背景中的上下文信息提供分割指导，从而提高分割性能. 这些方法的主要目标是通过改进特征表示来更好地捕捉细粒度的类别信息，以提高小样本分割的性能. 通过引入输入图像与掩码相乘的方式，它们能够更好地处理细节丢失的问题，从而在小目标分割任务中取得更好的效果.

LC-CAN[109]第一次利用数据增强来优化小样本分割框架，提出一个 instance-aware 数据增强机制提高支持和查询图像分布的一致性. 此外，LC-CAN 利用支持和查询图像之间的密集相关性，设计了局部一致性引导的交叉关注来细化查询特征表示.

基于匹配的方法直接计算支持和查询图像像素间的相似度对应关系，计算量较大，并且在整合高维关系得分时缺乏灵活性，收敛速度慢. 此外，在特征像素级别计算相似度，由于其特征可能存在错误激活导致噪声产生，对分割结果产生不良影响.

4.2.3　融合原型和匹配的小样本分割方法

基于原型的方法通过池化操作可以减少噪声对分割性能的影响，但在这个过程中可能会损失一些

空间细节信息.相比之下,基于匹配的方法可以更好地保留空间信息,但在计算相似度时容易受到噪声的干扰,从而对分割结果产生较大的影响.为了克服这些问题,一些方法综合原型和相似度计算,以提供空间细节信息并减少噪声的影响.

初始融合原型和匹配,是在基于原型的方法中,计算支持和查询高层语义特征间的相似度,得到先验掩码来辅助分割.Tian 等人[51]提出先验指导特征增强网络(PFENet).因为高层特征主要包含语义信息,PFENet 利用特征提取器输出的高层语义特征,直接计算掩码后的支持特征 $\boldsymbol{F}^S$ 和查询特征 $\boldsymbol{F}^Q$ 间的余弦相似度生成无需训练的先验掩码:

$$\boldsymbol{Y}_Q=\text{Cos}-\text{Similarity}(\boldsymbol{F}^Q,\boldsymbol{F}^S\times\boldsymbol{M}^S) \tag{47}$$

与支持原型一起用于指导分割,提升预测准确度和泛化性能.PFENet 中提出的先验掩码在后来很多方法中[82,110-112]都得到了使用,这种无需学习的掩码生成方式,能够避免产生基类偏好获得与类无关的知识,提高模型泛化性,同时又能充分利用支持和查询图像间的语义关系.在 PFENet 的基础上,Luo 等人[113]利用支持特征 $\boldsymbol{\mathcal{X}}_s$ 和查询特征 $\boldsymbol{\mathcal{X}}_q$ 间的上下文信息,使用邻近语义线索更好地定位目标:

$$\boldsymbol{\mathcal{R}}^c(i,j)=\sum_o[\boldsymbol{\mathcal{X}}_q(i+o),\boldsymbol{\mathcal{X}}_s(j+o)] \tag{48}$$

从而获得更准确的先验掩码,其中 $o\in[-m,m]\times[m,m]$ 表示 $m\times m$ 区块内的偏移量.此外,通过整体和局部计算上下文信息,可以减轻支持特征中冗余的、与查询图像不相关的关系响应,并将查询特征与先验掩码相乘来消除部分噪声.Wang 等人[111]指出许多方法忽视了数据间的语义关联,提出了一种分支网络结构(FFNet).一个分支计算支持和查询特征间的相似度,获得 K 个与查询特征最相似的支持特征像素,用于辅助分割.另一个分支计算类原型进行分割.PhotoFormer[110]直接利用 PFENet 中先验掩码辅助分割,并使用 Transformer 结构实现支持原型和查询特征间的交互,将支持原型作为 Transformer 模块的 Query,将查询特征作为 Key 和 Value,捕捉查询特征中的空间细节信息和目标类的语义信息.MIANet[112]则是计算不同尺度下支持和查询特征间的余弦相似度,得到多尺度的先验掩码辅助分割,多尺度下的特征交互能够为不同大小、不同形状的物体提供丰富的指导.并且通过三元组损失优化语义标签提供的通用类别信息与支持原型的融合,从而缓解类间差异大的问题.

随着基于匹配方法的快速发展,一些方法开始利用原型消除相似度度量产生的噪声,通过原型和匹配两个分支实现特征交互.Huang 等人[114]提出一个联合类无关和类有关的对齐网络(JC^2A),将来自支持图像中与类有关的信息和来自目标区域挖掘得到的与类无关的信息结合起来,消除类变体和背景混淆导致的类偏差.该方法在匹配分支对支持和查询特征建立点对点和点对块的关系矩阵,获得与类别最相关的目标空间信息和上下文信息:

$$\boldsymbol{P}^d_{aw}=\text{ReLU}(\boldsymbol{Q}_q)\times(\text{Softmax}(\boldsymbol{K}_S)^{\text{T}}\boldsymbol{V}_s) \tag{49}$$

$$\boldsymbol{P}^s_{aw}=\frac{1}{m^2}\sum_{m\times m}\text{ReLU}(\boldsymbol{Q}^p_q)\times(\text{Softmax}\,\mathcal{B}(\boldsymbol{K}^p_s))^{\text{T}}\mathcal{B}(\boldsymbol{V}^p_s)) \tag{50}$$

其中 $\boldsymbol{Q}_q$,$\boldsymbol{K}_S$ 和 $\boldsymbol{V}_s$ 分别是经过线性投影的查询和掩码支持特征,q 和 s 分别指支持和查询特征.$\mathcal{B}$ 是选取前 k 个最相关特征区域的稀疏化操作.将得到的点对点关系矩阵 $\boldsymbol{P}^d_{aw}$ 和点对块关系矩阵 $\boldsymbol{P}^s_{aw}$ 相加,然后与原型分支得到的支持原型输入解码器进行分割.除此之外,JC^2A 通过对基类某类中不同实例的原型加权得到该类的特征原型,来挖掘查询图像中与类无关的信息和潜在目标信息辅助分割.MSANet[115]和 MSHNet[116]均通过两个分支来分别计算支持和查询图像间的余弦相似度和支持原型与查询特征间的相似度进行分割.MSANet 还利用一个注意力模块处理支持特征,使其更加关注类别有关的信息.MSHNet 利用不同尺度的预测掩码优化模型,并通过实验证明基于全局特征生成的局部原型相似度与基于局部特征生成的全局余弦相似性在逻辑上是互补的.MCE[117]提出 Masked Cross-Image Encoding 来学习支持和查询特征中目标物体共享的视觉表示,利用对称的 cross-attention 结构来双向处理不同图像间的关系获得掩码的支持和查询图像特征图 $\boldsymbol{R}_s$ 和 $\boldsymbol{R}_q$,即

$$\boldsymbol{R}_s=\text{Softmax}\,\frac{(\boldsymbol{W}_q\boldsymbol{F}_s+\boldsymbol{\mathcal{M}})(\boldsymbol{W}_k\boldsymbol{F}_s)^{\text{T}}}{\sqrt{d}}(\boldsymbol{W}_v\boldsymbol{F}_q) \tag{51}$$

$$\boldsymbol{R}_q=\text{Softmax}\,\frac{(\boldsymbol{W}_q\boldsymbol{F}_q)(\boldsymbol{W}_k\boldsymbol{F}_q)^{\text{T}}}{\sqrt{d}}(\boldsymbol{W}_v\boldsymbol{F}_s+\boldsymbol{\mathcal{M}}) \tag{52}$$

其中 $\boldsymbol{\mathcal{M}}$ 是利用支持掩码经过变换得到的,$\boldsymbol{W}_*$ 是可学习参数,d 特征维度.将经过 MLP 处理过的 $\boldsymbol{R}_q$,$\boldsymbol{R}_s$,支持原型以及相似度度量一起输入解码器进行分割.

还有一些方法将原型和相似度度量串联起来,进一步提升分割性能.AMFormer[118]旨在从查询图像获得指导分割的类别信息.AMFormer 首先利用支持原型获得查询图像的初始伪掩码,并用其过滤

查询图像的背景干扰特征；然后利用 cross-attention 机制计算查询特征和过滤掉背景的查询目标特征间的相似度用于分割. 为了进一步提升分割性能，该方法采用对抗训练，通过计算真实掩码和预测掩码过滤得到的查询前景特征和多个局部代理间的相似度来鉴别真实掩码和预测掩码，从而不断提升“生成器”模块的分割性能. RiFeNet[119] 首先利用局部查询原型和全局支持原型对查询特征进行增强，然后将增强后的查询特征与支持原型增强后的支持特征输入 cross-attention 模块中计算相似度实现分割. 为了保持前景语义的一致性消除类内分布差异，RiFeNet 提出一个无标签分支，通过数据增强来增加样本多样性，使模型避免学习标注输入的样本偏差.

不同于上述提到的方法，Zhang 等人[52] 提出一种简单的“特征提取器＋线性分类器”框架(FPTrans)，引入一种新的提示机制使支持特征 $\boldsymbol{x}_s$ 和查询特征 $\boldsymbol{x}_q$ 在底层利用交叉注意力机制实现不同网络层 l 间的交互：

$$[\boldsymbol{x}_q^l,\boldsymbol{P}_q^l]=\mathcal{B}_l([\boldsymbol{x}_q^{l-1},\boldsymbol{P}^{l-1}]) \tag{53}$$

$$[\boldsymbol{x}_s^l,\boldsymbol{P}_s^l]=\mathcal{B}_l([\boldsymbol{x}_s^{l-1},\boldsymbol{P}^{l-1}]) \tag{54}$$

$$\boldsymbol{P}_l=(\boldsymbol{P}_q^l+\boldsymbol{P}_s^l)/2 \tag{55}$$

其中 $\boldsymbol{P}_*$ 是提示 token，$*$ 表示 s 或 q，$\mathcal{B}_l$ 是 Transformer 模块，基于提示获得多个前景和背景原型

$$\boldsymbol{v}_*=\frac{1}{G}\sum_{j=1}^{G}\boldsymbol{P}_{*j}^L \tag{56}$$

其中 $*$ 代表前景 f 或背景 b，G 是提示 token 的个数，L 是特征提取器层数. 除此之外，该方法还利用支持掩码过滤出的前景特征生成基于特征的代理，并使用多个局部背景代理提高模型的泛化性能. Wang 等人[120] 通过 Transformer 的编码器-解码器结构(AAFormer)将自适应原型作为代理融合到基于匹配的方法中. 特征表示编码器通过自注意力机制来学习包含全局上下文信息的像素特征；代理学习解码器利用交叉注意力机制蒸馏支持信息，得到几个具有空间感知性、上下文感知性和差异性的代理标记，并使用代理标记将像素级支持-查询相似度矩阵分解为两个低维的矩阵实现分割. FPTrans 通过提示学习和多个局部背景代理来增强特征表示，提高模型泛化性能. AAFormer 通过学习富含类别信息的代理标记，降低像素级匹配的计算复杂度.

这些方法的引入为分割任务带来了新的思路和改进，进一步提高了分割结果的准确性和鲁棒性.

4.2.4　其他方法

除了前三类方法外，还有一些方法从新的角度出发来解决小样本分割问题.

Yang 等人[53] 利用线性代数的知识来解决小样本分割问题，基于局部特征关系提出了一种新的转换模块(LTM). 它在高维嵌入空间计算余弦距离来构建关系矩阵：

$$\boldsymbol{R}_{ij}=\frac{\langle \boldsymbol{E}_{si},\boldsymbol{E}_{qj}\rangle}{\|\boldsymbol{E}_{si}\|_2\|\boldsymbol{E}_{qj}\|_2} \tag{57}$$

其中 $\boldsymbol{E}_{si}$ 和 $\boldsymbol{E}_{qj}$ 是支持和查询特征位置 i 和 j 的向量值，使用支持掩码 $\boldsymbol{G}_s$ 的广义逆进行线性转换，从而得到查询预测掩码：

$$\boldsymbol{A}=\boldsymbol{R}[(\boldsymbol{G}_s)^{\mathrm{T}}(\boldsymbol{G}_s(\boldsymbol{G}_s)^{\mathrm{T}})^{-1}] \tag{58}$$

不同于 LTM，Seo 等人[121] 提出一个与任务无关的特征 Transformer(TAFT)，学习一个线性转换矩阵 $\boldsymbol{P}=\boldsymbol{R}\boldsymbol{C}^+$ 处理支持图像特征，其中 $\boldsymbol{R}$ 是参考原型，$\boldsymbol{C}^+$ 是支持图像的前景和背景原型，该矩阵将特定任务的高层特征转换为一系列与任务无关且有利于分割的特征 $\boldsymbol{R}$. Xiong 等人[54] 则提出一种新的相似度度量，提出协方差矩阵的双形变聚合方法(DACM). 第一次利用基于高斯过程核学习方法得到的协方差核函数来代替传统的余弦相似度矩阵：

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sigma_0^2\exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{((\boldsymbol{x}_i)_d-(\boldsymbol{x}_j)_d)^2}{l_d^2}\right\} \tag{59}$$

其中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 指的是正则化后的支持和查询特征，σ_0^2、$\{l_d\}_{d=1}^D$ 是超参数. DACM 动态地采样容易分错的查询特征来学习高斯过程的协方差核函数，并提出双变形的 4D Transformer 处理得到的相似度矩阵. 而 Johnander 等人[55] 是从高斯过程回归的角度解决小样本分割问题. 通过对已知的支持特征 $\boldsymbol{x}_{sk}$ 和编码后的支持掩码 $\boldsymbol{M}_{sk}$ 进行高斯过程回归：

$$\mathcal{F}_{\text{gaussian}}=\Lambda(\{\boldsymbol{x}_{sk},\boldsymbol{M}_{sk}\}_k) \tag{60}$$

然后利用函数 $\mathcal{F}_{\text{gaussian}}$ 和查询特征得到查询掩码的概率分布实现分割. Zheng 等人[122] 提出从四元数的角度来进行关系矩阵学习，缓解高维关系张量的计算负担，并利用四元数代数的操作实现支持和查询图像内部的潜在交互. 利用四元数卷积实现对内部关系(即支持图像内部的关系)和外部关系(即查询集中的边界和形状信息)的联合学习，以充分利用支持和查询图像间的匹配得分进行交互. 四元数的代数特性为该方法的创新提供了理论基础，并为解决小样本分割问题提供了新的思路. 这些方法不同于传统的小样本分割方法，而是从概率、线性代数和高斯过程等数学角度出发，利用相关的数学理论和模型来解决小样本分割问题. 通过引入新的数学框架和思路，这些方法提供了一种不同的视角和解决方案，对小样本分割任务的改进具有积极的影响.

文献中提到了一些方法来补充卷积神经网络在

特征提取过程中缺少的细节纹理信息.其中,Azad等人[123]使用一系列高斯函数的差分(Difference of Gaussians)来减弱特征空间中的高频局部组成成分,从而改善模型的泛化性能.将去除高频部分的特征图输入到Bi-LSTM中有效地融合多尺度的空间表示.Min等人[124]利用纹理信息增强卷积神经网络输出的特征,引入纹理增强模块来丰富特征表示,更好地捕捉细节纹理信息.这些方法试图引入纹理信息来解决卷积神经网络在特征提取过程中存在的缺陷,通过补充丢失的细节纹理信息,可以提高模型的泛化能力和分割性能,从而改善小样本分割的性能.

目前的小样本分割方法将匹配和分割过程结合在一起,使用的分割模块笨重,限制了设计的灵活性,并增加了学习复杂度.为次,研究者们提出将匹配和分割解耦的方法.其中,MANet[125]和MMFormer[56]均采用先解耦后融合的方式,将小样本分割问题约束为一个掩码分类任务.首先解析查询特征生成一系列与类别无关的分割目标候选掩码,然后结合支持图像和掩码信息,对生成的候选掩码进行分类,最后根据掩码位置信息整合生成最终的查询掩码.MANet直接对支持原型和查询特征进行处理获得候选掩码的类别信息;MMFormer则先对支持和查询特征进行对齐,然后利用支持原型和候选掩码与查询特征生成的原型进行匹配实现候选掩码分类.这些方法通过将匹配和分割的过程分开来,增强了方法的灵活性和可解释性.解耦后的方法能够更好地适应不同的任务,并且提供了灵活的机制来处理小样本分割问题.这些方法为解决小样本分割问题提供了新的思路和解决方案.

Tan等人[126]利用目标内部的一致性,提出使用轮廓进行分割(CTANet),以解决目标边缘分割性能差和语义混淆的问题.Sun等人[127]提出了一种新的训练方式(SVF),利用奇异值分解预训练主干网络权重:

$$\boldsymbol{W}'=\boldsymbol{USV}^{\mathrm{T}} \tag{61}$$

其中$\boldsymbol{U}$和$\boldsymbol{V}$是酉矩阵,$\boldsymbol{S}$是奇异值.通过微调部分主干网络参数$\boldsymbol{S}$来调整新类的特征表示,同时保持预训练模型的语义线索.Liu等人[128]首次提出使用动态卷积来捕捉物体的内在细节,利用支持特征和掩码生成动态卷积核,实现支持和查询特征交互.这些方法提出了一系列创新思路和技术,旨在改善小样本分割的性能.通过引入轮廓信息、微调网络参数或采用动态卷积等方法,这些方法能够更好地捕捉目标的细节信息、适应新类别的特征表示,并提高分割的准确性和鲁棒性.

Hu等人[129]指出异质性是导致类内紧密度小的关键,提出多级异质性抑制网络(MuSH),从特征提取器设计的角度出发,将注意力从分割头转移到特征提取阶段,使得支持和查询图像在特征提取阶段进行交互.该方法利用Transformer中的注意力机制抑制样本间、区域间和图块间的异质性,通过计算交叉样本注意力关联token $\boldsymbol{x}_l^i$:

$$\boldsymbol{x}_*^{i+1}=\mathrm{Attn}(Q(\boldsymbol{x}_*^i),K(\{\boldsymbol{x}_*^i,\boldsymbol{x}_l^i\}),V(\{\boldsymbol{x}_*^i,\boldsymbol{x}_l^i\})) \tag{62}$$

其中$\boldsymbol{x}_*^i$表示第i个MuSH模块中的支持token $\boldsymbol{x}_s^i$或查询token $\boldsymbol{x}_q^i$,Attn是指交叉注意力的计算;跨区域$\boldsymbol{x}^{[c]}$和$\boldsymbol{x}^{[t]}\in\boldsymbol{X}^{[t]}$交互是比较两个区域间的余弦相似度:

$$\boldsymbol{x}^{[c]'}=\boldsymbol{x}^{[c]}+\mathrm{Softmax}\left(1-\frac{\boldsymbol{x}^{[c]}\boldsymbol{x}^{[t]}}{|\boldsymbol{x}^{[c]}||\boldsymbol{x}^{[t]}|}\right)\boldsymbol{X}^{[t]} \tag{63}$$

其中$\boldsymbol{x}^{[c]'}$表示更新的背景嵌入.同时MuSH还提出掩码图像分割任务训练模型来实现支持和查询样本间、不同区域和邻近区块间的交互并获得注意力权重,进一步提高类内的紧密度.

Peng等人[130]提出利用蒸馏思想解决小样本分割问题,提出分层密集蒸馏关系网络(HDMNet),利用蒸馏关系图来增强特征解析和位置细节的匹配.该方法指出,直接堆积自注意力和交叉注意力模块会破坏特征解析和匹配一致性的纯度.因此,该方法将特征解析和特征匹配过程解耦,利用自注意力机制捕捉支持和查询特征的自对应关系,利用交叉注意力机制进行匹配.同时,它对不同尺度、不同层级的关系图进行蒸馏,以补充高层特征缺失的位置细节信息.为了充分利用大量的未标注样本,Zhang等人[131]提出Image to Pseudo-Episode(IPE),首先使用Spectral Clustering算法为未标注样本生成伪掩码,然后利用图像增强技术获得新的支持-查询图像对帮助模型训练.BCPT[132]提出了一种新的预训练机制来解决合并背景问题,利用底层语义结构将新类与背景解耦,通过聚类方法来探索潜在的语义结构.为了缓解聚类分配的伪标签导致训练过程不稳定的情况,BCPT提出背景挖掘损失和使用基类来指导聚类过程,通过相似性计算,属于新类的聚类中心可以与背景的聚类中心更加分离.

随着计算机视觉领域一些预训练基础模型如CLIP[133]、SAM[134]、Stable Diffusion[135]等模型的横空出世,这些基础模型对小样本分割任务产生了一定的影响.Open AI在2021年1月份提出的CLIP

模型架起了文本和视觉信息之间的桥梁. 作为一个在 4 亿文本-图像对数据集上进行自监督学习预训练得到的多模态基础模型，CLIP 采用对比学习方式，其学习目标是实现图片和文字的匹配，在视觉任务上展现出卓越的零样本识别和泛化能力. 该模型提供了强大的文本和图像编码器，人们借助其预训练模型将其应用于小样本分割任务，并获得了出色的性能. CLIPSeg[136] 能够同时处理文本和图像信息，利用 CLIP 预训练好的文本和视觉编码器提取文本语义和视觉信息，实现多模态语义分割，通过一个架构实现三个任务：引用表达式分割、zero-shot 分割和 one-shot 分割. 在 PhraseCut 数据集[137] 的扩展版本上对模型进行训练后，CLIPSeg 会根据自由文本提示或支持集为查询图像生成二进制分割掩码. 当仅提供文本信息时，CLIPSeg 将描述或类别名称输入到 CLIP 的文本编码器得到语义提示来指导分割；当仅提供图像信息时，它利用 CLIP 的图像编码器对支持图像和掩码进行处理得到视觉提示指导分割. PGMA-Net[138] 则同时利用 CLIP 的文本提取器和视觉编码器，将与类别有关的文本和视觉特征转化为与类无关先验知识，并利用 CLIP 的视觉特征提取器获得支持和查询图像特征来计算它们间的相似度. 此外，它还提出了带有通道丢弃机制的分层解码器（HDCDM），灵活地利用得到的集成初始掩码和低级特征，不依赖于任何类别特定的信息. 借助于强大的多模态预训练基础模型，PGMA-Net 将小样本分割性能提升到了一个新的水平. PI-CLIP[139] 提出利用视觉文本对齐能力取代视觉先验表示，一方面利用 CLIP 的语义对齐能力来定位目标类别，另一方面利用支持和查询高层特征的匹配信息，通过构建注意力图的高阶关系来细化初始先验信息，以获取更可靠的指导并增强模型的泛化能力. DifFSS[34] 利用扩散模型来生成不同的支持样本增加支持集信息，模拟查询图像类别内的变化，如颜色、纹理变化、照明等，从而提升分割性能. SAM 模型是针对图像分割任务设计的预训练基础模型，在自然图像分割任务上获得了优异的性能，并表现出了强大的零样本识别能力. FSS-SAM[140] 利用 SAM 模型来解决轮廓分割不准确的问题. 该方法可以与任何 FSS 方法结合，首先使用 FSS 方法得到的预测掩码来生成提示，并将提示输入 SAM 来产生新的预测掩码. 为了避免 SAM 预测错误的掩码，FSS-SAM 提出了一种预测结果选择（PRS）算法，设置分割 IoU 阈值来选择 SAM 预测掩码或 FSS 预测掩码作为最终输出.

随着语言通用大模型的发展，人们开始研究在视觉领域提出一个通用大模型. 受到 NLP 领域中 in-context learning 的启发，Wang 等人提出了一个视觉通用模型 Painter[141]. Painter 以图像为中心，将任务提示和视觉任务的输出重新定义为图像. 它将视觉任务输出按照该任务类别转化为特定颜色的 RGB 图像，使用图像和目标转换图像一起作为输入，利用掩码图像建模（MIM）任务训练模型. 通过不同的任务提示，即不同的图像-目标对，如图像和深度图、图像和分割掩码等，Painter 可以执行不同的视觉任务，并且具有一定的泛化能力，即使是在训练过程中未出现过的任务也可以实现，在不进行调参时，在七个不同且具有挑战性的视觉任务（包括深度估计、关键点估计、语义分割、全景分割、图像去噪、图像去重和图像增强）获得了极具竞争力的性能. Painter 旨在将不同的视觉任务整合到一个模型中，从而提高模型的复用性和效率. 基于 Painter 的思想，Wang 等人提出了 SegGPT[33]，利用提示（prompt）完成任意目标的分割，专注于解决分割任务. SegGPT 依旧将不同类型的分割数据转换为相同的图像格式，从而使得不同分割任务可以统一到一个通用的模型架构中. 在 Painter 框架中，每个任务的颜色空间是预定义的，这导致模型容易陷入多任务学习的解决方案中. SegGPT 在将目标转换为图像时采用随机着色的方案，使得每个数据样本具有随机的颜色映射，使模型根据上下文完成多样化的任务，而不是依赖特定的颜色. SegGPT 提出 in-context coloring，在相似的语境下对颜色进行重新映射，这样可以避免在预定义的颜色空间中受限制，使得模型学习到更多的上下文信息. Painter 和 SegGPT 利用通用语言模型中 in-context learning 思想，将多个视觉任务整合到一个模型中，实现任务间的知识共享和迁移. 这种通用模型不仅减少了模型的复杂度和存储开销，还提高了模型的灵活性和泛化能力.

5　小样本分割算法性能比较

表 3、表 4 和表 5 综合比较了当前小样本分割算法在 1-way 1-shot 和 1-way 5-shot 设置下的性能，并汇总了它们在 FSS-1000、PASCAL-5i 和 COCO-20i 三个数据集上的结果. 其中，在 PASCAL-5i 和

COCO-20i 数据集上我们列出了本综述中提到的各算法在均交并比(*mIoU*)和前景-背景交并比(*FBIoU*)这两个性能评价指标下的结果;对于 FSS-1000 数据集,由于大多数算法未提供前景背景交并比的结果,我们仅展示了各算法在均交并比这个性能评价指标下的结果.此外,值得注意的是,许多算法使用了在 ImageNet 上预训练的模型进行初始化或直接进行特征提取.在 PASCAL-5i 和 COCO-20i 数据集中,由于某些类别与 ImageNet 中的类别存在重复,这可能对模型性能产生一定的影响.然而,FSS-1000 数据集中的新类别与 ImageNet 数据集完全不同.此外,我们可视化了部分算法的分割结果,如图 10 所示.接下来,我们对当前算法在不同性能评价指标下的发展水平、同一个算法在不同数据集上的表现及不同算法在同一数据集上的性能进行介绍分析.

表 3 小样本分割算法在 FSS-1000 数据集上性能
(RN 表示 ResNet) (单位:%)

方法	分类	1-shot	5-shot
OSLSM(VGG16)[4]	优化分类器	70.3	73.0
VPI(RN101)[89]	单个原型	84.3	87.7
BAM(RN50)[61]	单个原型	79.7	82.2
SSP[84]	多个原型	87.3	88.6
CMNet(RN50)[95]	图结构	82.5	83.8
DAN(RN101)[93]	图结构	85.2	88.1
HSNet(RN101)[31]	相似度	86.5	88.5
DCAMA(Swin-B)[32]	相似度	90.1	90.4
VAT-HM(RN101)[107]	相似度	90.2	90.5
VAT(RN101)[97]	相似度	90.3	90.8
MSI+VAT[108]	相似度	90.6	91.0
PFENet(RN50)[51]	融合原型和匹配	80.8	81.4
PFENet++(RN101)[113]	融合原型和匹配	88.6	89.2
DoG-LSTM(VGG16)[123]	其他方法(频域)	80.8	83.4
CTANet(RN50)[126]	其他方法(轮廓)	83.1	87.9
DACM(RN101)[54]	高斯过程	90.8	91.7
Painter[141]	其他方法(MIM)	61.7	62.3
SegGPT[33]	其他方法(MIM)	85.6	89.3

表 4 小样本分割算法在 PASCAL-5i 数据集上性能 (单位:%)

方法	分类	1-shot						5-shot					
		fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*	fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*
OSLSM(VGG16)[4]	优化分类器	33.6	55.3	40.9	33.5	40.8	61.3	35.9	58.1	42.7	39.1	44.0	61.5
CWT(RN50)[50]	优化分类器	56.3	62.0	59.9	47.2	56.4	—	61.3	68.5	68.5	56.6	63.7	—
RePRI(RN50)[49]	优化分类器	59.8	68.3	62.1	48.5	59.7	—	64.6	71.4	71.1	59.3	66.6	—
FWB(RN101)[43]	优化支持原型	51.3	64.5	56.7	52.2	56.2	—	54.9	67.4	62.2	55.3	59.9	—
SST(RN50)[48]	优化支持特征	54.4	66.4	57.1	52.5	57.6	—	58.6	68.7	63.1	55.3	61.4	—
CRNet(RN50)[59]	单个原型	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.6
SS-PFENet(RN50)[60]	单个原型	58.9	69.9	66.4	57.7	63.2	—	61.4	75.0	70.5	67.7	68.6	—
BAM(RN50)[61]	单个原型	69.0	73.6	67.6	61.1	67.8	79.7	70.6	75.1	70.8	67.2	70.9	82.2
BAM++(RN50)[62]	单个原型	69.5	74.2	69.2	61.5	68.6	79.7	70.8	75.3	73.0	69.0	72.1	82.8
PMMs(RN50)[30]	多个原型	55.2	66.9	52.6	50.7	56.3	—	56.3	67.3	54.5	51.0	57.3	—
SimPropNet[71]	多个原型	54.9	67.3	54.5	52.0	57.2	73.0	57.2	68.5	58.4	56.1	60.0	72.9
DPNet(RN50)[83]	多个原型	60.7	69.5	62.8	58.0	62.7	—	64.7	70.8	69.0	60.1	66.2	—
CMN(RN50)[73]	多个原型	64.3	70.0	57.4	59.4	62.8	72.3	65.8	70.4	57.6	60.8	63.7	72.8
DCP(RN50)[69]	多个原型	63.8	70.5	61.2	55.7	62.8	75.6	67.2	73.2	66.4	64.5	67.8	79.7
NERTNet(RN101)[74]	多个原型	65.5	71.8	59.1	58.3	63.7	75.3	67.9	73.2	60.1	66.8	67.0	78.2
CELP(RN50)[81]	多个原型	67.9	73.0	61.8	58.4	65.3	73.2	69.4	73.7	61.6	60.8	66.4	75.3
HPA(RN101)[91]	多个原型	67.2	73.1	64.3	59.8	66.1	76.6	68.3	75.2	66.4	67.8	69.4	80.4
IPMT(RN50)[78]	多个原型	72.8	73.7	59.2	61.6	66.8	77.1	73.1	74.7	61.6	63.4	68.2	81.4
IPRNet(RN101)[90]	多个原型	67.8	74.6	65.7	62.2	67.5	—	70.0	75.9	71.8	65.8	70.9	—
SAGNN(RN50)[96]	图结构	64.7	69.6	57.0	57.2	62.1	73.2	64.9	70.0	57.0	59.3	62.8	73.3
CMNet(RN50)[95]	图结构	65.4	71.5	55.2	58.1	62.5	73.5	67.0	71.7	55.8	59.9	63.6	74.1
ASNet(RN50)[100]	相似度	68.9	71.7	61.1	62.7	66.1	77.7	72.6	74.3	65.3	67.1	70.8	80.4
HSNet(RN101)[31]	相似度	67.3	72.3	62.0	63.1	66.2	77.6	71.8	74.4	67.0	68.3	70.4	80.6
FECANet(RN50)[102]	相似度	69.2	72.3	62.4	65.7	67.4	78.7	72.9	74.0	65.2	67.8	70.0	80.7
CATrans(Swin-T)[99]	相似度	68.0	73.5	64.9	63.7	67.6	—	75.9	79.1	78.3	75.6	77.3	—
VAT(RN101)[97]	相似度	70.0	72.5	64.8	64.2	67.9	79.6	75.0	75.2	68.4	69.5	72.0	83.2
DCAMA(Swin-B)[32]	相似度	72.2	73.8	64.3	67.1	69.3	78.5	75.7	77.1	72.0	74.8	74.9	82.9
MSI+VAT(RN101)[108]	相似度	73.1	73.9	64.7	68.8	70.1	82.3	73.6	76.1	68.0	71.3	72.2	82.3
TBS+DCAMA(Swin-B)[104]	相似度	74.7	74.4	66.1	69.5	71.2	80.0	76.5	76.5	71.4	76.3	75.2	83.4
FFNet(RN50)[111]	融合原型和匹配	60.3	69.1	56.0	56.3	60.4	71.1	61.7	69.1	57.1	58.3	61.6	72.5
PFENet(RN50)[51]	融合原型和匹配	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
AAFormer(RN101)[120]	融合原型和匹配	69.9	73.6	57.9	59.7	65.3	74.9	75.0	75.1	59.0	63.2	68.1	77.3
MSHNet(RN101)[116]	融合原型和匹配	68.6	73.1	61.3	63.9	66.7	77.7	73.0	76.3	68.1	71.9	72.3	82.7
MIANet(RN50)[112]	融合原型和匹配	68.5	75.8	67.5	63.2	68.7	—	70.2	77.4	70.0	68.8	71.6	—
FPTrans(DeiT-B)[52]	融合原型和匹配	72.3	70.6	68.3	64.1	68.8	—	76.7	79.0	81.0	75.1	78.0	—
MSANet(RN101)[115]	融合原型和匹配	70.8	75.2	67.3	64.3	69.1	80.4	73.8	77.8	73.1	71.2	74.0	84.3
AMFormer(ResNet101)[118]	融合原型和匹配	71.3	76.7	70.7	63.9	70.7	—	74.4	78.5	74.3	67.2	73.6	—

（续　表）

方法	分类	1-shot						5-shot					
		fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*	fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*
Adaptive FSS+DCAMA(Swin-B)[70]	融合原型和匹配	74.3	74.9	70.5	69.8	72.4	—	76.8	78.8	74.3	77.1	76.7	—
DoG-LSTM(RN101)[123]	其他方法(频域)	57.0	67.2	56.1	54.3	58.7	—	57.3	68.5	61.5	56.3	60.9	—
DGPNet(RN101)[55]	其他方法(高斯)	63.9	71.0	63.0	61.4	64.8	—	74.1	77.4	76.7	73.4	75.4	—
QCLNet(RN101)[122]	其他方法(四元数)	67.9	72.5	64.3	63.4	67.0	—	72.5	74.8	68.5	68.9	71.2	—
DPCN(RN50)[128]	其他方法(动态卷积)	65.7	71.6	69.1	60.6	66.7	78.0	70.0	73.2	70.9	65.5	69.9	80.7
FSS-SAM+BAM(RN50)[140]	其他方法 (SAM)	70.2	74.0	67.9	62.0	68.5	80.2	71.9	75.3	71.1	68.3	71.6	82.7
BAM+SVF(RN50)[127]	微调主干	69.4	74.5	68.8	63.1	69.0	80.1	72.1	76.2	72.0	68.9	72.3	83.2
DACM(RN101)[54]	其他方法(相似度)	69.9	74.1	66.2	66.0	69.1	79.4	74.2	76.4	71.1	71.6	73.3	83.1
MuSH(DeiT-B)[129]	特征提取器	71.2	71.5	67.0	66.6	69.1	—	75.9	77.8	78.6	74.7	76.7	—
HDMNet(RN50)[130]	其他方法(蒸馏)	71.0	75.4	68.9	62.1	69.4	—	71.3	76.2	71.3	68.5	71.8	—
PI-CLIP+HDMNet(RN50)[139]	其他方法(文本)	76.4	83.5	74.7	72.8	76.8	—	76.7	83.8	75.2	73.2	77.2	—
PGMA-Net(CLIP-RN101)[138]	其他方法(文本)	76.8	82.3	75.7	75.7	77.6	86.2	77.7	82.7	76.9	77.0	78.6	86.9
SegGPT[33]	其他方法(MIM)	—	—	—	—	83.2	—	—	—	—	—	89.8	—

表 5　小样本分割算法在 COCO-20i 数据集上性能　（单位：%）

方法	分类	1-shot						5-shot					
		fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*	fold0	fold1	fold2	fold3	*mIoU*	*FBIoU*
FWB(RN101)[43]	优化支持原型	17.0	18.0	21.0	28.9	21.1	—	19.1	21.5	23.9	30.1	21.4	—
CWT(RN50)[50]	优化分类器	32.2	36.0	31.6	31.6	32.9	—	40.1	43.8	39.0	42.4	41.3	—
RePRI(RN50)[49]	优化分类器	31.2	38.1	33.3	33.0	34.0	—	38.5	46.2	40.0	43.6	42.1	—
SS-PFENet(VGG16)[60]	单个原型	35.6	39.2	37.6	37.3	37.5	—	40.4	45.8	40.3	40.7	41.8	—
BAM(RN50)[61]	单个原型	43.4	50.6	47.5	43.4	46.2	—	49.3	54.2	51.6	49.6	51.2	—
BAM++(RN50)[62]	单个原型	44.4	52.0	47.0	45.2	47.2	—	52.5	57.0	51.0	49.5	52.5	—
QGNet(RN50)[77]	多个原型	36.7	41.4	38.7	36.6	38.3	62.4	41.5	48.1	46.3	43.6	43.6	66.2
CELP(RN101)[81]	多个原型	36.9	40.9	39.2	39.1	39.0	61.4	41.4	42.9	43.3	43.5	42.8	62.7
NERTNet(RN101)[74]	多个原型	38.3	40.4	39.5	38.1	39.1	67.5	42.3	44.4	44.2	41.7	43.2	69.6
CMN(RN50)[73]	多个原型	37.9	44.8	38.7	35.6	39.3	61.7	42.0	50.5	41.0	38.9	43.1	63.3
SSP(RN101)[84]	多个原型	39.1	45.1	42.7	41.2	42.0	—	47.4	54.5	50.4	49.6	50.2	—
IPMT(RN50)[78]	多个原型	41.4	45.1	45.6	40.0	43.0	—	43.5	49.7	48.7	47.9	47.5	—
IPRNet(RN101)[90]	多个原型	42.9	50.6	46.8	47.4	46.9	—	50.7	58.3	52.8	51.3	53.3	—
HPA(RN101)[91]	多个原型	43.2	50.5	45.5	46.3	46.9	68.8	49.4	58.4	52.5	50.9	52.8	74.4
MENUA+HDMNet(RN50)[92]	多个原型	—	—	—	—	50.9	—	—	—	—	—	57.1	—
CMNet(RN50)[95]	图结构	48.7	33.3	26.8	31.2	35.0	—	49.5	35.6	31.8	33.1	37.5	—
SAGNN(RN101)[96]	图结构	36.1	41.0	38.2	33.5	37.2	60.9	40.9	48.3	42.6	38.9	42.7	63.4
HSNet(RN101)[31]	相似度	37.2	44.1	42.4	41.3	41.2	69.1	45.9	53.0	51.8	47.1	49.5	72.4
VAT(RN50)[97]	相似度	39.0	43.8	42.6	39.7	41.3	68.8	44.1	51.1	50.2	46.1	47.9	72.4
FECANet(RN50)[102]	相似度	38.5	44.6	42.6	40.7	41.6	69.6	44.6	51.5	48.4	45.8	47.6	71.1
HSNet-HM(RN101)[107]	相似度	41.2	50.0	48.8	45.9	46.5	71.5	46.5	55.2	51.8	48.9	50.6	72.9
CATrans(Swin-T)[99]	相似度	47.9	52.3	49.2	48.0	49.4	—	59.3	64.1	59.6	57.3	60.1	—
MSI+VAT(RN101)[108]	相似度	44.8	54.2	52.3	48.0	49.8	—	49.3	58.0	56.1	52.7	54.0	—
DCAMA(Swin-B)[32]	相似度	49.5	52.7	52.8	48.7	50.9	73.2	55.4	60.3	59.9	57.5	58.3	76.9
TBS+DCAMA(Swin-B)[104]	相似度	49.6	54.3	54.1	51.3	52.3	74.2	53.7	62.9	59.3	58.2	58.5	76.8
AAFormer(RN50)[120]	融合原型和匹配	39.8	44.6	40.6	41.4	41.6	67.7	42.9	50.1	45.5	49.2	46.9	68.2
PFENet++(RN50)[113]	融合原型和匹配	40.9	46.0	42.3	40.1	42.3	65.7	47.5	53.3	47.3	46.4	48.6	70.3
MSHNet(RN101)[116]	融合原型和匹配	41.3	49.9	46.4	46.6	46.1	—	51.7	62.6	55.9	53.8	56.0	—
ProtoFormer(RN101)[110]	融合原型和匹配	42.9	50.7	48.4	45.8	47.0	70.0	49.6	59.7	56.4	53.0	54.7	74.6
FPTrans(DeiT-B)[52]	融合原型和匹配	44.4	48.9	50.6	44.0	47.0	—	54.2	62.5	61.3	57.6	58.9	—
MIANet(RN50)[112]	融合原型和匹配	42.5	53.0	47.8	47.4	47.7	—	45.8	58.2	51.3	51.9	51.7	—
AMFormer(RN101)[118]	融合原型和匹配	44.9	55.8	52.7	50.6	51.0	—	52.0	61.9	57.4	57.9	57.3	—
MSANet(RN101)[115]	融合原型和匹配	47.8	57.4	48.7	50.5	51.1	—	53.2	62.3	55.4	56.3	56.8	—
Adaptive+DCAMA(Swin-B)[70]	融合原型和匹配	51.4	53.1	54.4	51.9	52.7	—	57.8	61.0	60.9	58.5	59.6	—
DPCN(RN50)[128]	动态卷积	42.0	47.0	43.2	39.7	43.0	63.2	46.0	54.9	50.8	47.4	49.8	67.4
MMFormer(RN50)[56]	其他方法(解耦)	40.5	47.7	45.2	43.3	44.2	—	44.0	52.4	47.4	50.0	48.4	—
FSS-SAM+BAM(RN50)[140]	其他方法(SAM)	39.1	50.4	48.4	43.1	45.3	—	47.3	54.1	48.2	47.1	49.2	—
QCLNet(RN101)[122]	四元数	40.0	45.5	45.1	43.6	43.6	69.9	46.9	55.8	53.6	51.1	51.9	73.5
MuSH(DeiT-B)[129]	特征提取器	44.0	50.0	49.1	46.3	47.4	—	53.6	60.5	57.3	55.2	56.7	—
BAM+SVF(RNe0)[127]	微调主干	46.9	53.8	48.4	44.8	48.5	—	52.3	57.8	52.0	53.4	53.9	—
HDMNet(RN50)[130]	其他方法(蒸馏)	43.8	55.3	51.6	49.4	50.0	72.2	50.6	61.6	55.7	56.0	56.0	77.7
PGMA-Net(CLIP-RN101)[138]	其他方法(文本)	55.2	62.7	60.3	59.4	59.4	78.5	55.9	65.9	63.4	61.9	61.8	79.4
PI-CLIP+HDMNet(RN50)[139]	其他方法(文本)	49.3	65.7	55.8	56.3	56.8	—	56.4	66.2	55.9	58.0	59.1	—
SegGPT[33]	其他方法(MIM)	—	—	—	—	56.1	—	—	—	—	—	67.9	—

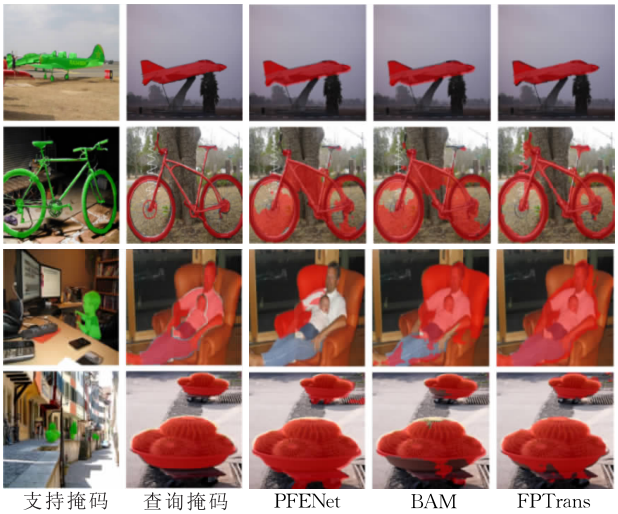


图 10 小样本分割部分算法分割结果可视化

5.1 小样本分割算法在不同性能评价指标下的发展现状

在本小节,我们在均交并比和前景-背景交并比性能评价指标下来介绍当前小样本分割算法的发展水平.

在均交并比($mIoU$)性能评价指标下,我们分析各个算法在 FSS-1000、PASCAL-5i 和 COCO-20i 三个数据集上分别在 1-shot 和 5-shot 设置下的性能比较. 在 FSS-1000 数据集上,目前在 1-shot 和 5-shot 设置下获得最佳性能的算法是 DACM,其性能分别达到了 90.8%和 91.7%. 该方法将小样本分割过程建模为高斯过程,通过核函数计算支持和查询图像之间的相似度关系来度量性能. 由于 FSS-1000 数据集中图像包含的分割目标单一,背景相对简单,因此即使是简单的分割网络也能取得不错的性能,如最初的 OSLSM 方法在 1-shot 和 5-shot 设置下其均交并比分别获得了 70.3%和 73.0%. 在 PASCAL-5i 数据集上,当前在 1-shot 和 5-shot 设置下获得最佳性能的算法均是今年北京智源人工智能研究院推出的通用分割大模型 SegGPT,其 1-shot 性能达到了 83.2%,5-shot 性能为 89.8%,小样本分割算法的性能获得大幅度提升. 在 1-shot 设置下,与之前传统小样本分割方法中获得最佳性能的 Adaptive FSS 方法相比,SegGPT 性能提升了 10.8%;在 5-shot 设置下,与之前方法中获得最佳性能的 FPTrans 相比,其性能提升了 11.8%. SegGPT 采用了 in-context learning 和提示学习的思想,通过建模掩码图像和随机着色方案来训练模型,旨在解决各种分割任务,并取得了出色的性能. 在 COCO-20i 数据集上,不同设置下的获得最佳性能的算法是不同的. 在 1-shot 设置下获得最佳性能的算法是 PGMA-Net,其均交并比性能为 59.4%,与未借助于预训练基础模型的最佳算法 Adaptive FSS 相比,其性能获得了 6.7%的提升;在 5-shot 设置下获得最佳性能的算法是 SegGPT,其均交并比性能为 67.9%,与之前传统小样本分割方法中的最佳算法 CATrans 方法相比,性能提升了 7.8%. 通过数据可以看出,预训练基础模型和通用大模型的出现使得小样本分割算法性能得到了很大提升,这对传统小样本分割任务产生了一定的冲击.

在前景-背景交并比($FBIoU$)性能评价指标下,我们只在 PASCAL-5i 和 COCO-20i 这两个数据集上分析提供该项指标的各算法的 1-shot 和 5-shot 性能比较. 借助 CLIP 预训练模型的 PGMA-Net 方法,在这两个数据集上均获得了最佳性能. 在 PASCAL-5i 数据集上,其 1-shot 性能为 86.2%,5-shot 为 86.9%. 在 COCO-20i 数据集上,其 1-shot 性能为 78.5%, 5-shot 是 79.4%.

5.2 小样本分割算法在不同数据集上的表现

我们以基于相似度计算的匹配方法 MSI 为例,比较它在三个数据集上的均交并比($mIoU$)性能. MSI 方法使用掩码与支持图像乘积以及原始图像获得的支持特征分别与查询特征计算余弦相似度,该思想可以用在所有基于相似度的匹配方法中. 本文将该方法应用在 HSNet 和 VAT 框架下进行实验验证,当将该方法应用在 VAT 架构上时,在三个数据集上均表现出了优异性能. 我们以 MSI 在 VAT 框架下的性能为例,说明同一算法在不同数据集下的性能表现.

在 FSS-1000 数据集上,该方法在 1-shot 和 5-shot 设置下的性能分别达到了 90.6%和 91.0%,相较于 VAT 方法性能分别提升了 0.3%和 0.2%;在 PASCAL-5i 数据集上,该方法分别达到了 70.1%和 72.2%,较 VAT 性能分别提升了 2.2%和 0.2%;在 COCO-20i 数据集上,该方法分别达到了 49.8%和 54.0%,性能较于 VAT 提升了 8.5%和 6.1%. 该方法在三个数据集上比较时可以看出,在 FSS-1000 数据集上获得了最好的结果,但由于该数据集性能本来就很好,所以提升幅度相对较小;而在 COCO-20i 数据集上的性能最低,表现出了很大提升空间;在 PASCAL-5i 数据集上的性能处于中间结果,获得了一定的性能提升.

此外,通过观察表中数据可以看出,各个算法在三个数据集上都表现出了相同的趋势,在 FSS-1000 数据集上获得最好的性能,正如前文提到的,FSS-1000 数据集每张图像只包含一个待分割目标,而且

背景相对简单，因此分割难度较小；在 COCO-20i 数据集上获得了最低的性能，是因为 COCO-20i 数据集中的每张图像都包含多个分割类别，分割目标较多，背景复杂，相对而言分割难度较大；而 PASCAL-5i 数据集性能处于中间，分割难度相对适中.

5.3　小样本分割算法在 PASCAL-5i 数据集上的表现

接下来我们以 PSACAL-5i 数据集为例，在我们的分类设置下来比较不同类型的算法性能.

基于优化的小样本分割方法中获得最佳性能的算法是 RePRI，其 1-shot 和 5-shot 均交并比（$mIoU$）性能分别达到了 59.7%和 66.6%. 基于优化的方法在推理过程中需要利用少量标注样本微调模型参数，极易出现过拟合现象，因此对模型结构和优化算法要求更高. RePRI 放弃设计复杂的模型结构来避免过拟合，针对推理方式进行改进，推理时采用 Transductive 方式，并放弃了元学习的训练方式.

从表 4 可以看出，该方法仍与基于度量学习中的大部分算法存在较大差距，这是由于基于优化的方法在推理时优化模型极易出现过拟合现象导致其分割性能受到影响. 而基于度量学习的方法推理时无需重新优化模型，旨在训练过程中学习到更多与类无关的知识实现新类泛化，在推理过程中表现出了优异的性能，从算法数目可以看出后续研究主要集中在该方向.

在基于度量学习的方法中我们又对其进行进一步的分类，分成了五部分，每个部分算法都有不同的特点. 第 1 部分是基于单个原型的方法，在该方法中目前性能最佳的算法是 BAM++，其 1-shot 和 5-shot 均交并比（$mIoU$）性能分别达到了 68.6%和 72.1%，其前景-背景交并比（$FBIoU$）分别是 79.7%和 82.8%. 它利用一个预训练好的基类分支来消除基类对预测掩码的影响，并进一步增加了多尺度信息交互. 在基于单个原型的方法中，BAM 和 BAM++算法使得该类方法性能得到较大幅度提升. 此外，这两个算法在训练过程中对训练样本进行筛选，剔除包含测试类别目标的训练图像，使用剩余样本进行训练，这也是导致性能大幅提升的原因之一，尤其是在 fold2 和 fold3 数据集上.

第 2 部分是基于多个原型的方法，该类方法中获得最佳性能的算法是 IPRNet，其 1-shot 和 5-shot 均交并比性能分别达到了 67.5%和 70.9%. 通过表 4 数据可以看出，与同阶段基于单个原型的方法相比，有一定的程度的性能提升，如 2020 年提出的 PMMs、SimPropNet 等算法均比同年利用单个原型的 CRNet 方法 1-shot 性能要好. 基于多个原型的方法能够很好地利用图像信息，采用不同的手段生成多个原型，实现分割性能提升，一定程度上缓解局部结构丢失的问题.

第 3 部分是基于匹配的方法，从表 4 数据可以看出，除基于图结构建模的匹配方法外，基于相似度计算的匹配方法性能普遍要比同阶段基于原型的方法性能要好，如 2021 年提出的 HSNet 在 1-shot 设置下性能达到 66.2%，已经超过大部分基于原型的分割方法，无论是单个原型还是多个原型. 在该类方法中当前获得最佳性能的算法是 TBS，其 1-shot 和 5-shot 均交并比性能分别达到了 71.2%和 75.2%.

第 4 部分是融合原型和匹配的小样本分割方法，在该类方法中，不同设置下的获得最佳性能的算法是不同的. 在 1-shot 设置下获得最佳均交并比性能的方法是 Adaptive FSS，其性能为 72.4%，该方法使用新类样本微调部分模型参数并保存其类原型用于增强特征；在 5-shot 设置下是 FPTrans 方法，其均交并比性能为 78.0%.

第 5 部分是一些其他方法，它们从不同的角度出发解决小样本分割问题. 除提出的通用模型 SegGPT 外，获得最佳性能的算法是 PGMA-Net，其借助强大的多模态预训练基础模型 CLIP 和文本标签信息，在 1-shot 和 5-shot 设置下均交并比（$mIoU$）性能分别达到了 77.6%和 78.6%，前景-背景交并比（$FBIoU$）分别是 86.2%和 86.9%. 在未借助预训练基础模型的小样本分割方法中，在 1-shot 设置下获得最佳性能的是 HDMNet 方法，获得了 69.4%均交并比性能，它利用蒸馏的思想来补充高层特征缺失的位置细节信息；在 5-shot 设置下获得最佳性能的是 MuSH，其均交并比性能为 76.7%，该算法从特征提取器设计的角度出发，将注意力从分割头转移到特征提取阶段.

5.4　小　结

总体而言，根据不同算法在三个数据集上的结果表明，基于度量学习的小样本分割方法的性能普遍优于基于优化的方法. 基于优化的方法在推理时需要使用少量样本进行模型微调，容易出现过拟合现象，对模型结构和优化算法要求更高. 相比之下，基于度量学习的算法能够有效避免这个问题，推理时无需对模型进行优化，直接使用训练好的模型进行推理. 因此，训练阶段如何建立一个与类别无关的模型成为提高模型泛化能力的关键. 在基于度量学习的各种方法中，基于匹配的方法普遍优于基于原型的方法，融合原型和匹配的方法通过合理的设计会使得性能得到进一步提升. 此外，借助于预训练基础模型的方法能够使得性能获得较大提升. 值得注意的是，目前大多数算法在性能比较时主要采用均交并比作为性能评价指标. 由于图像中的背景类别

通常占据大部分区域,均交并比只考虑前景部分的交并比,对于类别不平衡具有一定的稳定性.而前景-背景交并比的结果对类别平衡性非常敏感,如果某个类别的样本数量远远超过其他类别,那么这个类别的得分可能会主导总体得分,从而掩盖其他类别的性能.从表格结果来看,各方法的前景-背景交并比相对于均交并比都获得了较好的结果.

6 小样本分割任务与相关任务的关系

小样本分割任务是语义分割任务的扩展,旨在减少对大规模密集标注数据的需求,并具有识别新类的泛化能力.除小样本分割外,还涌现出许多相关任务,包括小样本实例分割、广义小样本分割、增量小样本分割、弱监督小样本分割和跨域小样本分割任务.下面将对这些任务及其挑战进行介绍.

6.1 小样本实例分割

小样本实例分割任务[142-145]是将小样本学习思想应用于实例分割,旨在利用过去学到的知识实现新类泛化,将待分割图像中属于同一类别的不同实例区分开.即给定一个具有掩码和边界框标注的支持图像,我们需要得到查询图像中属于支持集类别的所有目标实例类别、掩码和边界框信息.在该任务中,测试类可以是新类,或者同时包含新类和训练类.当前大部分小样本实例分割框架延续了传统实例分割任务中常用的两阶段方法,第 1 阶段生成目标候选的掩码,第 2 阶段对候选掩码进行分类.该任务相较于小样本分割任务而言更具挑战性,它不仅要识别类别,还要对每个实例进行准确分割.

6.2 广义小样本分割

在小样本分割任务中我们通常采用 episode 测试模型,也就是测试前需要对提供的标注图像和未标注图像进行分类,这需要大量的先验知识才能实现.此外,当前小样本分割在设置中只对新类进行分割,无法很好地解决同时分割基类和新类的问题.为此,Tian 等人[146]提出广义小样本分割,不需要预先知道查询样本中包含的类别信息即可进行分割.在该任务中,首先在一个大规模的基类数据集上训练模型,然后利用少量新类标注样本来优化模型,以便模型能够同时分割出基类和新类目标.该任务的关键在于如何在保证基类分割性能的情况下,利用少量的新类标注样本实现新类泛化.相较于 FSS,这种任务设置更加常见实用,可以有效地解决实际应用中遇到的基类和新类分割的问题.

6.3 增量小样本分割

在实际应用中,我们希望模型在不断学习新知识的同时,不会遗忘过去学习的知识.为此,Cermelli 等人[147]提出增量小样本分割任务,旨在利用少量新类样本来不断扩充模型,使其能够分割新类目标,同时在不能获取旧数据的情况下,具有保留、整合和优化旧知识的能力.传统的增量学习存在灾难性遗忘,即在新数据集上进行训练时,旧数据集上的性能会急剧下降,这种现象在小样本设置下仍旧存在.此外,在仅提供少量新类标记样本的情况下,直接微调模型容易导致模型的过拟合.因此,该任务关键在于如何避免灾难性遗忘和解决新类过拟合问题,并在小样本设置下实现新旧知识的整合,获得不断学习的能力.

6.4 弱监督小样本分割

小样本分割任务仍需要少量有密集标注的样本实现新类泛化,为进一步降低标记需求,Lee 等人[148]提出弱监督小样本分割任务.弱监督小样本分割任务旨在使用少量具有图像级语义标签、涂鸦、点或边界框等标注的样本训练模型,实现新类分割,并降低标注难度.常见的弱监督信号有四种:图像级标注、点标注、目标框标注和涂鸦标注.在实际的弱监督小样本分割任务中,通常会同时使用多个弱监督标注提高分割性能.在训练和测试的时候都只有弱监督标签作为指导,相较于掩码标注,弱监督标注包含更多的噪声,无法像掩码信息一样提供完整准确的类别知识.另外,由于提供的训练样本数量有限,这使得任务更具挑战性,其关键在于如何充分利用这些弱监督信号提取准确可靠的类别信息实现分割.

6.5 跨域小样本分割

当前小样本分割方法都是在源域和目标域是相同的前提下进行的,这对于很难收集足够的数据进行训练的领域来说是不现实的,对此,Lei 等人[149]提出跨域小样本分割任务,将在具有大量标注训练数据的域上学习到的知识迁移到数据稀缺的领域,在新域上进行分割.该任务除了域不同外,训练和测试类别也是不同的.跨域小样本分割任务是一项具有挑战性和研究潜力的任务.当目标域和源域数据分布差距过大时,源域训练的模型便不能很好地泛化到目标域,如何解决跨域时目标域性能不理想的问题是解决跨域小样本分割的关键.

7 研究趋势展望

随着视觉基础模型的飞速发展和进步,语义分

割任务也进入了一个新的时代. 2023 年 4 月，Meta 提出了 SAM 视觉基础模型，其优异的分割性能和强大的零样本分割能力对计算机视觉领域产生了极大的冲击. 与传统的图像分割方法不同，SAM 是一种交互式分割方法，需要人类在待分割图像上进行交互来提供目标类别信息才能获得理想的分割结果. 根据分割时是否需要人类参与可以将图像分割任务分为交互式和非交互式两类，如图 11 所示. 交互式分割方法以 SAM 为代表，需要在待分割图像上提供交互信息实现分割. 交互方式可以有多种形式，从人工标注得到待分割图像的掩码到给定目标的不完整掩码、目标框、前景和背景点、涂鸦和文本描述或标签. 图中从左到右的图片展示了交互程度逐渐减弱的情况. 需要说明的是 SAM 也可以实现非交互式的分割任务，即“Automatic”模式，但在这种设置下，SAM 会将图像中所有目标都分割出来.

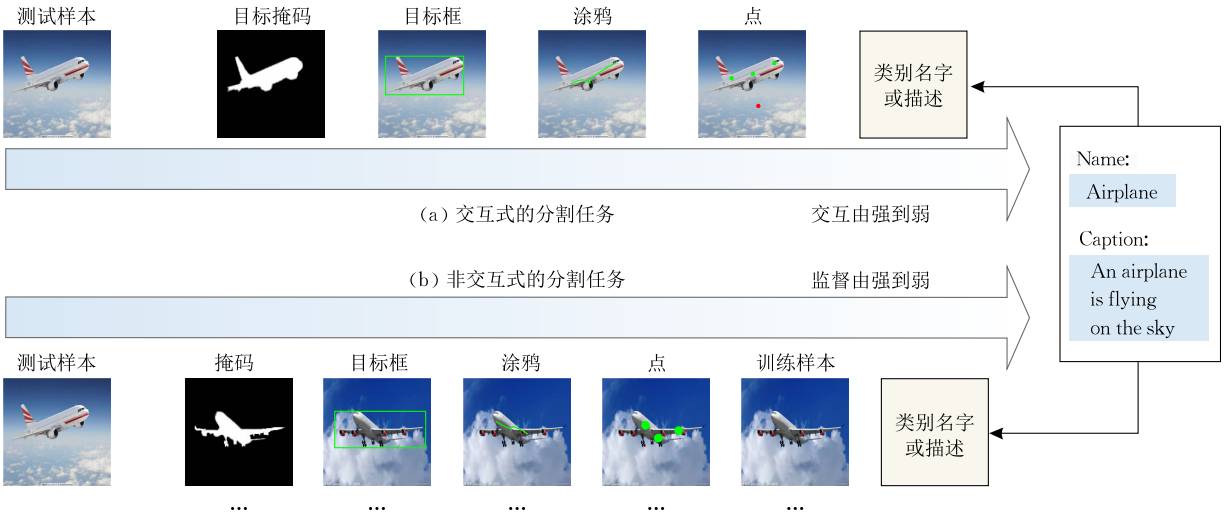


图 11　交互式分割和非交互式分割

非交互式分割方法不需要人工交互，能够自动实现分割，如传统分割、弱监督分割和小样本分割任务等，利用从其他图像中学到的知识来生成当前图像的分割掩码. 非交互式分割方法根据提供的样本数目又可以分为多样本和小样本设置. 在传统的语义分割任务中，提供大量同类图像-掩码对训练模型，可以分割出训练过程中出现过的类别，但不能识别训练过程中未出现过的类别. 为了减轻标注负担，人们提出弱监督语义分割，将密集标注掩码替换为目标框、涂鸦或点，在该设置下模型依然不能识别新类. 为进一步减少对样本数量的需求并提高模型泛化能力，人们提出小样本分割，仅利用少量标注实现新类分割. 弱监督小样本分割任务则是在小样本设置下，将掩码标注替换为弱监督信息，在极端情况下可以仅利用未标记的同类图像进行分割. 当仅利用类标签进行分割时，可以称之为零样本分割. 其中文本描述或标签这种监督既可以用于交互式分割，也可以用于非交互式分割，将这两种分割任务串联起来.

尽管小样本分割任务已经取得很大的进展，但仍存在许多问题和挑战. 本文总结当前小样本分割任务的发展现状，并对该领域进行展望：

(1) 小样本分割任务面临的挑战并没有得到很好解决.

① 样本稀缺的问题. 针对该问题我们可以从以下几点出发，来提升分割性能：(i) 数据生成. 相对于收集样本，直接生成样本更加简洁高效. 近年来图像生成模型取得了显著进展，例如 DALL-E2[150]、Stable Diffusion[135]、Imagen[151] 等，这些 AI 生成模型能够根据输入文本生成逼真的图像. 当前已有部分工作[34,152-153]进行相关探索. 探索如何利用这些生成模型生成大量具有类内差异性的图像训练小样本分割任务是十分有意义的，可以缓解样本收集和标注困难的问题；(ii) 主动学习与增量学习. 通过主动选择最具代表性和困难的样本进行标记，从而最大程度地利用有限的标记数据. 增量学习[147,154-155]可以持续学习样本知识，逐渐扩展样本空间，提高模型的性能. 这些方法可以在有限的标记数据下进行高效的训练；(iii) 探索无监督或弱监督学习：无监督学习可以利用未标记数据进行分割，通过自动生成目标分割掩码，并用其训练模型. 弱监督学习[148,156-157]利用弱标记数据学习实现分割，通过合理的设计可以获得与有监督方法相似的性能.

② 类内差异大的问题. 当前小样本分割算法在处理遮挡、干扰、多目标、小目标以及光照和视角等情况下的分割时仍面临较大的挑战，其分割实例如图 12 所示. 我们可以从以下几点出发解决该问题：

(i)特征提取.解决该问题的关键之一是获得更好的图像特征,可以选取合适的预训练模型作为特征提取器进行研究探索.当前的小样本分割任务主要使用分类任务预训练的主干网络来提取图像特征,然而,由于分类和分割关注图像的差异,导致分割性能受到影响.已经有些方法针对该问题来修正特征提取器,但他们使用当前任务的训练集训练模型,这样得到的特征提取器可能会对基类产生偏好,进而影响分割性能.目前有许多预训练好的主干网络可以更好地建模图像表示,例如 MAE[158]、扩散模型[159]、视觉基础模型 SAM[134]等,这些模型没有依赖训练数据,能够更加关注像素特征,获得更好的图像特征;(ii)语义等额外信息.探索融合语义标签信息等额外信息提升分割性能.相比于图像掩码,语义标签是相对容易获取的,同时具有高度概括的类别知识,能够提供部分利于分割的有用信息.同时,当前的自然语言模型取得了巨大成功,能够提供准确的语义嵌入帮助模型分割.

图 12 小样本分割中存在的问题

(2)当前的小样本分割任务研究主要集中在自然图像领域,而在实际应用领域如医疗图像、缺陷检测、遥感图像等,研究相对较少.

对于实际应用而言,数据收集和标注也是一个巨大的挑战,有些领域可能需要较强的专业知识才能进行标注.例如,在医疗图像等特殊领域,尽管小样本分割任务获得了一定的发展[160-163],但由于病人隐私和标注需要专业知识等原因,难以获得大规模有标记的数据.因此训练一个具体领域的垂类分割模型是十分困难的,要针对实际应用领域的小样本分割问题进行更深入的研究和探索.除此之外,实际应用中的数据图像与当前小样本分割任务中使用的自然图像可能存在较大差异,因此,当前的小样本分割模型只能用于学术研究,并不能直接部署到实际应用中.针对实际应用领域小样本分割任务中存在的数据稀缺及特殊的数据形态问题,可以考虑以下方向:

①迁移学习与领域自适应.利用已有大规模标记自然图像或其他领域数据训练模型,然后将其迁移到实际应用领域[164-167].迁移学习和领域自适应方法可以将模型从其他领域学习到的知识和特征,迁移到目标域数据上,从而降低数据需求,提高分割性能.

②自然图像领域中解决数据稀缺的方法可以扩展到特殊领域,如探索主动学习与增量学习、无监督或弱监督学习等方法.

(3)在交互式分割任务中,当前的模型并不能够完全解决当前小样本分割任务中存在的问题.

尽管 SAM 模型在许多图像数据集上展示了强大的分割能力和零样本识别能力,但仍存在一些局限性.

①SAM 模型需要人为交互才能获得理想的分割掩码,当采用非交互方式分割时,SAM 生成的掩码过于细粒化,不能得到用户想要的目标物体掩码.此外,在一些特殊领域,如医疗图像分割,需要专业知识和经验才能准确进行分割,一般的用户往往不具备这样的专业知识.因此,我们需要进一步研究和提出新的方法,使得这些视觉基础模型可以充分发挥其优势,实现准确且便捷的分割,而无需依赖人为交互.无论是自然图像还是特殊领域,均可以利用小样本分割任务中支持集来提供交互信息,如何利用支持集信息获得可靠的提示输入到 SAM 模型实现自动分割是一个值得探索的问题.

②SAM 提供了一种提示学习的思想.SAM 通过输入提示获得分割目标信息,改变提示可以获得不同的分割结果.小样本分割任务中的支持集也可以看作是一种提示信息,只不过提示信息来自于属于同类别的其他图像,而 SAM 模型的提示关于当

前分割图像信息. 因此，在小样本分割任务中，我们可以考虑利用提示学习作为信息补充，该提示可以是任务描述、支持集等来自于非分割图像本身的信息，或者是属于分割图像自身的语义标签等额外可用信息. 此外，提示学习可以作为微调大模型的一种方式，使其可以适应于特定的任务，微调 CLIP 或其他多模态基础模型，发挥其优势，使其适用于小样本分割任务.

③SAM 虽然在自然图像分割领域表现出了出色的性能，但在特殊领域中仍面临一些挑战. 当应用数据与训练数据具有较大差异时，大模型性能可能会受到影响. 一些实验[168-169]已经证明，SAM 模型并不是在所有类型的图像上都表现出色，它在应对与自然图像有较大差异的图像时表现不佳，例如医疗图像、遥感图像、红外图像和低对比度图像等特殊领域，需要进一步微调或优化. 因此，如何利用现有的视觉大模型，并通过少量样本的微调使其适应于特殊领域，值得进一步探索. 关于该问题，可以使用提示学习、Adapter 等方法在微调少量参数的情况下实现模型到特殊领域的适应.

总之，尽管小样本语义分割在自然图像分割中表现出了强大的性能，但仍然面临一些挑战. 需要从交互性、模型整合和适应性等方面着手进行研究，以进一步提高分割的准确性和应用性.

8　结　论

从 2017 年到今，小样本语义分割任务引起了学术界和工业界的广泛关注，并在短短几年涌现出了大量的算法，取得了许多出色的研究成果. 小样本语义分割任务可以在仅提供少量有标记样本的前提下实现对训练过程中未出现过的新类的分割，从而大大减少了对大规模密集标注数据的需求，同时具备较好的泛化能力. 本文较为全面地调研、归纳和分析了小样本语义分割任务的国内外研究现状，并对现有的小样本语义分割方法进行分类，讨论了今后的发展方向，希望能够对相关领域的研究人员提供有益的帮助.

参　考　文　献

[1] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6881-6890

[2] Strudel R, Pinel R G, Laptev I, et al. Segmenter: Transformer for semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 7242-7252

[3] Li L, Zhou T, Wang W, et al. Deep hierarchical semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Kuala Lumpur, Malaysia, 2022: 1246-1257

[4] Shaban A, Bansal S, Liu Z, et al. One-shot learning for semantic segmentation//Proceedings of the British Machine Vision Conference. London, UK, 2017: 1-13

[5] Hosna A, Merry E, Gyalmo J, et al. Transfer learning: A friendly introduction. Journal of Big Data, 2022, 9(1): 1-19

[6] Hospedales T, Antoniou A, Micaelli P, et al. Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169

[7] Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 4037-4058

[8] Chen Qiong, Yang Yong, Huang Tian-Lin, et al. A survey on few-shot image semantic segmentation. Frontiers of Data and Computing, 2021, 3(6): 17-34(in Chinese)
(陈琼，杨咏，黄天林等. 小样本图像语义分割综述. 数据与计算发展前沿，2021，3(6)：17-34)

[9] Wei Ting, Li Xin-Lei, Liu Hui. Survey on image semantic segmentation in dilemma of few-shot. Computer Engineering and Applications, 2023, 59(2): 1-11(in Chinese)
(韦婷，李馨蕾，刘慧. 小样本困境下的图像语义分割综述. 计算机工程与应用，2023，59(2)：1-11)

[10] Ren W, Tang Y, Sun Q, et al. Visual semantic segmentation based on few/zero-shot learning: An overview. IEEE/CAA Journal of Automatica Sinica, 2024, 11(5): 1-21

[11] Zhang J, Yang P, Wang W, et al. Image editing via segmentation guided self-attention network. IEEE Signal Processing Letters, 2020, 27: 1605-1609

[12] Ling H, Kreis K, Li D, et al. EditGAN: High-precision semantic image editing//Proceedings of the Annual Conference on Neural Information Processing Systems. Virtual, 2021: 16331-16345

[13] Luo Z, Kang H, Yao P, et al. Chinese image caption based on deep learning//Proceedings of the International Conference on Audio, Language and Image Processing. Shanghai, China, 2018: 216-220

[14] Yu T, Feng R, Feng R, et al. Inpaint anything: Segment anything meets image inpainting. CoRR abs/2304.06790, 2023

[15] Sharma V, Bishnu A, Patel L. Segmentation guided attention networks for visual question answering//Proceedings of ACL, Student Research Workshop. Vancouver, Canada, 2017: 43-48

[16] Pham V Q, Mishima N, Nakasu T. Improving visual question answering by semantic segmentation//Proceedings of the International Conference on Artificial Neural Networks. Bratislava, Slovakia, 2021: 459-470

[17] Liu Y, Zhang Y, Wang Y, et al. Rethinking transformers for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15

[18] Li K, Cao X, Meng D. A new learning paradigm for foundation model-based remote-sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12

[19] Ye Y, Yang K, Xiang K, et al. Universal semantic segmentation for fisheye urban driving images//Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC). Toronto, Canada, 2020: 648-655

[20] Li J, Jiang F, Yang J, et al. Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing, 2021, 465: 15-25

[21] Armand T P T, Bhattacharjee S, Choi H K, et al. Transformers effectiveness in medical image segmentation: A comparative analysis of UNet-based architectures//Proceedings of the International Conference on Artificial Intelligence in Information and Communication. Osaka, Japan, 2024: 238-242

[22] Valanarasu J M J, Oza P, Hacihaliloglu I, et al. Medical transformer: Gated axial-attention for medical image segmentation//Proceedings of the International Conference of Medical Image Computing and Computer Assisted Intervention. Strasbourg, France, 2021: 36-46

[23] Fawakherji M, Youssef A, Bloisi D, et al. Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation//Proceedings of the IEEE International Conference on Robotic Computing. Naples, Italy, 2019: 146-152

[24] Bosilj P, Aptoula E, Duckett T, et al. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. Journal of Field Robotics, 2020, 37(1): 7-19

[25] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[26] Long J, Shelhamer E, et al. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431-3440

[27] Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4077- 4087

[28] Dong N, Xing E P. Few-shot semantic segmentation with prototype learning//Proceedings of the British Machine Vision Conference. Newcastle, UK, 2018: 1-13

[29] Zhang C, Lin G, Liu F, et al. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 9586-9594

[30] Yang B, Liu C, Li B, et al. Prototype mixture models for few-shot semantic segmentation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 763-778

[31] Min J, Kang D, Cho M. Hypercorrelation squeeze for few-shot segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 6941-6952

[32] Shi X, Wei D, Zhang Y, et al. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 151-168

[33] Wang X, Zhang X, Cao Y, et al. SegGPT: Segmenting everything in context. CoRR abs/2304.03284, 2023

[34] Tan W, Chen S, Yan B. DifFSS: Diffusion model for few-shot semantic segmentation. CoRR abs/2307.00773, 2023

[35] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[36] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[37] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[38] Rakelly K, Shelhamer E, Darrell T, et al. Conditional networks for few-shot semantic segmentation//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-4

[39] Wang K, Liew J H, Zou Y, et al. PANet: Few-shot image semantic segmentation with prototype alignment//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 9196-9205

[40] Tian P, Wu Z, Qi L, et al. Differentiable meta-learning model for few-shot semantic segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12087-12094

[41] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[42] Li X, Wei T, Chen Y P, et al. FSS-1000: A 1000-class dataset for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 2866-2875

[43] Nguyen K, Todorovic S. Feature weighting and boosting for few-shot segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 622-631

[44] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[45] Everingham M, Eslami S A, Van Gool L, et al. The PASCAL visual object classes challenge: A retrospective. International journal of Computer Vision, 2015, 111: 98-136

[46] Hariharan B, Arbeláez P, Bourdev L, et al. Semantic contours from inverse detectors//Proceedings of the International Conference on Computer Vision. Barcelona, Spain, 2011: 991-998

[47] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[48] Zhu K, Zhai W, Zha Z J, et al. Self-supervised tuning for few-shot segmentation//Proceedings of the International Joint Conference on Artificial Intelligence. Vienna, Austria, 2020: 1019-1025

[49] Boudiaf M, Kervadec H, Masud Z I, et al. Few-shot segmentation without meta-learning: A good transductive inference is all you need?//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13979-13988

[50] Lu Z, He S, Zhu X, et al. Simpler is better: Few-shot semantic segmentation with classifier weight transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 8741-8750

[51] Tian Z, Zhao H, 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

[52] Zhang J W, Sun Y, Yang Y, et al. Feature-proxy transformer for few-shot segmentation//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 6575-6588

[53] Yang Y, Meng F, Li H, et al. A new local transformation module for few-shot segmentation//Proceedings of the International Conference on MultiMedia Modeling. Daejeon, Republic of Korea, 2020: 76-87

[54] Xiong Z, Li H, Zhu X X. Doubly deformable aggregation of covariance matrices for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 133-150

[55] Johnander J, Edstedt J, Felsberg M, et al. Dense Gaussian processes for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 217-234

[56] Zhang G, Navasardyan S, Chen L, et al. Mask matching transformer for few-shot segmentation//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 823-836

[57] Zhang X, Wei Y, Yang Y, et al. SG-One: Similarity guidance network for one-shot semantic segmentation. IEEE Transactions on Cybernetics, 2020, 50(9): 3855-3865

[58] Zhang C, Lin G, Liu F, et al. CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5217-5226

[59] Liu W, Zhang C, Lin G, et al. CRNet: Cross-reference networks for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4165-4173

[60] Li Y, Data G W P, Fu Y, et al. Few-shot semantic segmentation with self-supervision from pseudo-classes//Proceedings of the British Machine Vision Conference. Virtual, 2021: 164-177

[61] Lang C, Cheng G, Tu B, et al. Learning what not to segment: A new perspective on few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 8047-8057

[62] Kayabaşı A, Tüfekci G, Ulusoy İ. Elimination of non-novel segments at multi-scale for few-shot segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2023: 2558-2566

[63] Liu Y, Zhang X, Zhang S, et al. Part-aware prototype network for few-shot semantic segmentation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 142-158

[64] Li G, Jampani V, Sevilla-Lara L, et al. Adaptive prototype learning and allocation for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8334-8343

[65] Wang H, Zhao X, Pang Y, et al. Few-shot segmentation via rich prototype generation and recurrent prediction enhancement//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Shenzhen, China, 2022: 287-298

[66] Zhang B, Xiao J, Qin T. Self-guided and cross-guided learning for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8312-8321

[67] Zhang X, Wei Y, Li Z, et al. Rich embedding features for one-shot semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6484-6493

[68] Rao X, Lu T, Wang Z, et al. Few-shot semantic segmentation via frequency guided neural network. IEEE Signal Processing Letters, 2022, 29: 1092-1096

[69] Lang C, Tu B, Cheng G, et al. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation//Proceedings of the International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 1024-1030

[70] Wang J, Li J, Chen C, et al. Adaptive FS: A novel few-shot segmentation framework via prototype enhancement//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, British Columbia, 2024: 5463-5471

[71] Gairola S, Hemani M, Chopra A, et al. SimPropNet: Improved similarity propagation for few-shot image segmentation//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2020: 573-579

[72] Pambala A K, Dutta T, Biswas S. SML: Semantic meta-learning for few-shot semantic segmentation. Pattern Recognition Letters, 2021, 147: 93-99

[73] Xie G S, Xiong H, Liu J, et al. Few-shot semantic segmentation with cyclic memory network//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 7273-7282

[74] Liu Y, Liu N, Cao Q, et al. Learning non-target knowledge for few-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11563-11572

[75] Yang L, Zhuo W, Qi L, et al. Mining latent classes for few-shot segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 8701-8710

[76] Wu Z, Shi X, Lin G, et al. Learning meta-class memory for few-shot semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 497-506

[77] Liu W, Wu Z, Ding H, et al. Few-shot segmentation with global and local contrastive learning. CoRR abs/2108.05293, 2021

[78] Liu Y, Liu N, Yao X, et al. Intermediate prototype mining transformer for few-shot semantic segmentation//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 38020-38031

[79] Cong R, Xion H, Chen J, et al. Query-guided prototype evolution network for few-shot segmentation. IEEE Transactions on Multimedia, 2024, 26: 6501-6512

[80] Yang B, Wan F, Liu C, et al. Part-based semantic transform for few-shot semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7141-7152

[81] Zhao X, Chen X, Gong Z, et al. Contrastive enhancement using latent prototype for few-shot segmentation. CoRR abs/2203.04095, 2022

[82] Yang X, Ma L, Zhou Y, et al. Prior semantic harmonization network for few-shot semantic segmentation//Proceedings of the International Conference on Image Processing. Bordeaux, France, 2022: 1126-1130

[83] Mao B, Zhang X, Wang L, et al. Learning from the target: Dual prototype network for few-shot semantic segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 1953-1961

[84] Fan Q, Pei W, Tai Y W, et al. Self-support few-shot semantic segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 701-719

[85] Tang Y, Yu Y. Query-guided prototype learning with decoder alignment and dynamic fusion in few-shot segmentation. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19: 1-20

[86] Guan H, Spratling M. Query semantic reconstruction for background in few-shot segmentation. CoRR abs/2210.12055, 2022

[87] Guan H, Michael S. CobNet: Cross attention on object and background for few-shot segmentation//Proceedings of the International Conference on Pattern Recognition. Montreal, Canada, 2022: 39-45

[88] Siam M, Oreshkin B, Jagersand M. Adaptive masked proxies for few-shot segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5248-5257

[89] Wang H, Yang Y, Cao X, et al. Variational prototype inference for few-shot semantic segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 525-534

[90] Okazawa A. Interclass prototype relation for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 362-378

[91] Cheng G, Lang C, Han J. Holistic prototype activation for few-shot segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4650-4666

[92] Chen X, Shi M. Memory-guided network with uncertainty-based feature augmentation for few-shot semantic segmentation. CoRR abs/2406.00545, 2024

[93] Wang H, Zhang X, Hu Y, et al. Few-shot semantic segmentation with democratic attention networks//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 730-746

[94] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany, 2015: 234-241

[95] Liu W, Zhang C, Ding H, et al. Few-shot segmentation with optimal transport matching and message flow. IEEE Transactions on Multimedia, 2023, 25: 5130-5141

[96] Xie G S, Liu J, Xiong H, et al. Scale-aware graph neural network for few-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 5475-5484

[97] Hong S, Cho S, Nam J, et al. Cost aggregation with 4D convolutional swin transformer for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 108-126

[98] Zhang G, Kang G, Yang Y, et al. Few-shot segmentation via cycle-consistent transformer//Proceedings of the Annual Conference on Neural Information Processing Systems. Virtual, 2021: 21984-21996

[99] Zhang S, Wu T, Wu S, et al. CATrans: Context and affinity transformer for few-shot segmentation//Proceedings of the International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 1658-1664

[100] Kang D, Cho M. Integrative few-shot learning for classification and segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9969-9980

[101] Cao Q, Chen Y, Yao X, et al. Progressively dual prior guided few-shot semantic segmentation. CoRR abs/2211.15467, 2022

[102] Liu H, Peng P, Chen T, et al. FECANet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. IEEE Transactions on Multimedia, 2023, 25: 8580-8592

[103] Wang X, Luo X, Zhang T. Target-aware bi-transformer for few-shot segmentation//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Xiamen, China, 2023: 440-452

[104] Park S, Lee S, Hyun S, et al. Task-disruptive background suppression for few-shot segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 4442-4449

[105] Xu Q, Zhao W, Lin G, et al. Self-calibrated cross attention network for few-shot segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 655-665

[106] Chen H, Dong Y, Lu Z, et al. Dense affinity matching for few-shot segmentation. Neurocomputing, 2024, 577: 1-13

[107] Moon S, Sohn S S, Zhou H, et al. HM: Hybrid masking for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 506-523

[108] Moon S, Sohn S S, Zhou H, et al. MSI: Maximize support-set information for few-shot segmentation. CoRR abs/2212.04673, 2022

[109] Guo L, Liu H, Xia Y, et al. Boosting few-shot segmentation via instance-aware data augmentation and local consensus guided cross attention. CoRR abs/2401.09866, 2024

[110] Cao L, Guo Y, Yuan Y, et al. Prototype as query for few shot semantic segmentation. CoRR abs/2211.14764, 2022

[111] Wang Y N, Tian X, Zhong G. FFNet: Feature fusion network for few-shot semantic segmentation. Cognitive Computation, 2022, 14(2): 875-886

[112] Yang Y, Chen Q, Feng Y, et al. MIANet: Aggregating unbiased instance and general information for few-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7131-7140

[113] Luo X, Tian Z, Zhang T, et al. PFENet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. CoRR abs/2109.13788, 2021

[114] Huang K, Cheng M, Wang Y, et al. A joint framework towards class-aware and class-agnostic alignment for few-shot segmentation//Proceedings of the Asian Conference on Computer Vision. Macao, China, 2022: 431-447

[115] Iqbal E, Safarov S, Bang S. MSANet: Multi-similarity and attention guidance for boosting few-shot segmentation. CoRR abs/2206.09667, 2022

[116] Shi X, Cui Z, Zhang S, et al. Multi-similarity based hyper-relation network for few-shot segmentation. IET Image Process, 2023, 17(1): 204-214

[117] Xu W, Huang H, Cheng M, et al. Masked cross-image encoding for few-shot segmentation//Proceedings of the IEEE International Conference on Multimedia and Expo. Brisbane, Australia, 2023: 744-749

[118] Wang Y, Luo N, Zhang T. Focus on query: Adversarial mining transformer for few-shot segmentation. Advances in Neural Information Processing Systems, 2023, 36: 31524-31542

[119] Bao X, Qin J, Sun S, et al. Relevant intrinsic feature enhancement network for few-shot semantic segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 765-773

[120] Wang Y, Sun R, Zhang Z, et al. Adaptive agent transformer for few-shot segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 36-52

[121] Seo J, Park Y H, Yoon S W, et al. Task-adaptive feature transformer with semantic enrichment for few-shot segmentation. CoRR abs/2202.06498, 2022

[122] Zheng Z, Huang G, Yuan X, et al. Quaternion-valued correlation learning for few-shot semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2102-2115

[123] Azad R, Fayjie A R, Kauffmann C, et al. On the texture bias for few-shot CNN segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 2674-2683

[124] Min H, Zhang Y, Zhao Y, et al. Hybrid feature enhancement network for few-shot semantic segmentation. Pattern Recognition, 2023, 137: 109291

[125] Ao W, Zheng S, Meng Y. Few-shot semantic segmentation via mask aggregation. CoRR abs/2202.07231, 2022

[126] Tan W, Ru G, Jiang Y, et al. Rethinking and improving few-shot segmentation from a contour-aware perspective. IEEE Transactions on Multimedia, 2023, 25: 6917-6929

[127] Sun Y, Chen Q, He X, et al. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning//Proceedings of the Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 37484-37496

[128] Liu J, Bao Y, Xie G S, et al. Dynamic prototype convolution network for few-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11543-11552

[129] Hu Z, Sun Y, Yang Y. Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2022

[130] Peng B, Tian Z, Wu X, et al. Hierarchical dense correlation distillation for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Vancouver. Canada, 2023: 23641-23651

[131] Zhang J, Li Y, Wang Y, et al. Image to pseudo-episode: Boosting few-shot segmentation by unlabeled data. CoRR abs/2405.08765, 2024

[132] Yu Z, Lin T, Xu Y. Background clustering pre-training for few-shot segmentation//Proceedings of the IEEE International Conference on Image Processing. Kuala Lumpur, Malaysia, 2023: 1695-1699

[133] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8748-8763

[134] Kirillov A, Mintun E, Ravi N, et al. Segment anything//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4015-4026

[135] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision, and Pattern Recognition. New Orleans, USA, 2022: 10674-10685

[136] Lüddecke T, Ecker A S. Image segmentation using text and image prompts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2021: 7076- 7086

[137] Wu C, Lin Z, Cohen S D, et al. PhraseCut: Language-based image segmentation in the wild//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10213-10222

[138] Shuai C, Meng F, Zhang R, et al. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. CoRR abs/2308.07539, 2023

[139] Wang J, Zhang B, Pang J, et al. Rethinking prior information generation with CLIP for few-shot segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 3941-3951

[140] Feng C B, Lai Q, Liu K, et al. Boosting few-shot semantic segmentation via segment anything model. CoRR abs/2401.09826, 2024

[141] Wang X, Wang W, Cao Y, et al. Images speak in images: A generalist painter for in-context visual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6830-6839

[142] Michaelis C, Ustyuzhaninov I, Bethge M, et al. One-shot instance segmentation. CoRR abs/1811.11507, 2018

[143] Gao N, Shan Y, Wang Y, et al. Sap: Single-shot instance segmentation with affinity pyramid//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 642-651

[144] Fan Z, Yu J G, Liang Z, et al. FGN: Fully guided network for few-shot instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9169-9178

[145] Le M Q, Nguyen T V, Le T N, et al. MaskDiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 2874-2881

[146] Tian Z, Lai X, Jiang L, et al. Generalized few-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11553-11562

[147] Cermelli F, Mancini M, Xian Y, et al. Prototype-based incremental few-shot semantic segmentation//Proceedings of the British Machine Vision Conference. Virtual, 2021: 484-498

[148] Lee Y H, Yang F E, Wang Y C F. A pixel-level meta-learner for weakly supervised few-shot semantic segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 1607-1617

[149] Lei S, Zhang X, He J, et al. Cross-domain few-shot semantic segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 73-90

[150] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents. CoRR abs/2204.06125, 2022

[151] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, 2022, 35: 36479-36494

[152] Saha O, Cheng Z, Maji S. GANORCON: Are generative models useful for few-shot segmentation?//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9991-10000

[153] Wu W, Zhao Y, Shou M Z, et al. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 1206-1217

[154] Shi G, Wu Y, Liu J, et al. Incremental few-shot semantic segmentation via embedding adaptive-update and hyper-class representation//Proceedings of the ACM International Conference on Multimedia. Lisboa, Portugal, 2022: 5547-5556

[155] Zhou Y, Chen X, Guo Y, et al. Advancing incremental few-shot semantic segmentation via semantic-guided relation alignment and adaptation//Proceedings of the International Conference on Multimedia Modeling. Amsterdam, The Netherlands, 2024: 244-257

[156] Siam M, Doraiswamy N, Oreshkin B N, et al. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2020: 860-867

[157] Wang H, Liu L, Zhang W, et al. Iterative few-shot semantic segmentation from image label text//Proceedings of the International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 1385-1392

[158] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15979- 15988

[159] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models//Proceedings of the Annual Conference on Neural Information Processing Systems. Virtual, 2020: 6840-6851

[160] Tang H, Liu X, Sun S, et al. Recurrent mask refinement for few-shot medical image segmentation//Proceedings of the IEEE/ CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3898-3908

[161] Farshad A, Makarevich A, Belagiannis V, et al. MetaMedSeg: Volumetric meta-learning for few-shot organ segmentation//Proceedings of the Medical Image Computing and Computer-Assisted Intervention Workshop on Domain Adaptation and Representation Transfer. Singapore, 2022: 45-55

[162] Shen Q, Li Y, Jin J, et al. Q-Net: Query-informed few-shot medical image segmentation//Proceedings of the SAI Intelligent Systems Conference. Amsterdam, The Netherlands, 2023: 610-628

[163] Silva-Rodríguez J, Dolz J, Ayed I B. Transductive few-shot adapters for medical image segmentation. CoRR abs/2303.17051, 2023

[164] Tavera A, Cermelli F, Masone C, et al. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2022: 1959-1968

[165] Lu Y, Wu X, Wu Z, et al. Cross-domain few-shot segmentation with transductive fine-tuning. CoRR abs/2211.14745, 2022

[166] Wang W, Duan L, Wang Y, et al. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 7055-7064

[167] Zhao Z, Zhou F, Zeng Z, et al. Meta-hallucinator: Towards few-shot cross-modality cardiac image segmentation. CoRR abs/2305.06978, 2023

[168] Ji W, Li J, Bi Q, et al. Segment anything is not always perfect: An investigation of SAM on different real-world applications. CoRR abs/2304.05750, 2023

[169] Tang L, Xiao H, Li B. Can SAM segment anything? When SAM meets camouflaged object detection. CoRR abs/2304.04709, 2023

CHEN Shan-Juan, Ph. D. candidate. Her research interests mainly include machine learning and few-shot semantic segmentation.

YU Yun-Long, Ph. D., distinguished researcher. His research interests mainly include machine learning and computer vision.

LI Ying-Ming, Ph. D., associate professor. His research interests mainly include machine learning and multi-task learning.

Background

The semantic segmentation task is one of the important tasks and is the basis for many tasks in computer vision. However, due to its high dependence on data and poor generalization ability, its application in real life is limited. Few-shot semantic segmentation (FSS) task has been proposed to address these issues and aims to exploit the accumulated knowledge in the past to achieve novel class generalization guided by only a few densely labeled samples. FSS has attracted a lot of attention since it was proposed, and a large number of algorithms have emerged. Since its proposal in 2017, few-shot semantic segmentation task has made rapid development and progress, from pixel classification tasks to segmentation guided by prototypes, and then to the segmentation achieved through pixel-level relationship modeling. With the development of pre-trained foundation models, researchers began to explore the application of visual foundation models for few-shot semantic segmentation task, introducing the FSS task into a new stage of development. In addition, people began to move towards training a task universal model for few-shot learning task.

This paper investigates the research on few-shot semantic segmentation in the field of natural images. Although there have been research reviews on few-shot semantic segmentation methods in recent years, comprehensive reviews in this field are still rare. Moreover, with the rapid development of this field, a large number of algorithms have emerged, requiring reclassification for these novel algorithms. Compared with previous reviews, this article classifies current algorithms from the perspective of model optimization, based on whether gradient backpropagation exists in the inference process. The optimization-based methods need gradient backpropagation and optimize the trained models during inference, whereas metric-learning-based approaches do not require any optimization of the model during inference. Then we sort out the current algorithms and classify them into different categories according to their designs and the challenges they address in solving few-shot segmentation tasks. This classification can comprehensively cover current few-shot semantic segmentation algorithms, which is simple and intuitive. In addition to proposing a new classification scheme, this paper deeply analyzes existing methods and challenges, and summarizes their development trends. This will help researchers better understand the current research status, provide a broader perspective for future research, and promote further development and innovation in this field. Few-shot semantic segmentation task has a wide range of applications and development prospects, and it is hoped that through efforts, its performance can be improved and the development of the computer vision field can be promoted.