密集的交叉查询和支持注意力少数镜头的加权掩码聚合分割

新余市1 , 董伟2 , 张宇1 †, 路东环2 † , 穆南宁2 , 嘉顺陈1 , 马凯2和郑业峰2

东南大学计算机科学与工程学院,计算机网络与信息集成教育部重点实验室,南京,中国
{shixinyu,zhang yu,jiashunchen}@seu.edu.cn 腾讯贾维斯实验室,中国深圳
{ donwei,caleblu,masonning,kylekma,yefengzheng}@tencent.com

抽象的。少镜头语义分割 (FSS) 的研究引起了极大的关注,其目标是在仅给定目标类的少数带注释支持图像的情况下分割查询图像中的目标对象。这项具有挑战性的任务的关键是通过利用查询和支持图像之间的细粒度相关性来充分利用支持图像中的信息。然而,大多数现有方法要么将支持信息压缩为几个类别原型,要么在像素级别使用部分支持信息（例如,仅前景）,从而导致不可忽略的信息丢失。在本文中,我们提出了密集像素级交叉查询和支持注意力加权掩码聚合（DCAMA）,其中前景和背景支持信息通过成对查询和支持之间的多级像素相关性得到充分利用特征。通过 Transformer 架构中的缩放点积注意力实现,DCAMA 将每个查询像素视为一个标记,计算其与所有支持像素的相似性,并将其分割标签预测为所有支持像素标签的加法聚合 由相似之处。基于 DCAMA 的独特公式,我们进一步提出了用于 n-shot 分割的高效且有效的一次性推理,其中一次收集所有支持图像的像素用于掩码聚合。实验表明,我们的 DCAMA 在 PASCAL-5i COCO-20i和 FSS-1000 的标准 FSS 基准上显着提高了最新技术水平,例如,与以前相比,1-shot mIoU 的绝对改进为 3.1%、9.7% 和 3.6%最好的记录。烧蚀研究也验证了设计 DCAMA。

, ,

关键词:Few-shot 分割·密集交叉查询和支持注意力·注意力加权掩码聚合。

同等贡献,工作在腾讯贾维斯实验室完成。 †通讯作者。

1 简介

近年来,深度神经网络（DNN)取得了显着进展
在语义分割[25, 37]中,计算机视觉的基本任务之一。然而,DNN 的成功很大程度上依赖于大规模数据集,其中
丰富的训练图像可供每个目标类进行分割。在里面
在极低数据的情况下,DNN 的性能可能会比以前快速下降
由于泛化能力差,看不见的类只有很少的例子。人类,在
相比之下,能够利用从生活经验中积累的先验知识在低数据场景中快速学习新任务[16]。少量学习
(FSL) [11, 12] 是一种机器学习范式,旨在模仿人类学习者的这种泛化能力,其中模型可以快速适应
新颖的任务只给出了几个例子。具体来说,一个支持集包含
为模型适应给出了样本有限的新类,即
随后在包含相同类样本的查询集上进行评估。
FSL在计算机视觉领域得到了积极探索,例如图像
分类 [42]、图像检索 [39]、图像字幕和视觉问题 [8] 和语义分割 [7, 20, 23, 24, 26, 33, 35, 38, 43–46, 48–50] .
在本文中,我们解决了少镜头语义分割（FSS)的问题。

FSS 的关键挑战是充分利用包含在
小支撑集。以前的大多数作品都遵循原型设计的概念 [33],
其中支持图像中包含的信息通过类平均池化 [44, 48, 50] 或聚类 [20, 45] 被抽象为类智能原型,
查询特征与之匹配以进行分割标签预测。
然而,最初提出用于图像分类任务的原型可能会导致已经包含的宝贵信息的大量丢失。

应用于 FSS 时的稀缺样本。鉴于分割的密集性质
任务,[24,43,47] 最近提出探索像素之间的相关性
查询特征和前台支持特征,避免信息
原型设计中的压缩。然而,这些方法完全忽略了支持图像背景区域中包含的丰富信息。

张等人。 [49] 还考虑了后台支持功能
计算像素级相关性;然而他们只考虑了具有均匀采样支持像素的稀疏相关性,导致潜在的信息丢失
另一种。因此,以前的工作没有充分研究密集的像素方式
查询特征与前景和背景的相关性
支持 FSS 的功能。

在这项工作中,我们提出 Dense pixel-wise Cross-query-and-support Attention
FSS 的加权掩码聚合 (DCAMA),它充分利用了支持图像中所有可用的前景和背景特征。如图所示。
1,我们做出一个关键的观察,一个查询像素的掩码值可以是
通过支持掩码值的加法聚合按比例预测
通过其与相应支持图像像素的相似性加权 包括
前景和背景。这很直观:如果查询像素在语义上接近前景支持像素,则后者将投票支持前景为

前者的掩码值,反之亦然 度量学习的体现

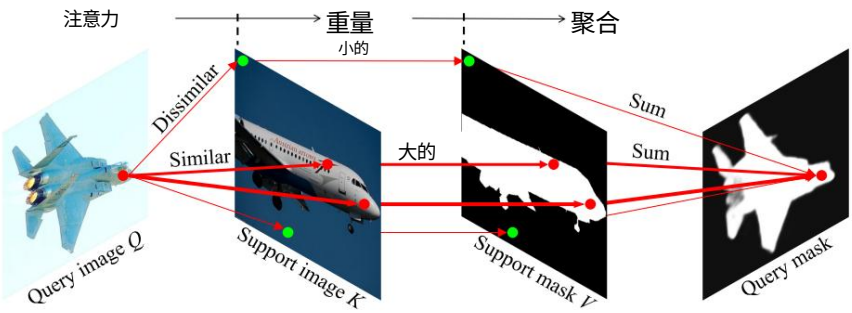

图 1. 我们方法的概念概述。查询掩码由支持掩码值的逐像素加法聚合直接预测,由密集的交叉查询和支持注意力加权。

荷兰国际集团 [15]。此外,我们注意到查询图像的所有像素的 DCAMA 计算管道可以通过 Transformer 架构 [40] 中的点积注意机制轻松实现,其中每个像素都被处理

作为记号,所有查询像素的扁平化特征构成查询矩阵 Q,所有支持图像像素的特征构成关 [1] 键矩阵 K,支持掩码像素的扁平化标签值构成值矩阵 V。然后,查询掩码可以通过 softmax(QKT)V 轻松有效地计算。对于实际实现,我们遵循多头注意力[40]、多尺度[18]和多层[24]特征相关的常见做法;此外,聚合门控掩码与跳过连接的支持和查询特征混合在一起,以进行精细的查询标签预测。正如我们将展示的那样,所提出的方法不仅比以前的最佳性能方法[24]产生了更好的性能,而且还展示了更高的训练效率。

此外,以前采用像素级相关管道的工作很少关注从 1-shot 到few-shot 分割的扩展:它们要么执行单独的 1-shot 推理,然后进行集成 [24],要么使用均匀采样支持的子集用于推理的像素 [49]。由于集成之前的独立推断和潜在有用像素的丢弃,这两种解决方案都导致了像素级信息的丢失。相比之下,我们充分利用了支持集,通过使用所有支持图像和掩码的所有像素分别构成K和V矩阵。同时,我们使用相同的 1-shot 训练模型在不同的小样本设置中进行测试。这不仅在计算上是经济的,而且是合理的,因为模型实际上从训练中学到的是一个用于交叉查询和支持注意力的度量空间。只要度量空间被很好地学习,从 1-shot 扩展到 few-shot 只是从更多支持像素中聚合查询掩码。

总之,我们做出以下贡献:

– 我们创新地将 FSS 问题建模为密集像素交叉查询和支持注意力加权掩码聚合的新范式

（DCAMA）,它充分利用了前台和后台支持

[1] 注意:不要将 FSL 中的查询与 Transformer 中的查询混淆。

形成。作为度量学习的一个体现,该范式在掩码聚合中是非参数的,并且有望很好地泛化。

– 为了简单和高效,我们在 Transformer 中使用成熟的点积注意力机制实现了新的 FSS 范式。

– 基于 DCAMA,我们提出了一种 n-shot 推理方法,该方法不仅充分利用了像素级别的可用支持图像,而且在计算上也很经济,无需针对不同的少镜头设置训练 n 特定模型。

在 PASCAL-5i [31]、COCO-20i [26] 和 FSS-1000 [17]三个标准 FSS 基准上的比较实验结果表明,我们的 DCAMA 在所有三个基准和两个基准上都设置了新的技术水平1 - 和 5-shot 设置。此外,我们进行了彻底的烧蚀实验来验证 DCAMA 的设计。

2 相关工作

语义分割。语义分割是计算机视觉中的一项基本任务,其目标是将图像的每个像素分类为预定义的对象类别之一。在过去十年中,随着 DNN [25, 37] 的进步,取得了令人瞩目的进展。基石全卷积网络（FCN)[22]提出用卷积层替换分类网络中的全连接输出层,以有效地输出像素级密集预测以进行语义分割。从那时起,流行的分割 DNN 模型 [3, 30, 52] 已经演变为由具有通用编码器-解码器架构的 FCN 主导,其中通常采用跳过连接 [30] 和多尺度处理 [4, 52] 等技术以获得更好的性能。

最近,受到 Vision Transformer (ViT) [9] 成功的启发,我们见证了将 Transformer 架构应用于语义分割的积极尝试 [34, 53]。值得注意的是,Swin Transformer 是一种通用的计算机视觉主干,具有分层架构和移动窗口,在 ADE20K 语义分割基准测试 [21] 上实现了新的最先进 (SOTA) 性能。尽管这些方法在丰富的训练数据下证明了它们的能力并启发了我们的工作,但它们都无法对低数据机制进行泛化。

少量学习。少样本学习 [11] 是一种范式,旨在提高机器学习模型在低数据状态下的泛化能力。受开创性工作 ProtoNet [33] 的推动,FSS 上的大多数先前工作 [7、20、44、45、48、50] 都遵循度量学习 [7] 管道,其中支持图像中包含的信息被压缩为抽象原型,并且查询图像是根据与度量空间中原型的一定距离进行分类的。 Dong 和 Xing [7] 扩展了 FSS 原型的概念,通过对蒙面支持图像的特征进行全局平均池化计算类别原型。 PANet [44] 不是对输入图像进行屏蔽,而是对用于原型设计的屏蔽支持特征执行平均池化,并引入原型对齐作为正则化。 CANet [48] 也依赖于 masked

特征池,但遵循关系网络 [36] 来学习深度度量
使用 DNN。 PFENet [38] 还进一步提出了一种免训练的先验掩码
作为多尺度特征丰富模块。意识到单个原型的有限表示能力,[5,20,45,46]都提出来表示一个类

有多个原型。这些基于原型的方法共同推进了
FSS研究;但是,在支持中压缩所有可用信息
形象化为一个或几个集中的原型不可避免地导致
信息丢失很大。

最近,研究人员开始利用 FSS 的像素级信息,
更好地利用支持信息并与
任务。 PGNet [47] 和 DAN [43] 建模了像素到像素之间的密集连接
具有图注意力的查询和支持图像 [41],而 HSNet [24]
构造 4D 相关张量来表示之间的密集对应关系
查询和支持图像。值得注意的是,HSNet 提出了用于高效高维卷积的 center-pivot 4D 卷积,并在三个
公共 FSS 基准上大幅实现了 SOTA 性能。然而,这些

方法都掩盖了支持图像中的背景区域,忽略
从而获得丰富的信息。相比之下,我们的 DCAMA 平等地利用了前景和背景信息。此外,实施简单

具有多头注意力的度量学习[40],我们的 DCAMA 更容易训练
与 HSNet 相比,在更少的时期和更多的时间内收敛到更高的性能
更短的时间。最后,而不是单独的 1-shot 推理的集合 [24] 或
训练 n 特定模型 [48] 进行 n-shot 推理,DCAMA 构建关键
和具有所有支持图像和掩码的像素的值矩阵,并推断
仅重用 1-shot 训练模型一次。

FSS 的视觉变形金刚。受最近在计算机视觉 [9, 21] 中取得成功的 Trans 前架构的启发,研究人
员最近也开始探索他们在 FSS 中的应用。孙等人。 [35] 提出采用标准的多头自注意力 Transformer
块进行全局增强。鲁

等。 [23] 设计了分类器权重转换器 (CWT) 以动态地
为每个查询图像调整分类器的权重。不过,两人还是
遵循原型设计流程,因此没有充分利用细粒度支持
信息。循环一致的变压器（CyCTR)[49]可能是最
在以下方面与我们的相关工作：(i)使用点积注意力机制进行像素级交叉查询和支持相似性计算,以及
（ii)两者的使用
前台和后台支持信息。主要区别在于
CyCTR 使用相似度来指导从
支持特征,然后通过常规方法将其分类为查询标签
FCN。相比之下,我们的 DCAMA 可以通过聚合由这种相似性加权的支持标签来直接预测查询标签,这
是度量学习
并有望很好地推广其非参数形式。另一个区别
是 CyCTR 对支持像素进行了二次采样,因此受到依赖于采样率的潜在地层损失的影响,而我们的
DCAMA 使
使用所有可用的支持像素。

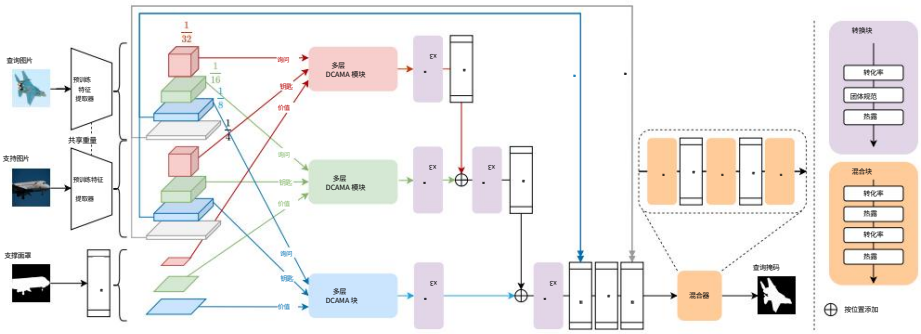

图 2. 提议框架的流程,以 1-shot 设置显示。 DCAMA: Dense Cross-query-and-support Attention weighted Mask Aggregation。

3 方法论

在本节中,我们首先介绍 Few-shot Semantic Segmentation (FSS) 的问题设置。然后,我们在 1-shot 设置中描述了我们的密集交叉查询和支持注意力加权掩码聚合 (DCAMA) 框架。最后,我们扩展了 n-shot 推理的框架。

3.1 问题设置

在正式定义中,一个单向 n-shot FSS 任务 T 包含一个支持集 S = {(I s , Ms)},其中 I s和Ms是支持图像及其基本真值掩码， Mq)},其中 S 和 Q 是从同一类中和 |S| = n; 和类似的查询集 Q = {(I q , 采样的。目标是学习一个模型来预测给定支持集 S的每个 I q的Mq ,前提是 n 对于少数镜头来说很小。对于方法开发,假设我们有两个图像集Dtrain和Dtest分别用于模型训练和评估,其中Dtrain和Dtest在类中不重叠。我们采用了广泛使用的元学习范式,称为情景训练 [42],其中每一集都旨在通过对训练集中的类和图像进行二次采样来模拟目标任务。具体来说,我们反复从Dtrain中抽取新的情节任务 T进行模型训练。预计情节的使用将使训练过程更忠实于测试环境,从而提高泛化能力[29]。为了进行测试,训练后的模型也使用情节任务进行评估,但从Dtest 中采样。

3.2 DCAMA 1-Shot 学习框架

概述。我们的 DCAMA 框架的概述如图 2 所示。为简单起见,我们首先描述我们的 1-shot 学习框架。框架的输入是查询图像、支持图像和掩码。首先,查询和支持图像都由预训练的特征提取器处理,产生多尺度查询和支持特征。同时,支持掩码被下采样到与图像特征匹配的多个尺度。二、

每个尺度的查询特征、支持特征和支持掩码被输入到与 Q、K 和 V 相同尺度的多层 DCAMA 块中，用于多头注意力 [40] 和查询掩码的聚合。在多个尺度上聚合的查询掩码被处理并与卷积、上采样（如果需要)和元素相加相结合。第三，前一阶段的输出（多尺度 DCAMA)通过跳跃连接与多尺度图像特征连接，随后由混合器混合以产生最终的查询掩码。在下文中，我们依次描述这三个阶段中的每一个，重点是第二个阶段 这是我们的主要贡献。

特征提取和掩码准备。首先，查询和支持图像都输入到预训练的特征提取器，以获得它们的多尺度多层特征图 {F i,l}和 {F si,l} 的集合，其中 i 是尺度输入图像的特征图和我们使用的特征提取器的 i ∈ { } 和 l ∈ {1, . . . , Li}是特定尺度 i 的所有层的索引。与大多数以前使用的作品不同，我们遵循 Min 等人。 [24]到每个尺度的最后一层特征图，即Fi,Li，也充分利用了所有中间层特征。同时，通过双线性插值从原始支持掩码生成不同尺度的支持掩码，作为所述尺度上的多尺度多层密集注意力加权掩码聚合。

$\frac{1}{8}, \frac{1}{16}, \frac{1}{32}$

缩放的点积注意力是 Transformer [40] 架构的核心，并被表述为：

$$ \quad (1)$$

其中 Q、K、V 是打包成矩阵的查询、键和值向量的集合，d 是查询和键向量的维度。在这项工作中，我们采用 Eqn。 (1) 在查询和支持特征中计算密集的像素级注意力，然后用注意力值对来自支持掩码的查询掩码聚合过程进行加权。不失一般性，我们用一对通用查询和支持特征图 F q描述该机制，其中 h、w 和 c 分别是高度、宽度和通道数，以及通用支持掩码Ms ∈ R大小相同。如图 3 所示，我们首先将二维（2D)输入展平以将每个像素视为一个标记，然后在添加位置编码和线性投影后从展平的 F q和 F s生成 Q 和 K 矩阵。我们按照原来的 Transformer [40] 使用不同频率的正弦和余弦函数进行位置编码，并采用多头注意力。至于支撑面罩，只需展平即可构造 V 。在那之后，标准缩放了点积注意力 Eqn. (1) 多头输出进行平均，并将张量重塑为 2D 以获得 $\hat{M}_q \in R^{高\times宽\times1}$，这是聚合查询掩码。

评论。值得解释 DCAMA 过程的物理意义。对于特定的查询像素，QKT测量其与所有支持像素的相似性，随后与 V 的乘法将其掩码值从

[2] 这 $\frac{1}{4}$ 由于硬件限制，比例特征没有交叉参与。

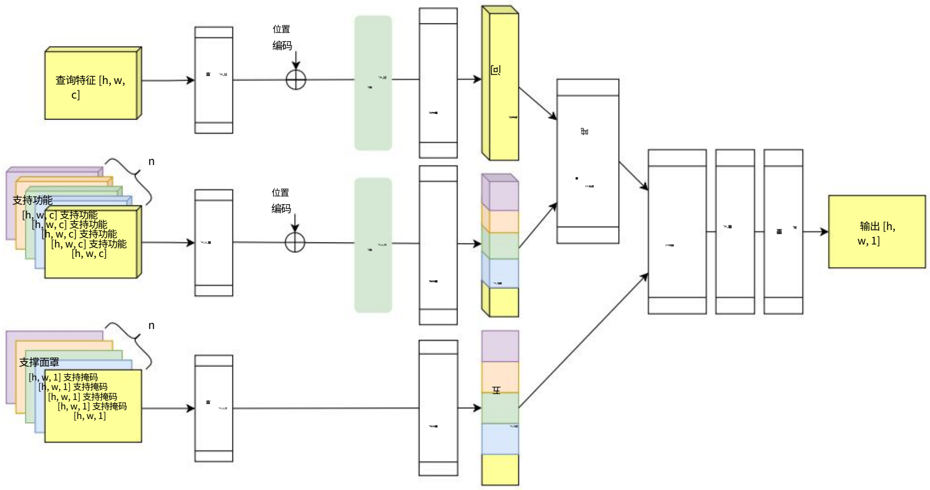

图 3. 用于通用 n-shot 设置 (n ⩾ 1) 的密集交叉查询和支持注意力加权掩码聚合 (DCAMA)。

支持面罩，由相似性加权。直观地说，如果它与前景比背景像素更相似（更接近），则加权聚合过程将为该像素投票给前景，反之亦然。通过这种方式，我们的 DCAMA 利用所有支持像素 前景和背景 进行有效的度量学习。

在实际实现中，我们对特定尺度 i 的所有中间层和最后一层的查询支持特征对 (F s Fq i,l)分别进行 DCAMA，并将独立聚合的查询掩码集连接起来得到 M^q = concat (M^ q i,l|l = ... 。对特定尺度的所有层特征进行 DCAMA 操作，然后进行连接组成一个多层 DCAMA 块（见图 2），我们为尺度 i ∈ { } 有三个这样的块。那么，M ^q

1 8, 1 16, 32

由三个 Conv 块处理，其通道数从Li逐渐增加到128，通过双线性插值进行上采样，通过逐元素加法与大一倍的对应物相结合，并再次由恒定通道的另外三个 Conv 块处理数字。

q前三个 Conv 块为有效的跨尺度集成做准备 M ^
后三个 Conv 块的规模更大。这个过程从 i = 8开始重复，产生一组中间查询掩码，以与跳过连接的图像特征融合以进行最终预测。 1 32

面具功能混合器。受通用语义分割 [30, 52] 中跳过连接设计成功的启发，我们还建议通过连接将图像特征跳过连接到前一阶段的输出（需要时上采样）（图 2）。具体来说，我们根据我们的经验实验（包括在补充材料中)跳过连接最后一层特征的尺度。然后，连接的中间查询掩码和图像特征由三个掩码特征混合器块融合，每个块包含两个系列的卷积和 ReLU 操作。混合器块逐渐减少

$\frac{1}{4}$ 和 $\frac{1}{8}$

输出通道数为 2（分别用于前景和背景）
用于 1 路分割,用两个交错的上采样操作来恢复
输出大小到输入图像的大小。

3.3 扩展到 n-Shot 推理

到目前为止,我们已经介绍了用于 1-shot 分割的 DCAMA 框架，
接下来我们也将其扩展为 n-shot 设置。虽然可以开发
并为每个不同的 n 值训练一个特定的模型（例如,[48]），这样做在计算上是禁止的。相比之下,之前的很多作品都扩展了
用于 n-shot 推理的 1-shot 训练模型,无需重新训练 [24, 38, 46]。这
最常见的方法是分别执行 n 个 1-shot 推理
支持图像的集合,然后是个别推理的某些集合 [24, 46]。然而,这种方法不可避免地丢失了像素级的细微线索
跨支持图像,因为它独立处理每个支持图像以进行推理。在这项工作中,我们还重用了 1-shot 训练模型进行 n-shot 推理
为了计算效率,但同时利用所有像素信息
在推理过程中同时包含所有支持图像。

由于 DCAMA 的问题表述,扩展是直截了当的。首先,我们获得所有图像的多尺度图像特征和掩码
支持图片。接下来,我们简单地处理额外支持中的额外像素
特征和掩码作为 K 和 V 中的更多标记,并对 ,
张量（图 3）。然后,整个 DCAMA 过程（交叉注意力、掩码聚合等)保持与 1-shot 设置相同。这是可行的核心
DCAMA 是等式中的缩放点积注意力。 (1),无参数。因此,DCAMA 过程实际上与 n 无关，可以应用于
用任意 n 推断。 [3] 直观地说,查询像素的标签是联合确定的
一次由所有可用的支持像素挖掘,而不管确切的数量
支持图片。这种一次性推理不同于单个推理的集合,其中首先使用每个支持获得图像级预测

图像独立,然后合并以产生集合级的最终预测。它
也不同于一些基于原型的 n-shot 方法 [7, 45],其中所有支持图像的特征被同时处理但被压缩成
一个或几个原型,失去像素级粒度。最后,一个小小的改编
是在支持图像中最大池化跳过连接的支持特征，
从而使图 2 所示的整个 DCAMA 框架适用于
通用 n-shot 设置,其中 n ⩾ 1。

4 实验与结果

数据集和评估。我们在三个标准上评估所提出的方法
FSS 基准测试:PASCAL-5i [31]、COCO-20i [26] 和 FSS-1000 [17]。 PASCAL 5^{i} 是从 PASCAL VOC 2012 [10] 和 SDS [13] 数据集创建的。它包含 20

[3] 尽管如此,训练 n 特定模型
n > 1,考虑到时间和 GPU 内存。

类,平均分为四折,即每折五类。 COCO-20i
是从 COCO [19] 数据集创建的更大且更具挑战性的基准。
它包括 80 类,再次平均分为四折。对于 PASCAL-5i
和 COCO-20i ，评估是通过交叉验证完成的,其中每个折叠都是 se
依次选为Dtest,其他三折为Dtrain； 1,000 个测试集
从Dtest中随机抽样进行评估 [24]。 FSS-1000 [17] 包括
1,000 个类,分为 520 个训练、验证和测试拆分，
分别有 240 和 240 个类别,从
用于评估的测试拆分 [24]。对于指标,我们采用平均交集
联合（mIoU)和前景-背景IoU（FB-IoU)[24]。用于 PASCAL-5i
和 COCO-20i ，单个折叠上的 mIoU,以及平均 mIoU 和 FB
报告跨折叠的 IoU;对于 FSS-1000,mIoU 和 FB-IoU
报告测试拆分。请注意,我们尝试遵循常见做法
以前的作品 [24, 26, 38, 49] 采用以进行公平比较。

实施细节。所有实验均使用 PyTorch 进行
[28] 框架（1.5.0）。对于主干特征提取器,我们使用 ResNet-50
和 ResNet-101 [14] 在 ImageNet [6] 上进行了预训练,因为它们在
以前的作品。4此外,我们还对基础 Swin Transformer 进行了实验
在 ImageNet-1K [21] 上预训练的模型（Swin-B）,以评估泛化性
我们在非卷积主干上的方法。我们使用三层多层
DCAMA 块用于刻度,分别 $\frac{1}{8}$、$\frac{1}{16}$ 和 $\frac{1}{32}$
交叉注意力加权掩码聚合的级别。除非另有规定，
尺度的最后一层特征是跳跃连接的。的输入大小 $\frac{1}{4}$ 和 $\frac{1}{8}$
支持和查询图像都是 384 × 384 像素。采用平均二元交叉熵损失： LBCE = -
$\frac{1}{n}$ P[y log p + (1 − y) log(1 − p)],其中 N 是
像素总数,y ∈ {0, 1} 是像素标签（0 代表背景,1
对于前景）,p是预测概率。我们只对
最终输出来训练我们的模型,并冻结主干参数。新元
使用优化器,设置学习率、动量和权重衰减
至10−3, 0.9 和10−4 ， 分别。批量大小设置为 48、48 和 40
帕斯卡-5i ， COCO-20i， 和FSS-1000,分别。我们按照 HSNet [24] 来
在没有数据增强的情况下训练我们的模型,直到收敛,以获得公平
与以前性能最佳的方法进行比较。培训是在
四个 NVIDIA Tesla V100 GPU,推理在 NVIDIA Tesla T4 上
显卡。我们的代码位于 https://github.com/pawn-sxy/DCAMA.git。

4.1 与现有技术的比较

在表 1 和表 2 中,我们比较了我们提出的 DCAMA 的性能
自 2020 年以来在 PASCAL i上发布的 FSS 的 SOTA 方法框架
5 ， COCO-20i， 和FSS-1000,分别。除非另有说明,否则报告
其他方法的数量来自原始论文;当有不同主干的结果可用时,我们只报告较高的主干以节省空间。

[4] 虽然一些作品（例如,[38])报告了 ResNet-50 和
ResNet-101、VGG-16 [32] 的性能大多不如 ResNet 系列
在以前关于 FSS 的作品中。因此,我们的实验中不包括 VGG-16。

表 1. PASCAL-5i （上)和 COCO-20i （下)的性能。 HSNet † :我们的
基于官方代码重新实现;注意[40]。粗体和带下划 *: 实现点积的方法
线的数字分别突出显示每个主干的最佳和次佳性能（如有必要）。

骨干	方法	类型	shot 5-shot											
			Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
PASCAL-5 1- [31]														
ResNet-50	PPNet [20]	原型	52.7	62.8	57.4	47.7	55.2 (5.6)	52.0	67.5	51.5	49.8	-	60.3	70.0	69.4	60.7	65.1 (4.6)	55.0	68.2	52.9	51.1	-
	PMM [45]		55.2 (7.2)	55.2	66.9	52.6	50.7	56.3 (5.3)	59.8	68.3		-	56.8 (6.7)	56.3	67.3	54.5	51.0	57.3 (6.1)	64.6	71.4		-
	转速 [45]		62.1	48.5	59.7 (7.2)	61.7	69.5	55.4	56.3	60.8 (5.6)		-	71.1	59.3	66.6 (5.0)							-
	RePRI [2]		73.3	63.1	70.7	55.8	57.9	61.9 (5.7)	73.9			-										-
	PEFNet [38]																					
	SCL [46]		63.0	70.0	56.5	57.7	61.8 (5.3)	71.9	64.5	70.9	57.3	58.7	62.9 (5.4)	72.8								
	TRFS* [35]		62.9	70.7	56.5	57.5	61.9 (5.6)					-	65.0	71.2	55.5	60.9	63.2 (5.7)					-
	DCAMA (Ours)像素级		67.5	72.3	59.6	59.0	64.6 (5.6)	75.7	70.5	73.9	63.7	65.8	68.5 (4.0)	79.5								
ResNet-101	CWT* [23]	原型	56.9	65.2	61.2	48.8	58.0 (6.1)	57.0	67.2	56.1	54.3	-	62.6	70.2	68.8	57.2	64.7 (5.2)	57.3	68.5	61.5	56.3	-
	DoG-LSTM [1]		58.7	(5.0)	54.7	68.6	57.8	51.6	58.2	(6.4)	71.9	57.9	60.9	(4.8)								-
	丹 [43]	逐像素	69.0	60.1	54.9	60.5	(5.3)	7														
	CyCTR* [49]		69.3	72.7	56.5	58.6	64.3	(6.9)	67.3	72.3	62.0	63.1	-	73.5	74.0	58.6	60.2	66.6	(7.2)			-
	高速网络 [24]		66.2	(4.1)	77.6	71.8	74.4	67.0	68.3	70.4	(2.9)	80.6										
	DCAMA（我们的)*		65.4	71.4	63.2	58.3	64.6	(4.7)	77.6	70.7	73.7	66.8	61.9	68.3	(4.4)	80.8						
Swin-B	HSNet † [24]	逐像素	67.9	74.0	60.3	67.0	67.3	(4.9)	77.9	72.2	77.5	64.0	72.6	71.6	(4.8)	81.2						
	DCAMA（我们的)*		72.2	73.8	64.3	67.1	69.3	(3.8)	78.5	75.7	77.1	72.0	74.8	74.9	(1.8)	82.9						
COCO-20 [26]																						
ResNet-50	PPNet [20]	原型	36.5	26.5	26.0	19.7	27.2	(6.0)	29.3	34.8	27.1	27.3	-	48.9	31.4	36.0	30.6	36.7	（7.3)	33.0	40.6	30.3	33.3	-
	PMM [45]		29.6	(3.1)	29.5	36.8	28.9	27.0	30.6	(3.7)	31.8	34.9	-	33.3	34.3	（3.8)	33.8	42.0	33.0	33.0	33.3	33.3	35.3	-
	转速 [45]		36.4	31.4	33.6	(2.1)	31.2	38.1	33.3	33.0	34.0	(2.6)	-	35.5	（3.7)	35.4	41.7	42.3	36.1	38.1	38.9	（3.1)	（3.1)	-
	TRFS* [35]		38.9	43.0	39.6	39.8	40.3	(1.6)	41.9	45.1	44.4	41.7	-	(3.1)										-
	RePRI [2]		43.3	(1.5)	69.5	45.9	50.5	50.7	46.0	48.3	(2.3)	71.7	-											-
	CyCTR* [49]	逐像素											-											-
	DCAMA（我们的)*																							
ResNet-101	CWT* [23]	原型	30.3	36.6	30.5	32.2	32.4	(2.5)	36.4	38.6	37.5	35.4	-	38.5	46.7	39.4	43.2	42.0	(3.3)	38.9	40.5	41.5	38.7	-
	SCL [46]		37.0	(1.2)	36.8	41.8	38.7	36.7	38.5	(2.1)	63.0	40.4	-	39.9	(1.2)									-
	PEFNet [38]		46.8	43.2	40.5	42.7	(2.6)	6																
	丹 [43]	逐像素	-	-	-	-	-				24.4 (-)		62.3	-	-	-	-				29.6 (-)		63.9	
	高速网络 [24]		37.2	44.1	42.4	41.3	41.2	(2.5)	69.1	45.9	53.0	51.8	47.1	49.5	(3.0)	72.4								
	DCAMA（我们的)*		41.5	46.2	45.2	41.3	43.5	(2.2)	69.9	48.0	58.0	54.3	47.1	51.9	(4.5)	73.3								
Swin-B	HSNet † [24]	逐像素	43.6	49.9	49.4	46.4	47.3	(2.5)	72.5	50.1	58.6	56.7	55.1	55.1	(3.2)	76.1								
	DCAMA（我们的)*		49.5	52.7	52.8	48.7	50.9	(1.8)	73.2	55.4	60.3	59.9	57.5	58.3	(2.0)	76.9								

最重要的是,我们的方法非常有竞争力:它达到了最好的性能
就几乎所有主干组合的mIoU和FB-IoU而言
网络（ResNet-50、ResNet-101 和 Swin-B)和少样本设置（1- 和 5-
拍摄)在所有三个基准数据集上。唯一的例外是 PASCAL-5i
使用 ResNet-101,我们的 DCAMA 和 HSNet [24] 共享前两名
用于具有可比性能的 1 次和 5 次拍摄设置中的 mIoU 和 FB-IoU。
当使用 Swin-B 作为主干特征提取器时,DCAMA 在所有三个基准上都显着提升了系统级 SOTA

之前的三基准 SOTA HSNet (ResNet-101),例如提高 3.1% (1-shot)
在 PASCAL-5i 上为 4.5%（5 发）, COCO-20i 上分别为 9.7% 和 8.8% , 和 3.6% 和
就 mIoU 而言,FSS-1000 分别为 1.9%。其次,虽然 HSNet 的性能随着 Swin-B 主干网络的提高而提高,但
在 mIoU 上仍然存在 2-3.6%（1-shot)和 1.5-3.3%（5-shot)的明显劣势。

DCAMA 的。此外,我们的 DCAMA 也比 HSNet 更稳定
褶皱,观察较低的标准偏差。这些结果表明
DCAMA 适用于基于卷积和注意力的主干。
第三,基于像素相关性的方法的性能（除了
对于 DAN [43])通常比基于原型的更好,这证实了利用细粒度像素级信息的直觉

用于 FSS 的任务。最后但同样重要的是,值得注意的是我们的 DCAMA

表 2. FSS-1000 [17] 上的性能。 HSNet † :我们基于
官方代码。粗体和带下划线的数字分别突出显示每个主干的最佳和次佳性能（如有必要）。

骨干	方法	类型	1发		5发	
			mIoU	FB-IoU	mIoU	FB-IoU
ResNet-50	DCAMA（我们的）	Pixel-wise	88.2	92.5	88.8	92.9
ResNet-101	DoG-LSTM [1]	原型	80.8	83.4	-	-
	DAN [43]	逐像素	85.2	88.1	-	-
	HSNet [24]	逐像素	86.5	88.5	87.1	88.8
	DCAMA（我们的）	逐像素	88.0	91.3	88.7	92.4
Swin-B	HSNet † [24]	逐像素	86.7	91.8	88.9	93.2
	DCAMA（我们的）	逐像素	90.1	93.8	90.4	94.1

图 4. (a) PASCAL-5i在 1-shot 设置中的定性结果,存在类内变化、大小差异、复杂背景和遮挡。
(b) 多尺度
由多层 DCAMA 块聚合的中间查询掩码,用于 1-shot
从 PASCAL-5i 采样的任务。

与其他三种方法（CWT [23]，
TRFS [35] 和 CyCTR [49])也实现了点积注意力
变压器[40]。图 4(a) 可视化了 DCAMA 的一些分割结果
在具有挑战性的情况下。补充材料中给出了更多的结果和可视化,包括区域分割的过度和不足的测量
[51]。

评论。虽然表 1 中表现最好的三种方法（HSNet [24]，
CyCTR [49] 和我们的)都依赖于像素级的交叉查询和支持相似性,它们的查询标签推断的基本概念是完全不同的,并且
值得澄清。 HSNet 根据相似度预测查询标签
查询像素到所有前景支持像素（同时忽略背景）;直观上,查询像素与前景支持越相似

像素,越有可能是前景。 CyCTR首先重构查询
基于与两者子集的相似性的支持特征中的特征
前景和背景支持像素,然后在重建的查询特征上训练分类器。我们的 DCAMA 直接从

由查询像素与所有支持像素的相似性加权的支持掩码，
代表一个完全不同的概念。

训练和推理效率。我们比较训练效率
我们的方法与 HSNet [24] 的方法相似。由于这两种方法都会产生 O(N2) 的像素相关性计算复杂度,因此它们也具有可比性
每个 epoch 的训练时间（例如,我们的
COCO-20i 上的硬件和培训设置）。然而,如图 5 所示,我们的
方法需要更少的训练时间来收敛。因此,我们的 DCAMA

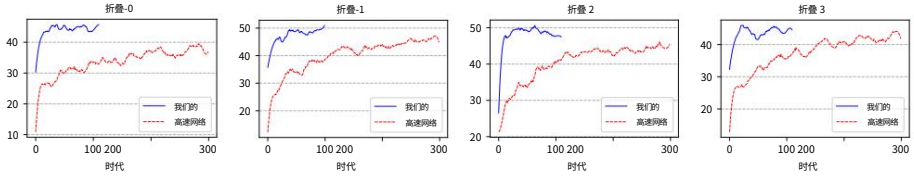

图 5. COCO-20i 训练期间验证集上的 mIoU 曲线。这
HSNet [24] 的曲线是使用作者发布的官方代码生成的。

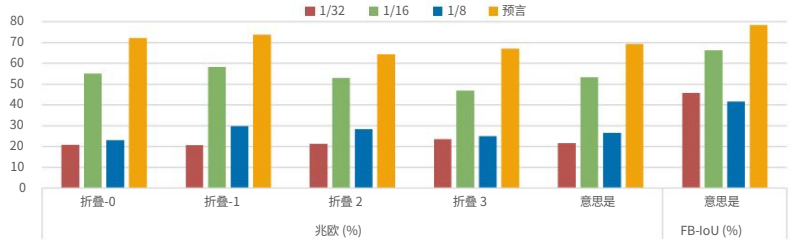

图 6. 聚合的多尺度中间查询掩码的性能
多层 DCAMA 块和最终预测（PASCAL-5i 上的 1-shot ）。

在训练时间方面也更有效率，除了达到更高
性能优于之前的 SOTA 方法 HSNet。至于推理，DCAMA
使用相同的骨干网以与 HSNet 相当的速度运行，例如，大约
分别使用 Swin-B 和 ResNet-101 每秒 8 帧和 20 帧 (FPS)，
用于入门级 NVIDIA Tesla T4 GPU 上的 1-shot 分割。相比之下，
CyCTR [49] 在 ResNet-50 主干上以大约 3 FPS 的速度运行。

4.2 消融研究

我们对 PASCAL-5i [31] 数据集进行彻底的消融研究，以获得
更深入地了解我们提出的 DCAMA 框架并验证其设计。
Swin-B [21] 被用作消融研究的主干特征提取器。

由多层 DCAMA 聚合的中间查询掩码
块。我们首先通过验证多层 DCAMA 块的输出来验证所提出的掩码聚合范式的物理意义

（图 3）确实是有意义的分割。为此，我们总结
尺度 i ∈ { } 的层维度，用 Otsu 对总和进行二值化
方法[27]，并调整生成的掩码大小以根据基本事实进行评估
用于 1-shot 分割。如图 6 所示，量表的 mIoU 和 FB-IoU
掩码相当高，接近表 1 中一些比较方法的掩码。
同时，规模面具的那些要低得多，这可能是原因：

$\frac{1}{8}$ 尺度特征没有学习足够的高级语义，并且
$\frac{1}{32}$ 比例特征过于抽象/特定于类且粗糙。最后的预测
有效整合多尺度中间掩膜，实现最优
表现。为了直观感知，我们还将中间掩码可视化
对于图 4(b) 中的特定 1-shot 任务，其中的叠加层清楚地展示了
它们是有效的分割，尽管不如最终预测准确。

表 3 消融效果研究
背景支持信息
（在 PASCAL-5i 上进行 1 次拍摄）。

背景	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FB-IoU
✓	72.2	73.8	64.3	67.1	69.3	78.5
	73.3	72.7	53.6	69.8	67.3	76.2

表 4. 消融策略的研究
n-shot 分割（PASCAL i上的 5-shot
5 ）。

策略	投票	平均	HSNet [24]	SCL [46]	我们的
mIoU	74.0	74.1	74.9	74.0	73.9
FB-IoU	82.0	82.0	82.9	82.0	81.8

这些结果验证了所提出的 DCAMA 确实按设计运行，
并且多尺度策略是有效的。

背景支持信息的影响。我们的 DCAMA 和之前的 SOTA HSNet [24] 之间的一个显着区别是 HSNet 忽略了
后台支持功能，而 DCAMA 使用它们来充分利用可用信息。为了评估差异的实际效果，

我们进行了一个烧蚀实验，其中背景像素被清零
在馈送到多层 DCAMA 块之前的支持特征图
训练和推理 类似于 HSNet，用于 1-shot 分割。结果
在表 3 中显示，忽略背景支持信息会导致减少
mIoU 和 FB-IoU 分别为 2.0% 和 2.3%，表明完全
利用支持集中的所有可用信息。

n-shot 推理策略。验证我们提议的有效性
one-pass n-shot 推理，我们将其 5-shot 性能与以下结果进行比较
五个 1-shot 预测的集合：朴素投票和平均，归一化
在 HSNet [24] 中投票，在 SCL [46] 中进行交叉引导平均。如表 4 所示，我们的策略优于所有集成，表明在
共同利用 FSS 的所有可用支持功能。

5 结论

在这项工作中，我们提出了一种基于度量学习的少镜头语义分割（FSS）的新范式：密集的交叉查询和支持注意
加权掩码聚合 (DCAMA)。此外，我们实施了 DCAMA
为了简单和高效，在 Transformer 结构中具有缩放的点生成注意力的框架。 DCAMA 框架不同于
以前的工作在三个方面：(i)它直接预测了一个掩码值
查询像素作为所有支持像素的掩码值的加法聚合，
由查询和支持特征之间的像素相似度加权； (二)
它充分利用了所有支持像素，包括前景和背景；
(iii) 它提出了高效且有效的一次性 n-shot 推理，该推理同时考虑了来自所有支持图像的像素。实验表明
我们的 DCAMA 框架为所有三种常用的
FSS 基准。对于未来的研究，很容易将范式调整为
其他涉及密集预测的小样本学习任务，例如检测。

确认。本工作得到国家重点研发计划（2018AAA0100104、2018AAA0100100）、江苏省自然科学基金（BK20211164)的资助。

参考

1. Azad, R.、Fayjie, AR、Kauffmann, C.、Ben Ayed, I.、Pedersoli, M.、Dolz, J.:关于少镜头 CNN 分割的纹理偏差。在:计算机视觉应用的 IEEE/CVF 冬季会议论文集。第 2674–2683 页（2021 年）

2. Boudiaf, M., Kervadec, H., Masud, ZI, Piantanida, P., Ben Ayed, I., Dolz, J.:没有元学习的小样本分割:你只需要一个好的转导推理吗?在:计算机视觉和模式识别 IEEE/CVF 会议论文集。第 13979-13988 页（2021 年）

3. Chen, LC, Papandreou, G., Kokkinos, I., Murphy, K., Yuille, AL:DeepLab:使用深度卷积网络、多孔卷积和全连接 CRF 进行语义图像分割。 IEEE 模式分析和机器智能汇刊 40(4), 834–848 (2017)

4. Chen, LC, Zhu, Y., Papandreou, G., Schroff, F., Adam, H.:用于语义图像分割的带有多孔可分离卷积的编码器-解码器。在:欧洲计算机视觉会议论文集。第 801-818 页（2018 年）

5. Cui, H., Wei, D., Ma, K., Gu, S., Zheng, Y.:用于稀缺数据的广义低样本医学图像分割的统一框架。 IEEE 医学影像汇刊 40(10), 2656–2671 (2021)

6. Deng, J., Dong, W., Socher, R., Li, LJ, Li, K., Fei-Fei, L.:ImageNet:大规模分层图像数据库。在:2009 年 IEEE 计算机视觉和模式识别会议。第 248-255 页。 IEEE (2009)

7. Dong, N., Xing, EP:带有原型学习的 Few-shot 语义分割。在:
英国机器视觉会议。卷。 3 (2018)
8. Dong, X.、Zhu, L.、Zhang, D.、Yang, Y.、Wu, F.:用于少镜头图像字幕和视觉问答的快速参数适应。在:ACM 国际多媒体会议论文集。第 54-62 页（2018 年）

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S ., Uszkoreit, J., Houlsby, N.:一张图像值 16x16 字:大规模图像识别的变形金刚。
国际学习代表大会（2021）
10. Everingham, M.、Van Gool, L.、Williams, CK、Winn, J.、Zisserman, A.:PASCAL 视觉对象类 (VOC) 挑战。国际计算机视觉杂志 88(2), 303–338 (2010)

11. Fei-Fei, L.、Fergus, R.、Perona, P.:对象类别的一次性学习。 IEEE 模式分析和机器智能汇刊 28(4), 594–611 (2006)
12. Fink, M.:利用类相关性度量的单个示例的对象分类。神经信息处理系统的进展 17, 449–456 (2005)

13. Hariharan, B.、Arbel´aez, P.、Girshick, R.、Malik, J.:同时检测和分割。在:欧洲计算机视觉会议论文集。第 297-312 页。施普林格 (2014)

14. He, K., Zhang, X., Ren, S., Sun, J.:用于图像识别的深度残差学习。在:IEEE 计算机视觉和模式识别会议论文集。第 770–778 页（2016 年）

15. Kulis, B. 等人:度量学习:一项调查。机器的基础和趋势®
学习 5(4), 287–364 (2013)
16. Lake, B.、Salakhutdinov, R.、Gross, J.、Tenenbaum, J.:简单视觉概念的一次性学习。在:认知科学学会年会论文集。卷。 33,第 2568-2573 页（2011 年）

17. Li, X., Wei, T., Chen, YP, Tai, YW, Tang, CK: FSS-1000: A 1000-class dataset
用于少镜头分割。在：计算机视觉和模式识别 IEEE/CVF 会议论文集。第 2869–2878 页（2020 年）

18. Lin, TY, Doll´ar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.：专题
用于目标检测的金字塔网络。在：计算机视觉和模式识别 IEEE/CVF 会议论文集。第 2117–2125 页（2017 年）

19. Lin, TY, Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P.,
Zitnick, CL：Microsoft COCO：上下文中的公共对象。在：诉讼程序
欧洲计算机视觉会议。第 740-755 页。施普林格 (2014)
20. Liu, Y., Zhang, X., Zhang, S., He, X.：用于少样本的部分感知原型网络
语义分割。在：欧洲计算机会议论文集
想象。第 142-158 页。施普林格 (2020)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transform: Hierarchical vision
Transformer using移动的窗户。在：诉讼程序
IEEE/CVF 国际计算机视觉会议。第 10012–10022 页
(2021)
22. Long, J., Shelhamer, E., Darrell, T.：语义的全卷积网络
分割。在：IEEE/CVF 计算机视觉会议论文集
和模式识别。第 3431–3440 页（2015 年）
23. Lu, Z., He, S., Zhu, X., Zhang, L., Song, YZ, Xiang, T.：越简单越好：使用分类器权重 Transformer 进行语义
分割。在：诉讼
IEEE/CVF 国际计算机视觉会议的成员。第 8741–8750 页
(2021)
24. Min, J., Kang, D., Cho, M.：用于少镜头分割的超相关挤压。
在：IEEE/CVF 计算机视觉国际会议论文集
(2021)
25. Minaee, S., Boykov, YY, Porikli, F., Plaza, AJ, Kehtarnavaz, N., Ter zopoulos, D.：使用深度学习进行图像
分割：一项调查。 IEEE
模式分析和机器智能交易第 1-1 页（2021 年）。
https://doi.org/10.1109/TPAMI.2021.3059968
26. Nguyen, K., Todorovic, S.：特征加权和提升少镜头分割。在：IEEE/CVF 国际计算机会议论文集

想象。第 622–631 页（2019 年）
27. Otsu, N.：灰度直方图的阈值选择方法。 IEEE Trans
对系统、人和控制论的行动 9(1), 62–66 (1979)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen,
T., Lin, Z., Gimelshein, N., Antiga, L. 等人：PyTorch：一种命令式风格的高性能深度学习库。神经信息处理系
统的进展 32, 8026–8037 (2019)

29. Ravi, S., Larochelle, H.：优化作为少样本学习的模型。在：国际米兰
全国学习代表大会（2017）
30. Ronneberger, O., Fischer, P., Brox, T.：U-Net：用于生物医学图像分割的卷积网络。在：医学图像计算国际会议

和计算机辅助干预。第 234-241 页。施普林格 (2015)
31. Shaban, A.、Bansal, S.、Liu, Z.、Essa, I.、Boots, B.：语义的一次性学习
分割。在：英国机器视觉会议论文集。第 167.1 页–
167.13 (2017)
32. Simonyan, K., Zisserman, A.：用于大规模的非常深的卷积网络
图像识别。 arXiv 预印本 arXiv:1409.1556 (2014)
33. Snell, J.、Swersky, K.、Zemel, R.：用于小样本学习的原型网络。在：
神经信息处理系统的进展。第 4080–4090 页（2017 年）

34. Strudel, R.、Garcia, R.、Laptev, I.、Schmid, C.:Segmenter:语义分割变压器。在:IEEE/CVF 计算机视觉国际会议论文集。第 7262–7272 页（2021 年）

35. Sun, G.、Liu, Y.、Liang, J.、Van Gool, L.:使用 Transformers 提升少镜头语义分割。 arXiv 预印本 arXiv:2108.02266 (2021)
36. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, PH, Hospedales, TM:学习比较:用于少样本学习的关系网络。在:计算机视觉和模式识别 IEEE/CVF 会议论文集。第 1199–1208 页（2018 年）

37. Taghanaki, SA, Abhishek, K., Cohen, JP, Cohen-Adad, J., Hamarneh, G.:自然和医学图像的深度语义分割:综述。人工智能评论 54(1), 137–178 (2021)

38. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.:用于少镜头分割的先验引导特征丰富网络。 IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1-1 (2020)

39. Triantafillou, E.、Zemel, R.、Urtasun, R.:通过信息检索镜头进行少量学习。在 :神经信息处理系统的进展。第 2252-2262 页（2017 年）

40. Vaswani, A.、Shazeer, N.、Parmar, N.、Uszkoreit, J.、Jones, L.、Gomez, AN、Kaiser, L.、Polosukhin, I.:注意力就是你所需要的。在 :神经信息处理系统的进展。第 5998–6008 页（2017 年）

41. Veli ˇckovi´c, P., Cucurull, G., Casanova, A., Romero, A., Li`o, P., Bengio, Y.:图注意力网络。在:国际学习表征会议（2018）,https://openreview.net/forum?id=rJXMpikCZ

42. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. 等人:一次性学习的匹配网络。神经信息处理系统的进展 29, 3630–3638 (2016)

43. Wang, H.、Zhang, X.、Hu, Y.、Yang, Y.、Cao, X.、Zhen, X.:带有民主注意力网络的 Few-shot 语义分割。在:欧洲计算机视觉会议论文集。第 730-746 页。施普林格 (2020)

44. Wang, K., Liew, JH, Zou, Y., Zhou, D., Feng, J.:PANet:Few-shot 图像语义分割与原型对齐。在:IEEE/CVF 计算机视觉国际会议论文集。第 9197–9206 页（2019 年）

45. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.:用于少镜头语义分割的原型混合模型。在 :欧洲计算机视觉会议论文集。第 763-778 页。施普林格 (2020)

46. Zhang, B., Xiao, J., Qin, T.:用于少镜头分割的自引导和交叉引导学习。在:计算机视觉和模式识别 IEEE/CVF 会议论文集。第 8312–8321 页（2021 年）

47. Zhang, C.、Lin, G.、Liu, F.、Guo, J.、Wu, Q.、Yao, R.:具有连接注意的金字塔图网络,用于基于区域的一次性语义分割。在:IEEE/CVF 计算机视觉国际会议论文集。第 9587–9595 页（2019 年）

48. Zhang, C.、Lin, G.、Liu, F.、Yao, R.、Shen, C.:CANet:具有迭代细化和细心小样本学习的类不可知分割网络。在:计算机视觉和模式识别 IEEE/CVF 会议论文集。第 5217–5226 页（2019 年）

49. Zhang, G.、Kang, G.、Wei, Y.、Yang, Y.:通过循环一致的 Transformer 进行少镜头分割。 arXiv 预印本 arXiv:2106.02320 (2021)
50. Zhang, X., Wei, Y., Yang, Y., Huang, TS:SG-One:用于一次性语义分割的相似性指导网络。 IEEE 控制论汇刊 50(9), 3855–3865 (2020)

51. Zhang, Y., Mehta, S., Caspi, A.:重新思考语义分割评估
可解释性和模型选择。 arXiv 预印本 arXiv:2101.08418 (2021)
52. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.:金字塔场景解析网络。在：
IEEE/CVF 计算机视觉和模式识别会议论文集。第 2881–2890 页（2017 年）

53. Zhu, F., Zhu, Y., Zhang, L., Wu, C., Fu, Y., Li, M.:一个统一的有效金字塔
用于语义分割的转换器。在：IEEE/CVF 国际计算机视觉会议论文集。第 2667–2677 页（2021 年）