◆特邀栏目◆

基于特征融合注意力的小样本语义分割算法*

李呓瑾[1],李少龙[1**],贺　彦[2],刘　炜[3]

(1.云南电网有限责任公司信息中心,云南昆明　650200;2.北京国科恒通科技股份有限公司,北京　100085;3.清华大学电机工程与应用电子技术系,北京　100084)

摘要:针对小样本语义分割任务中对查询图片的信息利用不充分的问题,提出一种基于特征融合注意力的小样本语义分割算法。首先,利用共享主干网络编码支持图片和查询图片,从而获取图片的深度特征;然后,利用注意力机制获取支持特征和查询特征的强关联语义信息,从而构造任务注意力特征图;最后,提出一种多特征注意力融合模块,它能够自适应融合多种特征的深层语义信息并进行特征解码,从而获取目标物体的分割掩码。在 PASCAL-5^i 和 COCO-20^i 公开数据集进行了实验,结果表明,所提出模型比当前主流的小样本语义分割模型在 1-way 1-shot 和 1-way 5-shot 任务中分割得更加精准,尤其是在更具有挑战性的 COCO-20^i 数据集上,所提出模型在 1-shot 的设定下达到了 28.8%的 mIoU 和 62.1%的 FB-IoU,在 5-shot 设定下达到了 36.9%的 mIoU 和 64.8%的 FB-IoU。

关键词:小样本语义分割;多特征融合;注意力机制;深层语义信息;分割掩码

中图分类号:TP393　　**文献标识码**:A　　**文章编号**:1005-9164(2023)05-0951-10

DOI:10.13656/j.cnki.gxkx.20231121.014

语义分割的目的是为图像中的每个像素点分配一个类标签,在医疗诊断、无人驾驶、图片编辑等领域具有广泛的应用前景[1]。近年来,基于深度学习的语义分割模型取得了突破性的进展,如全卷积神经网络 FCN、DeepLab、UNet 和 PSPNet 等[2],并利用膨胀卷积来增大感受野[3],从而增强了模型的分割性能。然而,基于深度学习的主流语义分割模型需要大量逐像素的标注数据,这类标注数据的获取费时费力成本高。虽然弱监督学习方法可以缓解模型对像素级标注的依赖,但仍然需要大量的弱标注数据[4]。此外,基于弱监督学习方法的模型对于新类或标注不充分的目标类的泛化性能较差。

受小样本学习的启发,Shaban 等[5]提出了一种基于双分支结构的小样本语义分割模型,其中支持分支将支持图片及对应的标注掩码作为输入,学习指导信息;查询分支以查询图片为输入,预测对应的分割

收稿日期:2022-10-15　　修回日期:2022-12-07

*国家自然科学基金项目(62222209)资助。

【第一作者简介】

李呓瑾(1991-),女,硕士,工程师,主要从事深度学习、计算机及其自动化研究。

【**通信作者简介】

李少龙(1985-),男,高级工程师,主要从事图像处理研究,E-mail:lujy_1979@163.com。

【引用本文】

李呓瑾,李少龙,贺彦,等.基于特征融合注意力的小样本语义分割算法[J].广西科学,2023,30(5):951-960.

LI Y J,LI S L,HE Y,et al.Few-shot Semantic Segmentation Based on Feature Fusion Attention Mechanism [J].Guangxi Sciences,2023,30(5):951-960.

掩码。之后,众多研究者基于该双分支结构去构造各种变体分割模型,提高模型对目标物体的分割性能[6,7]。

现有的小样本语义分割模型主要包括度量学习和元学习两大类[8]。度量学习方法首先利用支持分支获得每一像素的特征表示,然后通过计算查询图片中每一像素与特征表示之间的相似度,给出查询图片中目标物体的分割。Wang 等[9]提出了一种特征对齐的小样本语义分割模型,该模型利用全局平均池化策略获取支持图片的全局特征并将其作为目标物体的特征表示。Zhang 等[10]针对全局平均池化策略极易造成目标背景对前景干扰的问题,提出了一种掩码平均池化策略,该策略利用支持图片的掩码来分离目标的前景和背景,充分挖掘前景信息从而提高分割的性能。然而,仅利用全局平均特征构造单一原型不足以充分表示目标的不同部分。为此,Liu 等[11]提出了一种多特征表示的小样本语义分割模型,其通过均等划分支持分支中目标物体的多个区域并借助掩码平均池化获取每个区域的平均特征,从而构造目标物体的多个特征表示。类似地,Li 等[12]提出了一种自适应多特征表示的小样本语义分割模型,其借助支持分支的掩码信息获取多个不同大小的目标区域,并利用掩码池化策略提取每个区域的特征表示。贾熹滨等[13]提出了一种金字塔原型对齐的小样本语义分割模型,它通过提取不同尺度的特征构造目标物体的特征表达,并通过计算特征表达与查询分支目标物体之间的相似度分割目标区域。Liu 等[14]设计了一种动态原型卷积网络来构造类的特定多原型表示,该网络能够充分捕获目标的细节特征从而提高分割性能,并且可以被应用在小目标、多目标等复杂场景中。

元学习又称“学习如何学习”,指的是通过在多个任务中学习分割先验知识(一组参数)来指导新任务的过程,旨在强化模型的泛化性能。Liu 等[15]提出了一种基于 Transformer 的小样本语义分割模型,该模型利用 Transformer 编码块提供的注意力机制,构造动态权重的分类器,并采用预训练模型固化的策略,解决了小样本数据样本不足的问题。刘宇轩等[16]针对支持图像和查询图像共性信息利用不足的问题,提出了一种结合全局和局部特征的小样本语义分割模型,并在 PASCAL 数据集上验证了其有效性。Pambala 等[17]提出了一种基于元学习的小样本语义分割模型,在视觉信息的基础上通过引入文本语义信息来构造多尺度的融合特征,并利用解码器分割目标物体。虽然上述模型取得了一定的成功,但高质量的特征提取更能够进一步提升下游的分割性能。Tian 等[18]在特征编码阶段提出了一种强区分性的元学习模块,通过挖掘局部和全局特征图来进一步提高特征的表达能力。Wu 等[19]在支持分支和查询分支特征提取模块之后引入了一种元记忆学习模块,通过学习记忆支持图片和查询图片的相似性语义信息,来强化特征的表达能力。

虽然基于度量学习的小样本语义分割模型结构简单,参数较少,但其分割性能过度依赖于原型的质量。此外,仅利用无参数的相似度度量计算极易导致信息丢失或歧义。基于元学习的小样本语义分割模型虽然可以解决上述性能不佳的问题,但现有的该类分割模型主要聚焦于支持分支中支持图片的前景和背景的信息挖掘,而无法有效利用查询图片的信息。因此,本文提出的多特征融合的小样本语义分割模型仍然采用元学习的架构并在充分利用支持图片的前景、背景信息的同时,进一步挖掘查询分支中查询图片的信息,从而解决了现有的基于元学习模型对于查询图片信息利用不充分的问题。具体来说,所提出模型首先通过引入注意力模块从支持图片和查询图片中学习语义关联特征图;然后,借助掩码平均池化方法提取支持图片中前景和背景信息的全局特征;最后,利用特征融合模块将前景和背景特征进行尺度融合,并将其融合特征作为解码块的输入实现查询图片的掩码预测。

1 任务定义

本文的目标是利用少量带标注的图像学习一个小样本语义分割模型 θ,实现对未知类的准确分割。首先,从两个互不相交的集合 C_{seen} 和 C_{unseen}($C_{\text{seen}} \cap C_{\text{unseen}} = \varnothing$)中构造训练集 D_{train} 和测试集 D_{test}。其中训练集和测试集由多个 episode 组成,即:$D_{\text{train}} = \{(S_i, Q_i)\}_{i=1}^{N_{\text{train}}}$,$D_{\text{test}} = \{(S_i, Q_i)\}_{i=1}^{N_{\text{test}}}$。

θ 的训练和测试采用 episodic 机制,每个 episodic 可以简化为一个 C-way K-shot 的学习任务。支持集 S_i 由 K 张图片及其对应的掩码组成,即 S = (image, mask);类似地,查询集 Q_i 由与支持集语义类相同的 L 张图片组成 Q = (image, mask)。然而,测试阶段中的查询图片没有对应的 mask,即 image$\in \mathbb{R}^{3\times h\times w}$,mask$\in \mathbb{R}^{h\times w}$。

2 小样本语义分割模型

2.1 模型结构

图1给出了所提出模型的结构，包括特征提取、语义关联注意力模块、多特征融合模块和解码器等4部分。首先，使用同一个特征提取器编码查询图片和支持图片；其次，利用注意力机制来映射支持特征和查询特征之间的强语义相关；最后，通过特征融合注意力模块融合多个不同来源的特征，并利用解码块分割查询图片。

2.2 特征提取

特征提取的质量直接影响分割的效果。虽然浅层特征携带颜色、边缘等低层次线索，但其在语义层面上的区分性不强；而高层特征尽管具有较强的语义类区分能力，但难迁移到对未知类的分割任务中。此外，小目标物体由于下采样操作，其信息极易丢失，从而影响分割效果。为此，本文通过构造多尺度特征提取网络来强化深度特征空间的语义表达能力。特征提取网络结构如图2所示。

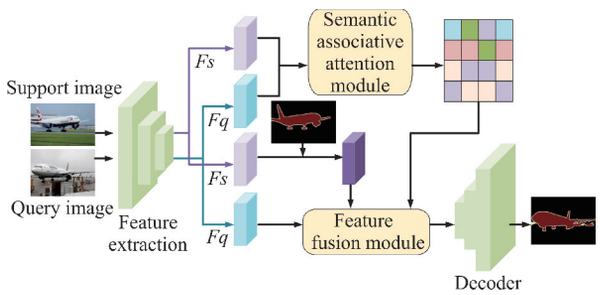

图1 模型架构

Fig. 1 Structure of proposed model

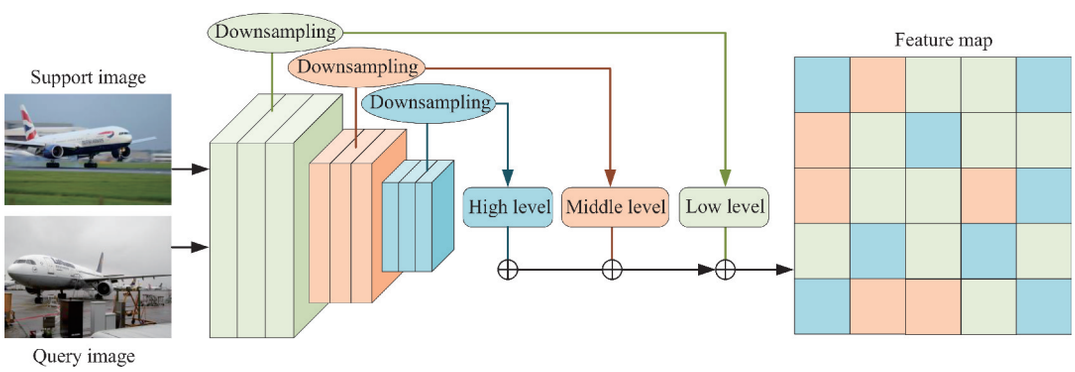

图2 多尺度特征提取流程

Fig. 2 Process of the multi-scale feature extraction

具体来说，首先利用在ImageNet上预训练的VGG-16、ResNet-50和ResNet-101分别作为主干网络进行多尺度特征提取。多尺度特征表示如公式(1)所示。

$$Fs = fs_l \oplus fs_m \oplus fs_h, \tag{1}$$

其中，Fs 表示融合后的特征；fs_l 表示低层特征，fs_m 表示中间层特征，fs_h 表示高层特征；⊕表示向量的连接操作。

考虑到深度卷积网络对于硬件要求较高，此处采用深度可分离卷积代替主干网络中的部分卷积块，通过减少模型参数计算量来降低运行时间开销，深度可分离卷积的结构如图3所示。具体来说，首先利用主干网络的第一个block将输入的支持图片 I_s 和查询图片 I_q 映射到浅层特征空间，获得特征图 F_o；然后，作为深度卷积(Depth-Wise Convolution，DWC)的输入沿着通道维度进行分解。将标准卷积过程分解成 K 个等效的DWC和 N 个逐点卷积(Pointwise Convolution，PC)，值得一提的是PC能够在保持特征提取质量不变的基础上降低计算复杂度。此处，DWC的输出特征图表示为 F_{dw}，支持图片 I_s 和查询图片 I_q 的深度融合特征表示为 F_{pc}。最后，将主干网络中每个block替换成图3所示的卷积结构，从而获得最终的支持图片融合特征 Fs 和查询图片融合特征 Fq。值得注意的是，此处特征包括前景区域特征和背景区域特征。

2.3 语义关联注意力

现有的小样本语义分割模型大多使用从支持分支中提取类的特征表达来指导查询图像的分割[9,12]。然而当带标注的支持图片数量有限时，极易导致从中提取的特征不足以表达类的强语义相关性，从而使得模型对于查询图片的预测不精确。为了解决该问题，受注意力机制在视觉和文本领域成功应用的启发，本文提出一种语义关联注意力模块(图4)。

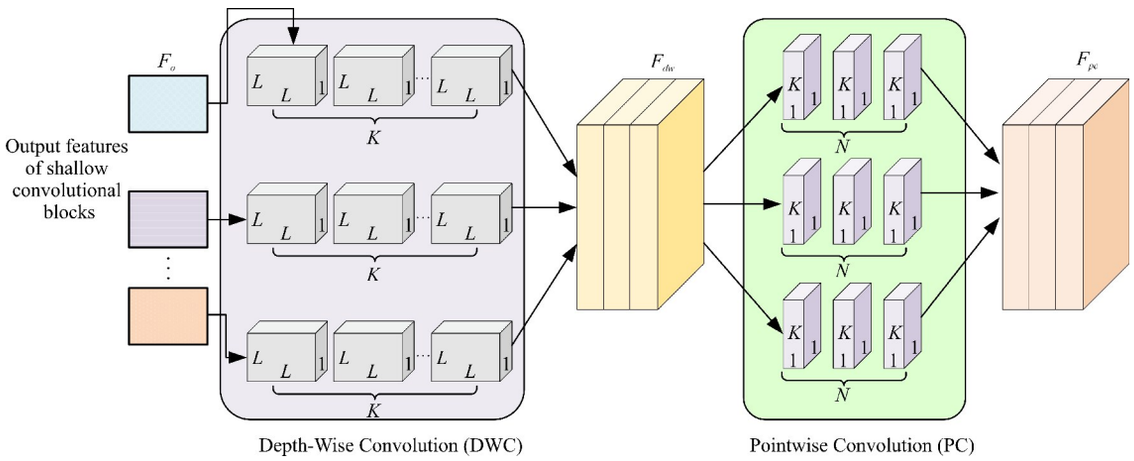

图 3　深度可分离卷积

Fig. 3　Deep separable convolution

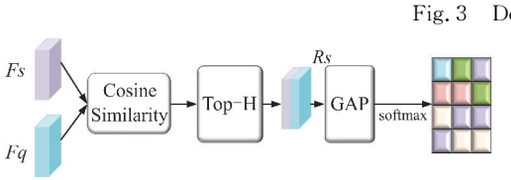

图 4　语义关联注意力模块

Fig. 4　Semantic associative attention module

考虑到支持分支和查询分支中的图片具有相同的语义信息,因此利用图 4 所示的语义关联注意力模块获取新的强语义特征。首先,将 $\{Fs, Fq\} \in \mathbb{R}^{512\times56\times56}$ 作为注意力模块的输入;然后,利用余弦相似度计算支持特征和查询特征的 Top-H 得分,从而选取最相似的 H 个特征构造新的语义特征。利用以上语义关联注意力模块获取的语义特征能够在特征融合阶段帮助定位查询图片的目标区域,从而有效解决现有模型对于目标区域定位不精确的问题。

图 4 中,模块首先计算了支持特征 Fs 和查询特征 Fq 在每一位置(x,y)处的余弦相似度[11],计算公式如下:

$$Ds = \frac{Fs^{x,y} \cdot Fq^{x,y}}{\| Fs^{x,y} \| \cdot \| Fq^{x,y} \|}。\tag{2}$$

然后,利用公式(3)选择得分最高的 H 个相似性特征构造语义关联特征 $Rs \in \mathbb{R}^{K\times56\times56}$。

$$Rs = \mathrm{argmax}_{Top-H}(Ds)。\tag{3}$$

最后,利用全局平均池化(Global Average Pooling,GAP)[20]获取每个特征区域的平均特征 $Rs' \in \mathbb{R}^{K\times1}$,并利用 softmax 函数沿通道方向计算每个位置的语义注意力 As_i,计算公式如下:

$$As_i = \frac{e^{wR'}}{\sum_i^{H,W} e^{wR'}}。\tag{4}$$

将注意力与对应特征进行关联,得到支持特征与查询特征强语义关联的整张注意力图 $\{As\}_i^K \in \mathbb{R}^{56\times56}$。

2.4　多特征注意力融合模块

由于拍摄角度、颜色变化和遮挡等外界因素的影响,来自同一语义类的支持图片和查询图片会存在很大差异,这极大地增加了分割的难度。因此,本文设计了一种多特征注意力融合模块,在支持图片和查询图片语义注意力生成图的基础上,该模块通过融合支持图片掩码过滤后的前景信息和查询图片的深度语义信息来进一步强化支持图片和查询图片之间的语义关联。具体来说,特征注意力融合模块的输入端包括语义关联注意力特征 $Rs \in \mathbb{R}^{K\times56\times56}$、掩码后的支持特征 $Fs' \in \mathbb{R}^{512\times448\times448}$ 和查询特征 $Fq \in \mathbb{R}^{512\times56\times56}$ 这 3 个分支。此处借助注意力机制的思想,融合(R, Fs')、(R, Fq)和(Fs', Fq)3 个不同尺度的特征,并将融合后的特征进行归一化操作,得到最终的融合特征 $FF \in \mathbb{R}^{512\times56\times56}$。特征融合流程如图 5 所示。此处,以$\{Fs', Fq\}$为例说明特征融合的流程。

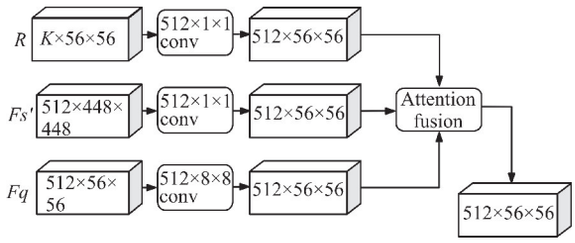

图 5　特征融合注意力

Fig. 5　Feature fusion attention

首先,将$\{Fs', Fq\} \in \mathbb{R}^{512\times56\times56}$ 利用 reshape 函数转换为$\{Fs', Fq\} \in \mathbb{R}^{512\times N}$,$N = 56\times56$,并利用公式(5)计算单特征之间的相似性矩阵 $M_1 \in \mathbb{R}^{512\times512}$。

$$M_1 = Fs'Fq^{\mathrm{T}}。\tag{5}$$

然后利用 softmax 函数按行计算注意力权重 W，并与单特征矩阵相乘获得融合特征 F_1。计算如公式(6)所示。

$$F_1 = \text{softmax}(M_1) \times Fs' + \text{softmax}(M_1) \times Fq。 \tag{6}$$

接着利用点乘运算将所有特征两两融合，得到最终的强语义特征 $FF = F_1(Fs', Fq) \cdot F_2(R, Fq) \cdot F_3(R, Fs)$。最后，将强语义特征 $FF \in \mathbb{R}^{512\times56\times56}$ 作为解码器的输入，得到查询图片的分割掩码，并利用交叉熵损失函数实现模型端到端的优化。

3 实验与结果分析

3.1 数据集

采用经典的 PASCAL-5^i[21] 和 COCO-20^i[22] 数据集进行模型训练与测试。其中 PASCAL-5^i 包含 5 953 张训练图片和 1 449 张测试图片，共包含 20 个类，划分为 4 折，其中 3 折(15 类)用于训练，剩余用于测试；COCO-20^i 每张图片中包含的类别更多，包括 82 081 张训练图片和 40 137 张测试图片，共包含 80 类，其中 3 折(60 类)用于训练，剩余用于测试。数据集详细信息如表 1 所示，部分样本如图 6 所示。

表 1 数据集描述

Table 1 Dataset description

数据集 Dataset	类 Categories
PASCAL-5^0	Areoplane, Bicycle, Bird, Boat, Bottle
PASCAL-5^1	Bus, Car, Cat, Chair, Cow
PASCAL-5^2	Diningtable, Dog, Horse, Motorbike, Person
PASCAL-5^3	Potted plant, Sheep, Sofa, Train, TV/Monitor
COCO-20^0	Person, Airplane, Boat, Parking meter, Dog, Elephant, Hat, Backpack, Suitcase, Sports ball, Skateboard, Chair, Wine glass, Spoon, Sandwich, Hot dog, Dining table, Mouse, Microwave, Scissors
COCO-20^1	Bicycle, Bus, Traffic light, Bench, Horse, Bear, Umbrella, Shoe, Teddy bear, Frisbee, Kite, Surfboard cup, Bowl, Orange, Pizza, Couch, Toilet, Remote, Oven, Book
COCO-20^2	Car, Train, Fire Hydrant, Bird, Sheep, Zebra, Handbag, Skis, Baseball bat, Tennis racket fork, Banana, Broccoli, Donut, Potted plant, TV, Keyboard, Sink, Toaster, Clock, Hair drier
COCO-20^3	Motorcycle, Truck, Stop Sign, Cat, Cow, Giraffe, Tie, Snowboard, Baseball glove, Bottle Knife, Apple, Carrot, Cake, Bed, Laptop, Cell phone, Refrigerator, Vase, Toothbrush

图 6 样本图片

Fig. 6 Example images

3.2 实验环境与评价指标

硬件：Nvidia A100 8×40 GB GPU。软件：Python 3.7，pytorch 深度学习框架。利用文献[7]中的 VGG-16 和文献[15]中的 ResNet-50 和 ResNet-101 等预训练模型作为特征提取器。输入图片的大小为 448 × 448；优化器采用 Adam；初始学习率为 0.000 1，权重衰减因子为 0.000 5，batch_size 为 8。

采用平均交并比 mIoU (mean Intersection over Union)和前景背景二分类交并比 FB-IoU(Foreground and Background IoU)[5] 作为评价指标。其中，mIoU 为所有类别真实值和预测值的交集和并集之比，FB-IoU 表示类别为 2 的二分类任务。计算公式如(7)所示。

$$\text{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}(p_{ji} - p_{ii})}。 \tag{7}$$

其中，$k+1$ 表示识别的总类别数，i 表示真实标记，j 表示预测标注，p_{ij} 表示将 i 预测为 j。

3.3 对比实验

3.3.1 PASCAL-5^i

为验证所提出模型的优越性，在 PASCAL-5^i 数据集上与当前主流模型进行对比实验，详细结果见表 2。当 VGG-16 作为主干网络时，所提出模型在 1-

shot 任务中可以获得 50.6%的 mIoU 和 69.2%的 FB-IoU。在 5-shot 任务中,虽然在 mIoU 评价指标下所提出模型略逊于 PANet,但在 FB-IoU 指标下所提出模型比 PANet 提高了 0.84%;当 ResNet-50 作为主干网络时,所提出模型在 1-shot 和 5-shot 任务上可以达到 59.4%和 60.1%的 mIoU 以及 72.8%和 73.4%的 FB-IoU,整体优势明显;当 ResNet-101 作为主干网络时,在 1-shot 和 5-shot 任务中,尽管在 mIoU 评价指标下所提出模型略逊于 GL 模型,但在 FB-IoU 指标下,所提出模型分别在 1-shot 和 5-shot 任务中比 GL 模型提高了 1.22%(73.8%→74.7%)和 0.94%(74.8%→ 75.5%)。

表 2　1-way 1-shot 和 1-way 5-shot 在 PASCAL-5^i 上的分割结果

Table 2　Segmentation results for 1-way 1-shot and 1-way 5-shot on PASCAL-5^i　　**Unit:%**

方法 Methods	主干网络 Backbones	1-shot						5-shot					
		P^0	P^1	P^2	P^3	mIoU	FB-IoU	P^0	P^1	P^2	P^3	mIoU	FB-IoU
OSLSM[5]	VGG-16	33.6	55.3	40.9	33.5	40.8	61.3	35.9	58.1	42.7	39.1	44.0	61.5
SG-One[11]		40.2	58.4	48.4	38.4	46.3	63.1	41.9	58.6	48.6	39.4	47.1	65.9
PANet[9]		42.3	58.0	51.1	41.2	48.2	66.5	51.8	64.6	59.8	46.5	**55.7**	70.7
Ours		43.9	56.4	55.8	46.3	**50.6**	**69.2**	49.3	59.2	59.7	46.9	53.8	**71.3**
CANet[23]	ResNet-50	52.5	65.9	51.3	51.9	55.4	67.2	55.5	67.8	51.9	53.2	57.1	69.6
LTM[24]		54.6	65.6	56.6	51.3	57.0		56.4	66.6	56.9	56.8	59.2	
PGNet[25]		56.0	66.9	50.6	50.4	56.0		57.7	68.7	52.9	54.6	58.5	
GL[16]		54.6	67.8	57.4	52.1	58.0	72.5	54.8	68.1	59.9	56.2	59.8	73.1
Ours		55.9	69.2	59.4	53.1	**59.4**	**72.8**	56.2	69.8	59.9	54.3	**60.1**	**73.4**
FWB[26]	ResNet-101	51.3	64.5	56.7	52.2	56.2		54.9	67.4	62.2	55.3	59.9	
DAN[27]		54.7	68.6	57.8	51.6	58.2		57.9	69.0	60.1	54.9	60.5	
GL[16]		57.5	68.7	58.7	54.5	**59.9**	73.8	58.1	69.8	60.8	58.9	**61.9**	74.8
Ours		54.3	67.6	61.2	55.1	59.6	**74.7**	56.9	68.3	62.4	56.0	60.9	**75.5**

Note:bold fonts represent the optimal results.

3.3.2　COCO-20^i

为进一步验证所提出模型的有效性,在更具有挑战性的 COCO-20^i 数据集上与当前主流模型进行对比实验,详细结果见表 3。当 VGG-16 作为主干网络时,所提出模型在 1-shot 任务中可以达到 35.8%的 mIoU 和 61.6%的 FB-IoU;在 5-shot 任务中,所提出模型略逊于 FFNet;当 ResNet-50 作为主干网络时,所提出模型在 1-shot 和 5-shot 任务上可以达到 38.6%和 41.3%的 mIoU 以及 64.2%和 66.8%的 FB-IoU;相比 1-shot 的分割任务,5-shot 的 mIoU 增加 7.0%(38.6%→41.3%),这表明所提出模型可以充分利用支持图片的信息来提高模型的分割性能。当 ResNet-101 作为主干网络时,所提出模型在 mIoU 和 FB-IoU 评价指标下,相比所有对比模型,整体优势明显。

3.3.3　2-way 1-shot 和 2-way 5-shot

为了验证所提出模型的鲁棒性,选择 ResNet-50 作为主干网络,PANet[9] 为对比模型,在 PASCAL-5^i 和 COCO-20^i 数据集上分别进行 2-way 1-shot 和 2-way 5-shot 的对比实验,结果详见表 4。所提出模型在 PASCAL-5^i 和 COCO-20^i 数据集上整体 2-way 的分割性能明显强于 PANet。具体地,在 PASCAL-5^i 数据集上,所提出模型在 1-shot 任务上可以达到 48.3%、55.1%和 56.8%的 mIoU,相比 PANet 至少提升 10.9%;在 5-shot 任务上可以达到 50.6%、57.1%和 58.3%的 mIoU。在更具有挑战性的 COCO-20^i 数据集上,所提出模型在 5-shot 任务上可以达到 36.8%、37.9%和 32.9%的 mIoU,相比 PANet 至少提升 9.3%。上述结果进一步验证了所提出模型更加鲁棒。

表 3　1-way 1-shot 和 1-way 5-shot 在 COCO-20^i 上的分割结果

Table 3　Segmentation results for 1-way 1-shot and 1-way 5-shot on COCO-20^i　　**Unit:%**

方法 Methods	主干网络 Backbones	1-shot						5-shot					
		C^0	C^1	C^2	C^3	mIoU	FB-IoU	C^0	C^1	C^2	C^3	mIoU	FB-IoU
PANet[9]	VGG-16					20.9	59.2					29.7	63.5
FFNet[28]		34.4	37.0	36.9	34.7	35.7	60.1	36.8	37.8	38.3	36.6	**37.6**	**64.0**
Ours		33.6	37.4	36.2	35.9	**35.8**	**61.6**	36.9	38.1	37.9	36.3	37.3	63.5
PANet[9]	Resnet-50	31.5	22.6	21.5	16.2	23.0		45.9	29.2	30.6	29.6	33.8	
PMMs[29]		29.3	34.8	27.1	27.3	29.6		33.0	44.5	30.3	33.3	34.3	
FFNet[28]		36.7	40.4	42.3	36.4	**39.0**	63.0	38.9	41.5	44.5	39.1	41.0	65.1
Ours		36.1	38.2	40.6	39.5	38.6	**64.2**	39.2	41.3	43.9	40.6	**41.3**	**66.8**
DAN[27]	Resnet-101					24.4						29.6	
FWB[26]		19.9	18.0	21.0	28.9	21.2		19.1	21.5	23.9	30.1	23.7	
PRNet[11]		41.1	26.3	24.7	21.9	28.5	61.4	50.4	32.8	31.8	29.1	36.0	64.6
Ours		39.6	27.4	26.1	21.3	**28.8**	**62.1**	46.2	36.7	34.2	30.5	**36.9**	**64.8**

Note: bold fonts represent the optimal results.

表 4　2-way 1-shot 和 2-way 5-shot 在 PASCAL-5^i 和 COCO-20^i 上的分割结果

Table 4　Segmentation results for 2-way 1-shot and 1-way 5-shot on COCO-20^i　　**Unit:%**

数据集 Datasets	设定 Setting	方法 Methods	mIoU			FB-IoU		
			VGG-16	ResNet-50	ResNet-101	VGG-16	ResNet-50	ResNet-101
PASCAL-5^i	1-shot	PANet	41.3	47.2	51.2	64.3	66.8	69.2
		Ours	**48.3**	**55.1**	**56.8**	**67.9**	**70.2**	**71.5**
	5-shot	PANet	47.2	49.5	53.1	66.4	67.8	71.1
		Ours	**50.6**	**57.1**	**58.3**	**70.0**	**71.4**	**73.6**
COCO-20^i	1-shot	PANet	21.8	29.2	**29.7**	59.2	60.0	61.8
		Ours	**32.4**	**34.2**	28.1	**60.2**	**63.1**	**63.6**
	5-shot	PANet	29.4	33.6	30.1	61.9	64.2	62.7
		Ours	**36.8**	**37.9**	**32.9**	**62.1**	**64.7**	**63.0**

Note: bold fonts represent the optimal results.

3.3.4　可视化结果

图 7 是所提出模型和 PANet 模型的分割可视化结果。从图 7 可以看出，虽然 PANet 可以准确地找到待分割目标的位置，但仍会出现丢失部分细节或误分割的情况。相比而言，所提出模型可以准确地将轮船、飞机、猫和自行车等类别从查询图片中分割出来，整体分割效果更好，可视化结果进一步验证了所提出的多特征融合的小样本语义分割模型的优越性。

图 7　分割可视化结果

Fig. 7　Segmentation visualization results

3.4 消融实验

首先,为了降低模型的参数量,所提出模型采用深度可分离卷积代替主干网络中的卷积块。为了探究深度可分离卷积对整体性能的影响,在 1-way 1-shot 任务上利用 PASCAL-5^i 和 COCO-20^i 数据集,采用 ResNet-50 作为主干网络设计消融实验。实验结果如表 5 所示,利用深度可分离卷积块代替标准卷积块的设计虽然对 mIoU 和 FB-IoU 评价指标的提升效果不明显,但是能够极大地减少模型参数量,从而有效地解决模型对硬件过于依赖的问题。

表 5 深度可分离卷积消融实验

Table 5 Ablation experiments of the deep separable convolution

模块 Module	PASCAL-5^i			COCO-20^i		
	参数量/MB Number of parameters/MB	mIoU/%	FB-IoU/%	参数量/MB Number of parameters/MB	mIoU/%	FB-IoU/%
Convolution	481	58.9	72.5	481	38.4	64.1
Deep separable convolution	**126**	**59.4**	**72.8**	**126**	**38.6**	**64.2**

Note: bold fonts represent the optimal results.

其次,为了探究语义关联注意力模块对模型性能的影响,在 1-way 1-shot 任务上利用 PASCAL-5^i 和 COCO-20^i 数据集,采用 ResNet-50 作为主干网络设计消融实验,实验结果如表 6 所示。结果表明引入语义关联注意力模块能够显著提升模型的分割性能,究其原因是语义关联注意力模块引入了支持图片和查询图片之间的强语义关联,这有助于捕获到目标物体的位置及其更多细节信息。

表 6 语义关联注意力模块消融实验

Table 6 Ablation experiment of the semantic associative attention **Unit: %**

变体方法 Variation methods	PASCAL-5^i		COCO-20^i	
	mIoU	FB-IoU	mIoU	FB-IoU
Baseline model	53.1	66.3	34.8	61.5
Semantic associative attention module	**59.4**	**72.8**	**38.6**	**64.2**

Note: bold fonts represent the optimal results.

最后,为了验证所提出设计的多特征注意力融合模块的优越性,在 1-way 1-shot 任务上利用 PASCAL-5^i 和 COCO-20^i 数据集,采用 ResNet-50 作为主干网络设计消融实验,基线模型采用简单的向量拼接方式将多个输入特征图进行融合,实验结果如表 7 所示。相比基线模型,将语义关联注意力特征图、掩码后的支持特征和查询特征利用多特征注意力融合模块进行特征融合,在两个数据集上分别提升了 10.4% mIoU 和 13.4% FB-IoU(1-shot),18.0% mIoU 和 8.3% FB-IoU(5-shot),分割性能明显提升。

表 7 多特征注意力融合模块消融实验

Table 7 Ablation experiments of the multi-feature attentional fusion **Unit: %**

模块 Module	PASCAL-5^i		COCO-20^i	
	mIoU	FB-IoU	mIoU	FB-IoU
Baseline model	53.8	64.2	32.7	59.3
Feature fusion module	**59.4**	**72.8**	**38.6**	**64.2**

Note: bold fonts represent the optimal results.

4 结论

针对现有模型对于支持分支查询图片信息利用不充分的问题,本文在元学习方法的基础上,提出了一种新的小样本语义分割方法。该方法利用深度可分离卷积代替原始卷积块,降低了其对硬件资源的要求。此外,受注意力机制的启发设计一种语义关联注意力模块,该模块通过计算支持特征和查询特征之间的相似性来引入上下文语义信息,从而提高了模型对未知目标物体分割时定位的准确性。在 PASCAL-5^i 和 COCO-20^i 数据集上进行测试,实验结果表明所提出模型比起对比模型,整体优势较为显著。在未来工作中,可以利用交叉注意力机制建立查询编码特征与支持前景特征之间的关联,从而强化目标前景原型的表达能力。

参考文献

[1] 田萱,王亮,丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报,2019,30(2):440-468.

[2] LIU B H,JIAO J B,YE Q X. Harmonic feature activation for few-shot semantic segmentation [J]. IEEE

Transactions on Image Processing,2021,30:3142-3153.
[3] ZHANG Y F,SIDIBÉ D,MOREL O,et al. Incorporating depth information into few-shot semantic segmentation [C]//Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). Milan,Italy:IEEE,2021:3582-3588.
[4] 袁铭阳,黄宏博,周长胜.全监督学习的图像语义分割方法研究进展[J].计算机工程与应用,2021,57(4):43-54.
[5] SHABAN A,BANSAL S,LIU Z,et al. One-shot learning for semantic segmentation [C]//Proceedings of the 25th British Machine Vision Conference (BMVC). London, UK: British Machine Vision Association, 2017: 1029-1038.
[6] LIU Y W,LIU N,YAO X W,et al. Intermediate prototype mining transformer for few-shot semantic segmentation [C]//36th Conference on Neural Information Processing Systems (NeurIPS 2022). New Orleans, USA:MIT Press,2022:38020-38031.
[7] YANG Y,CHEN Q,FENG Y,et al. MIANet:aggregating unbiased instance and general information for few-shot semantic segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (ICCV). Paris, France: IEEE, 2023: 7131-7140.
[8] CAO Z Y,ZHANG T F,DIAO W H,et al. Meta-seg:a generalized meta-learning framework for multi-class few-shot semantic segmentation [J]. IEEE Access, 2019,7:166109-166121.
[9] WANG K X,LIEW J H,ZOU Y T. PANet:few-shot image semantic segmentation with prototype alignment [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE,2019:9197-9206.
[10] ZHANG X L,WEI Y C,YANG Y,et al. SG-One:similarity guidance network for one-shot semantic segmentation [J]. IEEE Transactions on Cybernetics, 2020, 50(9):3855-3865.
[11] LIU Y F,ZHANG X Y,ZHANG S Y,et al. Part-aware prototype network for few-shot semantic segmentation [C]//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020:142-158.
[12] LI G,JAMPANI V,SEVILLA-LARA L,et al. Adaptive prototype learning and allocation for few-shot segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Online:IEEE,2021:8334-8343.
[13] 贾熹滨,李佳.金字塔原型对齐的轻量级小样本语义分割网络[J].北京工业大学学报,2021,47(5):455-462,519.
[14] LIU J,BAO Y Q,XIE G S,et al. Dynamic prototype convolution network for few-shot semantic segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans,USA:IEEE,2022:11553-11562.
[15] LIU Z H,HE S,ZHU X T,et al. Simpler is better: few-shot semantic segmentation with classifier weight transformer [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Online:IEEE,2021:8741-8750.
[16] 刘宇轩,孟凡满,李宏亮,等.一种结合全局和局部相似性的小样本分割方法[J].北京航空航天大学学报,2021,47(3):665-674.
[17] PAMBALA A K,DUTTA T,BISWAS S. SML:semantic meta-learning for few-shot semantic segmentation [J]. Pattern Recognition Letters,2021,147:93-99.
[18] TIAN P Z,WU Z K,QI L,et al. Differentiable meta-learning model for few-shot semantic segmentation [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York,USA:AAAI,2020:12087-12094.
[19] WU Z H,SHI X X,LIN G S,et al. Learning meta-class memory for few-shot semantic segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Online:IEEE,2021:517-526.
[20] BAO Y Q,SONG K C,WANG J,et al. Visible and thermal images fusion architecture for few-shot semantic segmentation [J]. Journal of Visual Communication and Image Representation,2021,80:103306.
[21] HU T,YANG P W,ZHANG C L,et al. Attention-based multi-context guiding for few-shot semantic segmentation [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu,USA:AAAI,2019: 8441-8448.
[22] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft COCO:common objects in context [C]//Proceedings of the European Conference on Computer Vision (ECCV). Zurich,Switzerland:Springer,2014:740-755.
[23] ZHANG C,LIN G S,LIU F Y. CANet:class-agnostic segmentation networks with iterative refinement and attentive few-shot learning [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019:

5217-5226.

[24] YANG Y W,MENG F M,LI H L. A new local transformation module for few-shot segmentation [C]//Proceedings of the 25th International Conference on Multimedia Modeling (MMM). Thessaloniki, Greece: Springer,2020:76-87.

[25] ZHANG C,LIN G S,LIU F Y,et al. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul,Korea:IEEE,2019:9587-9595.

[26] KHOI N,TODOROVIC S. Feature weighting and boosting for few-shot segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul,Korea:IEEE,2019:622-631.

[27] WANG H C,ZHANG X D,HU Y T,et al. Few-shot semantic segmentation with democratic attention networks [C]//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer,2020:730-746.

[28] WANG Y N,TIAN X T,ZHONG G Q. FFNet:feature fusion network for few-shot semantic segmentation [J]. Cognitive Computation,2022,14(2):875-886.

[29] YANG B Y,LIU C,LI B H,et al. Prototype mixture models for few-shot semantic segmentation [C]//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow,UK:Springer,2020:763-778.

Few-shot Semantic Segmentation Based on Feature Fusion Attention Mechanism

LI Yijin[1],LI Shaolong[1**],HE Yan[2],LIU Wei[3]

(1. Information Center of Yunnan Power Grid Co.,Ltd.,Kunming,Yunnan,650200,China;2. Beijing THPower Technology Co.,Ltd.,Beijing,100085,China;3. Department of Electrical Engineering and Applied Electronic Technology,Tsinghua University,Beijing,100084,China)

Abstract:Aiming at the problem of insufficient information utilization in query images for small sample semantic segmentation tasks,a few-shot semantic segmentation algorithm based on feature fusion attention is proposed. Firstly,it utilizes shared backbone networks to obtain deep features of both image and query images. Secondly,attention mechanisms are employed to capture strong semantic correlation information between support features and query features,constructing task attention feature maps. Finally,a multi-feature attention fusion module is proposed,which can adaptively fuse multiple features' deep semantic information and perform feature decoding,thereby obtaining target object segmentation masks. The proposed model is evaluated on PASCAL-5^i and COCO-20^i datasets,and experimental results show that the proposed model outperforms current mainstream small sample semantic segmentation models in terms of more precise segmentation in both 1-way 1-shot and 1-way 5-shot tasks. Especially on the more challenging COCO-20^i dataset,the proposed model achieves 28.8% mIoU and 62.1% FB-IoU under the setting of 1-shot,and 36.9% mIoU and 64.8% FB-IoU under the setting of 5-shot.

Key words:few-shot semantic segmentation;multi-feature fusion;attention mechanism;deep semantic information;segmentation mask

责任编辑:陆　雁,陈少凡