文章编号：1007-2780(2024)10-1421-10

基于双重聚合和自合并网络的小样本图像语义分割

刘　玉[1]，于　明[1*]，朱　叶[2]
(1. 河北工业大学 电子信息工程学院，天津 300401；
2. 河北工业大学 人工智能与数据科学学院，天津 300401)

摘要：小样本图像语义分割是一种非常具有挑战性的任务，它试图使用几个带标签的样本来分割新类对象。主流方法常会存在特征鉴别性不高和原型偏差等问题。为缓解这些问题，本文提出一种基于双重聚合和自合并网络的小样本图像语义分割方法，能够充分挖掘特征相似性并减小原型偏差。首先，提出一个特征-掩码双重聚合模块，在支持特征和查询特征之间构建覆盖所有空间位置的密集相似关系，为特征聚合和掩码聚合提供全局语义信息。具体来说，通过对特征相似矩阵进行特征和掩码双重聚合，可以为查询图像获取具有引导信息的增强特征和初始掩码。然后，提出自合并解码器，通过合并基于初始掩码的自原型和已知的支持原型来减小原型偏差，并通过融合增强特征与合并原型向解码器传递丰富的类别语义信息。最后，利用基类预测信息进一步优化来自解码器的预测结果。本文方法在数据集 PASCAL-5^i 上的 mIoU 在 1-shot 和 5-shot 情况下分别取得了 68.3% 和 71.5%，在数据集 COCO-20^i 上的 mIoU 在 1-shot 和 5-shot 情况下分别取得了 46.5% 和 51.4%，优于主流方法的分割性能，能够更准确地分割出新类的目标区域。

关　键　词：小样本图像语义分割；特征相似性；双重聚合；类内差异性；自合并
中图分类号：TP391.4　　**文献标识码：**A　　**doi：**10.37188/CJLCD.2024-0074

Bi-aggregation and self-merging network for few-shot image semantic segmentation

LIU Yu[1], YU Ming[1*], ZHU Ye[2]

(1. *School of Electronic and Information Engineering, Hebei University of Technology, Tianjin* 300401, *China*;
2. *School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin* 300401, *China*)

Abstract: Few-shot image semantic segmentation is a very challenging task that attempts to segment objects of new classes using only a few labeled samples. The mainstream methods often have problems of low discriminative feature and prototype deviation. To alleviate these problems, a new few-shot image semantic segmentation method based on a bi-aggregation and self-merging network is proposed, which can fully

收稿日期：2024-03-08；**修订日期：**2024-04-15.
基金项目：国家自然科学基金青年项目(No.62102129)；河北省自然科学基金(No.F2021202030)
Supported by Youth Program of National Natural Science Foundation of China (No.62102129); Natural Science Foundation of Hebei Province (No.F2021202030)
*通信联系人，E-mail：yuming@hebut.edu.cn

mine the similarity of features and reduce prototype bias. Firstly, we propose a feature-mask bi-aggregation module to provide global semantic information for the feature aggregation and mask aggregation by constructing a dense similarity relation between the support features and the query features covering all spatial locations. Specifically, an enhanced feature and an initial mask with guiding information can be obtained for the query image by performing feature and mask bi-aggregation on the similarity matrices. Then, a self-merging decoder is proposed, which reduces the prototype bias by adding the initial mask-based self-prototype with the known support prototypes, and conveys rich category semantic information to the decoder by fusing the merged prototype with the enhancement feature. Finally, the prediction results obtained by the decoder are further optimized by the prediction results of the base classes. The mIoU values of our method on the dataset PASCAL-5^i achieve 68.3% and 71.5% in the 1-shot and 5-shot cases, respectively, and on the dataset COCO-20^i achieve 46.5% and 51.4% in the 1-shot and 5-shot cases, respectively, which is superior to the segmentation performance of the mainstream methods, and can segment the target region of the new class more accurately.

Key words: few-shot semantic segmentation; similarity of features; bi-aggregation; intra-class diversity; self-merging

1 引 言

深度学习方法在图像语义分割[1-2]任务上取得了显著进展,但很大程度上依赖于大规模预定义类的带标签训练集,且不能泛化到新类任务上。为解决数据稀缺和新类问题,小样本学习被提出,其能够将从已知类的大规模元训练集中获得的元知识转移到新类任务上,且不需要过多数据量[3]。作为小样本学习任务的像素级拓展,小样本图像语义分割逐渐成为研究热点,其复杂性更高,应用范围更广,是有前途的计算机视觉研究方向[4-5]。

目前,小样本图像语义分割任务主要分为两个分支:支持分支和查询分支,分别包含带标签的支持图像和没有标签的查询图像。小样本图像语义分割任务的目标是学习一个模型,该模型可以在只有少量带标签支持图像的情况下,分割出查询图像中的新类。该任务面临两个难点,一是只有少量带标签数据,二是要分割出未见过的新类。

一些方法[6-8]采用孪生网络对支持图像和查询图像进行特征提取,并结合提供的支持掩码信息构建引导网络,以指导查询图像的分割。然而,这些网络常存在特征鉴别性不高的问题,如果没有提供足够的判别信息,模型就无法学习用于分割预测的关键特征。为提高特征鉴别性,Liu 等[9]继续使用引导网络并增加了支持和查询特征之间的共性信息,但基于全局平均池化的交叉参考模块忽略了局部的细节信息。Tian 等[10]搭建了多尺度特征增强模块以丰富上下文语义信息,但模型复杂度过高。

除了特征低鉴别性问题外,查询和支持图像之间的类别信息差距常被忽略。同一类别,从形状、大小、颜色或轮廓上都存在着巨大差异。强化支持原型[11]能够减少错误匹配,但当图像对之间存在较大的类内多样性时,强制将类别信息从支持分支传递至查询图像是无效的。Fan 等[12]试图使用查询原型降低类内差异性,但通过余弦相似计算的查询掩码往往不能较好地覆盖目标区域,故其查询原型常不具备代表性。

为了缓解上述问题,本文提出基于双重聚合和自合并网络(Bi-aggregation and Self-merging Network, BSNet)的小样本图像语义分割方法。该网络以查询图像和带标签的支持图像作为输入,用共享特征编码器来提取查询和支持特征,以便后续的双聚合和自合并过程。判断支持图像和查询图像中的对象是否属于同一类别,主要通过衡量它们之间共同特征的相似度。为了充分利用支持集,本文提出一个特征-掩码双重聚合模块,以全局的角度挖掘查询和支持特征之间的相似性,并通过有效地聚合支持特征的局部信息和支持掩码为查询图像保留相应的引导信息。具体来说,所提出模块从特征和掩码中生成键映射和值映射,分别用支持和查询的键映射构建相似密集关系,再

用支持特征的值映射和掩码的值映射进行聚合来为查询图像生成增强特征和初始掩码。为了缓解由类内差异导致的原型偏差问题，设计了自合并解码器，采用自合并策略为解码器传递更多的类别语义信息。具体地，先将基于初始掩码的查询原型与已知的支持原型合并，再将合并原型与增强特征融入解码器以分割出新类对象。最后，结合基类预测信息，将预测结果进行细节优化。

本文的主要贡献如下：

(1)提出一个基于BSNet的小样本图像语义分割框架，能够增强特征表达并丰富原型语义信息。

(2)提出特征-掩码双重聚合模块，构建了查询与支持特征之间的密集相似性，并通过特征和掩码双重聚合保留引导信息，为查询图像生成增强特征和初始掩码；

(3)提出自合并解码器，通过初始掩码为查询图像获取自原型，然后与支持原型合并，再与增强特征相融合来向解码器传递更多类别元知识。

(4)在两个基准数据集PASCAL-5^i和COCO-20^i上的实验结果表明，本文方法的分割精度优于主流方法。

2　任务定义

小样本图像语义分割的目的是学习一种类别不可知模型，该模型可以在只给出少量注释示例的情况下对新类进行密集预测。为了防止过拟合，模型采用元学习方法[13]，遵循情景训练范式，即在已知的基类上进行训练，并在未见的新类上进行性能测试。数据集被划分为训练集$\mathcal{D}_{\text{train}}$和测试集$\mathcal{D}_{\text{test}}$，分别对应基类和新类，且类别不相交。这两个集合由多个子集组成，每个子集由支持集$S=\left\{\left(I_i^s,\mathrm{M}_i^s\right)\right\}_{i=1}^{k}$和查询集$Q=\left(I^q,M^q\right)$组成，其中$I^*$、$M^*$分别表示图像及其对应的掩码标签，$k$表示支持集中包含支持图像的数量。在训练期间，模型通过从$\mathcal{D}_{\text{train}}$中采样基类片段来学习从$\left(I_i^s,M_i^s,I^q\right)$到查询掩码$M^q$的映射。当训练完成，模型就通过从$\mathcal{D}_{\text{test}}$中采样新类片段进行性能评估，而不需要进一步优化。1-shot分割模型常被优先展示，然后再将其推广到k-shot分割模型(k通常设置为5)。

3　网络结构

本文提出基于BSNet的小样本图像语义分割算法，如图1所示。首先，采用共享特征提取器对成对图像进行特征提取。其次，利用提出的特征-掩膜双重聚合模块对支持和查询特征之间的相关性建模，并通过特征聚合和掩码聚合两种聚合方式完成特征的增强和目标物体的初始定位。接着，采用掩码平均池化为支持和查询图像分别获取原型表达。然后，合并这些原型，并通过自

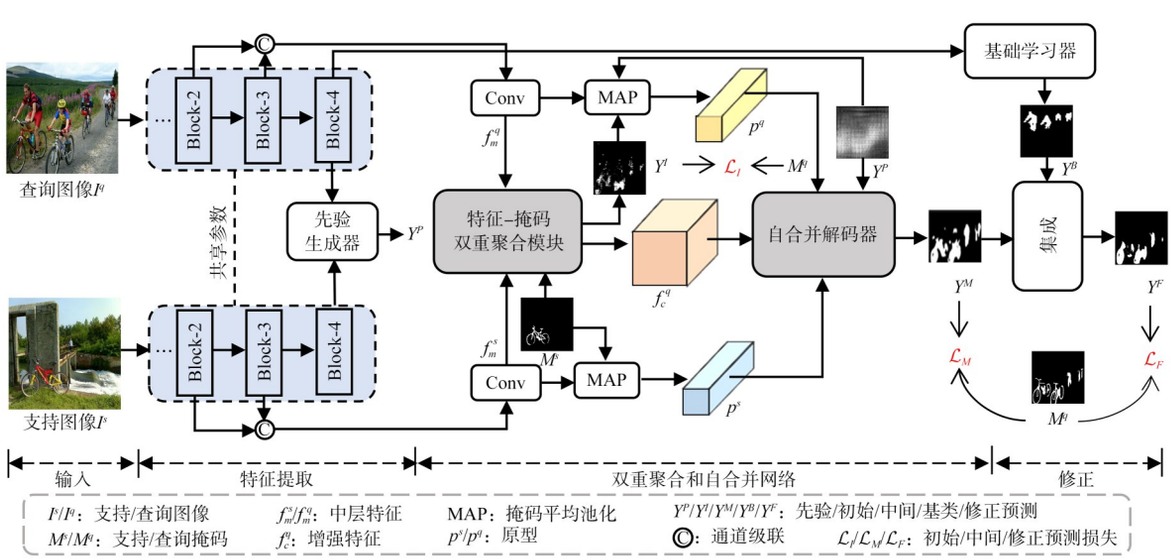


图1　基于双重聚合和自合并网络的小样本图像语义分割框架

Fig. 1　Structure of bi-aggregation and self-merging network for few-shot image semantic segmentation

合并解码器来产生中间预测值。最后，在基础学习器的辅助下，利用基类分割结果对中间预测值进行进一步修正，以优化预测的目标区域。

3.1 特征提取

在小样本分割任务中，通常使用孪生编码器对配对图像进行特征提取，其中卷积层的参数是权重共享的。给定查询图像 I^q 和支持图像 I^s，可以通过共享特征提取器生成 Block-i 的查询特征 $\boldsymbol{f}_i^q$ 和支持特征 $\boldsymbol{f}_i^s$。中层特征有助于不可知类目标对象的匹配[8]，因此，本文通过连接 Block-2 和 Block-3 的特征能为模型生成中层特征映射。中层特征可表示为：

$$\boldsymbol{f}_m^q = F_{1\times1}\left(\left[\boldsymbol{f}_2^q, \boldsymbol{f}_3^q\right]\right) \in \mathbb{R}^{C\times H\times W}, \tag{1}$$

$$\boldsymbol{f}_m^s = F_{1\times1}\left(\left[\boldsymbol{f}_2^s, \boldsymbol{f}_3^s\right]\right) \in \mathbb{R}^{C\times H\times W}, \tag{2}$$

其中：$[\cdot,\cdot]$ 为通道拼接，$F_{1\times1}$ 为 1×1 卷积和 ReLu 函数，C 为通道维数，H、W 分别为中层特征图的高度和宽度。另外，利用支持和查询高层特征之间余弦相似性获取的先验掩码能够增强模型的鲁棒性[10]。因此，从 Block-4 中提取的高层特征对 $\left\{\boldsymbol{f}_4^q, \boldsymbol{f}_4^s\right\}$ 被用来生成先验掩码 Y^P。为了与 BAM[14] 进行公平地比较，同样使用查询高层特征 f_4^q 来完成图像中基类对象的分割。

3.2 双重聚合和自合并网络

所提出的特征-掩码双重聚合网络主要由特征-掩码双重聚合模块和自合并解码器两部分构成。

3.2.1 特征-掩码双重聚合模块

通过构建特征之间的密集关系，能够获取相似性来聚焦目标物体。不同于文献[15-16]的单一聚合方式，本文提出一种双重聚合方式，通过特征和掩码的不同聚合为查询图像保留丰富的相似语义信息。如图 2 所示，在特征-掩码双重聚合(Feature-Mask Bi-Aggregation，FMBA)模块中，首先通过嵌入函数为两分支的中层特征分别学习键映射和值映射，并使用键映射来构建关系，然后通过特征聚合(Feature Aggregation，FA)和掩码聚合(Mask Aggregation，MA)为查询图像生成增强特征和初始掩码。

FMBA 的输入是查询和支持中层特征 $\boldsymbol{f}_m^q$、$\boldsymbol{f}_m^s$ 和下采样支持掩码 M^s。首先，3 个并行的 1×1 卷积层被用来学习每个输入特征的嵌入特征映射(即两个键映射和一个值映射)，这可以降低维数并增加输入特征的非线性。键映射用于度量查询特征和支持特征之间的相关性，值映射用于帮助检索相似语义信息。具体来说，对于查询特征，嵌入特征是两个键映射和一个值映射：$\boldsymbol{k}_1^q$，$\boldsymbol{k}_2^q \in \mathbb{R}^{C/2\times H\times W}$ 和 $\boldsymbol{v}^q \in \mathbb{R}^{C\times H\times W}$。对于支持特征，嵌入特征为 $\boldsymbol{k}_1^s, \boldsymbol{k}_2^s, \boldsymbol{v}^s \in \mathbb{R}^{C/2\times H\times W}$。对于支持掩码，只需对其进行下采样以产生 $\boldsymbol{v}^m \in \mathbb{R}^{1\times H\times W}$。

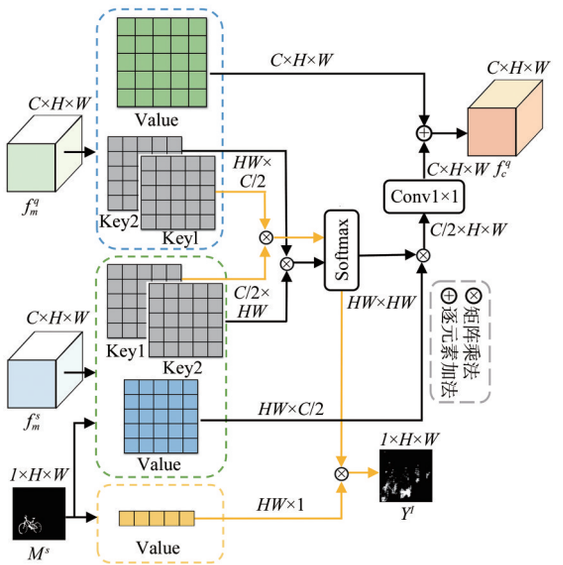


图 2 特征-掩码双重聚合模块结构

Fig. 2 Structure of feature-mask bi-aggregation module

然后，对生成的键、值映射进行重塑并构建密集关系。如图 2 所示，先计算两组查询和支持键映射之间的相关性，获取软权值后完成特征和掩码值映射的聚合。为了过滤支持背景的影响，支持键映射需要乘以相应的掩码值。具体地，通过重塑并以非局部方式可以获取两个逐像素相关矩阵，可表示为：

$$E\left(\boldsymbol{k}_1^q, \boldsymbol{k}_1^s\right) = \boldsymbol{k}_1^{q\mathrm{T}} \otimes \left(\boldsymbol{k}_1^s * \boldsymbol{v}^m\right), \tag{3}$$

$$E\left(\boldsymbol{k}_2^q, \boldsymbol{k}_2^s\right) = \boldsymbol{k}_2^{q\mathrm{T}} \otimes \left(\boldsymbol{k}_2^s * \boldsymbol{v}^m\right), \tag{4}$$

其中：$\otimes$ 表示矩阵乘法，$*$ 表示广播式元素乘法。为计算特征像素之间的相似度，需要进行 softmax 归一化以输出软权重 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$：

$$\boldsymbol{W}_1 = \frac{\exp\left(E\left(\boldsymbol{k}_1^q, \boldsymbol{k}_1^s\right)\right)}{\sum\exp\left(E\left(\boldsymbol{k}_1^q, \boldsymbol{k}_1^s\right)\right)}, \tag{5}$$

$$\boldsymbol{W}_2 = \frac{\exp\left(E\left(\boldsymbol{k}_2^q, \boldsymbol{k}_2^s\right)\right)}{\sum\exp\left(E\left(\boldsymbol{k}_2^q, \boldsymbol{k}_2^s\right)\right)}. \tag{6}$$

接着，用支持特征和掩码对软权重 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$

进行分别聚合以保留共性语义信息。特征聚合FA通过对生成的软权重 $\boldsymbol{W}_1$ 和支持值 $\boldsymbol{v}^s$ 进行矩阵乘来聚焦共性特征，然后用卷积层进行变换，并用残差连接来实现查询特征的增强。掩码聚合MA通过重塑的掩码值 $\boldsymbol{v}^m$ 与 $\boldsymbol{W}_2$ 的矩阵乘运算来定位相似目标区域。因此，增强特征 $\boldsymbol{f}_c^q$ 和初始掩码 Y^I 可表示为：

$$\boldsymbol{f}_c^q=\boldsymbol{v}^q+\varphi\left(\boldsymbol{W}_1\otimes\left(\boldsymbol{v}^s*\boldsymbol{v}^m\right)\right)\in\mathbb{R}^{C\times H\times W}, \tag{7}$$

$$Y^I=\boldsymbol{W}_2\otimes\boldsymbol{v}^m\in\mathbb{R}^{1\times H\times W}, \tag{8}$$

其中：+表示元素相加，φ 是1×1卷积层。

配备双重聚合模块，模型可以从支持特征及其掩码中保留像素级别的相似语义信息。只要存在一些共同特征，查询样本和支持样本之间共存对象的像素将被进一步激活，从而提供增强特征和初始掩码，便于后续预测。

为了优化初始掩码，采用二元交叉熵（Binary Cross Entropy，BCE）损失，则初始损失 L_I 可表示为：

$$Y_I^q=[1-Y^I,Y^I]\in\mathbb{R}^{2\times H\times W}, \tag{9}$$

$$L_I=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathrm{BCE}\left(Y_I^q(i,j),M^q(i,j)\right). \tag{10}$$

3.2.2　自合并解码器

由于同一类别的查询与支持图像间常存在着巨大的外观变化，这导致模型效果受限于少量支持样本提供的有限信息。受源于同一对象的像素比源于同一类别的不同对象的像素更相似的启发，SSP[12]提供了一种用查询原型去匹配查询特征的策略。本文也采用这一策略来提高分割精度，并利用初始预测掩码来获取查询原型。与SSP[12]中通过查询特征与支持原型余弦相似性得到初始预测掩码不同，本文则是利用由基于特征全局相关性的掩码聚合得到的初始预测掩码。

如图3所示，自融合解码器是以查询增强特征、支持原型、查询原型和先验掩码为输入，输出中间预测结果。由于前景原型能够直接获益，而背景常常无法共享全局语义的共性信息，因此只用前景原型来完成自原型合并的过程。

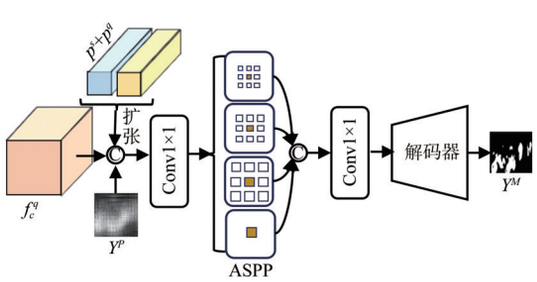


图3　自合并解码器结构

Fig. 3　Structure of self-merging decoder

在进入自合并解码器之前，先利用掩码平均池化（Masked Average Pooling，MAP）来为两分支获取原型表达。支持中层特征和支持掩码生成的支持原型过程如公式(11)所示：

$$\boldsymbol{p}^s=F_{\mathrm{pool}}\left(\boldsymbol{f}_m^s*D(M^s)\right), \tag{11}$$

其中：F_{pool} 为空间平均池化操作，D 是一个将 M^s 通过双线性差值技术重塑与 $\boldsymbol{f}_m^s$ 空间维度相同的函数。同理，查询原型也利用MAP来生成。为了保持支持原型与查询原型的语义一致性，选用查询中层特征 $\boldsymbol{f}_m^q$ 来完成。为了定位更多目标区域，本文将先验掩码 Y^P 加权至初始预测掩码 Y^I。自原型 $\boldsymbol{p}^q$ 获取过程如公式(12)和(13)所示：

$$Y=Y^I+Y^P, \tag{12}$$

$$\boldsymbol{p}^q=F_{\mathrm{pool}}\left(\boldsymbol{f}_m^q*\tau(Y)\right), \tag{13}$$

其中：$\tau(Y)=\begin{cases}1, Y\geqslant\tau\\0, \text{其他}\end{cases}$；$Y$ 是由公式(12)生成的查询掩码置信图；τ 是掩码阈值被用来控制查询特征采样范围，经验地将其设置为0.7。然后，将查询原型与支持原型共同用来匹配查询增强特征，以得到自融合预测值。

具体地，通过加权方式融合支持原型 $\boldsymbol{p}^s$ 和查询原型 $\boldsymbol{p}^q$：

$$\boldsymbol{p}^m=a_1\boldsymbol{p}^q+a_2\boldsymbol{p}^s, \tag{14}$$

其中，a_1 和 a_2 是原型融合权重，在实验中被设置为 $a_1=a_2=0.5$。

接着，对融合原型进行扩张，并与查询增强特征拼接后进行密集匹配以得到融合特征，然后再送入相应解码器以产生中间预测值：

$$Y^M=\mathcal{D}_{\mathrm{ASPP}}\left(F_{\mathrm{merge}}\left(\left[\boldsymbol{f}_c^q,\mathbb{E}\left(\boldsymbol{p}^m\right),Y^P\right]\right)\right), \tag{15}$$

其中：扩张函数 $\mathbb{E}(\cdot)$ 将原型 $\boldsymbol{p}^m$ 扩张至查询特征 $\boldsymbol{f}_c^q$ 相同维度，F_{merge} 是拼接后进行1×1卷积和Relu激活函数，$\mathcal{D}_{\mathrm{ASPP}}$ 是用于产生前景背景预测的基于ASPP[17]的解码器。解码器由两个卷积层构成，最后一层是1×1卷积和softmax组合，目的是生成前景背景预测。为优化中间预测结果，设置了中间预测损失：

$$L_M=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathrm{BCE}\left(Y^M(i,j),M^q(i,j)\right). \tag{16}$$

3.3 修正

模型在新类上的泛化能力常受基类信息的干扰。为了减轻基类信息对新类分割的干扰，本节利用基类信息修正中间预测结果。

BAM[14]引入了传统语义分割分支对查询图像的基类信息进行预测。沿用这一思路，通过基础学习器可以预测出基类信息：

$$Y^B=\mathcal{D}_{\text{Base\_learner}}\left(f_4^q\right)\in\mathbb{R}^{1\times H\times W}, \tag{17}$$

其中，$\mathcal{D}_{\text{Base_learner}}$是基础学习器的解码器，它以高层特征$f_4^q$为输入，最终生成所有基类信息的加权置信图$Y^B$。

由低层特征的格拉姆矩阵差值得到调整因子ψ，能够降低场景敏感度[14]。因此，依据ψ对中间预测值Y^M和基类预测值Y^B进行整合，能够得到最终预测值Y^F：

$$Y^F=\left[F_{\text{rec}}\left(\left[F_\psi\left(Y_0^M\right),Y^B\right]\right),F_\psi\left(Y_1^M\right)\right]\in\mathbb{R}^{2\times H\times W}, \tag{18}$$

其中：下标0和1分别代表背景和前景；F_ψ是调整中间预测结果的调整函数；F_{rec}是校正函数，用于将基类分割结果整合到背景中。

总之，最终损失为L：

$$L=L_F+\lambda_1L_M+\lambda_2L_I, \tag{19}$$

$$L_F=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\text{BCE}\left(Y^F(i,j),M^q(i,j)\right), \tag{20}$$

其中，λ_1、λ_2为平衡系数，实验中均设置为1.0。

当任务扩展到k-shot（$k>1$）设置时，模型可以使用多个支持图像对查询图像进行预测。本文采用调整因子ψ自适应估计每个支持图像的权重，其值越小表示支持和查询之间的风格差异越小，对应支持图像的贡献越大，反之亦然。

4 实验结果及分析

4.1 数据集与评价指标

为了评估方法的性能，本文在两个基准数据集PASCAL-5^i[6]和COCO-20^i[18]上进行实验。PASCAL-5^i是由数据集PASCAL VOC 2012[19]和SBD[20]构建的，将20个类别分为4个子集，每个子集的类别与OSLSM[6]保持一致。COCO-20^i是一个由MSCOCO[21]组成的大型数据集，其80个类别按照FWB[18]被平均分为4个子集。模型在3个子集上进行训练，并根据交叉验证协议在剩下的1个子集上进行测试。在推理过程中，从每个数据集的测试集中随机抽取1 000对支持和查询图像进行测试，并取5次随机种子的平均值作为最终测试结果。

本文采用平均交并比（mIoU）和前景-背景IoU（FB-IoU）作为评价指标。mIoU是所有类别交并比的平均值，FBIoU是前景交并比和背景交并比的平均值：

$$\text{IoU}=\frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}, \tag{21}$$

$$\text{mIoU}=\frac{1}{C}\sum_{c=1}^{C}\text{IoU}_c, \tag{22}$$

$$\text{FBIoU}=\frac{1}{2}\left(\text{IoU}_\text{F}+\text{IoU}_\text{B}\right), \tag{23}$$

其中：TP、FP和FN分别表示真阳性、假阳性和假阴性，C为总类别数，IoU_c为类别c的交并比值，IoU_F和IoU_B分别表示前景和背景的交并比值。

4.2 实验环境与训练设置

实验设备的软硬件环境：CPU：Intel Core i7-13700K CPU @ 5.40 GHz；显卡：NVIDIA GeForce GTX 4090；操作系统：Win11；深度学习框架：Pytorch。

实验的训练设置：在PASCAL-5^i上训练200个epochs，在COCO-20^i上训练50个epochs。采用SGD优化器，初始学习率设置为5e-3，批大小为4，学习率按照poly策略衰减[17]。在训练过程中，使用数据增强策略对输入图像进行随机缩放、水平翻转和旋转。本模型使用Resnet-50[22]作为骨干网，将多尺度特征增强模块替换为ASPP的PEFNet[10]变体作为基线。基础学习器参数与BAM保持一致。在训练过程中，骨干网的参数是固定的。

4.3 消融实验

为了验证模型的有效性，本节在常用于消融实验的数据集PASCAL-5^i上进行一系列的实验，并在主要评价指标mIoU下，研究每个组件对1-shot分割性能的影响。

4.3.1 支持背景对聚合模块的影响

为学习支持背景信息对聚合的影响，分别对特征聚合FA和掩码聚合MA中支持背景是否过滤分别进行实验。如表1所示，实验结果表明，当背景被滤除时，FA和MA都能取得较高的mIoU。在这里，MA的结果是用掩码聚合得到初始掩码并进行自支持融合后产生的。

表 1　支持背景信息对聚合模块的影响

Tab. 1　Influence of support background information on aggregation modules

方法	mIoU/%	
	不滤除	滤除
FA	65.5	65.9
MA	65.3	65.8

4.3.2　原型合并的有效性

为了证明模型中自合并策略的有效性，对原型合并进行消融学习。如表 2 所示，当使用先验掩码 Y^P 来获取查询原型时，mIoU 低于仅使用支持原型的基线。当使用初始掩码 Y^I 来获取查询原型时，mIoU 达到 65.6%。当同时使用先验掩码 Y^P 和初始掩码 Y^I 时，模型性能最佳，mIoU 与仅使用支持原型相比增加了 0.4%。这说明仅使用 Y^P 的原型不具备代表性，而使用初始掩码 Y^I 获取的原型能够在一定程度上缓解由类内差异导致的原型偏差问题，提高分割精度，且与 Y^P 同时使用后效果更好。

表 2　原型合并的有效性

Tab. 2　Effectiveness of prototype merging

$\boldsymbol{p}^s$	$\boldsymbol{p}^q/Y^P$	$\boldsymbol{p}^q/Y^I$	mIoU/%
√			65.4
√	√		65.2
√		√	65.6
√	√	√	65.8

4.3.3　初始掩码的获取

本小节对初始掩码的获取进行了消融实验。从表 3 可以看出，当 FA 与 MA 使用不同的关系矩阵时，模型的 mIoU 要高于使用相同的关系矩阵。因此，本文采用两组卷积来分别学习支持和查询特征的键映射，并构建两组特征关系矩阵。

表 3　不同初始掩码的比较

Tab. 3　Comparison of different initial mask

Method	mIoU/%
FMBA	66.2
FMBA_sc	65.5
Proto_sim	65.7

注：FMBA_sc 代表 FA 和 MA 的关系矩阵是同一个，Proto_sim 代表特征-原型的余弦相似性。

当模型使用特征-原型的余弦相似性得到初始掩码时，mIoU 达到了 65.7%，而本文方法 FMBA 的 mIoU 达到了 66.2%，本文方法超出了 0.5%。实验结果表明，FMBA 的设计是合理且有效的。

4.3.4　不同模块的消融实验

表 4 显示了 BSNet 中每个模块的有效性，其中除了对评价指标 mIoU 进行比较外，还展示了可学习参数量的变化。第一行显示了基线结果，其使用 ResNet-50 作为特征提取器并生成先验掩码。与基线相比，当执行特征聚合后，mIoU 增加了 0.5%。当再进行 MA 并执行原型合并后，mIoU 累计增加了 0.8%。经过基类分割结果校正后，mIoU 达到 68.3%，取得了较大的提升。与基线相比，本文模型的可学习参数只增加了 0.2M，效果却提升了 2.9%。

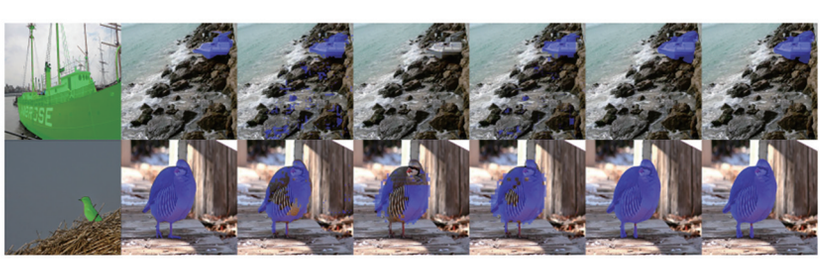
如图 4 所示，本文对各阶段产生的掩码进行

表 4　不同模块的消融实验

Tab. 4　Ablation study of different modules

Method				1-shot	#Learnable
Baseline	FA	MA	RF	mIoU/%	Params/M
√				65.4	4.9
√	√			65.9	5.0
√	√	√		66.2	5.1
√	√	√	√	68.3	5.1

注：FA 代表特征聚合，MA 代表掩码聚合，RF 代表修正。



(a) M^s　(b) M^q　(c) Y^I　(d) Y^P　(e) Y^I+Y^P　(f) Y^M　(g) Y^F

图 4　各阶段掩码的可视化结果

Fig. 4　Visualization results of each stage mask

了可视化。从图4可以看出,与初始掩码或先验掩码相比,它们的组合能够定位到更多的目标区域。因此,从组合掩码中获取的查询原型更具有代表性,这与表2的实验结果一致。通过观察中间掩码和最终掩码可以看出,即使在类内多样性较大的情况下,本文模型也可以获得准确的分割结果。

4.4 与主流方法对比

本节将所提出的方法在数据集PASCAL-5^i和COCO-20^i上与主流方法进行了比较。除了主流评价指标mIoU外,还额外对评价指标FBIoU和可学习参数量进行了对比。

表5展示了在1-shot和5-shot情况下,不同方法在mIoU、FBIoU和可学习参数量评价指标下的定量结果。由表5可知,与主流方法PFENet[10]、SSP[12]和DCAMA[15]相比,本文方法在两个数据集上都表现出了较大优势。且与最先进的方法BAM[14]相比,本文方法的分割性能仍有一定优势。具体地,在数据集PASCAL-5^i上,本文方法的主要评价指标mIoU在1-shot和5-shot情况下分别为68.3%和71.5%,与BAM相比,分别提升了0.5%和0.6%。在数据集COCO-20^i上,本文方法的主要评价指标mIoU在1-shot和5-shot情况下分别为46.5%和51.4%,与BAM相比,分别提升了0.3%和0.2%。另外,与主流方法相比,可学习参数量达到次最优,与BAM相比只增加了0.2M,在可接受范围内。

表5 本文方法与主流方法的定量结果比较

Tab. 5 Quantitative results of proposed method compared with the mainstream methods

方法	1-shot		5-shot		#Learnable params/M
	mIoU/%	FBIoU/%	mIoU/%	FBIoU/%	
Dataset PASCAL-5^i					
OSLSM[6]	40.8	—	43.9	—	272.6
SG-one[7]	46.3	63.1	47.1	65.9	19.0
CANet[8]	55.4	66.2	57.1	69.6	19.1
CRNet[9]	55.7	66.8	58.8	71.5	—
PFENet[10]	60.8	73.3	61.9	73.9	10.8
SSP[12]	60.9	—	68.8	—	8.7
DCAMA[15]	64.6	75.7	68.5	79.5	—
BAM[14]	67.8	79.7	70.9	82.2	4.9
本文方法	68.3	79.9	71.5	82.5	5.1
Dataset COCO-20^i					
FWB[18]	21.2	—	23.7	—	—
PFENet[10]	32.4	—	37.4	—	10.8
SSP[12]	37.4	—	44.1	—	8.7
DCAMA[15]	43.3	69.5	48.3	71.7	—
BAM[14]	46.2	—	51.2	—	4.9
本文方法	46.5	71.2	51.4	72.3	5.1

除了定量分析外,图5展示了在数据集PASCAL-5^i和COCO-20^i上的一些定性分割结果。如图5所示,前两行是包含掩码信息的支持图像和查询图像,后3行分别是基线、BAM和本文方法的可视化分割结果。从类别差异较小(如类别“瓶子”、“飞机”和“大象”等)的分割结果可以看出,与基线和BAM分割结果相比,本文方法能够更好地覆盖目标的细节信息。并且,对于外观差异较大的复杂图像对(如COCO-20^i的前两列),本文方法分割出了更多的目标区域。定量及定性的分析验证了本文方法优于主流方法,达到了最先进的水平。

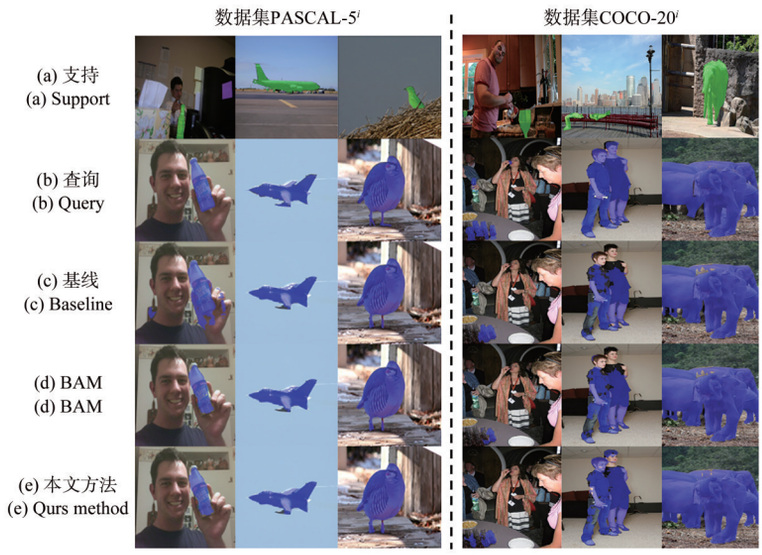


图 5　在两个数据集上的定性结果对比

Fig. 5　Qualitative comparison results on two benchmark datasets

5　结　　论

为了提高特征鉴别性并丰富原型表达,本文提出了基于双重聚合和自合并网络的小样本图像语义分割方法。首先,设计了一个特征-掩码双重聚合模块对支持和查询特征进行像素级关系建模,并通过特征和掩码聚合来完成特征的增强和初始掩码的定位。接着,利用初始掩码获取自原型,将其合并至已知支持原型以实现原型增强。然后,提出了一个自合并解码器,在增强特征和合并原型的共同作用和基类信息的校正下,能够精准地分割出新类对象。本文在两个基准数据集 PASCAL-5^i 和 COCO-20^i 上进行了广泛的实验,验证了所提出方法的有效性,且与主流方法相比,本文提出方法的性能达到了最优。

参　考　文　献:

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston: IEEE, 2015: 3431-3440.

[2] 蒋诗怡,徐杨,李丹杨,等. FRKDNet:基于知识蒸馏的特征提炼语义分割网络[J]. 液晶与显示,2023,38(11):1590-1599.
JIANG S Y, XU Y, LI D Y, *et al*. FRKDNet: feature refine semantic segmentation network based on knowledge distillation [J]. *Chinese Journal of Liquid Crystals and Displays*, 2023, 38(11): 1590-1599. (in Chinese)

[3] WANG Y Q, YAO Q M, KWOK J T, *et al*. Generalizing from a few examples: a survey on few-shot learning [J]. *ACM Computing Surveys*, 2021, 53(3): 1-34.

[4] 韦婷,李馨蕾,刘慧. 小样本困境下的图像语义分割综述[J]. 计算机工程与应用,2023,59(2):1-11.
WEI T, LI X L, LIU H. Survey on image semantic segmentation in dilemma of few-shot [J]. *Computer Engineering and Applications*, 2023, 59(2): 1-11. (in Chinese)

[5] REN W Q, TANG Y, SUN Q Y, *et al*. Visual semantic segmentation based on few/zero-shot learning: an overview [J]. *IEEE/CAA Journal of Automatica Sinica*, 2023, 11(5): 1106-1126.

[6] SHABAN A, BANSAL S, LIU Z, *et al*. One-shot learning for semantic segmentation [C]//*Proceedings of British Machine Vision Conference*. London: BMVA Press, 2017: 1-14.

[7] 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.

[8] ZHANG C, LIN G H, LIU F Y, *et al*. Canet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 5217-5226.

[9] LIU W D, ZHANG C, LIN G S, *et al*. CRNet: cross-reference networks for few-shot segmentation [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 4165-4173.

[10] TIAN Z T, ZHAO H S, SHU M, *et al*. Prior guided feature enrichment network for few-shot segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 44(2): 1050-1065.

[11] ZHANG B F, XIAO J M, QIN T. Self-guided and cross-guided learning for few-shot segmentation [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 8312-8321.

[12] FAN Q, PEI W J, TAI Y W, *et al*. Self-support few-shot semantic segmentation [C]//*Proceedings of* 17*th European Conference on Computer Vision*. Tel Aviv: Springer, 2022: 701-719.

[13] HOCHREITER S, YOUNGER A S, CONWELL P R. Learning to learn using gradient descent [C]//*Proceedings of International Conference on Artificial Neural Networks*. Vienna: Springer, 2001: 87-94.

[14] LANG C B, CHENG G, TU B F, *et al*. Learning what not to segment: a new perspective on few-shot segmentation [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 8057-8067.

[15] SHI X Y, WEI D, ZHANG Y, *et al*. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation [C]//*Proceedings of the* 17th *European Conference on Computer Vision* (*ECCV*). Tel Aviv: Springer, 2022: 151-168.

[16] HU H Z, BAI S, LI A X, *et al*. Dense relation distillation with context-aware aggregation for few-shot object detection [C]//*Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 10185-10194.

[17] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834-848.

[18] NGUYEN K, TODOROVIC S. Feature weighting and boosting for few-shot segmentation [C]//*Proceedings of IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 622-631.

[19] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The pascal Visual Object Classes (VOC) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.

[20] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, *et al*. Simultaneous detection and segmentation [C]//*Proceedings of the* 13*th European Conference on Computer Vision* (*ECCV*). Zurich: Springer, 2014: 297-312.

[21] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: common objects in context [C]//*Proceedings of the* 13*th European Conference on Computer Vision* (*ECCV*). Zurich: Springer, 2014: 740-755.

[22] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition [C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 770-778.

作者简介：

刘　玉，女，博士研究生，2017年于北方工业大学获得硕士学位，主要从事数字图像处理、模式识别的研究。E-mail: saralyliu@126.com

于　明，男，博士，教授，1999年于北京理工大学获得博士学位，主要从事图像数学变换、图像与视频编码的高效算法、视觉计算等方面的研究。E-mail: yuming@hebut.edu.cn