Computer Science and Application 计算机科学与应用, 2026, 16(2), 337-347
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162063

Hans汉斯

基于文本引导原型特征调制的小样本语义分割方法

庞云帆

合肥工业大学计算机与信息学院，安徽 合肥

收稿日期：2026年1月5日；录用日期：2026年2月5日；发布日期：2026年2月13日

摘　要

小样本语义分割旨在利用极少量标注样本实现对新类别的像素级分割。然而，受限于支持样本数量有限，传统方法主要依赖视觉特征构建类别原型，容易受到背景干扰及类内差异的影响，导致原型表达不稳定。为解决上述问题，本文提出一种基于文本引导原型特征调制的小样本语义分割方法。该方法引入类别级文本描述作为高层语义先验，通过文本－视觉相似性建模自适应聚合多条文本特征，并利用特征级线性调制机制对支持原型进行动态调节，从而增强类别判别性并抑制无关语义干扰。所提出的文本引导特征调制模块在不显著增加计算开销的前提下提升模型对新类别的泛化能力。实验结果表明，该方法在PASCAL-5^i和COCO-20^i数据集上均取得了优于基线模型的分割性能，验证了所提方法的有效性。

关键词

小样本语义分割，跨模态语义引导，特征调制

A Text-Guided Prototype Feature Modulation Method for Few-Shot Semantic Segmentation

Yunfan Pang

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui

Received: January 5, 2026; accepted: February 5, 2026; published: February 13, 2026

Abstract

Few-shot semantic segmentation aims to achieve pixel-level segmentation of novel classes with only a limited number of annotated samples. However, due to the scarcity of support samples, existing methods mainly rely on visual features to construct class prototypes, which are easily affected by background clutter and intra-class variations, leading to unstable prototype representations. To address

文章引用: 庞云帆. 基于文本引导原型特征调制的小样本语义分割方法[J]. 计算机科学与应用, 2026, 16(2): 337-347.
DOI: 10.12677/csa.2026.162063

this issue, this paper proposes a text-guided prototype feature modulation approach for few-shot semantic segmentation. The proposed method introduces category-level textual descriptions as high-level semantic priors, and adaptively aggregates multiple text features through text-visual similarity modeling. Furthermore, a feature-wise linear modulation mechanism is employed to dynamically adjust the support prototypes, thereby enhancing class discriminability and suppressing irrelevant semantic interference. The proposed text-guided feature modulation module improves the generalization ability to novel classes without introducing significant computational overhead. Experimental results on the PASCAL-5^i and COCO-20^i datasets demonstrate that the proposed method consistently outperforms the baseline models, validating its effectiveness.

Keywords

Few-Shot Semantic Segmentation, Cross-Modal Semantic Guidance, Feature Modulation

1. 引言

随着深度学习技术的快速发展，语义分割在自动驾驶、医学图像分析等领域取得了显著进展。传统语义分割方法通常依赖大规模精细标注数据进行监督学习，然而像素级标注成本高昂且耗时严重。为缓解对大量标注数据的依赖，小样本语义分割(Few-shot Semantic Segmentation, FSS)应运而生，其目标是在仅提供极少量带标注的支持样本条件下，实现对未见类别的准确像素级分割，具有重要的研究价值和实际意义。

现有小样本语义分割方法大多采用基于原型匹配的范式[1]，即通过支持集图像提取类别相关的视觉特征，并构建类别原型以指导查询图像的分割预测。但由于支持样本数量极其有限，支持原型往往难以充分刻画类别的整体语义特征，容易受到背景区域干扰以及类内外观差异的影响，从而导致原型表达不稳定，进而影响分割性能。尤其是在支持图像与查询图像存在显著视角变化、尺度差异或背景复杂度不一致的情况下，仅依赖视觉特征构建的原型往往难以准确对齐目标类别语义。为增强类别表示能力，部分研究尝试引入多尺度特征融合[2]或查询自适应更新[3]等策略，以缓解支持原型的噪声问题。然而，这些方法本质上仍局限于视觉模态内部的信息建模，难以从更高语义层面补充类别先验信息。

近年来，跨模态表示学习的兴起为小样本语义分割提供了新的思路。对比式语言－图像预训练模型(Contrastive Language-Image Pre-training, CLIP) [4]通过大规模图文对齐学习，具备强大的跨模态语义建模能力，其文本特征能够提供稳定且具有泛化性的类别级语义信息。相较于单一视觉实例，文本描述通常以类别层面的语义属性为表达核心，不依赖于具体实例的外观变化，因而能够提供更加稳定和一致的类别语义信息，可作为视觉原型的有效补充。然而，将文本语义有效引入小样本语义分割任务仍面临诸多挑战。首先，现有文本模板多源于图像分类任务，通常侧重于整体实例级描述，难以直接适配像素级分割场景。其次，同一类别可能对应多条文本描述，不同描述对具体实例的相关性存在差异，如何从多样化文本中提取对当前支持样本最有判别力的语义信息仍有待研究。此外，文本特征与视觉原型之间在特征分布和语义层级上存在差异，若直接进行融合，可能引入额外噪声，反而削弱模型判别能力。

基于上述分析，本文提出一种基于文本引导原型特征调制的小样本语义分割方法。该方法充分利用类别级文本描述所蕴含的高层语义先验，通过跨模态相似性建模机制，自适应地从多条文本特征中聚合

与当前支持原型最相关的语义表示。在此基础上，引入特征级线性调制机制，对支持原型进行动态调整，使其在保持视觉判别性的同时融入稳定的类别语义信息，从而有效缓解背景干扰和类内差异带来的影响。与直接特征拼接不同，该调制策略能够以轻量化方式实现对原型特征的精细控制，在不显著增加计算开销的前提下提升模型对新类别的泛化能力。

我们所提出的方法贡献如下：

1) 针对类别级文本描述数量多且语义相关性存在差异的问题，提出了一种基于文本–视觉相似度的文本特征自适应聚合策略，旨在从多条文本描述中动态筛选并融合与当前支持实例最相关的语义信息，有效缓解文本冗余带来的干扰。

2) 设计了一种文本引导的原型特征调制机制，将聚合后的文本语义作为高层先验，引导支持原型在特征层面进行动态调整，从而增强类别判别性并抑制背景和无关语义的影响。

3) 在 PASCAL-5^i 和 COCO-20^i 数据集上的大量实验结果表明，所提出方法在多种小样本设置下均取得了优于基线模型的性能提升，验证了方法的有效性与良好泛化能力。

2. 相关工作

2.1. 小样本学习

小样本学习(Few-Shot Learning, FSL) [5]旨在使模型在仅提供少量标注样本的条件下，仍能够有效学习并识别未见类别。现有小样本学习方法大致可分为三类。第一类为基于优化的方法[6]，该类方法通常采用元学习框架，在训练阶段显式模拟小样本任务的学习过程，从而使模型能够快速适应新任务并实现有效收敛。第二类为基于度量的方法[7]，通过学习判别性嵌入空间，缩小类内样本间距并扩大类间差异，进而基于相似度度量完成分类决策。第三类为基于数据增强的方法[8]，该类方法通过样本合成、特征重组或跨任务迁移等手段扩充训练数据规模，以缓解样本不足对模型泛化能力的限制。

2.2. 小样本语义分割

小样本语义分割是在小样本学习框架下，将研究目标从图像级分类拓展至像素级语义预测。现有方法大多采用支持–查询的匹配范式，典型方法如 PFENet [9]利用支持样本中目标区域的视觉特征构建类别原型，并通过特征相似度实现查询图像的像素级分类。然而，由于支持样本数量极为有限，基于视觉特征构建的原型往往难以充分表征目标类别的整体语义，易受背景噪声和类内差异影响。

近年来，跨模态学习的发展为小样本语义分割提供了新的思路。相关工作尝试引入文本信息作为高层语义先验，以增强模型对类别语义的理解能力。例如，PI-CLIP [10]重新审视了 CLIP 在小样本语义分割中生成先验信息的可靠性，并将文本语义作为辅助先验进行利用；PAT [11]通过类别级文本提示构建动态类感知增强机制，将文本语义引入特征迁移过程以提升分割性能。相比单一视觉实例，文本描述通常以类别级语义属性为核心，能够在一定程度上缓解类内差异带来的影响。

然而，现有基于文本引导的小样本语义分割方法仍存在不足：一方面，文本语义多依赖于固定描述模板，难以自适应建模不同类别间的语义相关性；另一方面，文本特征与视觉原型的融合方式较为粗粒度，尚未充分考虑文本语义对原型特征表达的精细调节能力。

3. 问题设定

小样本语义分割旨在利用极少量标注样本，引导模型对未见类别目标进行精确的像素级分割，其核心目标是在有限样本条件下实现从基类到新类的有效知识迁移。现有方法通常基于元学习框架，采用情景式训练策略以提升模型的泛化能力。

具体而言，给定数据集 D，其被划分为训练集 D_{train} 和测试集 D_{test}，且两者的类别集合满足 $C_{train} \cap C_{test} = \varnothing$。在训练阶段，模型通过从 D_{train} 中采样的情景任务进行优化。每个情景任务由支持集 S 和查询集 Q 构成。在 K-shot 设置下，支持集 $S = \left\{\left(I_i^s, M_i^s\right)\right\}_{i=1}^{K}$ 包含 K 组图像及其对应的二值掩码，其中 I^s 表示支持图像，M^s 表示对应的目标掩码；查询集 $Q = \left(I^q, M^q\right)$ 则包含待分割的查询图像 I^q 及其真实标注掩码 M^q。模型依据支持集提供的类别先验信息，对查询图像进行分割预测，并利用查询集的监督信号计算损失以更新模型参数。

在测试阶段，模型直接应用于从 D_{test} 中采样的情景任务，此时查询集的真实掩码 M^q 不可用，且不进行任何参数微调。通过上述训练与测试范式，模型能够将从 D_{train} 中学习到的知识有效迁移至 D_{test} 中的未见类别，实现高效的小样本语义分割。

4. 方法

4.1. 框架概述

本文在经典支持－查询匹配范式的小样本语义分割框架上，引入一种文本引导的原型特征调制模块，用于增强类别原型的语义表达能力。整体流程如图 1 所示。首先，支持图像与查询图像通过共享骨干网络提取多尺度视觉特征，并利用支持样本中目标区域特征构建初始类别原型。与此同时，为每个类别引入多条类别级文本描述作为高层语义先验，并通过文本编码器提取对应的文本特征。

为缓解多文本描述之间的语义冗余与相关性差异，本文根据视觉原型与文本特征之间的相似性，对多条文本特征进行自适应加权聚合，生成稳定的类别级文本表示。在此基础上，利用文本引导的特征调制机制对视觉原型进行动态调整，从特征通道层面增强与类别语义一致的表示并抑制无关干扰。最终，调制后的类别原型用于指导查询图像的像素级分类，从而获得分割结果。该模块可无缝嵌入现有小样本语义分割框架，在不显著增加计算开销的情况下提升模型对新类别的分割性能。

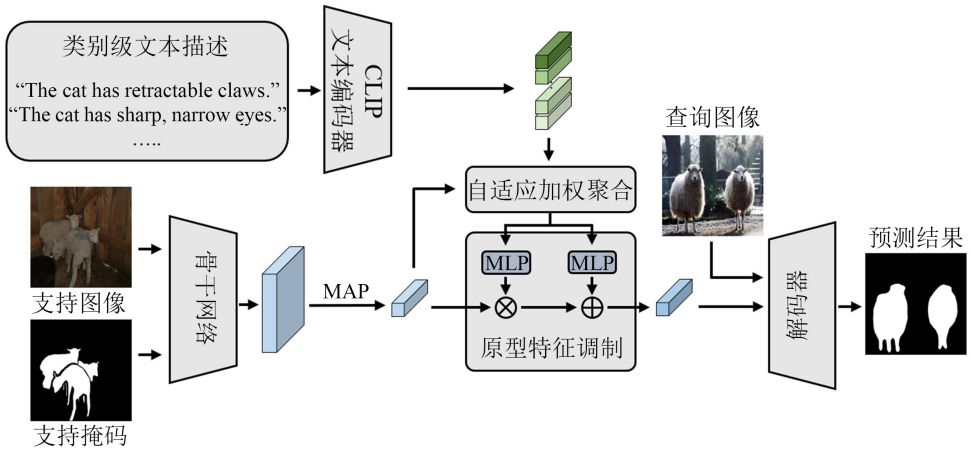

Figure 1. The framework of the proposed method
图 1. 我们所提出的方法框架

4.2. 类别文本描述生成与建模

为弥补小样本条件下仅依赖视觉支持样本构建类别原型所带来的语义不稳定问题，本文引入类别级文本描述作为高层语义先验，对目标类别进行辅助建模。文本描述以类别共性语义为核心，能够从抽象层面刻画类别的功能属性、结构特征或典型语义概念，从而在一定程度上弱化实例外观差异和背景干扰

对原型表达的影响。

4.2.1. 类别文本描述生成

为获得稳定且具有判别性的类别级文本先验，本文借助大语言模型生成多条类别文本描述，用于刻画目标类别的内在语义属性。与实例级文本描述不同，本文关注的是不随视角、尺度或具体场景变化的类别固有属性，以避免文本语义对具体视觉实例的过拟合。

具体而言，针对数据集中每一个目标类别 c，在已知全部类别集合 $C=\{c_1,c_2,\cdots,c_M\}$ 的条件下，向大语言模型提供统一的生成指令，要求其围绕目标类别的外观结构、形态特征、功能属性或典型组成要素生成多条简洁描述。文本生成过程遵循以下约束：(1) 描述应突出能够区分该类别与其他类别的 20 条关键属性；(2) 避免涉及具体实例、场景或视角变化相关的信息；(3) 尽量避免引入其他类别名称，以减少类间语义干扰；(4) 不同描述之间强调语义互补性，而非简单重复。

在此基础上，每个类别对应由多条属性描述共同构成的文本描述集合。所有文本描述经统一编码后作为类别级语义先验输入模型，并通过后续的自适应聚合机制提取与支持原型最相关的文本语义，用于指导原型特征调制。

4.2.2. 基于相似度的文本特征自适应聚合

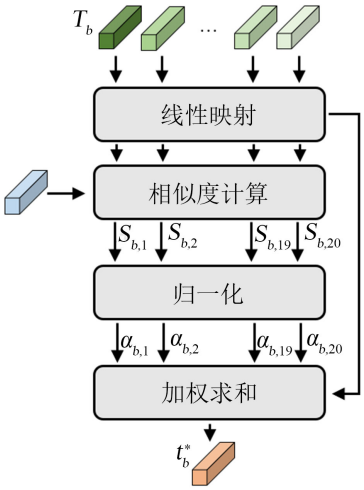

Figure 2. Adaptive aggregation process of textual features
图 2. 文本特征自适应聚合过程

在获得类别文本描述后，本文首先对其进行统一建模。对于每个类别，我们基于大语言模型生成多条互补的类别级语义描述，用以从不同语义属性角度刻画该类别的共性特征。具体而言，本文为每个类别构建由 $N=20$ 条文本描述组成的文本集合，并利用预训练的跨模态文本编码模型将文本映射至语义特征空间，得到对应的文本特征表示具体流程如图 2 所示。设第 b 个 episode 中的某一类别对应的文本特征集合表示为

$$T_b=\{t_{b,1},t_{b,2},\cdots,t_{b,N}\},t_{b,i}\in R^{d_t} \tag{1}$$

其中 d_t 表示文本编码器输出的特征维度。为了便于后续与视觉特征进行交互，本文采用线性映射将文本特征投影至与支持原型一致的特征维度：

$$\hat{t}_{b,i}=W_t t_{b,i},t_{b,i}\in R^d \tag{2}$$

其中 W_t 为可学习的线性映射矩阵。d 为支持原型的特征维度，与视觉特征空间保持一致。该操作不仅实现了跨模态特征维度对齐，也为后续相似度计算提供了统一的表示空间。

设支持样本构建得到的视觉原型为 $p_b \in R^d$，本文通过计算文本特征与视觉原型之间的相似度来衡量不同文本描述对当前类别实例的相关性。具体地，采用归一化后的点积相似度：

$$s_{b,i} = \frac{p_b^T \hat{t}_{b,i}}{\| p_b \| \ \| \hat{t}_{b,i} \|}. \tag{3}$$

该相似度度量能够反映文本语义与支持样本视觉特征在共享语义空间中的一致程度，从而为文本筛选提供依据。随后，通过 softmax 操作对相似度进行归一化，得到各文本描述的自适应权重：

$$\alpha_{b,i} = \frac{exp\left(s_{b,i}/\tau\right)}{\sum_{j=1}^{N} exp\left(s_{b,i}/\tau\right)}, \tag{4}$$

其中 τ 为温度系数。用于调节权重分布的平滑程度。相比于直接平均或固定选择文本描述，该自适应加权机制能够根据支持样本的视觉语义动态调整不同文本的贡献，从而突出与当前实例高度相关的语义信息，抑制冗余或不相关描述的干扰。

最终，类别级文本语义表示通过加权求和获得：

$$t_b^* = \sum_{i=1}^{N} \alpha_{b,i} \hat{t}_{b,i}. \tag{5}$$

通过上述文本特征自适应聚合过程，模型能够在类别级文本语义先验与实例级视觉信息之间建立有效联系，从多条文本描述中提取与当前支持原型最匹配的语义表示，为后续的文本引导原型特征调制提供更加稳定且具有针对性的语义支撑。

4.3. 文本引导的原型特征调制

本文通过文本特征自适应聚合获得了稳定的类别级文本语义表示 t_b^*，该表示刻画了目标类别的核心语义属性，为视觉特征提供了可靠的语义先验。在此基础上，本文进一步利用该文本语义对支持集视觉原型进行调制，从而构建语义一致性更强的类别原型表示。

在传统方法中，仅依赖视觉特征构建的原型容易受到支持样本遮挡、背景干扰及类内外观变化的影响，导致原型语义不稳定。为此，本文引入文本引导的原型特征调制机制，通过显式建模文本语义对视觉原型的影响，对原型特征进行语义层面的校正。

具体而言，首先将聚合后的文本语义特征 t_b^* 通过两个独立的线性映射，分别生成尺度调制向量 γ_b 与偏置调制向量 β_b：

$$\gamma_b = \varnothing_\gamma\left(t_b^*\right), \beta_b = \varnothing_\beta\left(t_b^*\right), \tag{6}$$

其中 $\varnothing_\gamma(\cdot)$ 与 $\varnothing_\beta(\cdot)$ 由轻量级多层感知机实现，用于建模文本语义与视觉特征之间的非线性映射关系。

随后，采用基于仿射变换的特征调制方式，对视觉原型 p_b 进行逐通道调制，得到最终的文本引导原型特征 $\tilde{p}_b$：

$$\tilde{p}_b = \gamma_b \odot p_b + \beta_b, \tag{7}$$

其中 $\odot$ 表示逐通道乘法。

上述调制过程从两个方面增强了原型表示的判别能力：一方面，尺度项能够根据类别语义自适应地放大或抑制不同通道的视觉响应，从而突出与目标类别高度相关的语义特征；另一方面，偏置项为原型

特征引入额外的语义补偿，有助于弥补支持样本中缺失或不显著的判别性信息。通过二者的协同作用，视觉原型在保持实例信息的同时，被有效对齐至文本所描述的类别语义空间。

通过文本引导的原型特征调制，模型能够在类别语义约束下抑制背景相关或噪声通道的干扰，缓解由支持样本不充分所引起的原型语义漂移问题，从而为后续查询特征的匹配与分割提供更加稳健且可泛化的类别表示。

5. 实验

5.1. 实验细节

本文在两个标准的小样本语义分割数据集 PASCAL-5^i [12]和 COCO-20^i [13]上对所提出的方法进行评估。PASCAL-5^i 基于 PASCAL VOC 2012 [14]，其中包括来自 SDS [15]的额外掩码注释，由 20 个类别组成，并划分为 4 个互不重叠的子集，每个子集包含 5 个类别。COCO-20^i 来源于 MS COCO [16]，共包含 80 个类别，同样划分为 4 个子集，每个子集包含 20 个类别。按照标准交叉验证协议，在第 i 个子集上进行测试，其余子集用于训练，确保训练与测试类别互不重叠。评价指标方面，本文采用平均交并比(mIoU)作为主要评价指标。本文采用预训练好的 ResNet50 [17]作为骨干网络，并在训练过程中固定其参数。文本特征由预训练的 CLIP ViT-B/16 [4]模型提取。输入图像尺寸统一为 473×473，并在训练过程中采用随机缩放和水平翻转等数据增强策略。在 PASCAL-5^i 和 COCO-20^i 上分别训练 100 轮和 50 轮，优化器选用 SGD，批大小为 8，初始学习率为 5e−3，动量为 0.9，权重衰减为 1e−4。所有实验均基于 PyTorch 框架实现。

5.2. 方法比较

我们将提出的方法与多种近期小样本语义分割方法进行了对比实验。表 1 给出了在 ResNet50 骨干网络下，模型在 PASCAL-5^i 数据集上的分割性能。可以看出，相较于基线方法 BAM [25]，本文方法在 1-shot 和 5-shot 设置下均取得了稳定的性能提升，表明所提出的文本引导原型特征调制方法能够在小样本条件下有效缓解原型表示不稳定的问题，从而显著提升模型对新类别的泛化能力。此外，我们还在更具挑战性的 COCO-20^i 数据集上进行了对比实验，结果如表 2 所示，实验结果同样验证了所提出方法在复杂场景和大规模类别设置下的有效性与鲁棒性。

Table 1. Comparison of mIoU performance on the PASCAL-5^i dataset
表 1. PASCAL-5^i 数据集上 mIoU 指标性能对比

	方法	1-shot					5-shot				
		5^0	5^1	5^2	5^3	mean	5^0	5^1	5^2	5^3	mean
骨干网络 ResNet50	CANet [18]	52.50	65.90	51.30	51.90	55.40	55.50	67.80	51.90	53.20	57.10
	PGNet [19]	56.00	66.90	50.60	50.40	56.00	57.70	68.70	52.90	54.60	58.50
	CRNet [20]	-	-	-	-	55.70	-	-	-	-	58.80
	PPNet [21]	48.58	60.58	55.71	46.47	52.84	58.85	68.28	66.77	57.98	62.97
	PFENet [9]	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
	HSNet [22]	64.30	70.70	60.30	60.50	63.90	70.30	73.20	67.40	67.10	69.50
	DCP [23]	63.81	70.54	61.16	55.69	62.80	67.19	73.15	66.39	64.48	67.80
	DAM [26]	67.30	72.00	62.40	59.90	65.40	**73.60**	74.60	69.90	67.20	**71.30**
	ABCNet [27]	68.80	73.40	62.30	59.50	66.00	71.70	74.20	65.40	67.00	69.60
	BAM [25]	68.97	73.59	67.55	**61.13**	67.81	70.59	75.05	70.79	**67.20**	70.91
	ours	**70.03**	**74.24**	**68.30**	60.92	**68.37**	71.03	**75.35**	**70.82**	67.16	71.09

Table 2. Comparison of mIoU performance on the COCO-20^i dataset
表 2. COCO-20^i 数据集上 mIoU 指标性能对比

	方法	1-shot					5-shot				
		5^0	5^1	5^2	5^3	mean	5^0	5^1	5^2	5^3	mean
骨干网络 ResNet50	PPNet [21]	28.09	30.84	29.49	27.70	29.03	38.97	40.81	37.07	37.28	38.53
	PFENet [9]	36.50	38.60	34.50	33.80	35.80	36.50	43.30	37.80	38.40	39.00
	HSNet [22]	36.30	43.10	38.70	38.70	39.20	43.30	51.30	48.20	45.00	46.90
	DCP [23]	40.89	43.77	42.60	38.29	41.39	45.82	49.66	43.69	46.62	46.48
	DPCN [24]	42.00	47.00	43.20	39.70	43.00	46.00	54.90	50.80	47.40	49.80
	DAM [26]	39.80	41.00	40.10	40.70	40.40	**50.10**	51.00	50.40	**49.60**	50.30
	ABCNet [27]	42.30	46.20	46.00	42.00	44.10	45.50	51.70	**52.60**	46.40	49.10
	BAM [25]	43.41	**50.59**	47.49	**43.42**	**46.23**	49.26	54.20	51.63	49.55	**51.16**
	ours	**44.32**	49.36	**47.80**	43.14	46.15	47.10	**54.30**	51.59	49.42	50.60

5.3. 消融实验

为系统分析本文各关键模块对模型性能的影响，我们在 PASCAL-5^i 数据集上开展了消融实验，并采用 1-shot 设置进行评估。所有实验均遵循标准的四折交叉验证协议，即分别在四个 flod 上进行测试，最终性能指标取结果的平均值。实验结果如表 3 所示。

具体而言，我们在基线模型的基础上逐步引入类别级文本语义、自适应文本特征聚合策略以及文本引导的原型特征调制模块，以验证各组成部分的有效性。首先，仅引入类别级文本语义先验即可带来一定的性能提升，表明高层文本语义能够为小样本语义分割提供有效的类别补充信息。然而，在未对多条文本描述进行筛选与加权的情况下，其性能增益相对有限。在此基础上，进一步引入基于相似度的文本特征自适应聚合机制后，模型性能得到持续提升，说明通过结合支持原型对多条文本描述进行动态加权，有助于突出与当前实例更加相关的语义信息，从而减轻文本冗余对模型判别能力的影响。相比之下，引入文本引导的原型特征线性调制模块能够带来更为明显的性能提升，验证了将文本语义以尺度与偏置参数的形式显式作用于视觉原型，在增强原型判别性、抑制无关语义干扰方面具有更直接且有效的作用。

当上述模块联合使用时，模型在 1-shot 场景下取得了最优性能，表明类别文本建模、自适应语义聚合与原型特征调制在功能上具有良好的互补性，共同提升了模型在少样本条件下的分割性能与泛化能力。

Table 3. Ablation study results of different modules
表 3. 各模块消融实验结果

类别文本先验	文本特征自适应聚合	原型特征调制	mIoU
			67.81
√			67.94
√	√		68.08
√		√	68.17
√	√	√	**68.37**

为分析类别文本描述数量对模型性能的影响，我们在 PASCAL-5^i 数据集的 fold-2 划分下，对不同文本描述数量 N 进行了消融实验，结果如表 4 所示。可以观察到，当 $N = 5$ 或 10 时，模型性能相对受限，

表明少量文本描述难以充分覆盖类别的多样化语义属性，所提供的类别级语义先验仍然较为有限。随着文本描述数量的增加，模型性能逐步提升，并在 $N = 20$ 时达到最优，说明适量且互补的文本描述有助于构建更加稳定、具有判别性的类别语义表示。

当进一步增大文本描述数量至 $N = 50$ 时，性能提升趋于饱和。这一现象可能是由于过多文本描述引入了较强的语义冗余与文本同质化，使得自适应聚合模块在筛选关键信息时面临更高的建模难度，从而限制了有效语义的进一步利用。

Table 4. Effect of the number of different text descriptions on model performance
表 4. 不同文本描述数量对模型性能的影响

文本描述数量 N	mIoU
5	67.94
10	68.17
20	68.30
50	68.23

5.4. 可视化

本小节通过可视化结果对模型的分割行为进行定性分析，具体结果如图 3 所示。其中，第(a)行展示支持集图像及其对应的掩码，第(b)行为查询图像及其真实掩码，第(c)行给出了基线模型 BAM [25]的分割结果，第(d)行为本文方法的分割结果。可以观察到，相比基线方法，本文方法在目标边界刻画和背景干扰抑制方面表现出更为明显的优势。

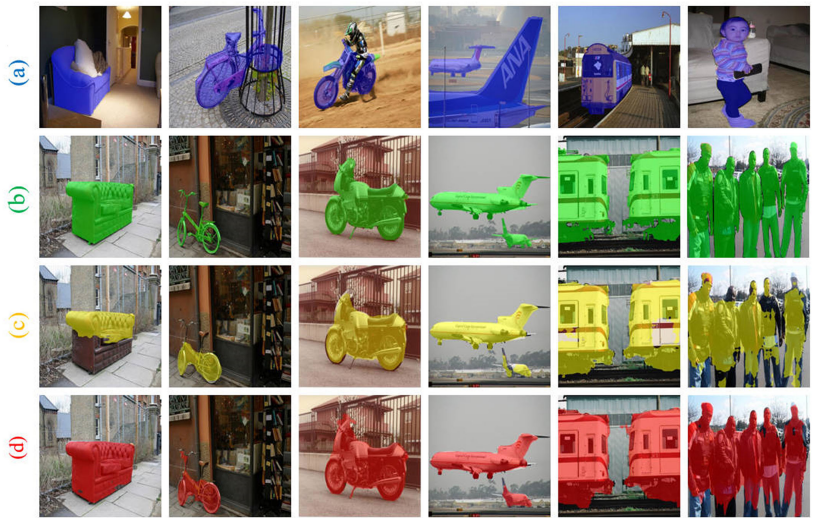

Figure 3. Diagram of visualization results
图 3. 可视化结果展示

6. 结论

本文针对小样本语义分割中支持样本有限导致的原型表达不稳定问题，提出了一种基于文本引导的原型特征调制方法，通过自适应聚合类别级文本语义实现对支持原型的动态调制，从而提升模型的类别判别能力；未来将进一步探索更灵活的文本建模策略及其在复杂场景和其他密集预测任务中的应用。

参考文献

[1] Zhang, X., Wei, Y., Yang, Y. and Huang, T.S. (2020) SG-One: Similarity Guidance Network for One-Shot Semantic Segmentation. *IEEE Transactions on Cybernetics*, **50**, 3855-3865. https://doi.org/10.1109/tcyb.2020.2992433

[2] Xie, G., Liu, J., Xiong, H. and Shao, L. (2021) Scale-Aware Graph Neural Network for Few-Shot Semantic Segmentation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 5471-5480. https://doi.org/10.1109/cvpr46437.2021.00543

[3] Zhu, L., Chen, T., Yin, J., See, S. and Liu, J. (2024) Addressing Background Context Bias in Few-Shot Segmentation through Iterative Modulation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 3370-3379. https://doi.org/10.1109/cvpr52733.2024.00324

[4] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, 18-24 July 2021, 8748-8763.

[5] Li, F.F., Fergus, R. and Perona, P. (2006) One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 594-611. https://doi.org/10.1109/tpami.2006.79

[6] Jamal, M.A. and Qi, G. (2019) Task Agnostic Meta-Learning for Few-Shot Learning. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 11711-11719. https://doi.org/10.1109/cvpr.2019.01199

[7] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S. and Hospedales, T.M. (2018) Learning to Compare: Relation Network for Few-Shot Learning. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1199-1208. https://doi.org/10.1109/cvpr.2018.00131

[8] Tan, W., Chen, S. and Yan, B. (2023) DifFSS: Diffusion Model for Few-Shot Semantic Segmentation. arXiv: 2307.00773.

[9] Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R. and Jia, J. (2022) Prior Guided Feature Enrichment Network for Few-Shot Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 1050-1065. https://doi.org/10.1109/tpami.2020.3013717

[10] Wang, J., Zhang, B., Pang, J., Chen, H. and Liu, W. (2024) Rethinking Prior Information Generation with CLIP for Few-Shot Segmentation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 3941-3951. https://doi.org/10.1109/cvpr52733.2024.00378

[11] Bi, H., Feng, Y., Diao, W., Wang, P., Mao, Y., Fu, K., *et al.* (2025) Prompt-And-Transfer: Dynamic Class-Aware Enhancement for Few-Shot Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**, 131-148. https://doi.org/10.1109/tpami.2024.3461779

[12] Shaban, A., Bansal, S., Liu, Z., Essa, I. and Boots, B. (2017) One-Shot Learning for Semantic Segmentation. *Proceedings of the British Machine Vision Conference* 2017, London, 4-7 September 2017, 167.1-167.13. https://doi.org/10.5244/c.31.167

[13] Nguyen, K. and Todorovic, S. (2019) Feature Weighting and Boosting for Few-Shot Segmentation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 622-631. https://doi.org/10.1109/iccv.2019.00071

[14] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2009) The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, **88**, 303-338. https://doi.org/10.1007/s11263-009-0275-4

[15] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S. and Malik, J. (2011) Semantic Contours from Inverse Detectors. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 991-998. https://doi.org/10.1109/iccv.2011.6126343

[16] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al.* (2014) Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Computer Vision—ECCV* 2014*e*, Springer International Publishing, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[17] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[18] Zhang, C., Lin, G., Liu, F., Yao, R. and Shen, C (2019) CANet: Class-Agnostic Segmentation Networks with Iterative Refinement and Attentive Few-Shot Learning. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5212-5221. https://doi.org/10.1109/cvpr.2019.00536

[19] Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q. and Yao, R. (2019) Pyramid Graph Networks with Connection Attentions for Region-Based One-Shot Semantic Segmentation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 9586-9594. https://doi.org/10.1109/iccv.2019.00968

[20] Liu, W., Zhang, C., Lin, G. and Liu, F. (2020) CRNet: Cross-Reference Networks for Few-Shot Segmentation. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 4164-472. https://doi.org/10.1109/cvpr42600.2020.00422

[21] Liu, Y., Zhang, X., Zhang, S. and He, X. (2020) Part-Aware Prototype Network for Few-Shot Semantic Segmentation. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 142-158. https://doi.org/10.1007/978-3-030-58545-7_9

[22] Min, J., Kang, D. and Cho, M. (2021) Hypercorrelation Squeeze for Few-Shot Segmenation. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 6921-6932. https://doi.org/10.1109/iccv48922.2021.00686

[23] Lang, C., Tu, B., Cheng, G. and Han, J. (2022) Beyond the Prototype: Divide-And-Conquer Proxies for Few-Shot Segmentation. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, 23-29 July 2022, 1024-1030. https://doi.org/10.24963/ijcai.2022/143

[24] Liu, J., Bao, Y., Xie, G., Xiong, H., Sonke, J. and Gavves, E. (2022) Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11543-11552. https://doi.org/10.1109/cvpr52688.2022.01126

[25] Lang, C., Cheng, G., Tu, B. and Han, J. (2022) Learning What Not to Segment: A New Perspective on Few-Shot Segmentation. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 8047-8057. https://doi.org/10.1109/cvpr52688.2022.00789

[26] Chen, H., Dong, Y., Lu, Z., Yu, Y., Li, Y., Han, J., *et al.* (2024) Dense Affinity Matching for Few-Shot Segmentation. *Neurocomputing*, **577**, Article ID: 127348. https://doi.org/10.1016/j.neucom.2024.127348

[27] Wang, Y., Sun, R. and Zhang, T. (2023) Rethinking the Correlation in Few-Shot Segmentation: A Buoys View. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 7183-7192. https://doi.org/10.1109/cvpr52729.2023.00694