arXiv:2304.01195v1 [cs.CV] 3 Apr 2023

并非所有特征都重要：通过自适应先验精炼增强小样本 CLIP

Xiangyang Zhu*[1]、Renrui Zhang*†[2, 3]、Bowei He[1],Aojun Zhou[2],Dong Wang[3],Bin Zhao[3],Peng Gao[3]

[1]香港城市大学 [2]香港中文大学[3]上海人工智能实验室

{xiangyzhu6-c, boweihe2-c}@my.cityu.edu.hk,{
zhangrenrui, gaopeng}@pjlab.org.cn

摘要

对比语言-图像预训练（*CLIP*）的普及推动了其在各类下游视觉任务中的应用。为了提升其在下游任务上的能力，少样本学习已成为广泛采用的技术。然而，现有方法要么性能有限，要么需要过多可学习参数。本文提出***APE***，一种自适应性先验细化方法，用于细化 *CLIP* 的预训练知识，以高计算效率实现更高准确率。通过先验细化模块，我们分析下游数据中的类间差异，并将领域特定知识与 *CLIP* 提取的缓存模型解耦。在此基础上，我们提出两种模型变体：免训练 *APE* 和需要训练的***APE-T***。我们探究测试图像、先验缓存模型与文本表征之间的三方关联，并仅训练一个轻量级类别残差模块。在 *11* 个基准数据集上的平均准确率方面，*APE* 和 *APE-T* 均达到最先进水平，在 *16-shot* 设置下，分别比次优方法高出 *+1.59%* 和 *+1.99%*，且使用的 ***×30*** 更少的可学习参数。代码见 https://github.com/yangyangyang127/APE。

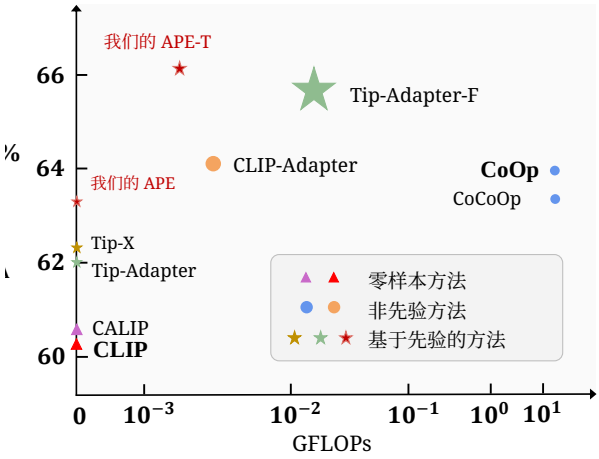


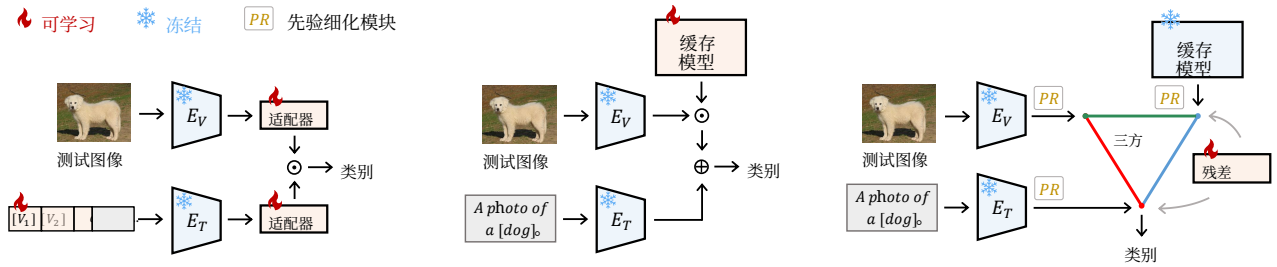
图1. **16样本 ImageNet 上准确率、训练 GFLOPs 和可学习参数的比较** [16] 分类。我们比较了包含梯度反向传播的训练 GFLOPs，图标大小表示可学习参数的数量。我们的 APE 和 APE-T 在实现效率较高的同时，取得了更优的性能。

1. 引言

对比视觉-语言预训练的兴起为多模态学习带来了新的范式 [46,48, 57]。它在各种下游视觉任务中广受欢迎，包括 2D 或 3D 分类 [38, 88, 92], 分割 [65, 83, 101], 和检测 [69, 87, 98]。CLIP [64] 是最受认可的对比视觉-语言模型之一，并已获得凭借其简洁性和卓越性能，CLIP 受到了广泛关注。它通过从互联网收集的大量图像-文本对进行预训练，展现出卓越的视觉-语言表征对齐能力，并在下游任务中取得了出色的零样本表现。为进一步提升 CLIP 在低数据场景下的性能，许多研究提出了少样本学习技术，在冻结的 CLIP 模型上添加可学习模块，以适应新的语义领域。

如图2 (a) 和 (b) 所示，根据是否利用 CLIP 的先验知识显式构建可学习模块，现有基于 CLIP 的小样本方法可分为两类。**1）非先验方法**在不使用 CLIP 先验知识的情况下随机初始化可学习模块，并在少样本训练期间对其进行优化。例如，CoOp 系列 [99, 100] 在 CLIP 文本编码器之前采用可学习提示

* 贡献相同。 †通讯作者。



(a) 非先验方法　　(b) 基于先验的方法　　(c) 我们提出的 APE-T

图2。**现有基于 CLIP 的小样本方法比较**。我们仅展示基于先验的方法和我们的 APE-T 中需要训练的模型变体。E_V，E_T 分别表示 CLIP 预训练的视觉编码器和文本编码器。

，而 CLIP-Adapter [24] 则在 CLIP 之后学习两个残差式适配器。这类网络只引入轻量级可学习参数，但小样本准确率有限，因为没有显式考虑预训练先验知识来构建额外模块。**2）基于先验的方法** 利用小样本数据中的 CLIP 提取特征构建键值缓存模型，并能够以免训练方式进行识别，包括 Tip-Adapter [91] 和 Tip-X [76]。随后，它们还可以将缓存模型视为性能良好的初始化，并微调缓存键以提高分类准确率。这些基于先验的方法将先验知识显式注入训练过程，但由于缓存大小庞大且包含大量可学习参数，因此较为繁琐。于是我们提出一个问题： 我们能否结合二者的优势，兼得两全之利，即既配备高效的可学习模块，又能利用 *CLIP* 的先验知识?

为此，我们提出了**自**适应**先**验细**化**，称为**APE**，通过细化视觉表征中的预训练知识，高效地将 CLIP 适配于小样本分类。如图1所示，APE 不仅能借助 CLIP 的先验取得优异性能，其计算资源消耗也低于非先验方法。我们发现，沿通道维度来看，并非 CLIP 的所有先验（即缓存模型或测试图像提取的视觉特征）对下游任务都至关重要。如图3所示，我们将 CLIP 提取的视觉表征中的特征通道分为两组，并分别将其与 ImageNet 中的文本表征的相似度图可视化 [16]。与第二组（b）相比，第一组（a）中的特征展现出更好的视觉-语言对齐。受此启发，我们提出了一种先验细化模块，通过类间相似度和方差这两项标准，自适应地选择最重要的特征通道。通过最大化少样本训练数据中的类间差异，精炼特征通道能够丢弃冗余信息，并以更低的内存开销减小缓存大小。

在此基础上，我们提出了该方法的两个变体，



图3。**视觉-语言对齐的相似度图**。我们采用搭载 ResNet-50 [32] 视觉编码器的 CLIP，并从 1024 个特征通道中精炼出 512 个，使精炼后的特征更加关注目标物体。

分别记为 APE 和 APE-T。前者是一种无需训练的模型，可直接利用精炼缓存模型进行推理。APE 创新性地探索测试图像、精炼缓存模型与文本表示之间的三方关联，从而实现稳健的无需训练识别。后者 APE-T（图2(c)）仅在此基础上训练轻量级类别残差，而非耗费大量成本对整个缓存模型进行微调。这些类别残差会进一步更新精炼缓存模型，并在不同模态间共享，以确保视觉—语言对应关系。我们的 APE 和 APE-T 在 11 个小样本基准上，分别取得了优于现有无需训练方法和需要训练方法的最先进性能，平均 16 样本准确率分别领先第二名 +1.59% 和 +1.99%。

我们的工作贡献总结如下：

•我们提出**自**适应**先**验细**化**（**APE**），这是一种 CLIP 适配方法，旨在显式利用其先验知识，同时保持计算效率。

• 先验细化后，我们探索 CLIP 提取的视觉—语言表征之间的三方亲和关系，以实现有效的少样本学习。

• 我们的免训练 APE 和 APE-T 在 11 个小样本基准数据集上展现了最先进的性能，证明了我们方法的优越性。

2. 相关工作

零样本 CLIP。 对于 C类数据集中的测试图像，CLIP [64] 利用其编码器提取 D维视觉和文本表征，分别记为$\mathbf{f} \in \mathbb{R}^{D}$ 和 $\mathbf{W} \in \mathbb{R}^{C\times D}$。随后，根据二者的相似度计算零样本分类 logits：

$$\mathbf{R}_{fW} = \mathbf{f}\mathbf{W}^{\top} \in \mathbf{R}^{1\times C}. \tag{1}$$

基于这一零样本范式，近期研究将 CLIP 的预训练能力扩展到了许多其他视觉任务，例如少样本图像分类 [63, 90,93, 99, 100], 视频识别 [52, 80]，3D 理解 [92, 97, 101], 和自监督学习 [25, 95]。其中，现有的少样本图像分类适配方法可分为两组。

非先验方法 在 CLIP 之上添加额外的可学习模块，并在没有明确利用 CLIP 先验知识的情况下对其进行随机初始化。这类方法包括 CoOp [100],CoCoOp [99], TPT [71], 和 CLIP-Adapter [24]。这些方法仅引入少量可学习参数，例如提示或适配器，但由于缺乏 CLIP 的先验知识，在下游任务中取得的准确率有限。

基于先验的方法 通过显式利用 CLIP 先验和缓存模型，可以获得更高的分类准确率，包括 Tip-Adapter [91], Causal-FS [51], 和 Tip-X [76]。对于包含 C 个类别、每类有 K 个样本的数据集，可以在其基础上构建键值缓存模型。缓存键和值分别使用 CLIP 提取的训练集特征 $\mathbf{F} \in \mathbb{R}^{CK\times D}$及其独热标签$\mathbf{L} \in \mathbb{R}^{CK\times C}$进行初始化。然后计算测试图像与训练图像之间的相似度$\mathbf{R}_{fF}$，公式如下：

$$\mathbf{R}_{fF} = \exp\left(-\beta(1-\mathbf{f}\mathbf{F}^{\top})\right) \in \mathbb{R}^{1\times CK}, \tag{2}$$

其中， β 是一个平滑标量。随后，将关系$\mathbf{R}_{fF}$ 视为权重，用于整合缓存值*i.e.*, 独热标签$\mathbf{L}$，并与零样本预测融合，得到少样本 logits,

$$\text{logits} = \mathbf{R}_{fW} + \alpha\mathbf{R}_{fF}\mathbf{L}, \tag{3}$$

其中， α 表示一个平衡因子。通过这种方式，基于先验的方法可以利用$\mathbf{R}_{fW}$ 和$\mathbf{R}_{fF}$ 的双边关系，实现无需训练的识别。在此基础上，还可以进一步使缓存模型具备可学习性，并在训练期间优化训练集特征$\mathbf{F}$。尽管可学习模块的初始化已明确融入CLIP的先验知识，但这些方法仍受缓存模型带来的过多参数所困扰。

与上述所有方法不同，我们的 APE 和 APE-T 不仅能凭借 CLIP 的先验知识取得有竞争力的表现，还能通过自适应先验细化模块引入轻量级参数和计算资源。

3. 方法

在第 3.1 节中，我们首先通过两个类间度量指标阐述 APE 中的先验细化模块。随后在第 3.2 节和第 3.3 节中，我们分别介绍基于精炼后表征的无需训练变体 APE 和需要训练变体 APE-T 的具体细节。

3.1. CLIP 的先验细化

对于下游数据集，CLIP 提取的视觉表征在通道维度上可能同时包含领域特定信息和冗余信息。前者在对下游图像进行分类时具有更强的判别性，后者则代表更通用的视觉语义。因此，我们提出两个标准——类间相似度和方差，以针对不同下游场景自适应地选择最重要的特征通道。

3.1.1 类间相似度

该准则旨在提取能最大限度降低类间相似度的特征通道，即最具分类判别性的通道。对于下游图像，我们将其经 CLIP 提取的特征表示为$\mathbf{x} \in \mathbb{R}^{D}$，其中 D 表示通道总数，我们希望从 D中筛选出 Q 个特征通道。接着，我们定义一个二值标记 $\mathbf{B} \in \{0,1\}^{D}$，其中 $B_k = 1 (k = 1, ..., D)$ 表示第 k^{th} 个元素 x_k 被选中，且$\mathbf{B}\mathbf{B}^{\top} = Q$。现在，我们的目标是找到最优的$\mathbf{B}$，以使下游数据的类间差异最大化。

对于一个包含 C个类别的下游数据集，我们计算所有训练样本中类别间的平均相似度 S。我们采用余弦相似度 $\delta(\cdot,\cdot)$ 作为度量，定义如下：

$$S = \sum_{i=1}^{C} P^{i} \sum_{\substack{j=1 \\ j\neq i}}^{C} P^{j} \frac{1}{M^{i}} \frac{1}{M^{j}} \sum_{m=1}^{M^{i}} \sum_{n=1}^{M^{j}} \delta(\mathbf{x}^{i,m}, \mathbf{x}^{j,n}), \tag{4}$$

其中 i, $j \in \{1, ..., C\}$ 表示两个不同的类别。P^{i}, P^{j} 表示这两个类别的先验概率，而 M^{i}, M^{j} 表示它们的训练样本总数。

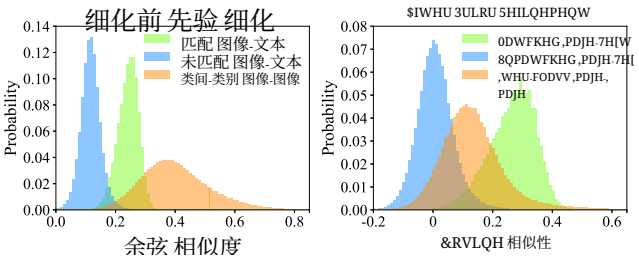


图 4. **先验细化模块的有效性，**该模块可降低类别间的视觉相似度，并改善文本与图像的对齐。

然而，针对整个数据集（即使只有少量样本）计算 S 的计算开销也很大。考虑到 CLIP 的对比式预训练已使视觉—语言表征充分对齐，下游类别的文本特征可视为一组视觉原型 [17, 37, 72]。这类原型能够近似嵌入空间中不同类别视觉特征的聚类中心 [27, 82]。为获得文本特征，我们只需使用模板“a photo of a[类别]”，并将所有类别名称填入 [类别] ，作为 CLIP 的输入。随后，我们将下游类别的文本特征记为 $\mathbf{x}^i \in \mathbb{R}^D$，其中 。

因此，对于每个类别，我们采用这些文本特征替代图像特征，这决定了$M^1 = ... = M^C = 1$。在开放世界设置下，我们也可以假设 $P^1 = ... = P^C = \frac{1}{C}$。随后，我们定义优化问题，以最小化类间相似度，

$$\begin{aligned} \min_{\mathbf{B}} \quad & S = \frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1 \\ j\neq i}}^{C}\delta(\mathbf{x}^i \odot \mathbf{B}, \mathbf{x}^j \odot \mathbf{B}), \\ s.t. \quad & \mathbf{B}\mathbf{B}^\top = Q, \end{aligned} \tag{5}$$

其中 $\odot$ 表示逐元素乘法，而$\mathbf{x} \odot \mathbf{B}$仅选择领域特定特征通道。我们进一步假设文本特征已进行 L2归一化，因此可以将余弦相似度简化为

$$S = \sum_{k=d_1}^{d_Q} S_k = \sum_{k=d_1}^{d_Q}\left(\frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1 \\ j\neq i}}^{C} x_k^i \cdot x_k^j\right), \tag{6}$$

其中 $k = \{d_1, d_2, ..., d_Q\}$ 表示所选特征通道的索引，且 $B_k = 1$，而 $S_k = \frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1 \\ j\neq i}}^{C} x_k^i \cdot x_k^j$ 表示 k^{th} 通道的平均类间相似度。根据公式 14，我们观察到，求解公式 5中的优化问题等价于选取平均相似度最小的 Q 个元素。也就是说，我们按照平均相似度对所有 D 个元素进行排序，并选取最小的前Q 个。这样，我们就能得到二值标记$\mathbf{B}$ ，并获得最适用于下游分类的判别性最强的特征通道。

3.1.2 类间方差

除了类间相似度之外，我们还引入另一项标准，用于剔除在不同类别之间几乎保持不变的特征通道。这些通道没有类间差异，对分类的影响也很小。为提高效率，我们还采用类别文本特征$\mathbf{x}^i \in \mathbb{R}^D$（其中 $i \in \{1, ..., C\}$）作为下游数据集的视觉原型。对于 k^{th} 特征通道，我们将其类间方差表示为

$$V_k = \frac{1}{C}\sum_{i=1}^{C}(x_k^i - \bar{x_k})^2, \tag{7}$$

其中 $\bar{x_k} = \sum_{i=1}^{C} x_k^i$ 表示k^{th} 通道在各类别间的平均方差。与公式14类似，方差准则也可以视为一个排序问题，但此处选择的是方差最高的前Q 个通道。通过这种方式，我们可以有效筛除 CLIP的先验知识中针对下游数据集冗余且信息较少的通道。

最后，我们通过平衡因子 λ 将相似度和方差标准融合，作为最终度量指标。对于 k^{th}特征通道，我们将其定义为

$$J_k = \lambda S_k - (1-\lambda)V_k, \tag{8}$$

其中 $k = 1, ..., D$。我们选取最小的前Q 个 J_k作为最终精炼特征通道，它们体现了最大的类间差异和判别能力。

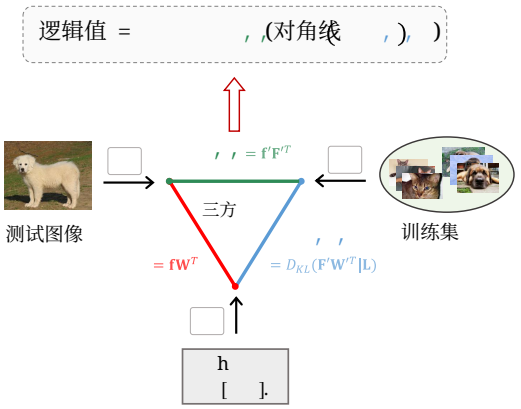
3.1.3 有效性

图 4 展示了自适应精炼模块带来的优势。我们使用 ImageNet [16]验证集上的文本特征进行精炼，并将统计结果可视化，其中类别数量 C 为 1000。我们采用 ResNet-50 [32] 作为 CLIP 的视觉编码器，并从全部$D = 1024$ 特征通道中精炼出个。我们比较了三种度量指标，参见 [76]。如图所示，对于精炼后的 512 个特征通道，图像之间的类间相似度（“类间图像-图像”）大幅降低，表明类别判别能力很强。同时，我们的精炼方法更好地对齐了配对的图像-文本特征（“匹配的图像-文本”），并拉远了未配对特征之间的距离（“不匹配的图像-文本”），从而增强了 CLIP 在下游识别中的多模态对应能力。

基于经过精炼的 CLIP 提取特征，我们提出了两种 CLIP 小样本适配方法：无需训练的 APE 和需要训练的 APE-T。

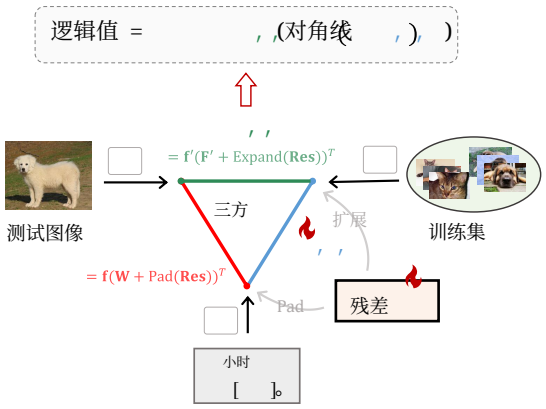
3.2. 免训练 APE

本质上，CLIP 是一种基于相似度的零样本分类器，它依赖于测试图像与嵌入空间中类别文本表示之间的距离。



免训练 APE

图 5. **APE 框架**。基于先验细化（PR），APE 以无需训练的方式探索视觉-语言表征之间的三方关系。



需训练 APE-T

图6. **APE-T 的框架**。我们的需训练变体在 APE 之上附加了可学习的类别残差以及$\mathbf{R}_{F'W'}$，用于小样本训练。

鉴于此，我们的 APE 基于经过优化的 CLIP 先验，并探索测试图像、下游类别文本和缓存模型中的训练图像之间的三方嵌入距离，如图5所示。

对于一个包含 K 个类别训练样本的 C 类 K 样本下游数据集，我们采用 CLIP 分别提取测试图像、类别文本和训练图像的 L2归一化特征，记作 $\mathbf{f} \in \mathbb{R}^{D}$、$\mathbf{W} \in \mathbb{R}^{C\times D}$ 和 $\mathbf{F} \in \mathbb{R}^{CK\times D}$。随后，我们执行自适应先验细化模块，为这三种特征筛选出最具信息量的通道，表示为 $\mathbf{f}' \in \mathbb{R}^{Q}$、$\mathbf{W}' \in \mathbb{R}^{C\times Q}$ 和 $\mathbf{F}' \in \mathbb{R}^{CK\times Q}$。这不仅能够丢弃预训练 CLIP 中的冗余信号，还能减少缓存模型的规模，从而降低推理时的计算成本。

对于三方关系，我们首先将$\mathbf{f}$ 与 $\mathbf{W}$ 之间的关系记为

$$\mathbf{R}_{fW} = \mathbf{f}\mathbf{W}^{\top} \in \mathbb{R}^{1\times C}, \tag{9}$$

它表示测试图像与类别文本之间的余弦相似度，即 CLIP 的零样本预测原始分类逻辑值，如第2节所述。接着，我们将$\mathbf{f}'$ 与$\mathbf{F}'$ 之间的亲和度表示为

$$\mathbf{R}_{f'F'} = \exp\left(-\beta(1-\mathbf{f}'\mathbf{F}'^{\top})\right) \in \mathbb{R}^{1\times CK}, \tag{10}$$

这表示缓存模型通过调节标量得到的图像-图像相似度 β，对应于基于先验的方法 [76, 91]。此外，我们考虑 $\mathbf{F}'$ 与$\mathbf{W}'$ 之间的关系，并将它们的余弦相似度表示为 $\mathbf{F}'\mathbf{W}'^{\top}$，该相似度表示 CLIP 对少样本训练数据的零样本预测。为了评估 CLIP 的下游识别能力，我们计算 CLIP 的预测与其独热标签$\mathbf{L}$之间的 KL散度，$D_{KL}(\cdot|\cdot)$。其公式为

$$\mathbf{R}_{F'W'} = \exp\left(\gamma D_{KL}(\mathbf{F}'\mathbf{W}'^{\top}|\mathbf{L})\right) \in \mathbb{R}^{1\times CK}, \tag{11}$$

其中，γ用作平滑因子。$\mathbf{R}_{F'W'}$可被视为缓存模型中每个训练特征的得分，表示其由 CLIP 提取的表征准确度，以及其对最终预测的贡献程度。

最后，整合所有三方关系后，我们得到 APE 的整体分类逻辑值:

$$\text{logits} = \mathbf{R}_{fW} + \alpha\mathbf{R}_{f'F'}\Big(\text{diag}(\mathbf{R}_{F'W'})\mathbf{L}\Big), \tag{12}$$

其中，α 表示平衡因子，diag(·)表示对角化。第一项表示 CLIP 的零样本预测，并包含其预训练先验知识。第二项表示来自缓存模型的小样本预测，该预测基于精炼特征通道和$\mathbf{R}_{F'W'}$的重新加权。因此，通过自适应先验精炼和三方关系分析，我们的 APE 能够高效且有效地增强小样本 CLIP。

3.3. 需训练 APE-T

为进一步提升 APE 的少样本性能，我们在图中提出了一种需要训练的框架 APE-T6。现有的基于先验的方法 [51, 91] 会直接对缓存模型中的所有训练特征进行微调，导致可学习参数规模庞大、计算成本高。相比之下，APE-T 会冻结缓存模型，仅训练一组额外的轻量级类别残差$\mathbf{Res}$ $\in\mathbb{R}^{C\times Q}$以及缓存分数$\mathbf{R}_{F'W'} \in \mathbb{R}^{1\times CK}$。

具体而言，类别残差 $\mathbf{Res}$ 由一组 C 可学习嵌入构成。每个嵌入对应一个下游类别，旨在少样本训练过程中针对不同类别优化细化后的 Q 特征通道。为保留嵌入空间中的视觉—语言对应关系，我们将$\mathbf{Res}$ 同时应用于文本特征$\mathbf{W}$ 和训练集特征 $\mathbf{F}'$。

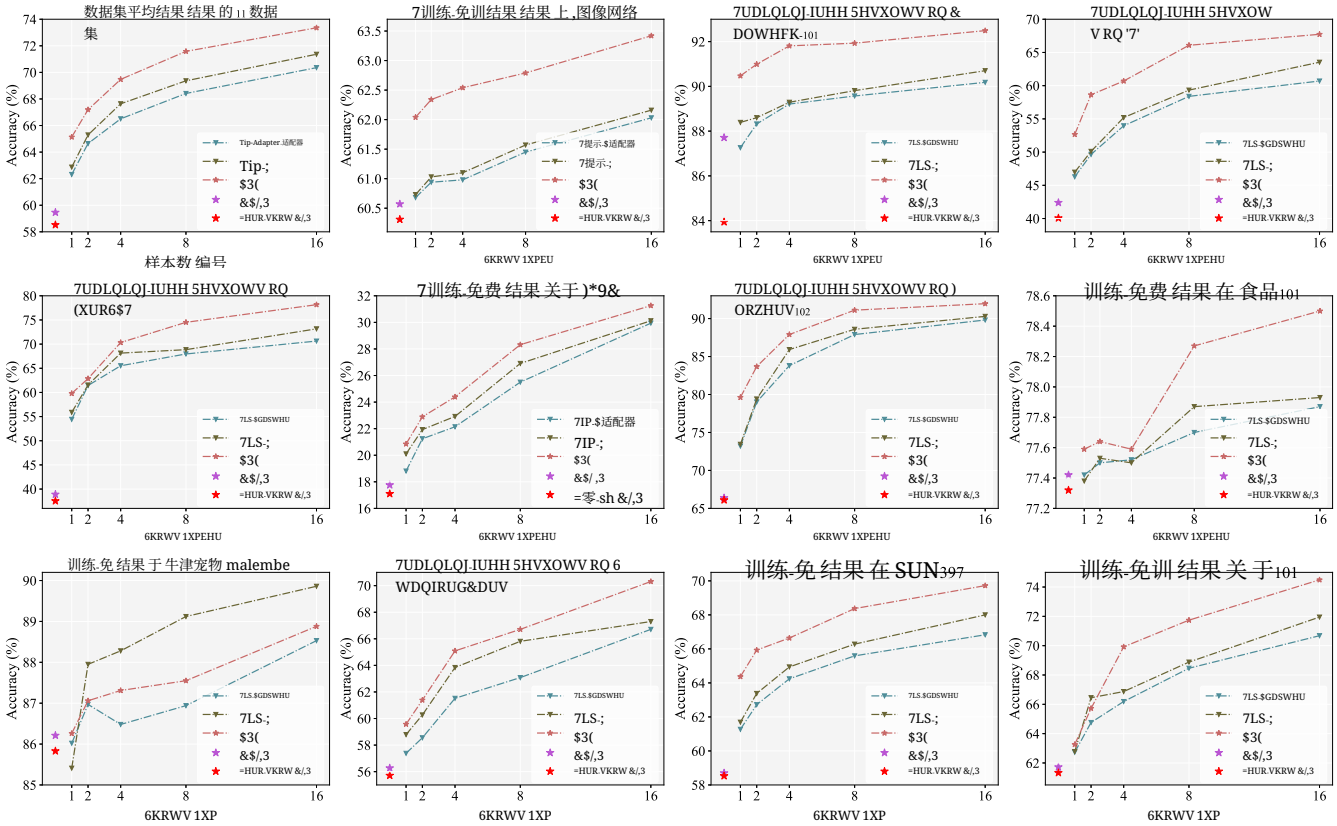


图 7. **少样本性能** **APE 与其他免训练方法**在 11 个图像分类数据集上的表现。

对于公式9，我们首先将 Q通道的 **Res** 填充至 D个通道，使其成为$\mathbf{W}$，方法是将冗余通道索引填充为零。然后，我们将填充后的**Res** 与$\mathbf{W}$逐元素相加，通过优化后的文本特征更新 CLIP 的零样本预测，表示为

$$\mathbf{R}_{fW} = \mathbf{f}\Big(\mathbf{W} + \text{Pad}(\mathbf{Res})\Big)^{\top}. \tag{13}$$

对于公式 10，我们首先将 C 嵌入**残差** CK 扩展到 CK 作为 $\mathbf{F}'$，方法是在每个类别内重复残差。然后，我们将扩展后的 **残差**与 $\mathbf{F}'$逐元素相加，通过优化训练集特征来提升缓存模型的少样本预测，公式如下：

$$\mathbf{R}_{f'F'} = \exp\Big(-\beta\big(1 - \mathbf{f}'(\mathbf{F}' + \text{Expand}(\mathbf{Res}))^{\top}\big)\Big).$$

对于公式 11，我们直接将 $\mathbf{R}_{F'W'}$设为可学习参数，使其在训练过程中无需手动计算。通过这种方式，APE-T 可以自适应地为不同的训练集特征学习最优缓存分数，并确定哪些特征对预测的贡献更大。

最后，我们还利用公式 12 获得 APE-T 的最终分类逻辑值。通过仅训练这类小规模参数，APE-T 无需对缓存模型进行成本高昂的微调，并通过更新两种模态的精炼特征取得了更优的性能。

4. 实验

在第 4.1 节中，我们首先介绍 APE 和 APE-T 的详细设置。随后在第 4.2 节中，我们在 11 个广泛采用的基准数据集上评估我们的方法。

4.1. 实验设置

数据集。 我们采用 11 个图像分类基准进行全面评估：ImageNet [16], Caltech-101 [22], DTD [14], EuroSAT [33], FGVCAircraft [56],花卉102 [58], 食物101 [10], OxfordPets [60], 斯坦福- 汽车 [41], SUN397 [85], 和 UCF 101 [73]。此外，我们采用 ImageNet-Sketch [79] 和 ImageNet-V2 [66] 来测试泛化能力。给定各数据集的少样本训练数据后，我们使用官方验证集调整模型，并在完整测试集上评估结果。

实验设置。 对于 APE 和 APE-T，我们默认采用 ResNet-50 [32] 作为 CLIP 的视觉编码器，其输出的视觉语言特征具有 $D = 1024$个通道。我们遵循现有工作 [24,91,100] 开展 1/2/4/8/16 样本学习，并使用 Tip-X 中的文本提示 [76]和 CuPL 中的文本提示 [62]。对于先验细化模块，我们将 λ 在公式 8中设为：APE 取 0.7，APE-T 取 0.2。训练 APE-T 时，我们采用 256 的批量大小和 AdamW [55]

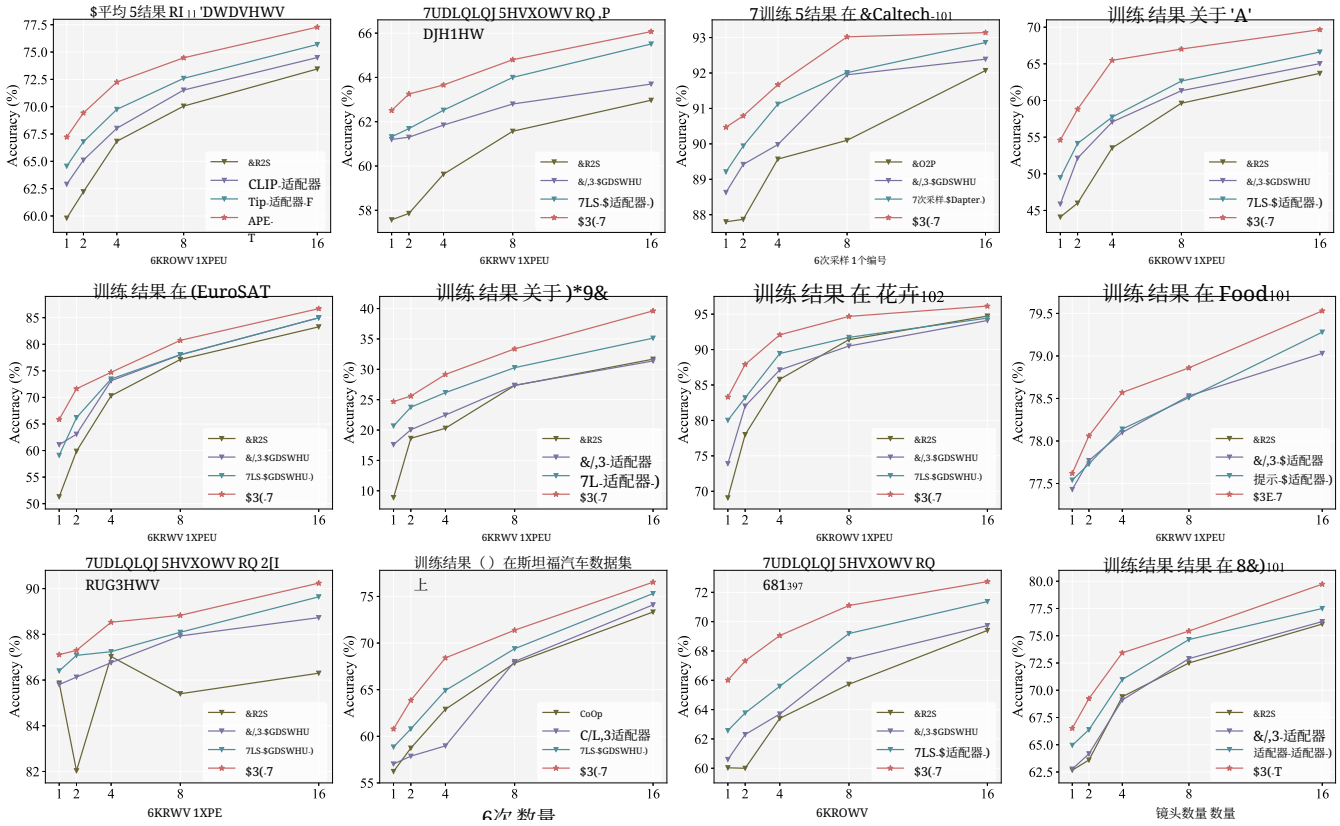


图8。**少样本 APE-T 与其他需要训练的方法**在11个图像分类数据集上的**性能** .

优化器，并使用余弦退火调度器 [54]。对于 ImageNet 和 Food101，我们使用 0.0001 的学习率；对于其余数据集，则使用 0.001 的学习率。

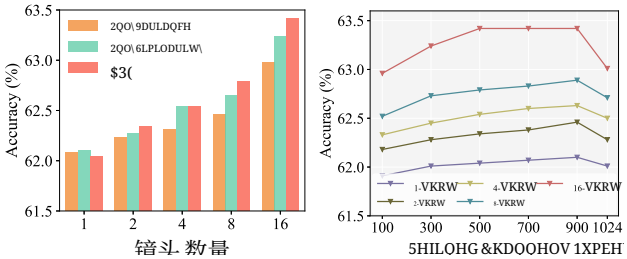
4.2. 性能分析

APE 结果。 在免训练设置下，我们将我们的 APE 与 Tip-Adapter [91] 和 Tip-X [76]在图 7中进行了比较。它们都是基于先验的方法，并且均为配备缓存模型的免训练方法。如11个数据集的平均结果所示，在1至16样本范围内，APE 相较于其他方法始终具有优势，表明其具备较强的少样本适应能力。尽管我们在 OxfordPets 上落后于 Tip-X，但在 DTD 和 Eu-roSAT 数据集上观察到了显著提升，即， +7.03% 和 +7.53%（在16样本设置下）分别超过 Tip-Adapter。这证明了针对不同下游场景细化领域特定知识并利用三方关系的有效性。

APE-T 结果。在图8中，我们将 APE-T 与另外三种需要训练的方法 CoOp [100], CLIP-Adapter [24], 和 Tip-Adapter-F [91] 进行了比较。我们的 APE-T 在每个基准测试上都优于现有方法，并在所有少样本设置下取得了最先进水平的结果。平均而言，APE-T 的 16样本准确率为 77.28%，超过 Tip-Adapter-F 达到 +1.59%。特别是，与 Tip-Adapter-F 相比，我们观察到 APE-T 在 DTD 和 FGV- CAircraft 上的分类准确率分别提升了 +3.05% 和 +4.50%。这些优异结果充分验证了通过我们可学习的类别残差更新精炼特征通道的重要性。

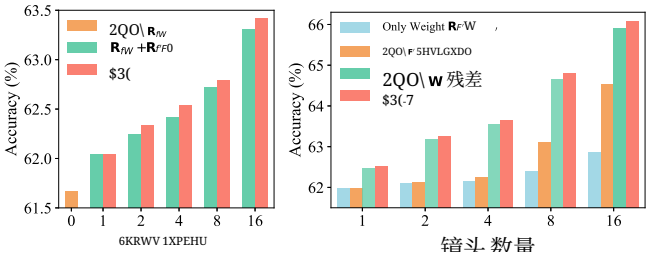
计算效率。 我们还在表1中比较了我们的方法与现有方法之间的计算开销。我们使用 NVIDIA RTX A6000 GPU 进行测试，并报告了16样本 ImageNet上的性能。如表所示，CoOp 的可学习参数最少，但需要较长的训练时间和较多的 GFLOPs，以通过整个文本编码器反向传播梯度。Tip-Adapter-F 缩短了训练时间，但通过微调整个缓存模型引入了大量可学习参数，同时梯度计算所需的 GFLOPs 也不少。相比之下，我们的 APE-T 不仅达到了最高准确率，还实现了更具优势的计算效率： **×5000 GFLOPs 少于 CoOp，且 ×30 参数少于 Tip-Adapter-F**。

泛化能力。 在表 2中，我们使用域内 ImageNet 训练模型，并在分布外数据集上测试其泛化能力。在域内性能最佳的情况下，



(a) 相似度与方差标准 (b) 待细化的通道数量

图9. **先验细化的消融研究**。



(a) 免训练 APE (b) 需训练 APE-T

图10. **关于 APE 和 APE-T 的消融研究**。

方法	训练	训练轮数	GFLOPs	参数量	Acc.
零样本					
CLIP [64]	-	-	-	-	60.33
CALIP [30]	-	-	-	-	60.57
无需训练					
Tip-Adapter [91]	0	0	-	0	62.03
Tip-X [76]	0	0	-	0	62.11
APE	0	0	-	0	**63.41**
需要训练					
CoOp [100]	14 h	200	>10	0.01 M	62.95
CLIP-Adapter [24]	50 分钟	200	0.004	0.52 M	63.59
Tip-Adapter-F [91]	5 分钟	20	0.030	16.3 M	65.51
APE-T	5 分钟	20	0.002	0.51 M	**66.07**

表1. **准确率（%）与效率的比较** 在16样本 ImageNet 上 [16]。"GFLOPs"是在训练过程中通过梯度反向传播计算得出的。

数据集	来源	目标	
	ImageNet [16]	-V2 [16]	-草图 [66]
零样本			
CLIP [64]	60.33	53.27	35.44
CALIP [30]	60.57	53.70	35.61
无需训练			
Tip-Adapter [91]	62.03	54.60	35.90
APE	**63.42**	**55.94**	**36.61**
训练			
CoOp [100]	62.95	54.58	31.04
CLIP-Adapter [24]	63.59	55.69	35.68
Tip-Adapter-F [91]	65.51	57.11	36.00
APE-T	**66.07**	**57.59**	**36.36**

表2. **APE 和 APE-T 的域泛化性能（%）**。我们使用16样本 ImageNet [16] 作为分布外测试前的训练数据。

我们的 APE 和 APE-T 在 ImageNet-V2 上均取得了显著的分布外性能。对于分布偏移更大的 ImageNet-Sketch，我们的免训练 APE 优于所有现有方法，包括需要训练的方法。然而，由于我们在域内 ImageNet 上训练了类别残差，APE-T 在 ImageNet-Sketch 上的测试表现不如 APE。

5. 消融研究

本节中，我们进行了大量消融实验，分别考察了我们的方法中先验细化模块、无需训练的 APE 和需要训练的 APE-T 所作的贡献。

先验细化模块。 在图9 (a)中，我们评估了两种细化标准——类间相似度和方差的影响，并采用基于 ResNet-50 [32] 的免训练 APE 作为基线。如图所示，缺少相似度或方差都会损害性能。此外，我们发现相似度准则比方差发挥着更重要的作用，能够更好地从 CLIP 提取的表征中选出最具判别性的通道。接着，在图9(b)中，我们考察了细化通道数 Q的影响。对于所有 shot，通道数处于 $[500, 900]$ 范围内时，性能更好。这表明，相较于其他冗余通道，我们精炼的特征通道更为重要。

无需训练的 APE。图10 （a）中，我们分解了所提出的三方关系，并分别揭示了它们的作用。对于 0-shot 结果，"仅 $\mathbf{R}fW$"表示零样本 CLIP 的性能，准确率为 61.64%。通过加入"$\mathbf{R}_{fW}$ + $\mathbf{R}f'F'$"，带有先验细化的缓存模型在少样本设置下能够取得更高的性能。最后，考虑全部三种关系（"APE"）构建了性能最佳的框架，这表明三边分析带来了有效提升。

需要训练的 APE-T。在图10(b) 中，我们比较了 APE-T 中不同可学习模块的影响，包括视觉$\mathbf{Res}$的类别残差$\mathbf{F}'$ 以及文本$\mathbf{W}$，还有缓存分数$\mathbf{R}_{F'W'}$。从所示结果可以看出，每个可学习组件对于充分释放 APE-T 的潜力都是必不可少的。我们观察到，调优 $\mathbf{W}$ 中的精炼特征通道比$\mathbf{F}'$更为重要。这表明，文本零样本预测比缓存模型发挥着更关键的作用，因为 CLIP 原本的预训练目标是视觉语言对比。

6. 结论

本文提出了一种**自**适应**先**验精**炼**方法（APE），以使 CLIP 适配下游数据集。我们的 APE 依据两项准则提取具有信息量的领域特定特征通道，并深入挖掘 CLIP 提取的三种表征之间的三方关系。在此基础上，我们提出了 APE 的两种模型变体，分别用于无需训练和需要训练的少样本学习。大量实验表明，我们的方法不仅能取得领先的少样本学习结果，还具有更高的效率。未来，我们将致力于把 APE 扩展到分类之外更广泛的 CLIP 下游任务，例如，例如开放世界目标检测、分割和三维点云识别，并进一步提高 APE-T 的训练效率，甚至实现无参数增强 [30, 94, 96]。

参考文献

[1] Maiwan Bahjat Abdulrazzaq 和 Jwan Najeeb Saeed。三种分类算法在手写数字识别中的比较。载于《IEEE高级科学与工程国际会议》，第58–63页，2019年。13[2] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds 等。Flamingo：用于少样本学习的视觉语言模型。arXiv 预印本 arXiv:2204.14198，2022年。13[3] M Augasta 和 Thangairulappan Kathirvalavakumar。神经网络剪枝算法——一项比较研究。《Open Computer Science》，3(3):105–115，2013年。13[4] Shaeela Ayesha、Muhammad Kashif Hanif 和 Ramzan Talib。高维数据降维技术的概述与比较研究。《Information Fusion》，59:44–58，2020年。13[5] Suresh Balakrishnama 和 Aravind Ganapathiraju。线性判别分析——简明教程。信号与信息处理研究所，18(1998):1–8，1998年。13[6] Alberto Baldrati、Marco Bertini、Tiberio Uricchio 和 Alberto Del Bimbo。结合基于 CLIP 的特征进行有效的条件式与组合式图像检索。载于《IEEE计算机视觉与模式识别会议》，第21466–21474页，2022年。13[7]Hangbo Bao、Wenhui Wang、Li Dong、Qiang Liu、Owais Khan Mohammed、Kriti Aggarwal、Subhojit Som 和 Furu Wei。VLMo：采用混合模态专家的统一视觉语言预训练。arXiv 预印本 arXiv:2111.02358，2021年。13[8] Jacek Biesiada 和 Wlodzisław Duch。面向高维数据的特征选择——基于 Pearson 冗余的过滤器。载于《Computer Recognition Systems 2》，第242–249页，2007年。13[9] Davis Blalock、Jose Javier Gonzalez Ortiz、Jonathan Frankle 和 John Guttag。神经网络剪枝的现状如何？《机器学习与系统会议论文集》，2:129–146，2020年。13

[10] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——使用随机森林挖掘判别性成分。载于欧洲计算机视觉会议，页码 446–461，2014。6[11] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakan- tan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是小样本学习器。《神经信息处理系统进展》，33:1877–1901，2020。14[12] Anthony Chen、Kevin Zhang、Renrui Zhang、Zihan Wang、Yuheng Lu、Yandong Guo 和 Shanghang Zhang。Pimae：用于三维目标检测的点云与图像交互式掩码自编码器。CVPR 2023，2023。13[13] Jiaben Chen、Renrui Zhang、Dongze Lian、Jiaqi Yang、Ziyao Zeng 和 Jianbo Shi。iquery：将乐器作为视听声音分离的查询。CVPR 2023，2022。13[14] Mircea Cimpoi、Subhransu Maji、Iasonas Kokkinos、Sammy Mohamed 和 Andrea Vedaldi。野外纹理描述。载于IEEE计算机视觉与模式识别会议论文集，页码 3606–3613，2014。6[15] Manoranjan Dash 和 Huan Liu。用于分类的特征选择。《智能数据分析》，1(1-4):131–156，1997。13[16] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于IEEE计算机视觉与模式识别会议，页码 248–255，2009。1, 2, 4, 6, 8, 14, 15, 16[17] Nanqing Dong 和 Eric P Xing。基于原型学习的小样本语义分割。载于BMVC，第3卷，2018年4月。[18] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过16×16个词：用于大规模图像识别的 Transformer。arXiv预印本 arXiv:2010.11929，2020。14, 16[19] Adel Sabry Eesa、Adnan Mohsin Abdulazeez Brifcani 和 Zeynep Orman。乌贼算法——一种新颖的仿生优化算法。《国际科学与工程研究杂志》，4(9):1978–1986，2013。13[20] Adel Sabry Eesa、Zeynep Orman 和 Adnan Mohsin Abdulazeez Brifcani。一种基于乌贼优化算法的入侵检测系统特征选择新方法。《专家系统及其应用》，42(5):2670–2679，2015。13[21] Pablo A Est´evez、Michel Tesmer、Claudio A Perez 和 Jacek M Zurada。归一化互信息特征选择。IEEE神经网络汇刊，20(2):189–201，2009。13[22] Li Fei-Fei、Rob Fergus 和 Pietro Perona。从少量训练样本中学习生成式视觉模型：在101个物体类别上测试的增量式贝叶斯方法。载于IEEE计算机视觉与模式识别会议研讨会，页码 178–178，2004。6

[23] Yulu Gan、Xianzheng Ma、Yihang Lou、Yan Bai、Renrui Zhang、Nian Shi 和 Lin Luo。装点新来者：用于持续测试时自适应的视觉域提示。AAAI 2023最佳学生论文，2022。13[24] Peng Gao、Shijie Geng、Renrui Zhang、Teli Ma、Rongyao Fang、Yongfeng Zhang、Hongsheng Li 和 Yu Qiao。CLIP-Adapter：通过特征适配器改进视觉-语言模型。arXiv预印本 arXiv:2110.04544，2021。2, 3, 6, 7, 8, 13[25] Peng Gao、Renrui Zhang、Rongyao Fang、Ziyi Lin、Hongyang Li、Hongsheng Li 和 Qiao Yu。先模仿，再重建：通过特征模仿增强掩码自编码器。arXiv预印本 arXiv:2303.05475，2023。3[26] Golnaz Ghiasi、Xiuye Gu、Yin Cui 和 Tsung-Yi Lin。开放词汇图像分割。arXiv预印本 arXiv:2112.12143，2021。13[27] Samantha Guerriero、Barbara Caputo 和 Thomas Mensink。DeepNCM：深度最近类均值分类器。2018。4[28] Ziyu Guo、Xianzhi Li 和 Pheng Ann Heng。Joint-MAE：用于三维点云预训练的2D-3D联合掩码自编码器。arXiv预印本 arXiv:2302.14007，2023。13[29] Ziyu Guo、Yiwen Tang、Renrui Zhang、Dong Wang、Zhigang Wang、Bin Zhao 和 Xuelong Li。ViewRefer：结合GPT与原型引导，掌握用于三维视觉定位的多视图知识。arXiv预印本 arXiv:2303 16894，2023 13。[30] Ziyu Guo*、Renrui Zhang*、Longtian Qiu*、Xianzheng Ma、Xupeng Miao、Xuming He 和 Bin Cui。CALIP：通过无参数注意力增强CLIP的零样本性能。AAAI 2023口头报告，2022。8, 9[31] Huy Ha 和 Shuran Song。语义抽象：利用2D视觉语言模型理解开放世界3D场景。载于2022年机器人学习会议论文集，2022。13[32] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议论文集，页码 770–778，2016。2, 4, 6, 8, 14, 15, 16[33] Patrick Helber、Benjamin Bischke、Andreas Dengel 和 Damian Borth。EuroSAT：用于土地利用与土地覆盖分类的新型数据集和深度学习基准。IEEE应用地球观测与遥感精选主题期刊，12(7):2217–2226，2019。6[34] Hengyuan Hu、Rui Peng、Yu-Wing Tai 和 Chi-Keung Tang。网络修剪：一种面向高效深度架构的数据驱动神经元剪枝方法。arXiv预印本 arXiv:1607.03250，2016。13[35] Peixiang Huang、Li Liu、Renrui Zhang、Song Zhang、Xinli Xu、Baichao Wang 和 Guoyi Liu。TIG-BEV：通过目标内部几何学习实现多视图BEV 3D目标检测。arXiv预印本 arXiv:2212.13979，2023。13[36] Xuan Huang、Lei Wu 和 Yinsong Ye。降维技术综述。《国际模式识别与人工智能杂志》，33(10):1950017，2019。13[37] Saumya Jetley、Bernardino Romera-Paredes、Sadeep Jaya- sumana 和 Philip Torr。原型先验：从改进分类到零样本学习。arXiv 预印本 arXiv:1512.01192，2015。4[38] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉语言表征学习。载于国际机器学习会议，第4904–4916页。PMLR，2021。1[39] Haojun Jia ng、Yuanze Lin、Dongchen Han、Shiji Song 和 Gao Huang。Pseudo-Q：为视觉定位生成伪语言查询。载于IEEE/CVF 计算机视觉与模式识别会议论文集，第15513–15523页，2022。13[40] Xin Jin、Anbang Xu、Rongfang Bie 和 Ping Guo。使用 SAGE 基因表达谱进行癌症分类的机器学习技术与卡方特征选择。载于生物医学应用数据挖掘：PAKDD研讨会，第106–115页，2006。13[41] Jonathan Krause、Michael Stark、Jia Deng 和 Li Fei-Fei。用于细粒度分类的三维对象表征。载于 IEEE 国际计算机视觉会议研讨会，第554–561页，2013。6[42] Benno Krojer、Vaibhav Adlakha、Vibhav Vineet、Yash Goyal、Edoardo Ponti 和 Siva Reddy。从上下文描述中检索图像。arXiv 预印本 arXiv:2203.15867，2022。13[43] Yann LeCun、John Denker 和 Sara Solla。最优脑损伤。神经信息处理系统进展，第2卷，1989。13[44] Namhoon Lee、Thalaiyasingam Ajanthan 和 Philip HS Torr。SNIP：基于连接敏感性的单次网络剪枝。arXiv 预印本 arXiv:1810.02340，2018。13[45] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型引导语言图像预训练。arXiv 预印本 arXiv:2301.12597，2023。13[46] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：面向统一视觉语言理解与生成的语言图像预训练引导。载于国际机器学习会议，第12888–12900页。PMLR，2022。1，13[47] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。先对齐后融合：通过动量蒸馏进行视觉与语言表征学习。神经信息处理系统进展，第34卷：9694–9705，2021。13[48] Yanghao Li、Haoqi Fan、Ronghang Hu、Christoph Feicht- enhofer 和 Kaiming He。通过掩码扩展语言图像预训练。arXiv 预印本 arXiv:2212 00794，2022 1。 。[49] Yangguang Li、Feng Liang、Lichen Zhao、Yufeng Cui、Wanli Ouyang、Jing Shao、Fengwei Yu 和 Junjie Yan。监督无处不在：一种数据高效的对比语言图像预训练范式。arXiv 预印本 arXiv:2110.05208，2021。13

[50] Tailin Liang、John Glossner、Lei Wang、Shaobo Shi 和 Xiaotong Zhang。深度神经网络加速的剪枝与量化：综述。Neurocomputing，461：370–403，2021。13[51] Guoliang Lin 和 Hanjiang Lai。从因果视角重新审视少样本学习。arXiv 预印本 arXiv:2209.13816，2022。3，5 [52] Ziyi Lin、Shijie Geng、Renrui Zhang、Peng Gao、Gerard de Melo、Xiaogang Wang、Jifeng Dai、Yu Qiao 和 Hong- sheng Li。冻结的 CLIP 模型是高效的视频学习器。ECCV 2022，2022。3[53] Zhuang Liu、Ji anguo Li、Zhiqiang Shen、Gao Huang、Shoumeng Yan 和 Changshui Zhang。通过网络瘦身学习高效的卷积网络。载于 IEEE 国际计算机视觉会议论文集，第2736–2744页，2017。13[54] Ilya Loshchilov 和 Frank Hutter。SGDR：带热重启的随机梯度下降。arXiv 预印本 arXiv:1608.03983，2016。7[55] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101，2017。6[56] Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew Blaschko 和 Andrea Vedaldi。飞机的细粒度视觉分类。arXiv 预印本 arXiv:1306.5151，2013。6[57] Norman Mu、Alexander Kirillov、David Wagner 和 Saining Xie。SLIP：自监督与语言图像预训练相结合。arXiv 预印本 arXiv:2112.12750，2021。1，13[58] Maria-Elena Nilsback 和 Andrew Zisserman。大规模类别花卉的自动分类。载于2008年第六届印度计算机视觉、图形学与图像处理会议，第722–729页，2008。6[59] D Lakshmi Padmaja 和 B Vishnuvardhan。科学数据降维中特征子集选择方法的比较研究。载于 IEEE 高级计算国际会议，第31–34页，2016。13[60] Omkar M Parkhi、Andrea Vedaldi、Andrew Zisserman 和 CV Jawahar。猫和狗。载于 IEEE 计算机视觉与模式识别会议，第3498 3505页，2012 6 –。[61] Hanchuan Peng、Fuhui Long 和 Chris Ding。基于最大依赖、最大相关和最小冗余互信息准则的特征选择。IEEE 模式分析与机器智能汇刊，27(8)：1226–1238，2005。13[62]Sarah Pratt、Rosanne Liu 和 Ali Farhadi。鸭嘴兽长什么样？为零样本图像分类生成定制化提示。arXiv 预印本 arXiv:2209.03320，2022。6，14[63] Longtian Qiu、Renrui Zhang、Ziyu Guo、Ziyao Zeng、Yafeng Li 和 Guangnan Zhang。VT-CLIP：利用视觉引导文本增强视觉语言模型。arXiv 预印本 arXiv:2112.02399，2021。3[64] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。学习可从自然语言迁移的视觉模型，

视觉。在国际机器学习会议上，页码 8748–8763，2021。1, 3, 8, 13, 14[65] Yongming Rao、Wenliang Zhao、Guangyi Chen、Yansong Tang、Zheng Zhu、Guan Huang、Jie Zhou 和 Jiwen Lu。DenseCLIP：通过上下文感知提示实现语言引导的密集预测。收录于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 18082–18091，2022年。1 [66] Benjamin Recht、Rebecca Roelofs、Ludwig Sch midt 和 Vaishaal Shankar。ImageNet分类器能否泛化到 ImageNet？收录于国际机器学习会议，页码

5389 5400 2019 6, 8 – 。PMLR，。

[67] Russell Reed。剪枝算法综述。《IEEE神经网络汇刊》，4(5):740–747，1993。13[68] Fadil Santosa 和 William W Symes。带限反射地震图的线性反演。《SIAM科学与统计计算杂志》，7(4):1307–1330，1986。13[69] Hengcan Shi、Munawar Hayat、Yicheng Wu 和 Jianfei Cai。ProposalCLIP：利用 CLIP线索进行无监督开放类别物体提议生成。收录于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 9611–9620，2022年。1[70] Gyungin Shin、Weidi Xie 和 Samuel Albanie。ReCo：检索与协同分割，实现零样本迁移。arXiv 预印本 arXiv:2206.07045，2022年。13[71] Manli Shu、Weili Nie、De-An Huang、Zhiding Yu、Tom Goldstein、Anima Anandkumar 和 Chaowei Xiao。视觉-语言模型零样本泛化的测试时提示调优。arXiv 预印本 arXiv:2209.07511，2022年。3[72] Jake Snell、Kevin Swersky 和 Richard Zemel。用于少样本学习的原型网络。神经信息处理系统进展，

30, 2017 4 。Khurram

[73] Soomro、Amir Roshan Zamir 和 Mubarak Shah。UCF101：一个包含野外视频中 101 个人类动作类别的数据集。arXiv 预印本 arXiv:1212.0402，2012年。6[74] Jiliang Tang、Salem Alelyani 和 Huan Liu。用于分类的特征选择：综述。《数据分类：算法与应用》，第 37 页，2014年。13 [75] Robert Tibshirani。通过 LASSO 进行回归收缩与选择。《英国皇家统计学会杂志：B辑（方法论）》，

58 1 267 288, 1996 13 ( ): – 。Vishaal

[76] Udandarao、Ankush Gupta 和 Samuel Albanie。SUS-X：视觉-语言模型的免训练仅名称迁移。arXiv 预印本 arXiv:2211.16198，2022年。2, 3, 4, 5, 6, 7, 8, 13, 14[77] S Velliangiri、SJPCS Alagumuthukrishnan 等。高效计算的降维技术综述。Procedia计算机科学，

165 104 111, 2019 13 : – 。

[78] B Venkatesh 和 J Anuradha。特征选择及其方法综述。《控制论与信息技术》，19(1):3–26，2019年。13[79] Haohan Wang、Songwei Ge、Zachary Lipton 和 Eric P Xing。通过惩罚局部预测能力学习鲁棒全局表征。收录于神经信息处理系统进展，页码 10506–10518，2019年。6

[80] Mengmeng Wang、Jiazheng Xing 和 Yong Liu。Action-clip：视频动作识别的新范式。arXiv 预印本 arXiv:2109.08472，2021。3[81] Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhil-iang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mohammed、Saksham Singhal、Subhojit Som 等。图像作为外语：面向所有视觉及视觉-语言任务的 BEiT 预训练。arXiv 预印本 arXiv:2208.10442，2022年。13[82] Wenguan Wang、Cheng Han、Tianfei Zhou 和 Dongfang Liu。基于深度最近质心的视觉识别。arXiv 预印本 arXiv:2209.07383，2022年。4[83] Zhaoqing Wang、Yu Lu、Qiang Li、Xunqiang Tao、Yandong Guo、Mingming Gong 和 Tongliang Liu。Cris：由 CLIP 驱动的指代表达图像分割。收录于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 11686–11695，2022年。1[84] Yanmin Wu、Xinhua Cheng、Renrui Zhang、Zesen Cheng 和 Jian Zhang。EDA：面向三维视觉与语言学习的显式文本解耦与密集对齐。计算机视觉与模式识别会议。2023，2022 13[85] Jianxiong Xiao、James Hays、Krista A Ehinger、Aude Oliva 和 Antonio Torralba。SUN数据库：从修道院到动物园的大规模场景识别。收录于 IEEE计算机学会计算机视觉与模式识别会议论文集，页码 3485–3492，2010年。6[86] Runxin Xu、Fuli Luo、Chengyu Wang、Baobao Chang、Jun Huang、Songfang Huang 和 Fei Huang。从密集到稀疏：通过对比剪枝实现更好的预训练语言模型压缩。arXiv 预印本 arXiv:2112 07198，2021 13 。 。[87] Lewei Yao、Jianhua Han、Youpeng Wen、Xiaodan Liang、Dan Xu、Wei Zhang、Zhenguo Li、Chunjing Xu 和 Hang Xu。DetCLIP：通过词典增强的视觉概念并行预训练实现开放世界检测。arXiv 预印本 arXiv:2209.09407，2022年。1[88] Xin Yuan、Zhe Lin、Jason Kuen、Jianming Zhang、Yilin Wang、Michael Maire、Ajinkya Kale 和 Baldo Faieta。用于视觉表征学习的多模态对比训练。收录于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 6995–7004，2021。1[89] Rizgar Zebari、Adnan Abdulazeez、Diyar Zeebaree、Dilovan Zebari 和 Jwan Saeed。面向特征选择与特征提取的降维技术综述。《应用科学与技术趋势杂志》，1(2):56–70，2020年。13[90] Renrui Zhang、Hanqiu Deng、Bohao Li、Wei Zhang、Hao Dong、Hongsheng Li、Peng Gao 和 Yu Qiao。预训练模型协作造就更好的小样本学习器。arXiv 预印本 arXiv:2209.12255，2022年。3[91] Renrui Zhang、Rongyao Fang、Wei Zhang、Peng Gao、Kunchang Li、Jifeng Dai、Yu Qiao 和 Hongsheng Li。Tip-Adapter：无需训练的 CLIP-Adapter，实现更好的视觉-语言建模。arXiv 预印本 arXiv:2111 03930，2021 2, 3, 5, 6, 7, 8, 13, 14 。 。[92] Renrui Zhang、Ziyu Guo、Wei Zhang、Kunchang Li、Xu-peng Miao、Bin Cui、Yu Qiao、Peng Gao 和 Hongsheng Li。Pointclip：通过 CLIP 实现点云理解。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 8552–8562 页，2022年。1, 3[93] Renrui Zhang、Xiangfei Hu、Bohao Li、Siyuan Huang、Hanqiu Deng、Hongsheng Li、Yu Qiao 和 Peng Gao。提示、生成，然后缓存：基础模型级联造就强大的少样本学习器。CVPR 2023，2023。3[94] Renrui Zhang、Liuhui Wang、Ziyu Guo 和 Jianbo Shi。最近邻与深度神经网络相结合，助力点云分析。载于 WACV 2023，2022。9[95] Renrui Zhang、Liuhui Wang、Yu Qiao、Peng Gao 和 Hongsheng Li。通过图像到点云掩码自编码器，从二维预训练模型中学习三维表示。CVPR 2023，2023。3[96] Renrui Zhang、Liuhui Wang、Yali Wang、Peng Gao、Hongsheng Li 和 Jianbo Shi。参数并非全部所需：从非参数网络入手进行三维点云分析。CVPR 2023，2023。9[97] Renrui Zhang、Ziyao Zeng 和 Ziyu Guo。语言能理解深度吗？ACM多媒体会议 2022，2022。3[98] Yiwu Zhong、Jianwei Yang、Pengchuan Zhang、Chunyuan Li、Noel Codella、Liunian Harold Li、Luowei Zhou、Xiyang Dai、Lu Yuan、Yin Li 等。RegionCLIP：基于区域的语言-图像预训练。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 16793–16803 页，2022。1[99] Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。面向视觉-语言模型的条件提示学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022。1, 3[100] Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。学习如何为视觉-语言模型设计提示。国际计算机视觉期刊（IJCV），2022。1, 3, 6, 7, 8[101] Xiangyang Zhu、Renrui Zhang、Bowei He、Ziyao Zeng、Shanghang Zhang 和 Peng Gao。PointCLIP v2：调整 CLIP，以实现强大的三维开放世界学习。arXiv 预印本 arXiv:2211.11682，2022。1, 3

A. 相关工作

A.1. 特征选择

所提出的先验细化模块本质上沿通道维度进行特征选择，这是广受认可的一种降维过程。本节将全面回顾特征选择及其与我们方法的联系。

特征选择通过高效地选取能够准确描述或定义数据的特征子集，来尽量降低维度对数据集的影响 [1, 59, 74, 89]。特征选择的主要目标是构建一个规模小但全面的特征子集，以捕捉输入数据的关键方面 [19,20,77]。特征选择有助于模型避免过拟合，并简化训练和推理的计算 [4, 36]。它还通过提炼特定任务的特征来提高模型的可解释性。在机器学习中，降低维度以及由此进行的特征选择，是最常见的噪声消除技术之一 [15,50]。

传统的选择方法依赖统计度量来选择特征 [78]。这些方法独立于学习算法，且所需计算量较少。方差阈值、Fisherscore、Pearson 相关系数 [8], 线性判别分析（LDA） [5], 卡方 [40], 和互信息 [21,61], 等经典统计准则，常用于评估特征的重要性。

在深度神经网络中，通道剪枝是提升内存占用和计算效率的一项关键技术 [3,9,50,67]。剪枝会移除对结果准确率没有显著贡献的冗余参数或神经元。当权重系数接近于零或存在重复时，可能会出现这种情况。传统的剪枝方法，例如最小绝对收缩与选择算子（LASSO） [68, 75], 、岭回归 [75],和最优脑损伤（OBD） [43], 得到了广泛应用。此外，近期的研究工作还将通道剪枝融入各种视觉或语言编码器 [34,44,53,86]。

与它们相比，所提出的自适应先验细化方法考虑视觉与语言表征之间的一致性，以减少冗余，并针对不同下游领域自适应地细化任务感知特征。它不仅降低了计算量和参数量，还提升了少样本性能。

A.2. 视觉-语言模型

近年来，多模态学习领域对视觉-语言（VL）预训练的关注度日益提高，并为各种任务提供了基础模型 [12, 13, 23, 28, 29,35, 84], [45,46]。在网络规模数据集上训练的现有视觉-语言模型（VLMs）展现出卓越的迁移能力，可应用于各种下游任务 [7, 47, 49, 57, 81]。例如，BLIP 训练了一个用于文本-图像检索、视觉问答及其他跨模态生成任务的多模态编码器-解码器网络 [46]。SLIP 将自监督融入视觉-语言对比学习，从而确保高效的预训练 [57]。Flamingo 仅通过少量输入/输出示例，便增强了视觉-语言模型在跨模态任务上的少样本能力 [2]。近期提出的 BLIP-2 则高效利用预训练的视觉-语言模型，并通过跨模态Transformer执行生成任务 [45]。这些方法通过大规模对比训练，显著提升了预训练模型在下游任务上的泛化能力。视觉-语言数据之间的对齐已成为监督视觉-语言训练的成熟原则。训练完成后，现成模型展现出卓越的特征提取能力。其中的代表模型是 CLIP [64]。

CLIP 在 4 亿组来源于互联网的图像-文本对上训练后，展现出卓越的视觉-语言表征对齐能力。它已被广泛应用于分类 [24, 76, 91], 视觉定位 [31, 39], 图像检索 [6, 42], 语义分割 [26, 70], 等任务，且只需进行有限的适配。在这项工作中，我们提出了一种基于 CLIP 的新型小样本框架，该框架也可扩展至其他视觉-语言模型。

B. 方法

本节中，我们详细推导正文公式 S 5 和公式14 中的优化目标。通道 通道 S_k 的平均类间相似度表示为

$$
\begin{aligned}
S &= \frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\delta(\mathbf{x}^i\odot\mathbf{B},\mathbf{x}^j\odot\mathbf{B}) \\
&= \frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{d=1}^{D} x_d^i\cdot x_d^j\cdot B_d \\
&= \frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{k=d_1}^{d_Q} x_k^i\cdot x_k^j \\
&= \sum_{k=d_1}^{d_Q}\left(\frac{1}{C^2}\sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C} x_k^i\cdot x_k^j\right) \\
&= \sum_{k=d_1}^{d_Q} S_k,
\end{aligned}
\tag{14}
$$

其中我们使用与正文相同的符号。

之后，我们在图50 [32]中展示了 ResNet-11骨干网络全部 1024 个通道的类间相似度和方差标准。我们在 Ima- geNet [16] 上使用文本表征进行统计。我们将主论文公式 λ 中的平衡因子8 设为 0.7。我们根据混合准则 J_k按升序排列通道。从图中可以看出，如果选取前 $Q = 500$ 个通道，类间相似度较低，而方差较大。此外，我们还观察到部分通道（约 $k = 900$）未被激活，因为它们的 S_k 和 V_k 均为 0。这些可视化结果表明，所提出的准则能够有效识别冗余通道。

更多相似度图见图12。从这些示例中可以看出，经过筛选的通道主要包含目标类别的信息，而其余通道则包含更多环境噪声和冗余信息。

C. 实验

设置。 对于先验细化，每个数据集所选的通道数为 Q，如表 3所示。对于文本提示，我们遵循 [76]，将 CuPL [62] 与基于模板的提示 [64]结合使用。我们在表 3 [11] 中展示了用于生成 CuPL 提示的 GPT-5和6语言指令。基于模板的提示列于表 7。

不同骨干网络。我们在图中采用不同的 CLIP编码器实现了我们的方法和现有模型——

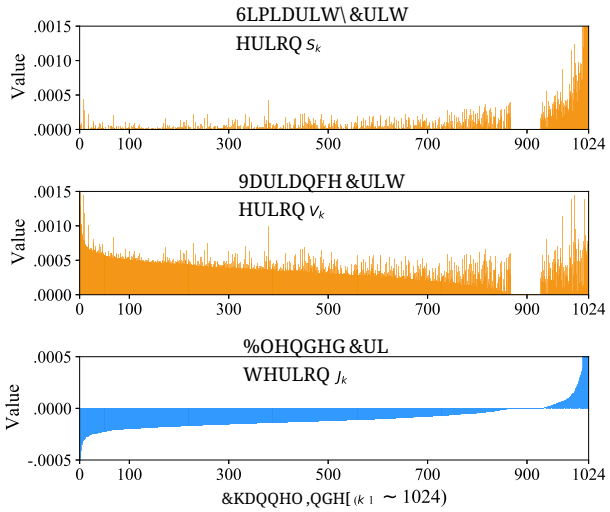


图11。 **ResNet-50 编码器 1024 个通道上的相似度与方差标准可视化。**

13。我们采用最佳设置，仅替换编码器网络。我们研究了 ResNet [32] 和视觉 Transformer（ViT） [18] 骨干网络，无论是在训练设置还是免训练设置下，我们仍然取得了最高准确率。

不同提示。 我们在图15中考察了提示的影响。涉及三种提示。模板提示是广泛使用的版本，例如，按照 [91]，为 ImageNet 集成 7 个不同的模板。CuPL 提示由 [62] 提出，通过 GPT-3 生成。我们在工作中将模板提示与 CuPL 提示集成，记为“CuPL+t”。从图 15可以看出，CuPL+t 提示能够提升所有少样本方法的表现。此外，我们的 APE 和 APE-T 在各种提示下都能保证达到最高准确率。

不同的通道数。 如图所示，我们还在采用 ViT-B/16 骨干网络的 APE 上验证了通道优化过程 14(a)，该过程输出 512 维表征。这证明了先验优化在其他骨干网络上的有效性。即使对于 512 维特征，我们的优化也能滤除冗余和噪声。

平衡因子 λ 和平滑因子 γ。 我们探究了 λ 在公式 8 中，以及 γ 在公式11中的取值所产生的影响。平衡因子 λ 控制相似度准则和方差标准在最终融合准则中的权重 J_k。我们发现，对于 APE， $\lambda = 0.7$ 能取得最佳准确率。这表明相似度准则对 APE 更为重要。平滑因子 γ 控制每个部分的贡献



输入图像



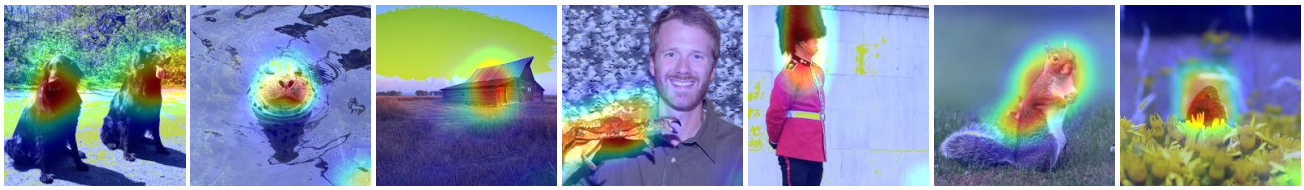
(a) 精炼特征通道



(b) 其他特征通道



输入图像



(c) 精炼特征通道



(d) 其他特征通道

图12。 **更多相似性映射示例**。 我们使用 ResNet-50 [32] 视觉编码器，并从 1024 个特征通道中精炼出 512 个。这些示例来自 ImageNet 验证集 [16]。

数据集	ImageNet	Caltech-101	DTD	EuroSAT	FGVC
Q	500	900	800	800	900
Food101	Flowers102	Pets	Cars	SUN397	UCF101
800	800	800	500	800	800

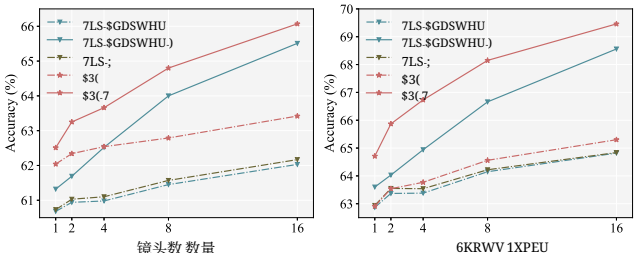
表 3. **精炼通道数** Q （针对每个数据集）。骨干网络为 ResNet-50 [32]，用于提取 1024 维表征。

平衡因子 λ	0.1	0.3	0.5	0.6	**0.7**	0.8
	63.02	63.15	63.27	63.33	**63.42**	63.37
平滑因子 γ	0.0	0.05	0.1	0.15	**0.2**	0.3
	62.64	63.04	63.31	63.34	**63.42**	63.06

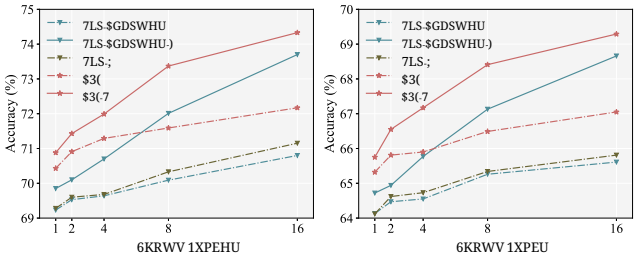
表 4. **超参数消融研究（%）**：ImageNet 上的 APE [16]。我们研究了公式 λ 8 中的混合平衡因子 γ，以及公式 11 中的平滑因子。实验采用 16样本 设置，并使用 ResNet-50 [32] 骨干网络。

从训练样本到最终预测。当 γ 从 0 增加到 0.1 时，我们观察到准确率显著提升，这表明关系 $\mathbf{R}_{F'W'}$ 的有效性。当 γ 从 0.2 增加到 0.3 时，即关系 *i.e.* $\mathbf{R}_{F'W'}$ 变得更加尖锐时，性能迅速下降，这表明少样本性能对超参数 γ 很敏感。

与主成分分析（PCA）的比较。最后，我们在图 14(b) 中将提出的显著通道精炼方法与主成分分析（PCA）进行比较，这两者都是降维方法。我们在 ImageNet 上使用 ResNet-50 骨干网络进行此实验。我们从文本表示中提取 $Q = 500$个主成分用于基于 PCA 的方法，与我们的精炼方法类似。我们分别在无需训练和需要训练的设置下进行实现。对于无需训练的变体（记为“PCA”），我们使用经 PCA 变换后的表示替代 APE 中的精炼版本。对于需要训练的版本（记为“PCA-T”），我们仅针对少样本学习优化主成分。结果如图 14 (b) 所示。我们观察到，我们的方法在无需训练和需要训练两种模式下均优于基于 PCA 的方法。我们认为，这与 PCA 算法中的基变换有关：变换后的表示中引入了意料之外的偏差，不利于预测。相比之下，我们的精炼方法完全继承 CLIP 的知识，不进行变换，因此适用于基于相似度的视觉语言方案。

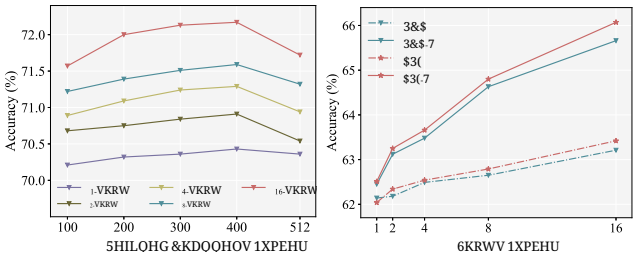


(a) 使用 ResNet-50 的结果 (b) 使用 ResNet-101 的结果



(c) 使用 ViT-B/16 的结果 (d) 使用 ViT-B/32 的结果

图13. **不同骨干网络的消融研究**。虚线和实线分别表示无需训练方法和需要训练的方法。共涉及四种网络结构：ResNet-50 [32], ResNet-101 [32], ViT-B/16 [18],和 ViT-B/32 [18]。



(a) ViT-B/16 的通道数 (b) 与主成分分析的比较

图14. **先验细化的更多消融研究**。

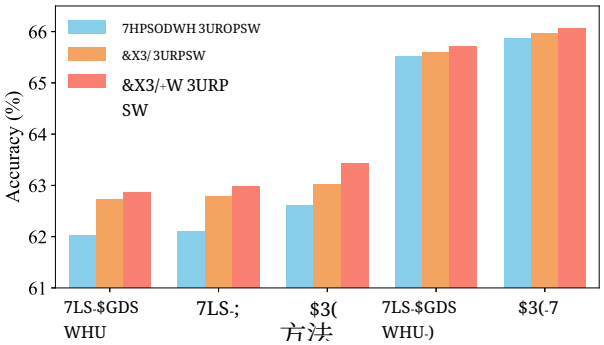


图15 . **用于 APE 和 APE- 的不同提示 T.**

数据集	GPT-3 指令
ImageNet	“描述一下 {}的外观” “如何识别 {}？” “ {}看起来是什么样的？” “描述一张来自互联网的 {}图像” “一张 {}图像的说明：”
Caltech101	“描述一下 {}的外观” “ {} 看起来是什么样的” “描述一张 {}的照片”
DTD	“ {}材料看起来是什么样的？” “ {}表面看起来是什么样的？” “ {}纹理看起来是什么样的？” “ {}物体看起来是什么样的？” “ {}东西看起来是什么样的？” “ {}图案看起来是什么样的？”
EuroSAT	“描述一幅关于 {}的航空卫星图像” “一张 {}的卫星照片看起来是什么样” “从视觉上描述一幅 {}的卫星图像”
FGVCAircraft	“描述一架 {}飞机”
Flowers102	“ {}花是什么样子的” “描述一下 {}的外观” “一张关于 {}的图像的说明文字” “从视觉上描述一种 {}花”
Food101	“描述一下 {}的外观” “从视觉上描述一下 {}” “你如何判断照片中的食物是 {}？”
OxfordPets	“描述一下 {}宠物的外观” “从视觉上描述一下 {}，一种宠物”
StanfordCars	“如何识别一辆 {}” “ {}的描述，一种汽车” “一张 {}照片的说明：” “ {}的主要特征是什么？” “ {}外观的描述” “ {}汽车有哪些特征？” “描述一张来自互联网的 {}图像” “ {} 看起来是什么样子？” “描述一下 {}这种汽车，看起来是什么样子”
SUN397	“描述一下 {}是什么样子” “如何识别 {}？” “描述一张 {}的照片”
UCF101	“描述一下正在进行 {}的人是什么样子” “描述 {}的过程” “一个人如何 {}”

表 5. **CuPL 中使用的 GPT-3 指令**（1/2）。

数据集	GPT-3 指令
ImageNet-V2	“描述一下 {}看起来是什么样子” “如何识别 {}？” “ {}看起来是什么样子？” “描述一张来自互联网的 {}图像” “一张 {}图像的说明：”
ImageNet-Sketch	“描述一下 {}的外观” “如何识别 {}？” “ {}看起来是什么样子？” “描述一张来自互联网的 {}图像” “一张 {}的图像说明：”

表 6. **GPT-3 指令（在 CuPL 中使用**（2/2）。

数据集	模板提示
ImageNet	“一张 {} 的照片。” “一张糟糕的 {} 照片。” “一件折纸 {}。” “一张大型 {}的照片。” “电子游戏中的 {}。” “ {}的艺术作品。” “一张小型 {}的照片。”
Caltech101	“一张 {}的照片。”
DTD	“{} 纹理。”
EuroSAT	“一张居中的 {}卫星照片。”
FGVCAircraft	“一张 {}的照片，它是一种飞机。”
Flowers102	“一张 {}的照片，一种花。”
Food101	“一张 {}的照片，一种食物。”
OxfordPets	“一张 {}的照片，一种宠物。”
StanfordCars	“一张 {}的照片。”
SUN397	“描述一下 {}的样子”
UCF101	“一张有人正在进行 {}的照片。”
ImageNet-V2	“一张 {}的照片。” “一张拍得很差的 {}照片。” “一个折纸 {}。” “一张大型 {}的照片。” “电子游戏中的一个 {}。” “ {}的艺术作品。” “一张小 {}的照片。”
ImageNet-Sketch	“一张 {}的照片。” “一张糟糕的 {}照片。” “一个折纸 {}。” “一张大 {}的照片。” “电子游戏中的一 {}。” “ {}的艺术作品。” “一张小型 {}的照片。”

表7。**基于模板的提示**，适用于每个数据集。