基于属性锚点改进的文本提示学习方法

李政[1], 宋奕兵[2], 程明明[1], 李翔[1]*, 杨健[1]*
[1] PCA Lab, VCIP, 计算机学院, 南开大学, [2] 阿里巴巴达摩院
zhengli97@mail.nankai.edu.cn, yibingsong.cv@gmail.com
{cmm, xiang.li.implus, csjyang}@nankai.edu.cn
翻译：过翔天，南开大学

Abstract

基于文本的提示学习方法 *(Textual-based prompt learning)* 主要采用多个可学习的软提示词元与硬类别词元以级联的方式作为文本输入，旨在实现图像空间与文本（类别）空间在下游任务中的特征对齐。然而，现有的训练仅限于使图像与预定义已知类别对齐，而无法关联到未知类别。在本文中，我们提出利用通用属性作为桥梁，以加强图像与未知类别之间的对齐程度。具体而言，我们为视觉-语言模型引入了一种名为 ***ATPrompt*** 的基于属性锚点的文本提示学习方法。该方法通过在可学习的软提示中加入多个属性词元，将软提示的学习空间从原始的一维类别层面扩展至多维属性层面。通过这种改进，我们将文本提示从以类别为中心的形式转变为属性-类别混合形式。此外，我们引入了一种直接可微的属性搜索方法，用以识别适用于下游任务且具有代表性的属性。作为一种易用的插件式技术，*ATPrompt* 能够无缝替换现有基于文本方法中的基础提示形式，以可忽略的计算开销实现普遍性能提升。在 *11* 个数据集上的广泛实验验证了我们方法的有效性。开源代码可见 *https://github.com/zhengli97/ATPrompt*。

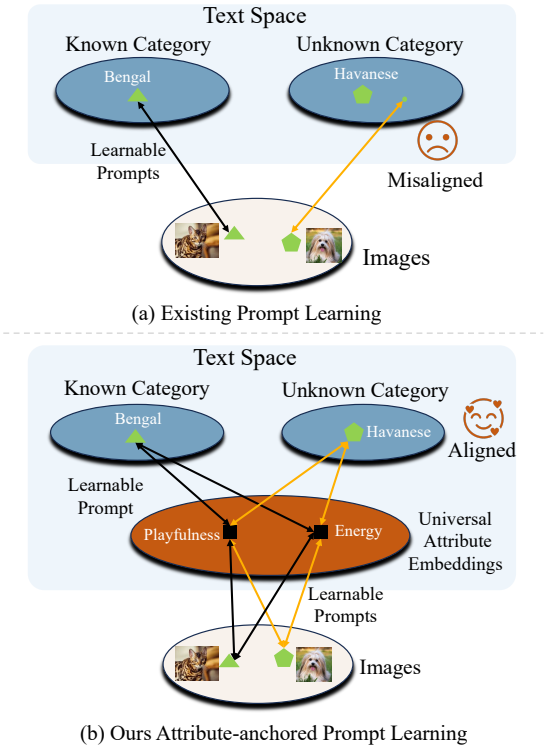


图 1. 通过可学习提示方法的图像与文本（类别）的对齐过程间的对比。(a) 目前的提示学习方法将图像与预定义的类别对齐，但无法与未知类别建立准确的关联。(b) ATPrompt 利用通用属性作为媒介，在图像和未知类别之间实现更准确的对齐。

1. 引言

视觉-语言模型 (VLMs) [15, 26, 27, 33, 40, 44, 45]，例如 CLIP [40] 和 ALIGN [15]，近年来表现优异。这些模型使用对比损失进行训练，以建立图像和文本（类别）空间之间的对齐。受到自然语言处理 (NLP) 领域成功经验的启发 [24, 25]，提示学习 [16, 60, 61] 已成为一种参数高效的解决方案，能够将强大的视觉-语言模型 (VLMs)

*通讯作者

适配到下游任务中。具有一些可学习软提示的模型经过训练可以达到与完全微调的模型相当的性能，甚至优于它们 [16]。根据软提示词元的应用方式，现有的方法大致可以分为基于文本的方法 [18, 19, 29, 51, 60, 61] 和基于视觉的方法 [1, 2, 16, 22, 57]。其中，基于文本的方法是最基础和最直接的，占据多数。

在典型的图像分类任务中，现有的基于文本的方法 [18, 19, 60, 61] 主要采用可学习的软提示词元与硬类别词元级联的传统形式，以替代人工构建的文本提示（如 “a photo of a class”）作为编码器的输入。虽然这类文本提示展现出很强的性能，但它在训练过程中限制图像只能与预定义的已知类别进行对齐，因此无法与未知类别准确关联起来，如图 1(a) 所示。直觉上，当人类面对陌生类别时，常会将其与附加属性（如色彩、形态或纹理）建立关联以增强可理解性与清晰度，而非仅仅是陈述物体名称。例如，人类可能将猎豹描述为：“猎豹是头部娇小的猫科动物，黄色短毛，带有黑色斑点。”或将苹果称为：“那个带有橘色条纹的红色球形水果是苹果。”而非笼统地说“这是猎豹”或“那是苹果”。属性可以作为连接未知类别和已知知识的桥梁。

基于上述观察，我们提出了一种创新方法，将属性作为增强图像与未知类别对齐能力的桥梁。具体而言，本研究为视觉语言模型 (VLMs) 引入了名为 ATPrompt 的基于属性锚点的文本提示学习方法。该方法通过整合多个固定通用属性词元到可学习的软提示中，将软提示的学习空间从原始的一维类别层面扩展至多维属性层面。在锚定属性的引导下，软词元在训练过程中不仅习得类别特定表征，还获取与属性关联的通用表征。如图 1(b) 所示，相较于原始方法，该策略显著提升了图像与未知类别的对齐效果。此外，根据软提示的应用深度，我们分别提出 ATPrompt 的两种架构：浅层版与深层版，以确保与现有不同深度方法 [18, 19, 61] 的兼容性。为确定最优属性组合，我们提出了一种简洁高效的可微分属性搜索方法，该方法从大语言模型构建的候选项池中学习识别合适的属性。该搜索操作只需要执行一次，一旦完成，所选择出的属性就可以直接被 ATPrompt 用于模型训练。

作为一种易用的插件式技术，ATPrompt 能够无缝替换基于文本的提示学习方法中使用的现有形式，从而在无需额外计算开销的情况下实现普遍性能提升。

我们的贡献可总结如下：

- 我们提出了一种高效的基于锚点属性的文本提示学习方法，将软提示的学习空间从一维类别层面扩展至多维属性层面。
- 我们引入了一种有效的可微搜索方法来为下游任务选择合适的属性。
- 我们提出了浅、深两种版本的 ATPrompt，以确保与现有不同深度的提示学习方法兼容。
- 广泛实验表明，ATPrompt 可以无缝集成到现有的基于文本的方法中，从而以可忽略不计的计算成本带来普遍的性能提示。

2. 相关工作

视觉-语言模型的提示学习。受自然语言处理 (NLP) 最新进展的启发 [24, 25]，提示学习 [18, 19, 29, 42, 58, 60–62] 在视觉研究领域受到广泛关注，研究者们致力于将该技术应用于视觉-语言模型 (VLMs) [15, 40, 53]，例如 CLIP。CoOp [61] 是一种开创性的基于文本的方法，它提出了使用软文本词元和硬类别词元组合作为输入的思想。随后的研究 [18, 19, 23, 29, 55, 60] 都主要采用这种文本提示形式。然而，这种形式将软提示约束在一维预定义的类别空间内进行图像对齐，限制了它们在未知类别上的适用性。因此，采用当前文本形式进行训练将更有可能过拟合已有类别，削弱模型对未知类别的零样本泛化能力。为了解决这一限制，学界已提出多种方法 [17, 19, 29, 55, 62]. 例如，KgCoOp [55] 使用人工构建的硬提示在训练过程中对可学习的软提示进行正则化。PromptSRC [19] 利用 CLIP [40] 的原始特征对图像与文本分支的软提示学习做正则化约束。PromptKD [29] 利用预先训练好的强教师模型，去引导带有可学习的提示学生模型的学习 [13, 28, 30, 54]。尽管取得了这些进展，但上述方法均未能解决形式本身固有的限制。在这篇文章中，我们为视觉-语言模型 (VLM) 引入了基于属性锚点的文本提示形式，该形式利用属性作为桥梁，在图像和未知类别之间建立更准确的关联。

视觉-语言模型 (VLMs) 中的属性。在实际应用中，类别通常包含多重属性特征。当人们遇到陌生类别时，往往会使用额外的属性来描述它，以提升表述清晰度，而非仅陈述其名称。受到这一观察的启发，许多研究 [4, 35, 47, 49, 56] 开始利用属性特征以实现其目

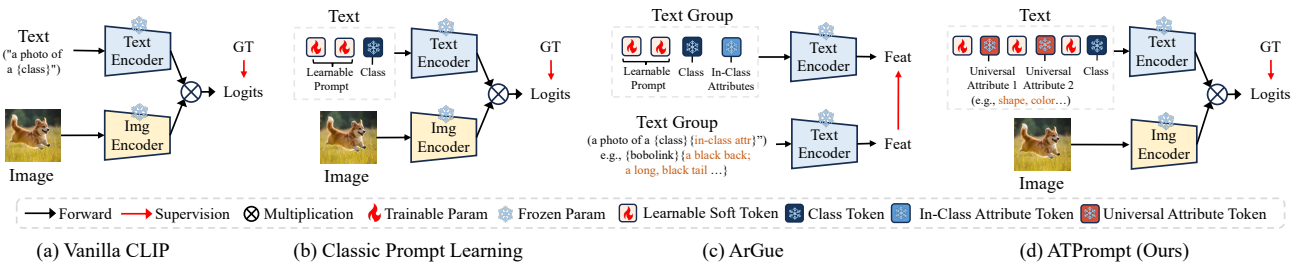


图 2. 现有方法之间的架构比较。(a) 原始 CLIP (Vanilla CLIP) 采用人工构建的文本模板作为文本编码器输入。(b) 经典提示学习 (Classic Prompt Learning) 提出将多个可学习的软词元与类别词元相级联的新文本提示形式。(c) ArGue [47] 利用由大型语言模型挖掘的多个类内属性作为补充信息。利用这些属性构建不同的文本组作为学习目标，从而正则化软词元的学习。通过集成所有组获得最终的预测结果。(d) 我们的 ATPrompt 将通用属性作为学习组件并将它们锚定到现有的软提示模板中。通过这一操作，我们将软词元的学习空间扩展到多维属性层面，并增强图像与未知类别文本的对齐能力。

标。VCD [35] 作为开创性工作，率先提出运用大语言模型 (LLMs) 将类别名称分解为多个类内属性 (如鸟类分解为鸟喙与尾羽) 以提升分类性能。AAPL [20] 提出一种元网络架构，用来提取编码后的图像特征中的视觉属性特征，以增强图像-文本对齐能力。TAP [7] 提出了一种结构化“属性树”方法，通过利用特定属性的知识图来强化视觉-语言模型的性能。ArGue [47] 利用大语言模型挖掘多个类内属性，并将其整合到软提示与固定模板中构建多个文本组。该方法采用固定模板生成的原始文本特征对软词元学习实施约束，具体流程如图 2(c) 所示。现有研究多聚焦于利用类内属性提供辅助信息以提升模型性能，然而在未知类别场景下，需重新获取新类别的属性特征这一过程，往往非常复杂且成本高昂。

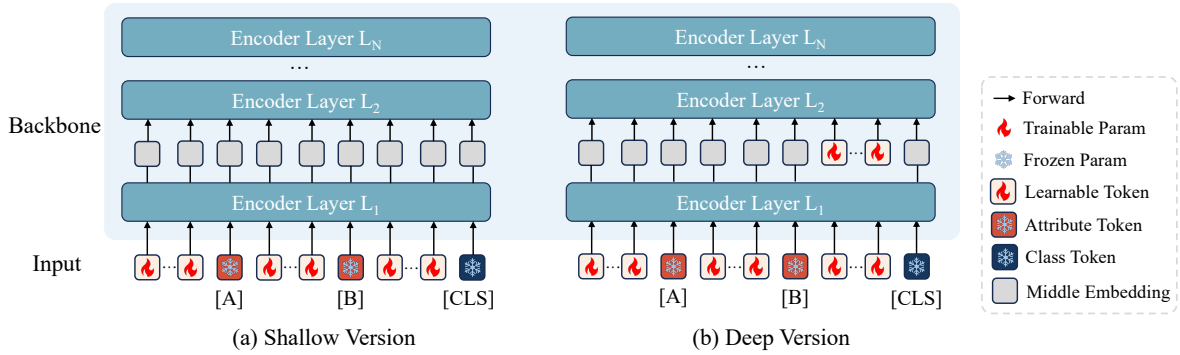
在本文中，我们认为通用（类间）属性相较于先前工作中采用的类内属性更加有效和鲁棒。不同于将属性作为学习目标，我们视其为学习的组件，并提出将通用属性锚定到软提示模板中，使现有的以类为中心的形式 [61] 转换为混合属性-类可学习的文本提示形式。该方法可无缝集成于现有的基于文本的方法中，在不产生额外计算成本的情况下提高已有基线方法的性能。

3. 方法

提示学习 [18, 19, 60–62] 旨在通过训练插入的可学习软词元来增强预训练的视觉-语言模型（如 CLIP）对下游任务的泛化能力。现有的基于文本的方法都遵循经典的提示范式，将软提示词元和硬类别词元拼接起来作为文本编码器的输入，如图 2(b) 所示。本文提出了一种简单有效的文本提示学习方法 ATPrompt，该方法将多个固定的通用属性词元锚定到原始软提示中，如图 2(d) 所示。在属性特征的引导下，软提示通过训练不仅能习得类别特定表征，更能学习属性相关的通用表征。当遇到未知类别时，这些学习到的与属性相关的词元可以提供额外的信息，以优化图像-文本对齐。此外，我们提出了了包含序列化步骤的自动化流程来识别通用属性。首先，我们运利用大语言模型为当前下游任务类别构建属性池；其次，我们提出了一种可微属性搜索方法，旨在从池中筛选出最适配属性锚定提示形式的属性。对于每个任务，此搜索操作只执行一次。一旦属性最终确定，它们将被整合到我们的 ATPrompt 中，以进行针对性的模型微调。

3.1. 背景

视觉-语言模型。现有的视觉-语言模型 [15, 40]，例如 CLIP，在经过 4 亿对图像文本对的训练后已经展现出显著的零样本学习泛化性能。这些模型的核心目标在于习得由各自编码器生成的图像与文本模态之间的对齐关系。给定一个带标签的图像分类数据集 $\mathrm{D} = \{(x, c)\}$，其中包括 N 个类标签 $C = \{c_i\}_{i=1}^{N}$，CLIP 通过计算图像特征与每个类相应的文本特征之间的余弦相似度来进行预测。具体来说，对于每个输入图像 x，通过图像编码器 $h_I(x)$ 进行特征提取，得到特征向量 $u = h_I(x)$。同时，针对每个类别，通过使用人工设计的模板生成一系列文本描述 t。然后，将这些文本描述输入文本编码器 $h_T(x)$，得到文本特征 $w = h_T(t)$。最终，图像 x 被



(a) Shallow Version (b) Deep Version

图 3. 浅层版和深层版的计算流程示意图。以两个属性 [A] 和 [B] 为例。(a) 浅层版将硬属性词元、软提示词元与类别词元拼接后输入编码器计算。(b) 深层版采用相同输入结构，但在完成自注意力计算后丢弃类别相关软提示词元，并在下一层计算前重新引入。两种形式可兼容现有不同深度的提示方法，包括输入级方法如 CoOp [61]、CoCoOp [60] 和深度级方法如 MaPLe [18].

分类为 c 的输出概率按如下公式计算：

$$p(c|x) = \frac{\exp(\cos(u, w_c)/\tau)}{\sum_{i=1}^{N} \exp(\cos(u, w_i)/\tau)}. \quad (1)$$

其中 τ 为温度参数，$\cos(\cdot,\cdot)$ 代表余弦相似度函数。

视觉-语言模型中的提示学习。不同于用于手工设计硬提示用于图文匹配，其存在精度不足、灵活性欠缺等局限性，近期的提示学习研究 [18, 19, 55, 60, 61] 如 CoOp，提出为下游任务学习合适的软文本提示以进行替代。具体来说，将 M 个可学习的软词元 $[T_i]_{i=1}^{M}$ 与硬类别词元 [CLS] 级联后作为文本编码器的输入，如图 2(b) 所示。其形式如下：

$$P_T = [T_1][T_2]...[T_M][\text{CLS}], \quad (2)$$

其中，M 表示软词元的长度。为了简化，我们省略了输入中的前缀和后缀词元。

除了在输入层嵌入软词元外，现有的研究 [16, 18, 19, 29] 还探索了在模型更深层引入软词元。这是通过在 Transformer 块中添加软词元接着在自注意力机制计算后将其移除来实现的。对于第 i 块，这个过程可以描述为：

$$[\text{CLS}_i] = L_i([\text{T}_{i-1}, \text{CLS}_{i-1}]). \quad (3)$$

其中，L_i 表示第 i 个 Transformer 块，T_i 表示可学习软词元的集合，定义为 $\text{T}_i = \{[T_1]_i, ..., [T_M]_i\}$。

3.2. 学习具有通用属性的软提示

我们的方法引入了两种变体，根据应用软词元的层数，可以分为浅层版和深层版，如图 3 所示。

浅层版。我们首先介绍浅层版本，其中硬属性词元仅在输入层被锚定，如图 3(a) 所示。给定两个通用属性，A 和 B，根据式 (2)，输入文本编码器的浅层文本提示 P_T 可表示为：

$$P_T = [T_{a_1}]...[T_{a_m}][\text{A}][T_{b_1}]...[T_{b_m}][\text{B}][T_1]...[T_M][\text{CLS}]. \quad (4)$$

其中，a_m 和 b_m 是指定属性 A 和 B 的软词元长度的超参数。在本方法中，这些参数默认设置为相同值。

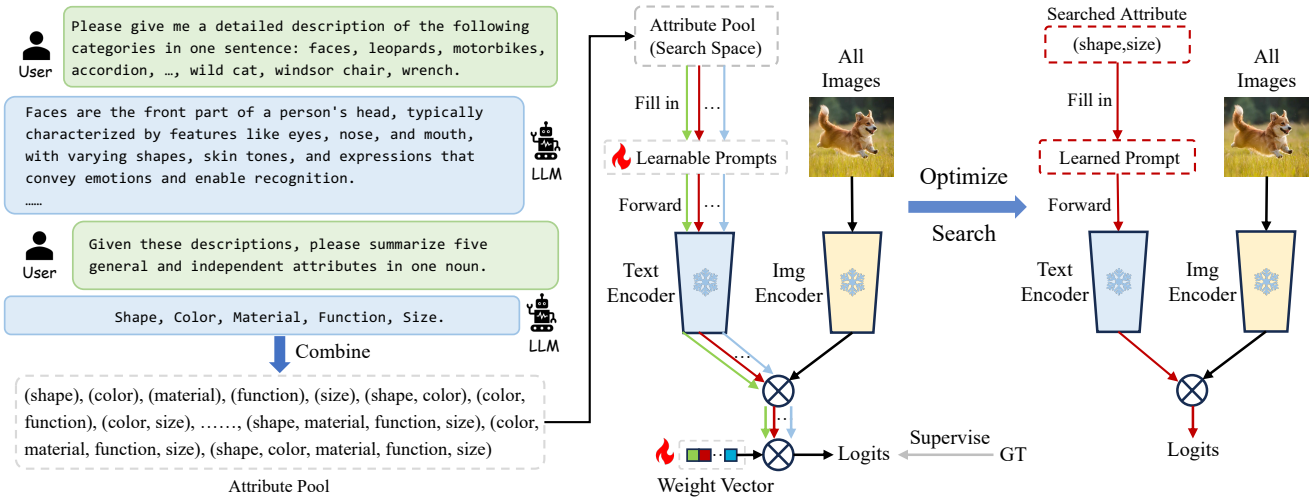
除了将硬类别词元置于文本提示末尾外，我们还可以将它置于中间或前端位置。在表 5 中，我们验证了置于每个位置的性能表现并选择效果最优的末尾位置作为默认形式。

深层版。在这个版本中，可学习的软词元被引入到深层网络的输入中。先前的工作，如 VPT 和 MaPLe，会丢弃所有软词元，并在 Transformer 模块之后重新引入。当此操作应用于属性相关词元时，新引入的过多低级词元与现存的高级词元之间会出现语义差别，从而削弱了层间特征的连续性。在本研究中，我们的方法仅选择性地丢弃输入中与类别相关的软词元并在之后重新添加，具体为 $[T_1], ..., [T_M]$，如图 3(b) 所示。基于式 (3)，ATPrompt 的深层版本可改写如下：

$$[\text{F}_1, \_, \text{CLS}_1] = L_1([\text{T}_{a_0}, \text{A}, \text{T}_{b_0}, \text{B}, \text{T}_0, \text{CLS}_0]), \quad (5)$$

$$[\text{F}_i, \_, \text{CLS}_i] = L_i([\text{F}_{i-1}, \text{T}_{i-1}, \text{CLS}_{i-1}]). \quad i = 2, 3, ..., M. \quad (6)$$

其中，F_i 表示第 i 层 Transformer 计算出的特征。我们在表 6 中证明了该操作的有效性。



(a) Form an attribute pool. (b) Differentiable attribute search.

图 4. 我们的属性搜索流程概述。(a) 我们首先通过迭代查询大型语言模型获取多个独立属性，随后将这些属性聚合形成候选组合池，该组合池将作为搜索过程的输入。(b) 每个候选组合的前向计算过程由不同颜色标识的路径表示。为了找到最优属性，我们采用交替优化算法，该算法联合优化软词元和相应的路径权重向量。训练完成后，选择权重最高的路径所对应的属性组合作为最终输出。

训练。设 θ 表示所有软词元的权重，设 v 表示被选定的固定属性词元。在给定带有标签的数据集 $\mathrm{D}=\{(x,c)\}$ 进行训练，我们的目标是最小化预测值与真实值标签之间的交叉熵损失。这一过程可表述如下：

$$\min_{\theta} L_{train} = \min_{\theta} \sum_{x \in D} \mathrm{CE}(f(x\ v, \theta), c). \tag{7}$$

其中，$f(\cdot)$ 表示 CLIP 模型函数。

3.3. 属性搜索

选择属性需要考虑两个关键因素：属性的内容和数量。尽管直接使用大型语言模型查询是一种直接的方法，但其存在显著缺点。该方法无法为特定下游数据集确定最优属性数量，且仅通过类别名称进行查询可能引入语义偏差。为解决这些问题，我们提出一种自动化流程，为当前下游任务选择合适的属性内容与数量，如图 4 所示。

属性池。受思维链 (CoT) [50, 59] 启发，我们将整个流程分成多个步骤，以增强大型语言模型的推理能力。首先，我们通过提示大语言模型为每个已知类别生成描述性语句，从而丰富与类别相关的信息。接着，以这些描述为上下文，通过提示大语言模型总结出一组在这些类别中通用的独立属性基。然后通过创建这些基的所有可能组合形成一个属性池，如图 4(a) 所示。对于 N 个属性基，属性池中共有 $w = C_N^1 + C_N^2 + ... + C_N^N$ 个候选组合，构成了我们的搜索空间。需要注意的是，我们不考虑属性的顺序，因为排列通常不会引入显著的语义偏差或影响最终性能，我们在接下来的实验中验证了这一结论。

属性搜索。受 DARTS [32] 启发，我们引入了一种可微属性搜索方法，该方法旨在学习从搜索空间 $\mathcal{V}$ 中寻找具有代表性的属性 v，如图 4(b) 所示。为了使搜索空间连续，我们将离散的属性选择放宽至对所有 w 个可能候选的 Softmax 加权和：

$$f(x, v\ \alpha, \theta) = \sum_{i \in \mathcal{V}} \frac{\exp(\alpha_i)}{\sum_{i' \in \mathcal{V}} \exp(\alpha_{i'})} f(x, v_i\ \theta). \tag{8}$$

其中，α_i 表示属性组合 v_i 的权重。属性搜索的任务由此简化为学习候选池的权重向量 α。

放宽处理后，我们的目标是共同学习属性权重 α 和软提示词元 θ。遵循标准做法 [32, 39, 63]，我们通过最小化验证损失 L_{val} 来优化权重 α，同时通过最小化训练损失 L_{train} 来学习软词元。我们采用交替算法 [14, 28] 来解决这一双重优化问题，交替优化这两个子问题：

$$\hat{\alpha} = \arg\min_{\alpha} L_{val}(f(x, v\ \alpha, \hat{\theta}), c), \tag{9}$$

$$\hat{\theta} = \arg\min_{\theta} L_{train}(f(x, v\ \hat{\alpha}, \theta), c). \quad (10)$$

其中 L_{train} 和 L_{val} 都采用交叉熵损失函数。经过搜索，选择权重最高的属性组合 (α_i)。

代价分析。与传统的神经架构搜索 (NAS) 方法 [8, 31, 46] 在计算开销巨大的网络级参数中搜索不同，我们的方法聚焦于轻量的词元级搜索空间。这种设计使我们的方法比以前的方法更高效。实际中，对于某些数据集情况下，我们的搜索方法能在约 5 个轮次内收敛，在单张 A800 GPU 上通常耗时不足 5 分钟。这凸显了我们的方法相较于传统参数架构搜索的实用优势。此外，筛选更小的属性基集合可以缩小搜索空间，还能进一步提升搜索效率。

4. 实验

4.1. 实验设置

基类到新类 (Base-to-Novel) 的泛化能力。遵循 [19, 29, 60, 61] 的做法，我们将数据集划分为基类和新类。模型在基类训练集上进行训练，并在测试集上评估。本实验中应用的属性是基于基类搜索得到的。对于没有专用验证集的数据集例如 ImageNet，我们将 16 样本标记 (16-shots) 数据分成两半，一半用于训练，另一半用于验证。

跨数据集 (Cross-dataset) 实验。与之前的工作 [19, 60, 61] 一致，我们首先在 ImageNet-1K 源数据集上训练模型，然后在多个分布外数据集上评估其泛化能力。实验中使用的属性从源数据集中获取。

属性搜索。我们在属性池中选取了 5 个独立属性作为属性基，这为搜索过程生成了 31 个候选属性组合。我们使用 ChatGPT-4o 进行属性查询。表 3 展示了部分查询到的属性基以及经我们的搜索算法筛选出的最终组合。附录中提供了关于我们搜索过程的更多细节。

实现细节。我们在 15 个常用的识别数据集上评估模型性能。我们报告了基类和新类的准确率，以及这两者在 3 次运行中取平均后的调和平均值 (HM)。每个数据集的详细信息见附录。

4.2. 基类到新类 (Base-to-Novel) 的泛化能力

如表 1 所示，我们在 11 个不同的识别数据集上评估了五种基线方法的基类到新类别泛化性能，包括集成和未集成 ATPrompt 两种情况。值得注意的是，ATPrompt 一致的提升了所有基线方法的平均性能。

在某些情况下改善效果有限的原因。(1) 可学习文本提示是早期研究中的核心部分，通过 ATPrompt 对其进行优化能够显著提升性能。(2) 近年来的研究已超越可学习文本提示的范畴，通过引入额外的可学习模块进行了扩展。由于我们的工作仅涉及可学习文本提示部分的优化，因此提升效果变得不够显著。

4.3. 跨数据集 (Cross-dataset) 评估

表 2 展示了三种基线方法的跨数据集泛化结果。我们的方法表现出优异的性能，使得 CoOp、CoCoOp 和 MaPLe 分别提高了 1.38%、0.85% 和 0.45%。

4.4. 域泛化 (Domain Generalization) 实验

表 4 展示了三种基线方法的域泛化结果。结果表明我们的方法比 CoOp、CoCoOp 和 MaPLe 方法分别提高了 0.90%、0.49% 和 0.33%。

4.5. 进一步分析

实验默认在 ImageNet 上进行。为了最小化其他组件的影响，我们主要采用 CoOp 作为基线方法。我们的 ATPrompt 中使用了两个属性（颜色和形状）。更多实验结果请参见附录。

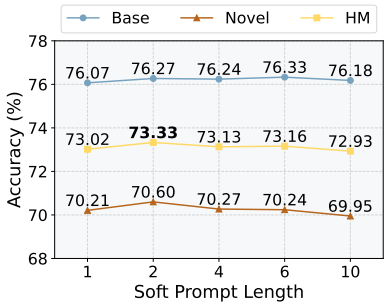


图 5. ImageNet 上不同软词元长度的示意图。增加词元可能导致对基类的过度拟合，并削弱对新类的泛化。

软提示长度。在图 5 中，我们研究了属性词元和类别词元的最佳软词元长度。通过将长度从 1 变化到 10，我们观察到较长的提示会稀释属性词元的引导作用，从而减少对新类的泛化能力。

类别词元位置。在我们的方法中，属性词元与类别词元的相对位置需要仔细考量。在表 5 中，我们考察了类别词元位于两个属性词元中间或两侧的多种配置。结果表明，当类别词元置于末尾时可获得最佳性能，这与 CoOp 的研究发现一致。

方法	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV 22)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp (CVPR 22)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
MaPLe (CVPR 23)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC (ICCV 23)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ArGue (CVPR 24)	83.69	78.07	80.78	76.92	72.06	74.41	98.43	95.20	96.79	95.36	97.95	96.64
DePT (CVPR 24)	83.66	71.82	77.29	77.13	70.10	73.45	98.33	94.33	96.29	94.70	97.63	96.14
CoPrompt (ICLR 24)	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87
PromptKD (CVPR 24)	86.96	80.73	83.73	80.83	74.66	77.62	98.91	96.65	97.77	96.30	98.01	97.15
CoOp + ATPrompt	82.68	68.04	**74.65** (+2.99)	76.27	70.60	**73.33**	97.95	93.63	**95.74**	94.77	96.59	**95.67**
CoCoOp + ATPrompt	81.69	74.54	**77.95** (+2.12)	76.43	70.50	**73.35**	97.96	95.27	**96.60**	95.46	97.89	**96.66**
MaPLe + ATPrompt	82.98	75.76	**79.21** (+0.66)	76.94	70.72	**73.70**	98.32	95.09	**96.68**	95.62	97.63	**96.61**
DePT + ATPrompt	83.80	73.75	**78.45** (+1.16)	77.32	70.65	**73.83**	98.48	94.60	**96.50**	94.65	97.99	**96.29**
PromptKD + ATPrompt	87.05	81.82	**84.35** (+0.62)	80.90	74.83	**77.75**	98.90	96.52	97.70	96.92	98.27	**97.59**

方法	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV 22)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp (CVPR 22)	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
MaPLe (CVPR 23)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
PromptSRC (ICCV 23)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
ArGue (CVPR 24)	75.64	73.38	74.49	98.34	75.41	85.36	92.33	91.96	92.14	40.46	38.03	39.21
DePT (CVPR 24)	79.67	72.40	75.86	98.20	72.00	83.08	90.43	91.33	90.88	42.53	22.53	29.46
CoPrompt (ICLR 24)	76.97	74.40	75.66	97.27	76.60	85.71	90.73	92.07	91.40	40.20	39.33	39.76
PromptKD (CVPR 24)	82.80	83.37	83.13	99.42	82.62	90.24	92.43	93.68	93.05	49.12	41.81	45.17
CoOp + ATPrompt	77.43	66.55	**71.58**	97.44	67.52	**79.77**	88.74	87.44	**88.09**	40.38	27.22	**32.52**
CoCoOp + ATPrompt	74.50	73.47	**73.98**	96.52	73.59	**83.51**	90.59	91.74	**91.16**	37.30	33.15	**35.10**
MaPLe + ATPrompt	75.39	73.84	**74.61**	97.82	75.07	**84.95**	90.65	92.00	91.32	37.61	36.15	**36.87**
DePT + ATPrompt	79.29	73.47	**76.27**	98.20	73.69	**84.20**	90.42	91.69	**91.05**	43.19	33.23	**37.56**
PromptKD + ATPrompt	82.51	84.03	**83.26**	99.15	82.03	89.78	92.48	93.86	**93.22**	49.63	42.35	**45.70**

方法	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp (IJCV 22)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp (CVPR 22)	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
MaPLe (CVPR 23)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC (ICCV 23)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ArGue (CVPR 24)	81.52	80.74	81.13	81.60	66.55	73.31	94.43	88.24	91.23	85.56	79.29	82.31
DePT (CVPR 24)	82.37	75.07	78.55	83.20	56.13	67.04	88.27	66.27	75.70	85.43	72.17	78.24
CoPrompt (ICLR 24)	82.63	80.03	81.30	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
PromptKD (CVPR 24)	83.69	81.54	82.60	85.84	71.37	77.94	97.54	82.08	89.14	89.71	82.27	86.10
CoOp + ATPrompt	80.84	68.64	**74.24**	80.83	45.49	**58.22**	90.34	59.79	**71.96**	84.49	64.96	**73.45**
CoCoOp + ATPrompt	80.50	76.86	**78.64**	78.63	56.89	**66.02**	87.95	74.15	**80.46**	82.74	76.40	**79.44**
MaPLe + ATPrompt	80.98	78.15	79.54	80.50	58.28	67.61	94.84	77.59	**85.35**	84.08	78.88	**81.40**
DePT + ATPrompt	82.42	76.48	**79.34**	82.64	56.77	**67.30**	89.60	69.50	**78.28**	85.60	73.15	**78.89**
PromptKD + ATPrompt	83.87	81.35	82.59	86.92	72.34	**78.96**	97.05	92.07	**94.49**	89.29	82.44	85.73

表 1. 在 11 个数据集上，我们针对五个基线模型开展了是否使用 ATPrompt 的基类到新类别 (Base-to-Novel) 泛化实验。结果显示，我们的方法在不同基线方法上均实现了一致的平均性能提升。

方法	源数据集	目标数据集										Average
	Image Net	Caltech 101	Oxford Pets	Stanford Cars	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
+ATPrompt	71.67	93.96	90.65	65.01	70.40	85.86	20.97	65.77	43.44	46.59	69.92	**65.26** (+1.38)
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
+ATPrompt	71.27	93.79	90.62	65.90	71.17	86.03	23.22	66.63	44.44	48.70	70.71	**66.59** (+0.85)
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
+ATPrompt	70.69	94.04	91.03	66.06	71.99	86.33	24.42	67.05	45.21	48.63	69.15	**66.75** (+0.45)

表 2. 在 11 个数据集上，我们针对三种基线方法开展了是否使用 ATPrompt 的跨数据集 (Cross-dataset) 泛化实验。结果显示，ATPrompt 在目标数据集上实现了一致的平均性能提升。

数据集	属性基	搜索结果
ImageNet	color, size, shape, habitat, behavior	(color, shape)
Caltech101	shape, color, material, function, size	(shape,size)
OxfordPets	loyalty, affection, energy, playfulness, intelligence	(playfulness, energy)
StanfordCars	design, engine, performance, luxury, color	(luxury)
Flowers102	color, flower, habitat, growth, season	(color, habitat, growth)
Food101	flavor, texture, origin, ingredients, preparation	(flavor, preparation)

表 3. 经过可微属性搜索得到的部分结果。完整的结果请参考附录。

方法	源数据集	目标数据集				Average
	ImageNet	-V2	-S	-A	-R	
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
+ATPrompt	71.67	64.43	49.13	50.91	76.24	**60.18** (+0.90)
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
+ATPrompt	71.27	64.66	49.15	51.44	76.33	**60.40** (+0.49)
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
+ATPrompt	70.69	64.40	49.10	51.77	77.11	**60.60** (+0.33)

表 4. 在四个数据集上针对三种基线方法开展了是否使用我们 ATPrompt 的域泛化实验。ATPrompt 的集成带来了更优的泛化性能。

深度版本的提示操作。在 ATPrompt-Deep 中，我们在经过块后仅丢弃类别软词元，同时保留硬属性词元和软属性词元。在本部分中，我们对比了部分丢弃（即移除软属性词元同时保留硬词元）和完全丢弃（即同时移除软属性词元和硬属性词元）两种操作的性能，结果如表 6 所示。

位置	Base	Novel	HM
Front	76.12	70.50	73.20
Middle	76.13	70.29	73.09
End	**76.27**	**70.60**	**73.33**

表 5. ImageNet 上不同类词元位置的比较。末尾位置的效果最好。

对属性词元的操作	Base	Novel	HM
保留所有硬、软词元	**76.94**	**70.72**	**73.70**
部分丢弃和重新添加	76.87	70.44	73.51
完全丢弃和重新添加	76.83	70.10	73.31

表 6. 基于 MaPLe+ATPrompt 的深层软、硬属性词元操作比较。在深层保留硬、软属性词元比其他操作效果更好。

结果表明，在前向过程中同时保留硬、软属性词元可获得最佳性能。相反，丢弃并重新引入硬属性词元或软属性词元会损害性能，可能是因为这会破坏跨层属性表征的连续性，并使优化过程复杂化。

属性	Base	Novel	HM
(shape, color)	76.32	70.39	73.24
(color, shape)	76.27	70.60	**73.33**
(size, habitat)	76.44	70.23	73.20
(habitat, size)	76.46	70.16	73.14

表 7. 在 ImageNet 上对不同顺序的比较。属性的顺序不会显著影响模型，且性能波动在合理范围内。

属性顺序。在本研究中，我们没有特别关注属性的顺

序，因为在现实中，改变顺序通常不会导致语义偏差。表 7 定量评估了属性顺序对提示学习性能的影响。从该表中我们观察到，尽管顺序存在差异，但始终能得到相似的结果，且不同顺序下的性能波动均在合理范围内。

与其他属性的比较。在表 8 中，我们探究了通过其他方法获取的属性的有效性，具体而言，即通过手动选择与类别无关的属性和通用属性。结果表明，手动选择的无关属性在训练阶段表现出相当的性能；然而，当应用于新类别时，其性能较差。这表明，不正确的属性词元会导致软词元形成有偏差的表征，进而降低其零样本泛化能力。

种类	属性	Base	Novel	HM
Common	(shape, size)	82.83	67.13	74.16
	(color, texture)	82.73	67.56	74.38
Irrelevant	(plane, engines)	82.81	66.22	73.59
	(football, sport)	82.77	67.14	74.14
Serched	-	82.68	**68.04**	**74.65**

表 8. 在 11 个数据集上不同属性配置的平均性能对比。我们方法得到的属性取得了最佳性能。

5. 结论

在本文中，我们提出了 ATPrompt，一种以属性为锚点的文本提示学习方法，该方法以通用属性为桥梁，提升模型对已见类别到未见类别的泛化能力。我们的方法通过将固定属性词元与提示锚定，把软提示的学习空间从一维的以类别为中心的结构拓展至多维属性空间。为了确保选出最优属性，我们提出了一套自动化流程，旨在为任意下游任务识别最适配的候选属性。ATPrompt 设计有浅层和深层架构变体，能与现有提示学习方法广泛兼容。大量实验验证了该方法的有效性，我们相信这项工作为提示学习领域中可学习提示的基础结构研究提供新的方向。

局限性与未来工作。这项工作是对基本提示形式的初步研究，它在单独运用时无法达到与基于正则化的方法相当的性能。此外，目前属性锚点的选择依赖于人工实验，通过基于学习的方法自动发现最优锚点位置仍然是未来研究的一个有前途的方向。

致谢。本研究受国家自然科学基金 (Nos. 62361166670、U24A20330) 的支持。本研究还得到国家自然科学基金青年科学基金项目 (Grant No.62206134)、中央高校基本科研业务费 (070-63253222) 以及天津视觉计算与智能感知重点实验室 (VCIP) 的支持，计算资源由南开大学超级计算中心 (NKSC) 提供。

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 35:25005–25017, 2022. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 13

[4] Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. Ovarnet: Towards open-vocabulary object attribute recognition. In *CVPR*, pages 23518–23527, 2023. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 13

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 13

[7] Tong Ding, Wanhua Li, Zhongqi Miao, and Hanspeter Pfister. Tree of attributes prompt learning for vision-language models. *arXiv preprint arXiv:2410.11201*, 2024. 3

[8] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *JMLR*, 20 (55):1–21, 2019. 6

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 13

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover

classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019. 13

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 13

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 13

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[14] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *NeurIPS*, 33:5632–5643, 2020. 5

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2, 3

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 1, 2, 4

[17] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *ICCV*, pages 15670–15680, 2023. 2

[18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2, 3, 4, 13

[19] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 2, 3, 4, 6

[20] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *CVPR Workshop*, pages 1572–1582, 2024. 3

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, pages 554–561, 2013. 13

[22] Nilakshan Kunananthaseelan, Jing Zhang, and Mehrtash Harandi. Lavip: Language-grounded visual prompting. In *AAAI*, pages 2840–2848, 2024. 2

[23] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, pages 1401–1411, 2023. 2

[24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 2

[25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1, 2

[26] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 2024. 1

[27] Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-CLIP: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *ICML*, pages 28243–28258. PMLR, 2024. 1

[28] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023. 2, 5

[29] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, pages 26617–26626, 2024. 2, 4, 6, 13

[30] Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. Dual teachers for self-knowledge distillation. *Pattern Recognition*, 151: 110422, 2024. 2

[31] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *ECCV*, pages 19–34, 2018. 6

[32] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 5, 13

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 1

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 13

[35] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2, 3

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 13

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 13

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 13

[39] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *ICML*, pages 4095–4104. PMLR, 2018. 5

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 13

[42] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023. 2

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 13

[44] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, pages 13019–13029, 2024. 1

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1

[46] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, pages 2820–2828, 2019. 6

[47] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *CVPR*, pages 28578–28587, 2024. 2, 3

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 13

[49] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *AAAI*, pages 5749–5757, 2024. 2

[50] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 5

[51] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *ECCV*, 2024. 2

[52] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 13

[53] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, and Yongjun Xu. Clip-kd: An empirical study of distilling clip models. *arXiv preprint arXiv:2307.12732*, 2023. 2

[54] Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowl-

edge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023. 2

[55] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 2, 4

[56] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *AAAI*, pages 6979–6987, 2024. 2

[57] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *CVPR*, pages 12924–12933, 2024. 2, 13

[58] Xiaoqin Zhang, Zhenni Yu, Li Zhao, Deng-Ping Fan, and Guobao Xiao. Comprompter: reconceptualized segment anything model with multiprompt network for camouflaged object detection. *Science China Information Sciences*, 68(1):112104, 2025. 2

[59] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 5

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2, 3, 4, 6, 13

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 2, 3, 4, 6, 13

[62] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669, 2023. 2, 3

[63] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018. 5

基于属性锚点改进的文本提示学习方法

Supplementary Material

6. 实现细节

6.1. 数据集

我们在 15 个识别数据集上评估了我们方法的性能。为了评估从基类别到新类别 (Base-to-Novel) 的泛化能力以及跨数据集 (Cross-dataset) 的性能，我们在 11 个不同的识别数据集上进行了测试。具体而言，这些数据集包括用于通用物体分类的 ImageNet-1K [6] 和 Caltech-101 [9]；用于细粒度分类的 OxfordPets [37]，StanfordCars [21]，Flowers-102 [36]，Food-101 [3]，和 FGVCAircraft [34]；用于场景识别的 SUN-397 [52]；用于动作识别的 UCF-101 [43]；用于纹理分类的 DTD [5]；以及用于卫星图像识别的 EuroSAT [10]。对于域泛化 (Domain Generalization) 实验，我们使用 ImageNet-1K [6] 作为源数据集 (Source Dataset)，并将其四个变体作为目标数据集 (Target Dataset)，包括 ImageNet-V2 [41]、ImageNet-Sketch [48]、ImageNet-A [12] 和 ImageNet-R [11]。

6.2. 属性搜索

受 DARTS [32] 启发，我们采用可微分搜索方法来确定所提出的属性锚定形式中属性的最优内容和数量。搜索过程进行 10 轮，批大小 (Batch Size) 为 32。我们使用 SGD 优化软提示 θ，初始学习率为 0.002。使用优化权重向量 α 初始学习率为 0.02。在实验中，我们使用 5 个属性基，这些属性基会生成 31 种（即 $C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5$）候选组合用于搜索过程。

表 12 给出了由大型语言模型生成的五个属性基，以及搜索后确定的最优属性组合。此外，表 13 显示了 Caltech-101 数据集上搜索阶段的所有候选组合的最终权重。

6.3. 基类到新类 **(Base-to-Novel)** 的泛化能力

基线方法。为了评估 ATPrompt，我们将其与几种主流的基于文本的提示学习方法相结合，包括 CoOp [61]，CoCoOp [60]，MaPLe [18]，DePT [57] 和 PromptKD [29]。实验设置详情如下：

实验设置。我们的框架基于 PyTorch [38] 实现，所有实验均在单卡 NVIDIA A800 GPU 上进行。遵循基线方法，我们采用标准的数据增强方案，包括随机调整大小裁剪和翻转。我们使用随机梯度下降 (SGD) 作为优化器。默认情况下，属性词元和类别词元的软词元长度设置为相同，因为属性词元和类别词元被视为同等重要。各基线方法的具体实现细节如下：

CoOp+ATPrompt：遵循该基线方法设置，我们采用 32 的批量大小和 0.002 的初始学习率。原始论文提到 ResNet-50 模型的可学习提示长度 $M = 16$，但未明确 ViT-B/16 模型的提示长度。在我们的实验中，属性和类别词元的软词元长度均设为 2。基线模型的训练轮次为 200，我们将其缩短至 100 轮，同时保持相同的余弦衰减调度策略。图 6 展示了原始 CoOp 与 CoOp+ATPrompt 在架构上的差异。

CoCoOp+ATPrompt：我们遵循基线方法的设置，采用 1 的批量大小和 0.002 的初始学习率。原始论文规定软类别词元的长度为 4，而我们将属性和类别词元的可学习词元长度均设为 2。我们采用与基线相同的训练策略：10 个轮次，采用余弦衰减。

CoCoOp 的原始设计使用元网络为所有软提示词元生成偏移量。我们保留该元网络，但修改了其应用方式：如图 7 所示，元词元现在仅作为类别软词元 $[T_1], ..., [T_M]$ 的偏移量。

MaPLe+ATPrompt：我们遵循基线方法的超参数。采用 4 的批量大小和 0.0035 的初始学习率。我们与原始的提示配置有所不同：基线将可学习提示长度设为 2，而我们的方法将属性词元和类别词元的软词元长度均设为 4。训练策略与基线保持一致。

MaPLe + ATPrompt 的架构修改主要在于投影机制。原始的 MaPLe 框架将所有文本软词元输入到一个投影层以生成相应的视觉词元，然后将这些视觉词元融合到图像编码器中。然而，我们的方法仅将类别软词元选择性地输入到该投影层，而属性词元则保持不变。图 8 直观地展示了这种架构差异。

DePT+ATPrompt：我们采用基线方法的训练配置，批大小为 32，初始学习率为 0.0035，平衡损失的超参数 λ=0.7，训练 10 个轮次。DePT+ATPromp 的主要

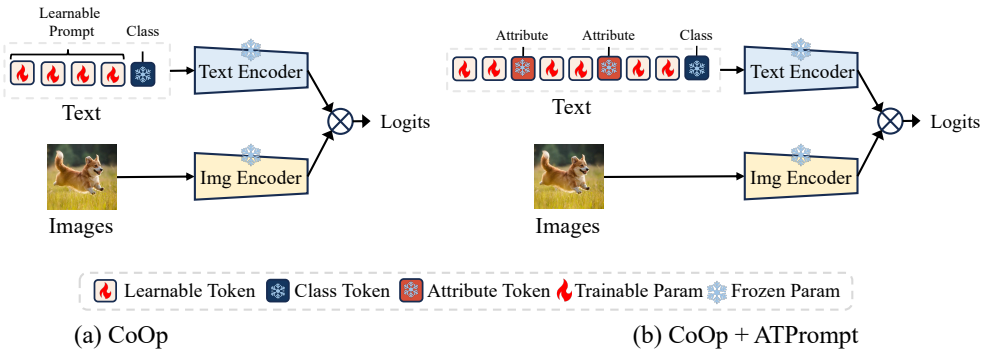


图 6. CoOp 和 CoOp+ATPrompt 的架构对比。

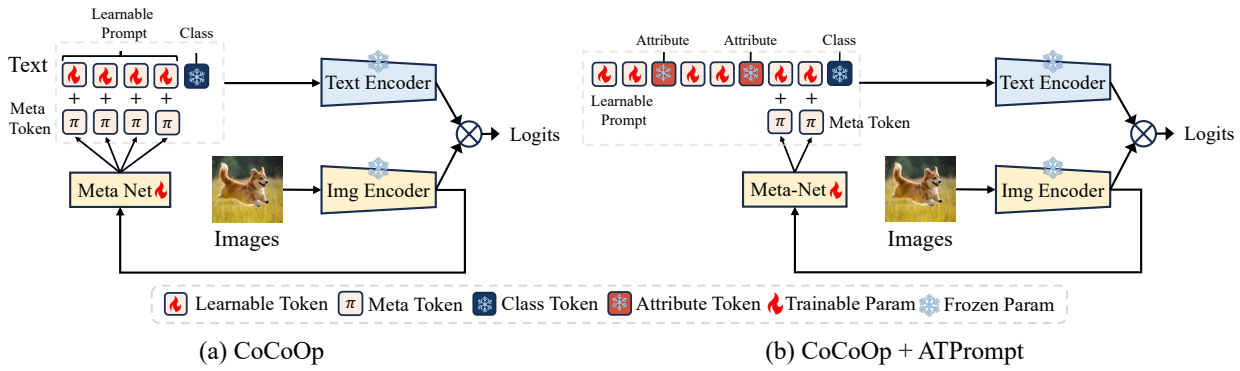


图 7. CoOp 和 CoOp+ATPrompt 的架构对比。在 CoCoOp+ATPrompt 中，元词元作为偏移量加到类别软词元中。

配置中，可学习词元长度设为 4。对于复杂度较低的数据集（即 Caltech-101、OxfordPets 和 StanfordCars），我们对这些参数进行了调整，将软词元长度设为 2，平衡超参数设为 0.6。DePT 与 DePT+ATPrompt 模型之间的架构差异详见图 9。

7. 附加实验

7.1. 消融实验

属性顺序。在正文中，我们的实验证实了属性的顺序不会显著影响模型性能，结果在可接受范围内波动。此处，我们在表 9 中提供补充实验以支持这一观察结果。

属性位置。我们还研究了属性词元在提示中的位置影响。图 10 展示了所测试的位置，表 10 呈现了结果。我们的研究结果表明，“间隔”配置（即属性词元置于类别词元之间）的性能最佳。

初始化。基线方法通常使用短语”a photo of a” 的嵌入值来初始化软词元。而属性词元的加入使得该策略不再适合我们的方法。因此，我们采用从均值为 0、标准差为 0.02 的高斯分布中进行采样来初始化类别软词元 ($[T_1], ..., [T_M]$)。如表 11 所示，这种随机初始化可为训练提供更优的起点。

8. 讨论

与直接向大型语言模型查询的比较。直接向大语言模型查询通用属性存在两个挑战：确定最优的属性内容和找到理想的属性数量。我们的实验表明，两个属性数量通常是最优的。因此，用户可以通过直接询问大语言模型总结两个通用属性来简化我们的搜索过程。这提供了一种更简单的方法，尽管可能会导致轻微的性能损失。

为什么在源数据集（ImageNet）上搜索属性可以很好地泛化？ 在 ImageNet 上识别出的属性（例如颜色、形状）是自然物体的基本属性。因此，在这些通用属性的指导下学习到的表征本质上具有可泛化性，并且能有效迁移到其他数据集和类别中。

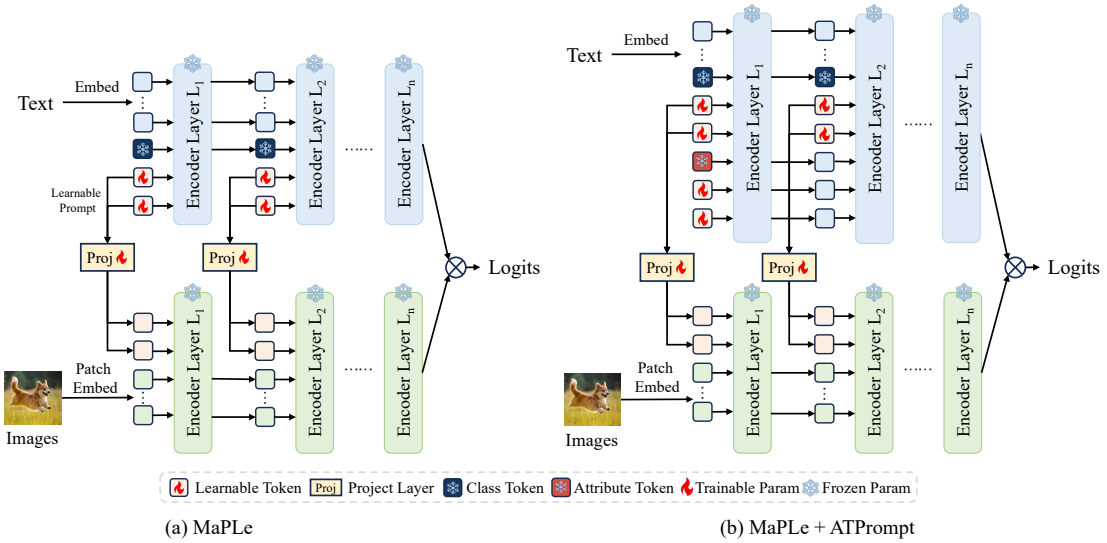


图 8. MaPLe 和 MaPLe+ATPrompt 的架构对比。

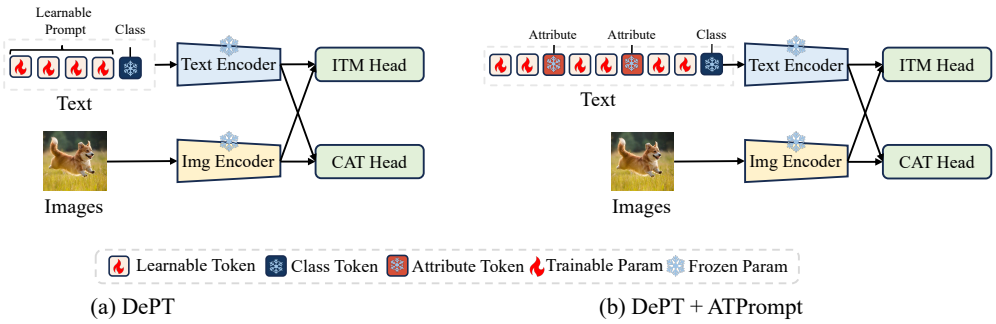


图 9. DePT 和 DePT+ATPrompt 的架构对比。

为什么 ATPrompt 单独使用时没有超过基于正则化的方法？ATPrompt 是一个旨在优化可学习提示形式的基础插件模块。相比之下，基于正则化的方法通常是一个综合的整体框架，会同时采用多个组件（例如可学习的视觉提示、多层感知器 MLPs 等）。虽然 ATPrompt 单独使用时可能无法超越这些多方面的方法，但其优势在于能够集成到其他方法中，将它们的性能提升到超过先前基线的水平。

9. 局限性与未来工作

除正文中讨论的局限性外，我们还探讨了以下未来研究方向：(1) 尽管我们的可微分搜索方法很高效，但我们旨在进一步改进属性发现过程。一个有前景的方向是利用多模态大型语言模型 (MLLMs)，可能采用思维链 (CoT) 等技术，以更好地自动选择最优属性的内容和数量。(2) 我们当前的方法将固定的、显式的属性嵌入到提示中。未来，我们计划探索向隐式的、可学习属性词元的过渡。这将使模型能够在训练过程中以数据驱动的方式发现最优属性词元，有望进一步提升性能。

属性	Base	Novel	HM
(shape, color)	76.32	70.39	73.24
(color, shape)	76.27	70.60	**73.33**
(size, habitat)	76.44	70.23	73.20
(habitat, size)	76.46	70.16	73.14
(material, function)	76.40	70.13	73.13
(function, material)	76.28	70.00	73.01
(growth, season)	76.46	70.18	73.19
(season, growth)	76.40	70.21	73.17
(color, size, shape)	76.27	69.95	72.97
(shape, size, color)	76.32	70.19	73.13
(habitat, size, shape)	76.50	70.21	73.22
(habitat, shape, size)	76.46	70.08	73.13
Searched Attributes (color, shape)	76.27	70.60	**73.33**

表 9. 在 ImageNet 上对不同属性顺序的比较。属性顺序的变化不会显著影响模型性能。

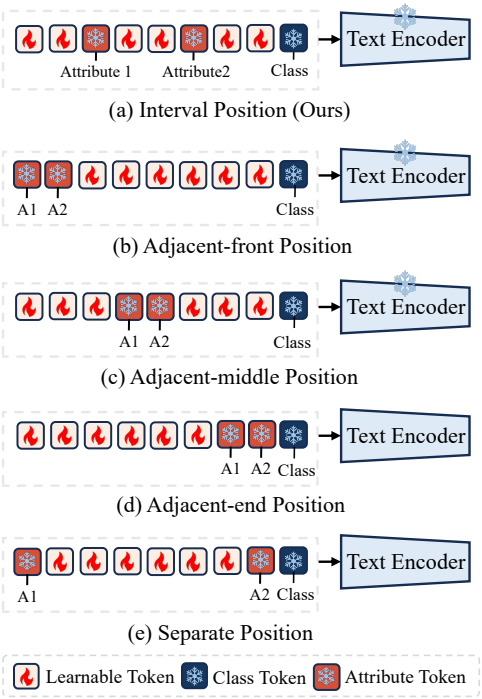


图 10. 以两个属性为例，对属性词元置于不同位置的比较。

Version	Base	Novel	HM
Baseline (CoOp)	76.47	67.88	71.92
(a) Interval (Ours)	76.27	70.60	73.33
(b) Adjacent-front	76.39	70.22	73.18
(c) Adjacent-middle	76.46	70.11	73.15
(d) Adjacent-end	76.34	70.31	73.20
(e) Separate	76.48	70.08	73.14

表 10. 在 ImageNet 上的 ATPrompt 中不同位置属性词元的性能结果。间隔版本的结果最佳。

属性	Base	Novel	HM
“a photo of a”	76.40	70.07	73.10
Random Normal Init	76.27	70.60	73.33

表 11. 在 ImageNet 上对不同初始化方式的比较。正态随机初始化性能更好。

数据集	属性基	搜索结果
ImageNet-1K	color, size, shape, habitat, behavior	(color, shape)
Caltech-101	shape, color, material, function, size	(shape,size)
Oxford Pets	loyalty, affection, playfulness, energy, intelligence	(playfulness, energy)
Stanford Cars	design, engine, performance, luxury, color	(luxury)
Flowers-102	color, flower, habitat, growth, season	(color, habitat, growth)
Food-101	flavor, texture, origin, ingredients, preparation	(flavor, preparation)
FGVC Aircraft	design, capacity, range, engines, liveries	(design, range)
SUN-397	architecture, environment, structure, design, function	(function)
DTD	pattern, texture, color, design, structure	(pattern, color, design)
EuroSAT	habitat, foliage, infrastructure, terrain, watercourse	(habitat)
UCF-101	precision, coordination, technique, strength, control	(precision)

表 12. 每个数据集的属性基和搜索结果。

属性	shape, color, material, function, size
属性组 & 对应的权重值	(shape), weight: 0.298 (color), weight: 0.004 (material), weight: 0.002 (function), weight: 0.002 (size), weight: 0.003 (shape, color), weight: 0.003 (shape, material), weight: 0.006 (shape, function), weight: 0.000 **(shape, size), weight: 0.565** (color, material), weight: 0.000 (color, function), weight: 0.001 (color, size), weight: 0.005 (material, function), weight: 0.000 (material, size), weight: 0.002 (function, size), weight: 0.002 (shape, color, material), weight: 0.002 (shape, color, function), weight: 0.002 (shape, color, size), weight: 0.000 (shape, material, function), weight: 0.001 (shape, material, size), weight: 0.085 (shape, function, size), weight: 0.001 (color, material, function), weight: 0.001 (color, material, size), weight: 0.000 (color, function, size), weight: 0.002 (material, function, size), weight: 0.001 (shape, color, material, function), weight: 0.001 (shape, color, material, size), weight: 0.001 (shape, color, function, size), weight: 0.001 (shape, material, function, size), weight: 0.005 (color, material, function, size), weight: 0.001 (shape, color, material, function, size), weight: 0.001

表 13. 在 Caltech101 数据集上经过 40 个轮次的属性搜索后的输出结果。