PromptKD: 用于视觉-语言模型的无监督提示蒸馏

李政[1], 李翔[2,1]*, 傅欣艺[3], 张鑫[1], 王维强[3], 陈硕[4], 杨健[1]*
[1] PCA Lab, VCIP, 计算机学院, 南开大学
[2] 南开国际先进研究院（深圳福田）, [3] 蚂蚁集团, [4] 日本理化学研究所
{zhengli97, zhasion}@mail.nankai.edu.cn, {xiang.li.implus, csjyang}@nankai.edu.cn
{fxy122992, weiqiang.wwq}@antgroup.com, shuo.chen.ya@riken.jp

摘要

提示学习（*Prompt Learning*）已成为一种用于增强诸如 *CLIP* 之类视觉-语言模型 *(VLMs)* 的重要技术，以提升其在特定领域下游任务中的表现。现有的研究工作主要集中在设计各种形式的提示上，却忽视了将提示作为有效蒸馏介质从更大规模的教师模型中进行学习的潜力。在本文中，我们提出了一种无监督的域提示蒸馏框架，该框架旨在通过使用未标注的域图像借助于由提示驱动的模仿，将大型教师模型的知识推理至轻量级目标模型中。具体而言，我们的框架包括两个阶段。在初始阶段，我们使用域（少样本）标签预训练一个大型 *CLIP* 教师模型。在预训练之后，我们利用 *CLIP* 独特的模态解耦特性，通过教师文本编码器预先计算并存储文本特征作为类别向量，这一过程仅需执行一次。在后续阶段，这些存储的类别向量在教师和学生模型的图像编码器之间共享，用于计算预测的逻辑值 *(Logits)*。此外，我们通过 *KL* 散度对齐教师和学生模型的逻辑值，促使学生图像编码器通过可学习的提示生成与教师相似的概率分布。我们所提出的提示蒸馏过程消除了对标注数据的依赖，使算法能够利用领域内大量未标注的图像。最终，训练良好的学生图像编码器和预先存储的文本特征（类别向量）被用于推理。据我们所知，我们的方法是首个 (*1*) 为 *CLIP* 执行无监督的域特定提示驱动的知识蒸馏算法，以及 (*2*) 建立了一种实用的文本特征预存储机制，作为教师和学生之间共享的类别向量的方法。我们在 *11* 个数据集上进行了广泛实验，结果验证了我们方法的有效性。公开代码可见于 *https://github.com/zhengli97/PromptKD*。

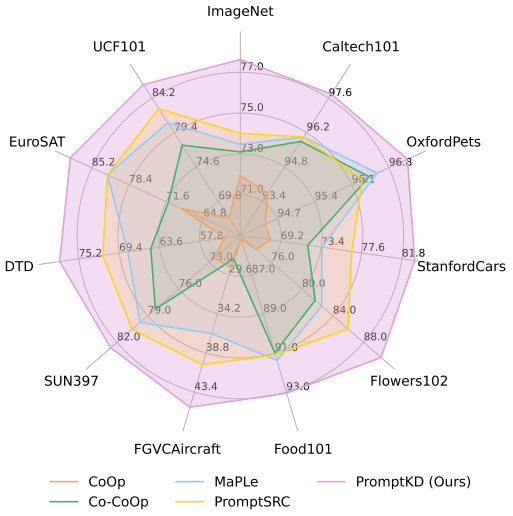

图 1. 在基类到新类的泛化能力上的调和平均数（HM）比较。所有方法均采用了预训练 CLIP 模型中的 **ViT-B/16 图像编码器**。PromptKD 在 11 个不同的识别数据集上实现了最先进的性能。

1. 引言

近年来，大规模预训练的视觉-语言模型 (VLMs)，如 CLIP [41, 68] 和 ALIGN [17]，在特定领域的下游任务中展现出卓越的泛化能力。与传统视觉框架不同，CLIP 等视觉-语言模型通常采用双塔架构，包括图像编码器和文本编码器。这些模型使用对比损失进行训练，以学习一个统一的嵌入空间，使多模态信号的表示能够对齐。

*通讯作者

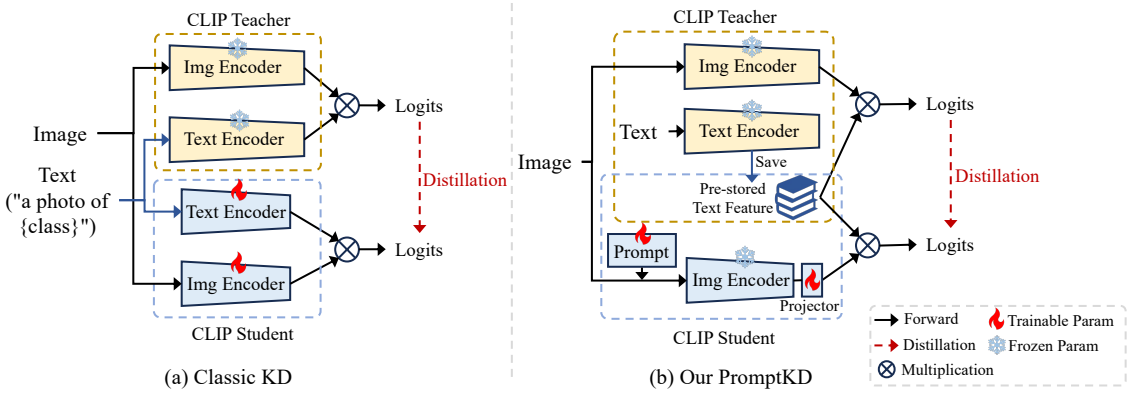

图 2. 经典 CLIP 知识蒸馏（如 CLIP-KD [63]）与我们提出的提示蒸馏框架的架构对比。(a) 经典知识蒸馏方法在独立的教师模型和独立的学生模型之间进行蒸馏，通常使用教师模型的软标签对学生模型进行全参数微调。(b) PromptKD 打破了教师-学生模型独立性的传统模式，我们提出复用教师预训练阶段获得的文本特征，并将其直接引入学生图像编码器中，以同时支持蒸馏和推理。

为了更好地优化模型以适应特定域的下游任务，已有研究 [10, 21, 65, 71, 72] 主要集中于保持原始模型的权重不变的情况下适配 CLIP 的表示。受自然语言处理 (NLP) [26, 28] 领域成功经验的启发，提示学习 [18, 71, 72] 被提出，以获取连续的提示表示，从而替代精心设计的离散提示。根据提示学习的信息来源，现有方法大致可分为三类：基于文本的、基于视觉的以及文本-视觉结合的提示学习方法。基于文本的提示学习方法 [71, 72] 旨在自适应学习适合下游任务的文本提示，而非采用固定形式。基于视觉的提示学习方法 [5, 18] 采用类似原则，并进一步扩展至视觉模态。文本-视觉结合的提示学习方法 [21, 22, 25, 52] 采用一种同时优化图像和文本分支的策略，而非分别处理它们。

现有研究主要关注于在有限标注数据的情况下学习有效的 *提示词元*，同时保持优秀的泛化能力。在本文中，我们提出了一种新的无监督框架（称为 “PromptKD”），仅使用大量无标签的域数据进行训练，其中我们将可学习提示作为域蒸馏媒介，使得 CLIP 学生模型能够从大型 CLIP 教师模型中学习知识。具体而言，我们的框架包括两个阶段：教师模型预训练阶段和学生模型蒸馏阶段。

在第一阶段，我们使用现有的先进方法 [21, 22] 在少量有标签的域数据上预训练一个大规模的 CLIP 教师模型。预训练完成后，我们利用 CLIP 的模态解耦特性，通过教师文本编码器预计算并存储文本特征作为类别向量，*这一过程仅需执行一次*。

在第二阶段，存储的类别向量在教师与学生的图像编码器之间共享，用于计算预测逻辑值，同时无需任何额外的文本分支计算成本。与传统知识蒸馏方案（如图 2(a)）中学生模型通常通过全参数微调来模仿教师模型的统计行为不同，我们提出利用学生的可学习视觉提示通过 KL 散度对教师与学生模型的逻辑值进行对齐，从而通过提示蒸馏，使学生图像编码器生成与教师模型相似的概率分布。此外，由于教师和学生模型特征维度不同，我们额外引入了一个投影模块，以调整特征维度间的差异。

借助教师-学生范式的优势，我们可以利用教师模型为目标领域的大量无标签图像生成软标签，从而实现学生模型的训练，而无需标注数据。最终，训练良好的学生图像编码器和预存的教师文本特征（类别向量）被用于推理。图 2 直观展示了经典 CLIP 蒸馏方法与我们提出的提示蒸馏框架的架构对比。

实验结果（图 1）表明，我们提出的 PromptKD 在 11 个不同的识别数据集上优于现有方法，并在 ViT-B/16 图像编码器 CLIP 模型上取得了最先进的性能。具体而言，我们的方法在 11 个数据集的基类别和新类别上分别提升了 2.70% 和 4.63% 的平均性能。

我们的贡献可总结如下：

- 我们的方法是首个在无标签域数据上进行基于提示的 CLIP 域特定知识蒸馏的方法。
- 我们利用 CLIP 独特的解耦模态特性，复用预存的文本特征，实现无额外文本分支计算成本的高效蒸馏和推理。
- 依托教师-学生范式，我们能够利用教师模型为大量

无标签域数据生成软标签，从而在无需标注数据的情况下训练学生模型。

- 在 11 个数据集上的广泛实验验证了我们方法的有效性。

2. 相关工作

视觉-语言模型中的提示学习 提示学习是一种可以在无需完全重新训练原始模型的情况下，将大规模预训练模型（如 CLIP [41]）推理到下游任务 [11, 42, 66] 的技术。该方法通过可学习的文本或视觉软提示适配特定任务的表示，取代了人工设计的硬提示（例如"a photo of a {classname}"）。软提示 [18, 25, 44, 71, 72] 可以通过冻结的预训练模型进行反向传播被优化，从而提升模型性能。现有研究主要集中在使用有限的域标注数据设计各种高效的提示形式。例如，MaPLe [21] 提出同时在图像和文本分支学习提示，而非分开学习。PromptSRC [22] 利用模型本身的原始特征对每个分支的提示学习进行正则化。以往的方法需要在图像 [8, 56] 和文本分支上对每个输入的软提示词元进行前向和反向计算。而在本工作中，我们利用 CLIP 独特的模态解耦特性，将训练良好的教师文本特征预存为类别向量，以供学生模型蒸馏。这种方式使得学生 CLIP 训练时，仅需进行图像分支的前向和反向计算，而无需涉及文本分支，从而简化了蒸馏和推理过程，提高了效率。

零样本学习 在已知类别的标注训练集上，零样本学习（ZSL） [32, 55, 58] 旨在学习一个分类器，以对新类别的测试样本进行分类。现有方法可根据测试图像是否可用，大致分为两类：归纳式（Inductive） [59, 67] 和推理式（Transductive） [49, 51] 零样本学习。以往的提示学习方法（如 MaPLe 和 PromptSRC）主要关注实例归纳式设定，仅使用带标签的训练样本。在本研究中，我们探索了推理式零样本学习设定，即在模型训练过程中同时利用已有和新类别的图像。具体而言，我们的教师模型采用与 PromptSRC 相同的训练策略，在已有类别的样本上进行训练，并使用真实标签。不同的是，我们的目标学生模型是在完整的无标签数据集（包括已有和未知类别的所有样本）上进行训练，而不使用任何真实标签。

知识蒸馏 知识蒸馏（Knowledge Distillation，KD）[15] 旨在通过一个大规模预训练教师模型的监督，训练一个轻量级学生模型。近年来，出现了多种蒸馏形式，以有效地将教师模型的知识传递给学生模型，包括逻辑值 (Logits) 对齐 [29, 31, 69, 70]、特征模仿 [4, 27, 64] 和样本关系匹配 [38, 61]。除了传统的图像分类任务，知识蒸馏还在多个视觉任务中取得了成功，例如目标检测 [2, 19, 54]、图像分割 [33, 62] 和姿态估计 [30]。近年来，许多研究 [24, 40, 57, 63] 将注意力转向 CLIP 模型，这些工作提出利用 CLIP 模型卓越的泛化能力来增强现有模型的学习能力。例如，CLIP-KD [63] 发现，在蒸馏预训练 CLIP 模型时，最简单的 MSE 损失特征模仿方法能取得最佳效果。TinyCLIP [57] 在教师和学生之间的仿射空间中执行跨模态特征对齐。我们的工作不同于以往的蒸馏方法，这些方法通常利用预训练的大型 CLIP 教师模型，对整个学生模型进行训练。而我们的方法采用了一种更高效的方式，通过 学生模型的可学习提示进行蒸馏，同时保持学生模型的原始 CLIP 权重冻结。这使得知识的有效推理得以实现，而无需对学生模型进行大量的重新训练。

3. 方法

提示学习 [18, 72] 旨在通过引入可学习的提示来增强现有视觉-语言模型（如 CLIP）的下游任务性能。现有研究主要关注于在有限的域标注数据上设计有效的提示学习形式，同时保持对未见图像的强泛化能力。在本文中，我们首次探讨将提示作为有效的知识蒸馏媒介，通过在无标签域图像上对齐模型间预测结果的方式使学生模型从大的 CLIP 教师模型中学习知识。图 3 展示了我们提出的提示蒸馏方法的整体框架。具体而言，我们的方法包含两个主要阶段：教师模型预训练阶段和学生模型提示蒸馏阶段。在初始阶段，我们首先使用现有的先进方法在少量标注数据上预训练一个大规模 CLIP 教师模型（见图 3(a)）。在预训练完成后，我们从教师文本编码器中提取并存储高质量的文本特征，作为类别向量。在随后的阶段，这些预存的类别向量被复用，与教师和学生图像编码器的输出相乘，以获得每个模型的预测结果。然后，我们通过提示模仿的方式来执行蒸馏过程，促使学生模型生成与教师模型相似的预测结果，如图 3(b) 所示。此外，我们引入一个额外的映射模块，以对齐教师文本特征和学生图像特征的维度。最终，训练好的学生图像编码器分支和预存的教师文本特征（类别向量）被用于推理（见图 3(c)）。

在下面的章节中，我们首先在 3.1 小节介绍视觉-

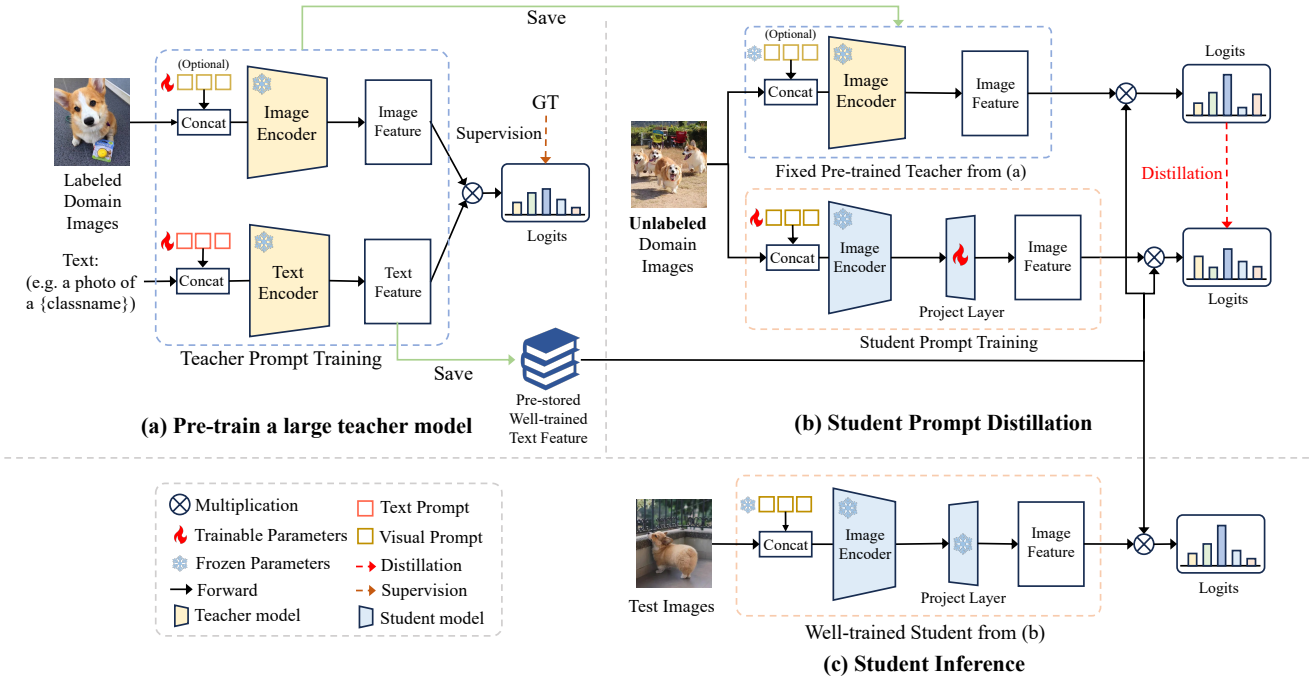

图 3. 我们提出的提示蒸馏（PromptKD）框架概述。(a) 首先，我们使用现有最先进的提示学习方法，在有标签的训练图像上预训练一个大规模 CLIP 教师模型，并存储所有可能类别的文本特征。(b) 在蒸馏阶段，训练仅涉及学生图像提示和映射层，由于使用了预存的类别向量作为类表示，文本编码过程无需额外计算开销。(c) 最终，训练良好的学生模型和预存的类别向量被用于推理。

语言模型的背景知识及知识蒸馏方法。然后，在 3.2 小节详细介绍我们的方法。

3.1. 背景

视觉-语言模型 现有的视觉-语言模型如 CLIP [41] 和 ALIGN [17] 旨在通过对齐图像和文本学习一个联合嵌入空间。遵循 [21, 22, 71]，我们选择 CLIP 作为基模型。CLIP 由两个编码器组成，一个用于图像，一个用于文本。给定一个带标签的视觉识别数据集 $D = \{x_j, y_j\}_{j=1}^{M}$，其中包含 N 个类别名称 $\mathrm{c} = \{c_i\}_{i=1}^{N}$，CLIP 使用模板 “a photo of a $\{c_i\}$” 生成文本描述 t_i。然后，每个文本描述 t_i 通过文本编码器 f_T 处理，以获得归一化的文本特征 $w_i = f_T(t_i)/||f_T(t_i)||_2 \in \mathbb{R}^d$，其中 d 表示特征维度。所有类别的文本特征 $\mathrm{W} = [w_1, w_2, ..., w_N] \in \mathbb{R}^{N\times d}$ 可视为图像分类的权重向量。对于数据集 D 中的输入图像 x，图像编码器 f_I 生成归一化的图像特征 $u = f_I(x)/||f_I(x)||_2 \in \mathbb{R}^d$。最终的输出概率计算如下：

$$p(y|x) = \frac{\exp(uw_y^{\mathsf{T}}/\tau)}{\sum_{i=1}^{N}\exp(uw_i^{\mathsf{T}}/\tau)}, \tag{1}$$

其中 uw^{T} 表示输出的 logits，τ 是温度参数。

与手工设计的硬提示（Hard Prompts）不同，近期研究（如 CoOp [72]）提出了一种自适应学习软文本提示（soft textual prompts）的方法，以适配下游任务。具体而言，M 个可学习的文本向量 $\{v_1, v_2, ..., v_M\}$（即前缀）被添加到类别（CLASS）标记之前，以构建场景化的表征。这样，类别 c_i 的提示 t_i 可表示为 $t_i = \{v_1, v_2, ..., v_M, c_i\}$，其中每个向量 v_i $(i \in 1, 2, ..., M)$ 具有与词嵌入相同的维度，而 M 是一个超参数，决定了前缀的长度。除了文本提示学习方法外，视觉提示也得到了广泛研究。一些研究 [18, 21, 22] 采用与文本提示方法类似的思路，在输入到图像编码器的图像块中添加多个可学习的视觉前缀。这些视觉提示旨在引导图像编码器提取更具语义意义和任务相关性的视觉特征。通过引入这些可学习的视觉前缀，模型能够利用额外的上下文信息和先验知识，以提升其在图像理解任务上的表现。

知识蒸馏 知识蒸馏（Knowledge Distillation, KD）[15] 旨在将预训练的大规模教师模型的知识推理到轻量级学生模型中。蒸馏后，学生模型可以学习教师模型的知识并用于最终部署。具体而言，Kullback-Leibler（KL）

散度损失用于匹配两个模型的输出分布：

$$L_{kd}(q^t, q^s, \tau) = \tau^2 KL(\sigma(q^t/\tau), \sigma(q^s/\tau)), \quad (2)$$

其中 q^t 和 q^s 分别表示教师和学生模型的 logits，$\sigma(\cdot)$ 是 softmax 函数，τ 是温度参数 [15, 31]，用于控制分布的平滑程度。

3.2. PromptKD: 用于视觉-语言模型的提示蒸馏

我们提出的提示蒸馏框架包括两个阶段：教师模型预训练和学生模型提示蒸馏，如图 3 所示。本节将详细介绍各个阶段的过程。

阶段 I: 教师模型预训练 在初始阶段，我们使用带标签的域数据预训练一个大规模 CLIP 教师模型，如图 3(a) 所示。为此，我们可以采用现有的提示学习方法（如 MaPLe [21] 和 PromptSRC [22]），或者直接使用公开可用的预训练 CLIP 模型。给定一个带标注的域数据集 $D_{labeled} = \{x_i, y_i\}_{i=1}^{M}$ 及其类别名称集合，教师 CLIP 模型以训练图像和包含类别名称的文本描述作为输入，并通过图像编码器 f_I^t 和文本编码器 f_T^t 生成归一化的图像特征 $u \in \mathbb{R}^d$ 和文本特征 $w \in \mathbb{R}^d$。最终的输出结果 p^t 由式 (1) 计算得到。通常，教师模型的软提示参数通过最小化预测概率 p 与真实标签 y 之间的交叉熵损失进行更新。

一旦文本编码器的训练完成，其输出的特征便保持固定，无需进一步更新。在此情况下，我们将所有 N 个类别的经过良好训练的教师文本特征 $W = [w_1, w_2, ..., w_N] \in \mathbb{R}^{N \times d}$ 保存为共享类别向量，这些向量将在后续处理阶段中被利用。这一操作消除了对学生 CLIP 文本分支的需求，从而在训练过程中节省了大量的计算成本。此外，通过我们的 PromptKD 方法，我们可以用学生轻量级图像编码器替换教师的大型图像编码器，在部署时减少计算成本，同时保持竞争力的性能。

阶段 II: 学生模型提示蒸馏 在此阶段，我们的目标是训练一个学生模型，使其通过提示模仿对齐教师的输出结果，如图 3(b) 所示。由于教师文本特征的复用，我们只需训练学生图像编码器分支 f_I^s 及其可学习的视觉提示和特征映射模块。对于无标签域数据集 $D_{unlabeled}$，将图像 x 输入教师模型和学生模型的图像编码器，我们可以得到归一化的教师图像特征 $u^t = f_I^t(x)/||f_I^t(x)||_2 \in \mathbb{R}^d$ 和学生图像特征 $u^s = P(f_I^s(x))/||P(f_I^s(x))||_2 \in \mathbb{R}^d$。其中，学生模型的映射模块 $P(\cdot)$ 以较小的计算代价调整特征维度以确保对齐。然后，我们将教师文本特征 $W \in \mathbb{R}^{N \times d}$ 与教师和学生图像特征相乘，得到各自的输出逻辑值：$q^t = u^t W^\top \in \mathbb{R}^N$ 和 $q^s = u^s W^\top \in \mathbb{R}^N$。最终，我们优化学生模型，使其在无标签域数据集 $D_{unlabeled}$ 上生成与教师模型相似的输出，可由下式表示：

$$L_{stu} = L_{kd}(q^t, q^s, \tau). \quad (3)$$

Algorithm 1 PromptKD 的伪代码（PyTorch）

```
# tea_t: text encoder of teacher CLIP
# tea_i: image encoder of teacher CLIP
# stu_i: image encoder of student CLIP
# l_tea: teacher output logits
# l_stu: student output logits
# Proj: Feature Projector

# init
f_txt_t = tea_t(txt_of_all_classes)

# forward
for img in unlabeled_dataset:
    f_img_t = tea_i(img)
    f_img_s = stu_i(img)

    f_img_s = Proj(f_img_s)

    # get output predictions
    l_tea = f_img_t * f_txt_t.t()
    l_stu = f_img_s * f_txt_t.t()

    # calculate distillation loss
    loss = KLDivergence(l_stu, l_tea)
    loss.backward()
```

算法 1 展示了 PromptKD 的 PyTorch 风格的伪代码。

推理 训练完成后，我们使用训练好的学生图像编码器 f_I^s 和预存的教师文本特征 W（类别向量）进行推理。

4. 实验

4.1. 实验设置

基类到新类的泛化能力 遵循 [21, 22, 71]，我们将训练和测试数据集划分为基类和新类。教师模型采用 PromptSRC [22] 方法进行预训练，并遵循与 PromptSRC 相同的训练设置。在蒸馏过程中，我们使用整个

ViT-B/16	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
PromptSRC	84.26	76.10	79.97
PromptKD	86.96	80.73	83.73
Δ	+2.70	+4.63	+3.76

(a) 在 11 个数据集上的平均值

ViT-B/16	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
PromptSRC	77.60	70.73	74.01
PromptKD	80.83	74.66	77.62
Δ	+3.23	+3.93	+3.61

(b) ImageNet

ViT-B/16	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
PromptSRC	98.10	94.03	96.02
PromptKD	98.91	96.65	97.77
Δ	+0.81	+2.62	+1.75

(c) Caltech101

ViT-B/16	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
PromptSRC	95.33	97.30	96.30
PromptKD	96.30	98.01	97.15
Δ	+0.97	+0.71	+0.85

(d) OxfordPets

ViT-B/16	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
PromptSRC	78.27	74.97	76.58
PromptKD	82.80	83.37	83.13
Δ	+4.53	+8.40	+6.55

(e) StanfordCars

ViT-B/16	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
PromptSRC	98.07	76.50	85.95
PromptKD	99.42	82.62	90.24
Δ	+1.35	+6.12	+4.29

(f) Flowers102

ViT-B/16	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
PromptSRC	90.67	91.53	91.10
PromptKD	92.43	93.68	93.05
Δ	+1.76	+2.15	+1.95

(g) Food101

ViT-B/16	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
PromptSRC	42.73	37.87	40.15
PromptKD	49.12	41.81	45.17
Δ	+6.39	+3.94	+5.02

(h) FGVCAircraft

ViT-B/16	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
PromptSRC	82.67	78.47	80.52
PromptKD	83.69	81.54	82.60
Δ	+1.02	+3.07	+2.08

(i) SUN397

ViT-B/16	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
PromptSRC	83.37	62.97	71.75
PromptKD	85.84	71.37	77.94
Δ	+2.47	+8.40	+6.19

(j) DTD

ViT-B/16	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
PromptSRC	92.90	73.90	82.32
PromptKD	97.54	82.08	89.14
Δ	+4.64	+8.18	+6.82

(k) EuroSAT

ViT-B/16	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
PromptSRC	87.10	78.80	82.74
PromptKD	89.71	82.27	86.10
Δ	+2.61	+3.47	+3.36

(l) UCF101

表 1. 与现有最先进的方法在基类到新类泛化任务上的比较。使用 CLIP 模型的 **ViT-B/16 图像编码器**，我们提出的 PromptKD 展现出强大的泛化能力，并在 11 个识别数据集上取得了显著提升。在我们的方法中，默认的教师模型为 ViT-L/14 CLIP 模型。Δ 符号表示相较于之前的 SOTA 方法 PromptSRC 的性能提升。实验结果表明，PromptKD 在所有数据集上均优于现有方法。

无标签训练集来训练学生模型。在蒸馏完成后，学生模型在测试集上的基类和新类性能进行评估。

跨数据集评估 与 PromptSRC [22] 相同，我们的教师模型在源数据集 (即 ImageNet) 上进行 16-shot 训练数据配置的预训练。然后，我们利用目标无标签数据集的训练集来训练学生模型，并在训练后在目标测试集上评估其性能。在 PromptKD 中，我们使用未见类别的无标签图像进行学生模型训练，这属于推理式 (*Transductive*) 零样本学习方法。而对于 CoOp、MaPLe 和 PromptSRC 等方法，它们的训练基于已有类别数

		目标数据集										
ZSL	ViT-B/16	Caltech 101	Oxford Pets	Standford Cars	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	Avg.
In-ductive	CoOp	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
	CoCoOp	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
	MaPLe	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
	PromptSRC	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
Trans-ductive	PromptKD	93.61	**91.59**	**73.93**	**75.33**	**88.84**	**26.24**	**68.57**	**55.08**	**63.74**	**76.39**	**71.33**
	Δ	+0.01	+1.34	+8.23	+5.08	+2.69	+2.34	+1.47	+8.21	+18.24	+7.64	+5.52

表 2. PromptKD 与现有先进方法在跨数据集基准评估上的比较。基于我们的流程，我们在无标签的域数据上分别执行无监督提示蒸馏（即推理式）。源模型在 ImageNet [7] 上进行训练。“ZSL” 表示零样本学习（Zero-Shot Learning）的设定类型。实验结果表明，PromptKD 在 10 个数据集中的 9 个上取得了更优的性能。

据，属于归纳式（*Inductive*）零样本学习方法。

数据集 我们在 11 个常用的视觉识别数据集上评估模型的性能。每个数据集的详细信息在附录中提供。

实现细节 我们采用 ViT-L/14 CLIP 作为教师模型，ViT-B/16 CLIP 作为目标学生模型。除非特别说明，我们默认使用 PromptSRC [22] 作为教师模型的预训练方法。我们列出基类和新类的准确率，并计算它们的调和平均数 (HM)，所有结果均取 3 次运行的平均值。由于篇幅限制，请参考附录以获取更多实现细节和实验结果。

4.2. 基类到新类 (Base-to-Novel) 的泛化能力

如表 1 所示，在使用相同的 ViT-B/16 CLIP 预训练图像编码器的基础上，我们将 PromptKD 与最新的提示学习方法，包括 CoOp、CoCoOp、MaPLe 和 PromptSRC 在 11 个识别数据集上进行比较。与以往方法相比，PromptKD 在所有 11 个数据集上均表现出更优的性能。我们提供了 ViT-L/14 作为教师模型时的每个数据集的教师模型准确率，详情见附录。

4.3. 跨数据集 (Cross-dataset) 评估

表 2 展示了 PromptKD 与 CoOp、CoCoOp、MaPLe 和 PromptSRC 的性能对比。与先前方法相比，我们的方法在 10 个数据集中有 9 个表现更优，并在整体上相比之前的方法平均提升 5.52%。这表明 PromptKD 在跨数据集泛化方面具有较强的优势。

4.4. 与其他方法的比较

在 PromptKD 中，我们使用无标签图像来训练目标学生模型。表 3 对比了我们的方法与其他利用无标签数据进行模型训练的方法。大多数方法采用伪标签（Pseudo-labeling）来利用无标签数据，而我们的方法采用教师-学生范式，其中教师模型通过提供软标签来指导学生模型在无标签数据上进行训练。为了公平比较，使用少量标注样本的方法（[35] 和 PromptKD）均基于 PromptSRC 框架实现。所有实验均采用 **ViT-B/16** 的 CLIP 模型，且少量标注样本的数量为 16。在 Flowers102 数据集上的结果凸显了我们方法相较于之前方法的明显性能优势。

方法	域数据	Base	Novel	HM
CLIP	Zero-shot	72.08	77.80	74.83
PromptSRC	Few-shot	98.07	76.50	85.95
CLIP-PR [20]		65.05	71.13	67.96
UPL [16]	Unlabeled	74.83	78.04	76.40
LaFTer [36]		79.49	82.91	81.16
FPL [35]		97.60	78.27	86.87
IFPL [35]	Few-shot	97.73	80.27	88.14
GRIP [35]	+	97.83	80.87	88.54
PromptKD	Unlabeled	99.42	82.62	90.24
Δ		+1.59	+1.75	+1.70

表 3. 与现有使用无标签数据的方法在 Flowers102 数据集上的比较。实验结果表明，我们的方法相较于先前方法取得了更优的性能。

4.5. 消融实验

本部分对框架中的不同组件进行消融实验，以验证其有效性。默认情况下，我们在 ImageNet 数据集上进行蒸馏实验，并默认使用每个类别 64 张图像（总计 64,000 张包含 1000 个类别的图像）作为无标签训练集

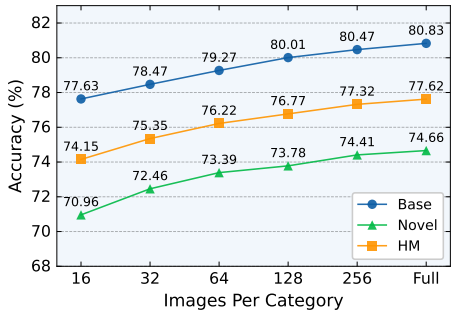

图 4. 随着每个类别用于蒸馏的无标签图像数量增加，学生模型在 ImageNet 数据集上的分类准确率提升。

进行蒸馏，除非另有说明。基类和新类的准确率在测试集上评估。

训练时使用的数据量 在本节中，我们的目标是评估训练数据量对蒸馏性能的影响，如图 4 所示，随着训练图像数量逐渐增加至整个训练数据集，模型的准确率持续提升。然而，需要注意的是，当训练图像数量进一步增加时，性能提升的速率开始趋于平稳。

蒸馏形式	损失	Base	Novel	HM
特征 (Feature)	L1	73.09	65.98	69.35
	MSE	71.89	66.17	68.91
逻辑值 (Logit)	KL	**79.27**	**73.39**	**76.22**

表 4. 不同蒸馏形式的比较。基于逻辑值的蒸馏方法效果最好。

蒸馏形式 表 4 比较了基于特征和基于逻辑值的蒸馏方法。在特征蒸馏中，我们对齐教师和学生的图像编码特征。通过细致地调整超参数，我们发现逻辑值蒸馏显著优于特征蒸馏。一个可能的原因是，由于教师和学生模型的结构存在差异，图像特征空间的对齐比逻辑值空间的对齐更困难。

蒸馏方法 表 5 对比了不同的蒸馏方式的表现。“Projector Only” 代表仅在学生模型中使用映射模块，不包含可学习的提示。“Full Fine-tune” 代表像 CLIP-KD [63] 那样对学生模型的所有参数进行微调。 “w/o Shared Text Feature” 代表训练学生模型时不使用教师的文本特征，而是用学生的文本编码器生成文本特征。实验结果表明，PromptKD 采用的提示蒸馏策略以及共享教师类别向量的机制对最终性能起到了关键作用。

教师预训练方法 表 6 比较了不同的教师预训练方法，我们采用多种方法对教师模型进行预训练实验，包括 vanilla CLIP 和 MaPLe。实验结果表明，教师模型通过预训练获得更高的准确率，与学生模型的蒸馏性能

方法	Base	Novel	HM
CLIP	72.43	68.14	70.22
Projector Only	78.48	72.79	75.53
Full Fine-tune	75.90	70.95	73.34
w/o Shared Text Feature	78.79	73.37	75.98
PromptKD	79.27	73.39	76.22

表 5. 不同蒸馏方法的消融实验。

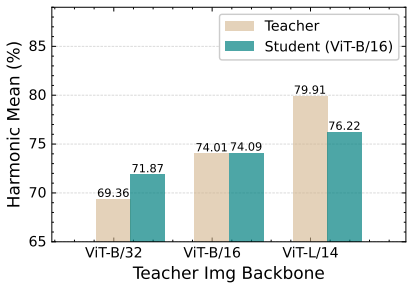

图 5. 不同容量教师模型的蒸馏结果对比。更强的教师模型能够带来更优的蒸馏性能。

提升相一致。值得注意的是，无论采用何种教师模型，都能对学生模型带来显著的性能提升。

Role (Method)	Img Backbone	Base	Novel	HM
CLIP	ViT-B/16	72.43	68.14	70.22
PromptSRC	ViT-B/16	77.60	70.73	74.01
Teacher (CLIP)	ViT-L/14	79.18	74.03	76.52
Student	ViT-B/16	76.53	72.58	74.50
Teacher (MaPLe)	ViT-L/14	82.79	76.88	79.73
Student	ViT-B/16	78.43	**73.61**	75.95
Teacher (PromptSRC)	ViT-L/14	83.24	76.83	79.91
Student	ViT-B/16	**79.27**	73.39	**76.22**

表 6. 不同预训练方法的比较。使用 PromptSRC 预训练的教师模型带来最佳的学生模型的表现。

不同教师模型的蒸馏性能 在本部分中，我们研究了不同容量的教师模型对学生模型性能的影响，如图 5所示。我们使用官方 PromptSRC 代码对 ViT-B/16 和 ViT-B/32 CLIP 模型进行预训练，并将其作为教师模型。实验结果表明，更强的教师模型能够在蒸馏过程中带来更好的性能提升。

推理计算成本分析 表 7 展示了推理计算成本的比较，并与 CoOp、CoCoOp 和 PromptSRC 等提示学习方法进行对比。所有方法的推理成本均在单张 A100 GPU 上进行计算，使用 SUN397 数据集。实验结果表明，我们的方法在推理阶段比之前的方法更高效，证实了其在实际应用中的实用性。

方法	GFLOPs (test)	FPS	HM
CoOp	162.5	1344	71.66
CoCoOp	162.5	15.08	75.83
PromptSRC	162.8	1380	79.97
PromptKD	42.5	1710	83.73

表 7. 现有方法在 SUN397 数据集上的计算成本对比。实验结果表明，我们的 PromptKD 在测试阶段比以往方法更高效。

5. 结论

在本文中，我们提出了一种用于视觉-语言模型的两阶段无监督提示蒸馏框架，旨在无标签的域数据上通过提示模仿，将大规模 CLIP 教师模型的知识推理到轻量级 CLIP 学生模型。我们的方法首先在少样本的域标注数据上预训练一个大规模教师模型，然后在大规模无标签的域数据上执行学生提示蒸馏。利用 CLIP 的独特模态解耦特性，我们提出复用预存的教师文本特征，并将其带入学生图像编码器中，以用于蒸馏和推理。我们在 11 个识别数据集上的大量实验验证了该方法的有效性。

局限性与未来工作 蒸馏方法的有效性与通过无标签域样本所传递的知识紧密相关。当蒸馏数据缺乏目标域的表示时，蒸馏后的学生模型在该特定域的泛化能力可能会受到影响，甚至产生偏差。未来，我们计划探索潜在的正则化方法，以缓解这些问题，并进一步提升蒸馏过程的泛化能力。

致谢 本研究受国家自然科学基金青年科学基金项目（No. 62206134）、国家自然科学基金（No. 62361166670）、中央高校基本科研业务费（070-63233084）及天津市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] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1

[2] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *arXiv preprint arXiv:2207.02039*, 2022. 3

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 3

[4] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, pages 11933–11942, 2022. 3

[5] Han Cheng, Wang Qifan, Cui Yiming, Cao Zhiwen, Wang Wenguan, Qi Siyuan, and Liu Dongfang. E2vpt: An effective and efficient approach for visual prompt tuning. In *ICCV*, 2023. 2

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[7] 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. 7, 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[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. 1

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, pages 1–15, 2023. 2

[11] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 3

[12] 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. 1

[13] 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. 1

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3, 4, 5, 2

[16] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 7

[17] 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, 4

[18] 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. 2, 3, 4

[19] Zihao Jia, Shengkun Sun, Guangcan Liu, and Bo Liu. Mssd: multi-scale self-distillation for object detection. *Visual Intelligence*, 2(1):8, 2024. 3

[20] Jonathan Kahana, Niv Cohen, and Yedid Hoshen. Improving zero-shot models with label distribution priors. *arXiv preprint arXiv:2212.00784*, 2022. 7

[21] 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, 5, 1

[22] 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, 5, 6, 7, 1

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, pages 554–561, 2013. 1

[24] Clement Laroudie, Andrei Bursuc, Mai Lan Ha, and Gianni Franchi. Improving clip robustness with knowledge distillation and self-training. *arXiv preprint arXiv:2309.10361*, 2023. 3

[25] 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, 3

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

[27] Mingchao Li, Kun Huang, Xiao Ma, Yuexuan Wang, Wen Fan, and Qiang Chen. Mask distillation network for conjunctival hyperemia severity classification. *Machine Intelligence Research*, 20(6):909–922, 2023. 3

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

[29] Zheng Li, Ying Huang, Defang Chen, Tianren Luo, Ning Cai, and Zhigeng Pan. Online knowledge distillation via multi-branch diversity enhancement. In *ACCV*, 2020. 3

[30] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *ICCV*, pages 11740–11750, 2021. 3

[31] 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. 3, 5, 2

[32] Xianglong Liu, Shihao Bai, Shan An, Shuo Wang, Wei Liu, Xiaowei Zhao, and Yuqing Ma. A meaningful learning method for zero-shot semantic segmentation. *Science China Information Sciences*, 66(11):210103, 2023. 3

[33] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, pages 2604–2613, 2019. 3

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[35] Cristina Menghini, Andrew Delworth, and Stephen H Bach. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. *arXiv preprint arXiv:2306.01669*, 2023. 7

[36] M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *NeurIPS*, 2023. 7

[37] 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. 1

[38] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 3

[39] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 1

[40] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, pages 18983–18992, 2023. 3

[41] 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, 3, 4

[42] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 3

[43] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 1

[44] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *arXiv preprint arXiv:2304.04704*, 2023. 3

[45] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023. 1

[46] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 3

[48] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35:14274–14289, 2022. 1

[49] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. In *CVPR*, pages 1024–1033, 2018. 3

[50] 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. 1

[51] Ziyu Wan, Dongdong Chen, Yan Li, Xingguang Yan, Junge Zhang, Yizhou Yu, and Jing Liao. Transductive zero-shot learning with visual structure constraint. *NeurIPS*, 32, 2019. 3

[52] Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alexander G Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. *ICLR*, 2023. 2

[53] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1

[54] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for dense object detection. *arXiv preprint arXiv:2306.11369*, 2023. 3

[55] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on In-*

telligent Systems and Technology (TIST), 10(2):1–37, 2019. 3

[56] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424, 2022. 3

[57] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, pages 21970–21980, 2023. 3

[58] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41(9):2251–2265, 2018. 3

[59] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 3

[60] 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. 1

[61] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. In *AAAI*, pages 3045–3053, 2022. 3

[62] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, pages 12319–12328, 2022. 3

[63] 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, 3, 8

[64] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021. 3

[65] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *NeurIPS*, 36, 2024. 2

[66] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *NeurIPS*, 35:9125–9138, 2022. 3

[67] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 2021–2030, 2017. 3

[68] Xuying Zhang, Bo-Wen Yin, Yuming Chen, Zheng Lin, Yunheng Li, Qibin Hou, and Ming-Ming Cheng. Temo: Towards text-driven 3d stylization for multi-object meshes. *arXiv preprint arXiv:2312.04248*, 2023. 1

[69] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 3

[70] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. 3

[71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2, 3, 4, 5, 1

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

PromptKD: 用于视觉-语言模型的无监督提示蒸馏

补充材料

6. 实验设置

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的泛化能力以及跨数据集的性能，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 [7] 和 Caltech101 [9]；用于细粒度分类的 OxfordPets [39]、StanfordCars [23]、Flowers102 [37]、Food101 [1] 和 FGVCAircraft [34]；用于场景识别的 SUN397 [60]；用于动作识别的 UCF101 [50]；用于纹理分类的 DTD [6]；以及用于卫星图像识别的 EuroSAT [12]。对于域泛化实验，我们使用 ImageNet-1K 作为源数据集，并将其四个变体作为目标数据集，包括 ImageNet-V2 [43]、ImageNet-Sketch [53]、ImageNet-A [14] 和 ImageNet-R [13]。

训练细节 对于 PromptKD，我们遵循与 PromptSRC 相同的设置，将提示深度设置为 9，视觉和语言提示长度设置为 4。我们使用随机梯度下降（SGD）作为优化器。所有学生模型均训练 20 个 epoch，批大小为 8，学习率为 0.005。我们遵循 PromptSRC 中的标准数据增强方案，即随机调整大小裁剪和随机翻转。当前蒸馏方法中的温度超参数 τ 默认设置为 1。第一层的文本提示使用 “a photo of a {classname}” 的词嵌入进行初始化。所有实验均在单张 Nvidia A100 GPU 上进行。

训练数据使用 在我们方法的初始阶段，我们使用 PromptSRC 预训练我们的 ViT-L/14 CLIP 教师模型。在此阶段，我们使用与 PromptSRC 相同的训练数据进行训练。在后续阶段，我们采用基于归纳式的零样本学习范式，并使用整个训练数据集来训练我们的学生模型。在表 8中，我们提供了在基类到新类 (Base-to-Novel) 泛化设置下用于训练的图像数量的详细信息。

7. 附加实验

域泛化 (Domain Generalization) 在我们的 PromptKD 方法中，教师模型首先使用 PromptSRC [22] 在源数据集（即 ImageNet）上进行预训练。然后，我们使用无标签的目标数据集训练学生模型，并在训练完成后评估其性能。

数据集	Train	Test Base	Test Novel
ImageNet	1,281,167	25,000	25,000
Caltech101	4,128	1,549	916
OxfordPets	2,944	1,881	1,788
StandfordCars	6,509	4,002	4,039
Flowers102	4,093	1,053	1,410
Food101	50,500	15,300	15,000
FGVCAircraft	3,334	1,666	1,667
SUN397	15,880	9,950	9,900
DTD	2,820	864	828
EuroSAT	13,500	4,200	3,900
UCF101	7,639	1,934	1,849

表 8. 每个数据集用于蒸馏和测试的图像数量。

		目标数据集				
ZSL	ViT-B/16	-V2	-S	-A	-R	Avg.
Inductive	CLIP	60.83	46.15	47.77	73.96	57.18
	CoOp	64.20	47.99	49.71	75.21	59.28
	CoCoOp	64.07	48.75	50.63	76.18	59.91
	MaPLe	64.07	49.15	50.90	76.98	60.27
	PromptSRC	64.35	49.55	50.90	77.80	60.65
Transductive	TPT	63.45	47.94	54.77	77.06	60.81
	CoOp+TPT	66.83	49.29	57.95	77.27	62.83
	CoCoOp+TPT	64.85	48.47	58.47	78.65	62.61
	PromptAlign	65.29	50.23	59.37	79.33	63.55
	PromptKD	**69.77**	**58.72**	**70.36**	**87.01**	**71.47**
	Δ	+4.48	+8.49	+10.99	+7.68	+7.92

表 9. PromptKD 与现有先进方法在域泛化设定下的表现比较。基于我们的流程，我们分别对无标签的域数据执行无监督的提示蒸馏，即采用推理式 (Transductive) 设定。源模型在 ImageNet [7] 上进行训练。“ZSL” 表示零样本学习 (Zero-Shot Learning) 设定类型。PromptKD 在所有目标数据集上均实现了稳定的性能提升。

在表 9 中，我们展示了 PromptKD 与其他最先进的方法（如 CoOp [72]、CoCoOp [71]、MaPLe [21]、PromptSRC [22]、TPT [48]、PromptAlign [45]）在四个不同数据集上的实验结果。在目标数据集上，我们的方法相较于其他方法表现出明显的性能优势。

教师模型准确率 在表 10 和表 11 中，我们展示了基于 ViT-L/14 预训练的 CLIP 教师模型在基类到新类任务和跨数据集实验中的准确率。

数据集	Base	Novel	HM
ImageNet	83.24	76.83	79.91
Caltech101	98.71	98.03	98.37
OxfordPets	96.86	98.82	97.83
StandfordCars	84.53	84.25	84.39
Flowers102	99.05	82.60	90.08
Food101	94.56	95.15	94.85
FGVCAircraft	54.44	43.07	48.09
SUN397	84.97	81.09	82.98
DTD	85.76	70.65	77.48
EuroSAT	94.79	83.15	88.59
UCF101	89.50	82.26	85.73

表 10. 预训练的 ViT-L/14 CLIP 教师模型在基类到新类 (Base-to-Novel) 泛化实验中的准确率。

ViT-L/14	数据集	准确率
源数据集	ImageNet	78.12
目标数据集	Caltech101	95.61
	OxfordPets	94.19
	StandfordCars	77.70
	Flowers102	77.54
	Food101	91.59
	FGVCAircraft	31.29
	SUN397	70.86
	DTD	56.32
	EuroSAT	47.55
	UCF101	76.20
Avg.		71.89

表 11. 预训练的 ViT-L/14 CLIP 教师模型在跨数据集 (Cross-dataset) 泛化实验中的准确率。

映射器 表 12 展示了使用不同 MLP 层数的映射器（Projector）对蒸馏性能的影响。结果表明，使用两层 MLP 便足以实现特征对齐。增加或减少 MLP 层数可能会导致训练过程中的过拟合或欠拟合问题。

MLP 层	Base	Novel	HM
1	78.97	72.90	75.81
2	**79.27**	**73.39**	**76.22**
3	79.10	72.72	75.78

表 12. 映射器（Projector）层数对比。2 层 MLP 表现最佳。

不同学生模型的蒸馏实验 为了验证 PromptKD 在不同容量学生模型上的有效性，我们进一步在使用 ViT-B/32 作为图像编码器的 CLIP 模型上进行实验，如表 13 所示。实验结果表明，学生模型通过 PromptKD 方法均获得了稳定的性能提升。

Role	Img Backbone	Base	Novel	HM
Teacher	ViT-L/14	83.24	76.83	79.91
Baseline		67.52	64.04	65.73
Student	ViT-B/32	74.29	69.29	71.70
Δ		+6.77	+5.25	+5.97
Baseline		72.43	68.14	70.22
Student	ViT-B/16	80.83	74.66	77.62
Δ		+8.40	+6.52	+7.40

表 13. 使用不同学生 CLIP 模型的提示蒸馏实验。Δ 表示相较于基线结果的性能提升。不同容量的学生模型均取得了稳定的提升。

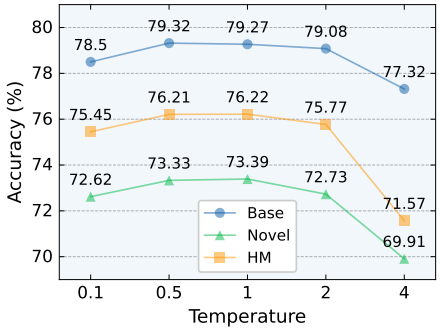

图 6. 温度超参数的选择。当温度超参数 $\tau = 1$ 时，模型性能最佳。

温度超参数 温度参数控制概率分布的平滑度 [15]，并影响蒸馏过程的学习难度 [31]。在传统蒸馏方法中，通常将温度参数 τ 设为 4，以适用于大多数教师-学生模型对和数据集。在图 6 中，我们评估了不同温度值对我们提出的提示蒸馏方法的影响。结果表明，传统的 $\tau = 4$ 设定不适用于当前任务。随着温度值的增加，模型性能迅速下降。有趣的是，当 $\tau = 1$ 时，模型达到了最佳性能。

更长训练周期的蒸馏 在 PromptKD 中，为了公平比较，我们采用了与 PromptSRC 相同的训练轮数，即 20 个 epoch。在本部分，我们探讨更长的训练周期是否会进一步提升学生模型的性能。如表 14 所示，我们分别进行了 20、40 和 60 轮训练实验。结果表明，训练时间越长，学生模型的性能越高。

Train Epoch	Base	Novel	HM
20	79.27	73.39	76.22
40	79.75	73.65	76.58
60	**79.89**	**73.68**	**76.66**

表 14. 更长训练周期的蒸馏。训练时间越长，学生模型的性能越高。

8. 讨论

全微调方法的实验结果 在正文的表 5 中，我们注意到全微调方法的结果比其他蒸馏方法低很多（>2%）。这主要有两个原因，第一个原因是我们训练中使用的数据集规模有限。它远小于通常用于训练 CLIP 的 CC3M [47]、CC12M [3] 或 LAION-400M [46] 数据集。第二个原因是训练时间较短。为了与其他实验设置保持一致，我们仅对学生模型训练了 20 个 epoch。总的来说，如果使用更大的数据集并采用更长的训练时间，全微调方法的性能将会有所提升。

使用较差教师模型的蒸馏 在正文的图 5 中，当选择一个比学生（ViT-B/16）更弱的教师模型（ViT-B/32）时，使用 PromptKD 训练的学生模型表现出比基线方法更好的性能（71.87%>70.22%）。这种情况与传统蒸馏方法不同，在传统方法中，较差的教师模型通常会导致学生性能显著下降。这种差异源于提示学习方法的特点，即只训练可学习的提示，而保持原始 CLIP 模型的权重不变。冻结的 CLIP 模型在预测过程中仍然具有重要影响，训练后的提示不会显著偏置模型的推理能力或结果。