MaPLe: Multi-modal Prompt Learning

MaPLe: 多模态提示学习

Muhammad Uzair Khattak [1] Hanoona Rasheed [1] Muhammad Maaz [1]
穆罕默德 · 乌扎尔 · 哈塔克 [1] 哈努娜 · 拉希德 [1] 穆罕默德 · 马阿兹 [1]
Salman Khan [1,2] Fahad Shahbaz Khan [1,3]
萨尔曼 · 汗 [1,2] 法哈德 · 沙巴兹 · 汗 [1,3]
[1] Mohamed bin Zayed University of AI [2] Australian National University [3] Linköping University
[1] 穆罕默德 · 本 · 扎耶德人工智能大学 [2] 澳大利亚国立大学 [3] 林雪平大学

Abstract

摘要

Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on three representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method Co-CoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets. Our code and pre-trained models are available at https://github.com/muzairkhattak/multimodal-prompt-learning.

预训练的视觉-语言 (V-L) 模型，如 CLIP，已显示出对下游任务的优秀泛化能力。然而，它们对输入文本提示的选择非常敏感，需要仔细选择提示模板以获得良好的性能。受到自然语言处理 (NLP) 文献的启发，最近的 CLIP 适应方法将提示作为文本输入，以微调 CLIP 以适应下游任务。我们注意到，在 CLIP 的单一分支 (语言或视觉) 中使用提示来适应表示是次优的，因为这不允许在下游任务中动态调整两个表示空间的灵活性。在本研究中，我们提出了多模态提示学习 (MaPLe)，旨在改善视觉和语言分支之间的对齐。我们的设计促进了视觉-语言提示之间的强耦合，以确保相互协同，并抑制学习独立的单模态解决方案。此外，我们在不同的早期阶段学习单独的提示，以逐步建模阶段间特征关系，从而允许丰富的上下文学习。我们在三个具有代表性的任务上评估了我们方法的有效性，这些任务涉及对新类别的泛化、新目标数据集和未见领域的转移。与最先进的方法 Co-CoOp 相比，MaPLe 展现出良好的性能，在新类别上获得了绝对增益 3.45% ，在整体调和均值上获得了 2.72% ，该结果是基于 11 个不同的图像识别数据集的平均值。我们的代码和预训练模型可在 https://github.com/muzairkhattak/multimodal-prompt-learning 获取。

1. Introduction

1. 引言

Foundational vision-language (V-L) models such as CLIP (Contrastive Language-Image Pretraining) [32] have shown excellent generalization ability to downstream tasks. Such models are trained to align language and vision modalities on web-scale data e.g., 400 million text-image pairs in CLIP. These models can reason about open-vocabulary visual concepts, thanks to the rich supervision provided by natural language. During inference, hand-engineered text prompts are used e.g., 'a photo of a <category>' as a query for text encoder. The output text embeddings are matched with the visual embeddings from an image encoder to predict the output class. Designing high quality contextual prompts have been proven to enhance the performance of CLIP and other V-L models [17,42].

基础的视觉-语言 (V-L) 模型，如 CLIP (对比语言-图像预训练) [32]，在下游任务中表现出优秀的泛化能力。这些模型在网络规模的数据上进行训练，例如，CLIP 中的 4 亿对文本-图像对。这些模型能够推理开放词汇的视觉概念，这得益于自然语言提供的丰富监督。在推理过程中，使用手工设计的文本提示，例如，' 一张 < 类别 > 的照片' 作为文本编码器的查询。输出的文本嵌入与来自图像编码器的视觉嵌入匹配，以预测输出类别。设计高质量的上下文提示已被证明可以提高 CLIP 和其他 V-L 模型的性能 [17,42]。

Despite the effectiveness of CLIP towards generalization to new concepts, its massive scale and scarcity of training data (e.g., few-shot setting) makes it infeasible to fine-tune the full model for downstream tasks. Such fine-tuning can also forget the useful knowledge acquired in the large-scale pretraining phase and can pose a risk of overfitting to the downstream task. To address the above challenges, existing works propose language prompt learning to avoid manually adjusting the prompt templates and providing a mechanism to adapt the model while keeping the original weights frozen [14, 25, 29, 48, 49] . Inspired from Natural Language Processing (NLP), these approaches only explore prompt learning for the text encoder in CLIP (Fig. 1:a) while adaptation choices together with an equally important image encoder of CLIP remains an unexplored topic in the literature.

尽管 CLIP 在对新概念的泛化方面表现出色，但其庞大的规模和训练数据的稀缺性 (例如，少量样本设置) 使得对下游任务进行全模型微调变得不可行。这种微调还可能会遗忘在大规模预训练阶段获得的有用知识，并可能导致对下游任务的过拟合风险。为了解决上述挑战，现有研究提出了语言提示学习，以避免手动调整提示模板，并提供了一种在保持原始权重不变的情况下适应模型的机制 [14, 25, 29, 48, 49] 。受自然语言处理 (NLP) 启发，这些方法仅探索了 CLIP 中文本编码器的提示学习 (图 1:a)，而适应选择以及同样重要的 CLIP 图像编码器仍然是文献中未被探索的话题。

Our motivation derives from the multi-modal nature of CLIP, where a text and image encoder co-exist and both contribute towards properly aligning the V-L modalities. We argue that any prompting technique should adapt the model completely and therefore, learning prompts only for the text encoder in CLIP is not sufficient to model the adaptations needed for the image encoder. To this end, we set out to achieve completeness in the prompting approach and propose Multi-modal Prompt Learning (MaPLe) to adequately fine-tune the text and image encoder representations such that their optimal alignment can be achieved on the downstream tasks (Fig. 1:b). Our extensive experiments on three key representative settings including base-to-novel generalization, cross-dataset evaluation, and domain generalization demonstrate the strength of MaPLe. On base-to-novel generalization, our proposed MaPLe outperforms existing prompt learning approaches across 11 diverse image recognition datasets (Fig. 1:c) and achieves absolute average gain of 3.45% on novel classes and 2.72% on harmonic-mean over the state-of-the-art method Co-CoOp [48]. Further, MaPLe demonstrates favorable generalization ability and robustness in cross-dataset transfer and domain generalization settings, leading to consistent improvements compared to existing approaches. Owing to its streamlined architectural design, MaPLe exhibits improved efficiency during both training and inference without much overhead, as compared to Co − CoOp which lacks efficiency due to its image instance conditioned design. In summary, the main contributions of this work include:

我们的动机源于 CLIP 的多模态特性，其中文本和图像编码器共存，并共同促进 V-L 模态的适当对齐。我们认为，任何提示技术都应该完全适应模型，因此，仅为 CLIP 中的文本编码器学习提示不足以建模图像编码器所需的适应性。为此，我们致力于在提示方法中实现完整性，并提出多模态提示学习 (MaPLe)，以充分微调文本和图像编码器的表示，从而在下游任务中实现它们的最佳对齐 (图 1:b)。我们在三个关键代表性设置上进行了广泛的实验，包括基础到新颖的泛化、跨数据集评估和领域泛化，展示了 MaPLe 的强大能力。在基础到新颖的泛化中，我们提出的 MaPLe 在 11 个不同的图像识别数据集上超越了现有的提示学习方法 (图 1:c)，在新类上获得了绝对平均增益 3.45% ，在与最先进的方法 Co-CoOp [48] 的调和平均上获得了 2.72% 的增益。此外，MaPLe 在跨数据集转移和领域泛化设置中表现出良好的泛化能力和鲁棒性，与现有方法相比，带来了持续的改进。由于其简化的架构设计，MaPLe 在训练和推理期间的效率得到了提升，相较于由于其图像实例条件设计而缺乏效率的 Co − CoOp ，没有太多开销。总之，这项工作的主要贡献包括:

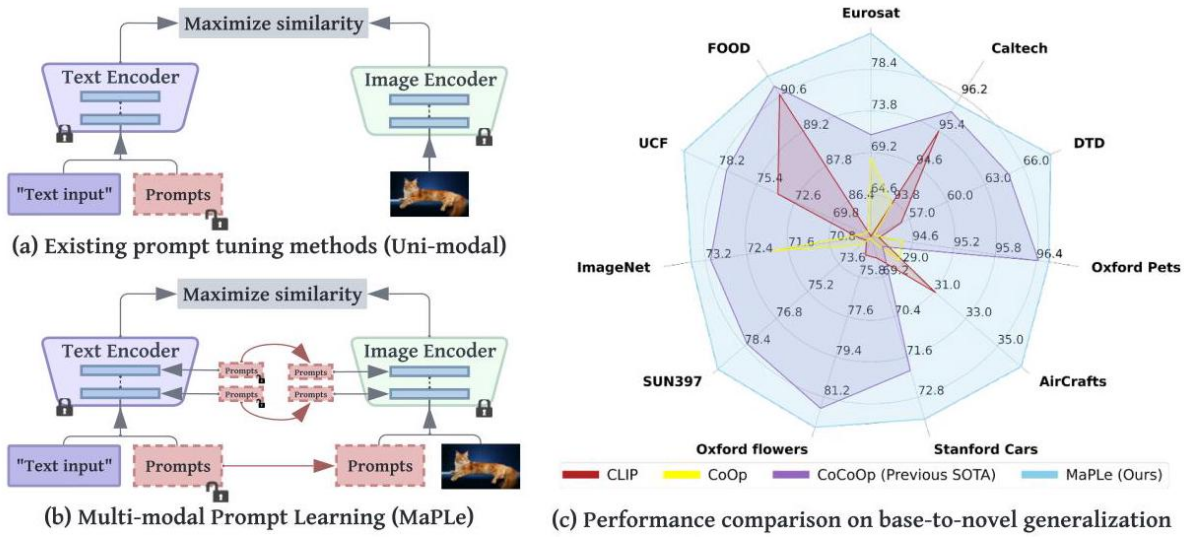


Figure 1. Comparison of MaPLe with standard prompt learning methods. (a) Existing methods adopt uni-modal prompting techniques to fine-tune CLIP representations as prompts are learned only in a single branch of CLIP (language or vision). (b) MaPLe introduces branch-aware hierarchical prompts that adapt both language and vision branches simultaneously for improved generalization. (c) MaPLe surpasses state-of-the-art methods on 11 diverse image recognition datasets for novel class generalization task.

图 1. MaPLe 与标准提示学习方法的比较。(a) 现有方法采用单模态提示技术来微调 CLIP 表示，因为提示仅在 CLIP 的单一分支 (语言或视觉) 中学习。(b) MaPLe 引入了分支感知的层次化提示，同时适应语言和视觉分支，以提高泛化能力。(c) MaPLe 在 11 个不同的图像识别数据集上超越了最先进的方法，以应对新类别泛化任务。

- We propose multi-modal prompt learning in CLIP to favourably align its vision-language representations. To the best of our knowledge, this is the first multimodal prompting approach for fine-tuning CLIP.
- 我们提出了在 CLIP 中进行多模态提示学习，以有利地对齐其视觉-语言表示。据我们所知，这是第一个用于微调 CLIP 的多模态提示方法。
- To link prompts learned in text and image encoders, we propose a coupling function to explicitly condition vision prompts on their language counterparts. It acts as a bridge between the two modalities and allows mutual propagation of gradients to promote synergy.
- 为了链接在文本和图像编码器中学习的提示，我们提出了一种耦合函数，以明确地将视觉提示条件化于其语言对应物。它充当了两种模态之间的桥梁，并允许梯度的相互传播，以促进协同作用。
- Our multi-modal prompts are learned across multiple transformer blocks in both vision and language branches to progressively learn the synergistic behaviour of both modalities. This deep prompting strategy allows modeling the contextual relationships independently, thus providing more flexibility to align the vision-language representations.
- 我们的多模态提示是在视觉和语言分支的多个变换块中学习的，以逐步学习两种模态的协同行为。这种深度提示策略允许独立建模上下文关系，从而提供更多灵活性以对齐视觉-语言表示。

2. Related Work

2. 相关工作

Vision Language Models: The combined use of language supervision with natural images is found to be of great interest in the computer vision community. In contrast to models learned with only image supervision, these vision-language (V-L) models encode rich multimodal representations. Recently, V-L

models like CLIP [32], ALIGN [15], LiT [45], FILIP [41] and Florence [43] have demonstrated exceptional performance on a wide spectrum of tasks including few-shot and zero-shot visual recognition. These models learn joint image-language representations in a self-supervised manner using abundantly available data from the web. For example, CLIP and ALIGN respectively use $\sim$ 400M and $\sim$ 1 B image-text pairs to train a multimodal network. Although these pre-trained V-L models learn generalized representations, efficiently adapting them to downstream tasks is still a challenging problem. Many works have demonstrated better performance on downstream tasks by using tailored methods to adapt V-L models for few-shot image-recognition [9, 19, 46] , object detection [8,10,27,34,44,50], and segmentation [5,22,26,33]. In this work, we propose a novel multi-modal prompt learning technique to effectively adapt CLIP for few-shot and zero-shot visual recognition tasks.

视觉语言模型: 将语言监督与自然图像结合使用在计算机视觉领域引起了极大的兴趣。与仅使用图像监督学习的模型相比，这些视觉-语言 (V-L) 模型编码了丰富的多模态表示。最近，像 CLIP [32]、ALIGN [15]、LiT [45]、FILIP [41] 和 Florence [43] 的 V-L 模型在包括少样本和零样本视觉识别在内的广泛任务上表现出色。这些模型以自监督的方式学习联合图像-语言表示，利用来自网络的大量可用数据。例如，CLIP 和 ALIGN 分别使用 $\sim$ 400M 和 $\sim$ 1 B 图像-文本对来训练多模态网络。尽管这些预训练的 V-L 模型学习了通用表示，但有效地将它们适应于下游任务仍然是一个具有挑战性的问题。许多研究表明，通过使用定制的方法将 V-L 模型适应于少样本图像识别 [9, 19, 46] 、目标检测 [8,10,27,34,44,50] 和分割 [5,22,26,33]，可以在下游任务上获得更好的性能。在本研究中，我们提出了一种新颖的多模态提示学习技术，以有效地将 CLIP 适应于少样本和零样本视觉识别任务。

Prompt Learning: The instructions in the form of a sentence, known as text prompt, are usually given to the language branch of a V-L model, allowing it to better understand the task. Prompts can be handcrafted for a downstream task or learned automatically during fine-tuning stage. The latter is referred to as 'Prompt Learning' which was first used in NLP [21,23,24] followed by the adaptation in V-L [48,49,51] and vision-only [16,38,39,47] models. Similar to [16] our design also uses deep 'vision' prompting. However, ours is the first multi-modal prompting design while [16] is uni-modal.

提示学习: 以句子的形式给出的指令，称为文本提示，通常提供给 V-L 模型的语言分支，使其更好地理解任务。提示可以为下游任务手工制作，也可以在微调阶段自动学习。后者被称为 "提示学习"，最初在自然语言处理 (NLP) 中使用 [21,23,24]，随后在 V-L [48,49,51] 和仅视觉 [16,38,39,47] 模型中进行了适应。与 [16] 类似，我们的设计也使用深度 "视觉" 提示。然而，我们的设计是第一个多模态提示设计，而 [16] 是单模态的。

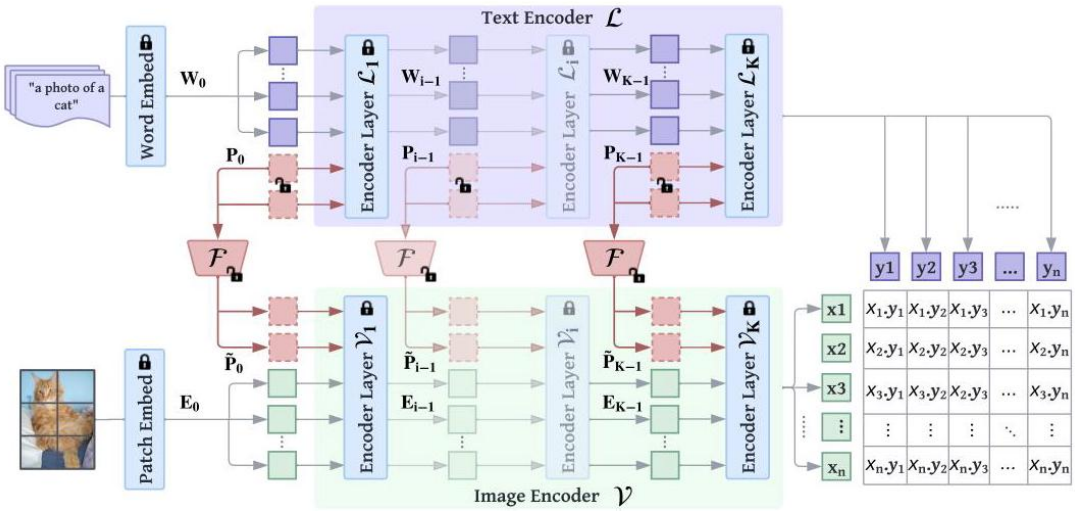


Figure 2. Overview of our proposed MaPLe (Multi-modal Prompt Learning) framework for prompt learning in V-L models. MaPLe tunes both vision and language branches where only the context prompts are learned, while the rest of the model is frozen. MaPLe conditions the vision prompts on language prompts via a V-L coupling function $\mathcal{F}$ to induce mutual synergy between the two modalities. Our framework uses deep contextual prompting where separate context prompts are learned across multiple transformer blocks.

图 2. 我们提出的 MaPLe(多模态提示学习) 框架在 V-L 模型中进行提示学习的概述。MaPLe 调整视觉和语言分支，其中仅学习上下文提示，而模型的其余部分保持不变。MaPLe 通过 V-L 耦合函数 $\mathcal{F}$ 将视觉提示与语言提示相结合，以促进两种模态之间的相互协同。我们的框架使用深度上下文提示，其中在多个变换器块中学习独立的上下文提示。

Prompt Learning in Vision Language models: Full fine-tuning and linear probing [9] are two typical approaches to adapt a V-L model (i.e. CLIP) to the downstream tasks. The complete fine-tuning results in degrading the previously learned joint V-L representation while linear probing limits the zero-shot capability of CLIP. To this end, inspired from prompt learning in NLP, many works have proposed to adapt V-L models by learning the prompt tokens in an end-to-end training. CoOp [49] fine-tunes CLIP for few-shot transfer by optimizing continuous set of prompt vectors at its language branch. Co-CoOp [48] highlights the inferior performance of CoOp on novel classes and solves the generalization issue by explicitly conditioning prompts on image instances. [25] proposes to optimize multiple set of prompts by learning the distribution of prompts. [18] adapt CLIP by learning prompts for video understanding tasks. [1] perform visual prompt tuning on CLIP by prompting on the vision branch. We note that the existing methods follow independent uni-modal solutions and learn prompts either in the language or in the vision branch of CLIP, thus adapting CLIP partially. In this paper, we explore an important question: given the multimodal nature of CLIP, is complete prompting (i.e., in both language and vision branches) better suited to adapt CLIP? Our work is the first to answer this question by investigating the effectiveness of multi-modal prompt learning in order to improve alignment between vision and language representations.

视觉语言模型中的提示学习: 完全微调和线性探测 [9] 是将 V-L 模型 (即 CLIP) 适应于下游任务的两种典型方法。完全微调导致先前学习的联合 V-L 表示退化，而线性探测限制了 CLIP 的零样本能力。为此，受到 NLP 中提示学习的启发，许多研究提出通过端到端训练学习提示令牌来适应 V-L 模型。CoOp [49] 通过优化其语言分支中的连续提示向量集来微调 CLIP，以实现少样本迁移。Co-CoOp [48] 突出了 CoOp 在新类别上的较差性能，并通过明确地将提示条件化于图像实例来解决泛化问题。[25] 提出了通过学习提示的分布来优化多个提示集。[18] 针对视频理解任务学习提示以适应 CLIP。[1] 通过在视觉分支上提示来对 CLIP 进行视觉提示调优。我们注意到，现有方法遵循独立的单模态解决方案，并在 CLIP 的语言或视觉分支中学习提示，因此仅部分适应 CLIP。本文探讨了一个重要问题: 考虑到 CLIP 的多模态特性，完全提示 (即在语言和视觉分支中) 是否更适合于适应 CLIP？我们的工作首次通过研究多模态提示学习的有效性来回答这个问题，以改善视觉和语言表示之间的对齐。

3. Method

3. 方法

Our approach concerns with fine-tuning a pre-trained multimodal CLIP for better generalization to downstream tasks through context optimization via prompting. Fig. 2 shows the overall architecture of our proposed MaPLe (Multimodal Prompt Learning) framework. Unlike previous approaches [48, 49] which learn context prompts only at the language branch, MaPLe proposes a joint prompting approach where the context prompts are learned in both vision and language branches. Specifically, we append learnable context tokens in the language branch and explicitly condition the vision prompts on the language prompts via a coupling function to establish interaction between them. To learn hierarchical contextual representations, we introduce deep prompting in both branches through separate learnable context prompts across different transformer blocks. During fine-tuning, only the context prompts along with their coupling function are learned while the rest of the model is frozen. Below, we first outline the pre-trained CLIP architecture and then present our proposed fine-tuning approach.

我们的方法涉及通过提示优化对预训练的多模态 CLIP 进行微调，以更好地推广到下游任务。图 2 显示了我们提出的 MaPLe(多模态提示学习) 框架的整体架构。与之前的方法 [48, 49] 仅在语言分支学习上下文提示不同，MaPLe 提出了联合提示的方法，其中上下文提示在视觉和语言分支中同时学习。具体而言，我们在语言分支中附加可学习的上下文标记，并通过耦合函数明确地将视觉提示与语言提示相条件，以建立它们之间的交互。为了学习层次上下文表示，我们通过在不同的变换器块中使用单独的可学习上下文提示，在两个分支中引入深度提示。在微调过程中，仅学习上下文提示及其耦合函数，而模型的其余部分保持不变。下面，我们首先概述预训练的 CLIP 架构，然后介绍我们提出的微调方法。

3.1. Revisiting CLIP

3.1. 重新审视 CLIP

We build our approach on a pre-trained vision-language (V-L) model, CLIP, which consists of a text and vision encoder. Consistent with existing prompting methods [48,49], we use a vision transformer (ViT)

[6] based CLIP model. CLIP encodes an image $I \in \mathbb{R}^{H\times W\times 3}$ and a corresponding text description as explained below.

我们的方法建立在一个预训练的视觉-语言 (V-L) 模型 CLIP 上，该模型由文本和视觉编码器组成。与现有的提示方法 [48,49] 一致，我们使用基于视觉变换器 (ViT) [6] 的 CLIP 模型。CLIP 编码图像 $I \in \mathbb{R}^{H\times W\times 3}$ 和相应的文本描述，如下所述。

Encoding Image: Image encoder $\mathcal{V}$ with K transformer layers $\{\mathcal{V}_i\}_{i=1}^{K}$, splits the image I into M fixed-size patches which are projected into patch embeddings $E_0 \in \mathbb{R}^{M\times d_v}$. Patch embeddings E_i are input to the $(i+1)^{\text{th}}$ transformer block $(\mathcal{V}_{i+1})$ along with a learnable class (CLS) token c_i and sequentially processed through K transformer blocks,

编码图像: 图像编码器 $\mathcal{V}$ 具有 K 变换器层 $\{\mathcal{V}_i\}_{i=1}^{K}$ ，将图像 I 切分为 M 固定大小的补丁，这些补丁被投影到补丁嵌入 $E_0 \in \mathbb{R}^{M\times d_v}$ 中。补丁嵌入 E_i 与可学习的类别 (CLS) 标记 c_i 一起输入到 $(i+1)^{\text{th}}$ 变换器块 $(\mathcal{V}_{i+1})$ 中，并通过 K 变换器块顺序处理，

$$[\mathrm{c}_i, E_i] = \mathcal{V}_i\left([\mathrm{c}_{i-1}, E_{i-1}]\right)\ i = 1, 2, \cdots, K.$$

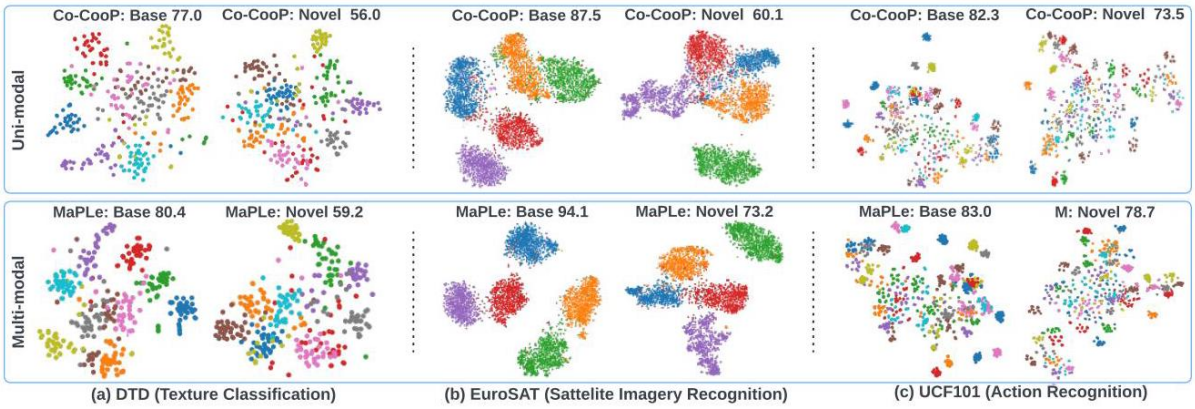


Figure 3. t-SNE plots of image embeddings in uni-modal prompting method Co-CoOp, and MaPLe on 3 diverse image recognition datasets. MaPLe shows better separability in both base and novel classes.

图 3. 在三种不同的图像识别数据集上，uni-modal 提示方法 Co-CoOp 和 MaPLe 的图像嵌入的 t-SNE 图。MaPLe 在基础类和新颖类中显示出更好的可分离性。

To obtain the final image representation x , the class token c_K of last transformer layer $(\mathcal{V}_K)$ is projected to a common V-L latent embedding space via ImageProj,

为了获得最终的图像表示 x ，最后一个变换器层的类别标记 c_K 通过 ImageProj 投影到一个共同的 V-L 潜在嵌入空间，

$$x = \text{ImageProj}\left(\mathrm{c}_K\right)\ x \in \mathbb{R}^{d_{vl}}.$$

Encoding Text: CLIP text encoder generates feature representations for text description by tokenizing the words and projecting them to word embeddings $W_0 = \left[w_0^1, w_0^2, \cdots, w_0^N\right] \in \mathbb{R}^{N\times d_l}$. At each stage, W_i is input to the $(i+1)^{\text{th}}$ transformer layer of text encoding branch $(\mathcal{L}_{i+1})$,

编码文本:CLIP 文本编码器通过对单词进行分词并将其投影到词嵌入 $W_0 = \left[w_0^1, w_0^2, \cdots, w_0^N\right] \in \mathbb{R}^{N\times d_l}$ 来生成文本描述的特征表示。在每个阶段，W_i 被输入到文本编码分支的 $(i+1)^{\text{th}}$ 变换器层 $(\mathcal{L}_{i+1})$ ，

$$[W_i] = \mathcal{L}_i\left(W_{i-1}\right)\ i = 1, 2, \cdots, K.$$

The final text representation z is obtained by projecting the text embeddings corresponding to the last token of the last transformer block $\mathcal{L}_K$ to a common V-L latent embedding space via TextProj,

最终文本表示 z 是通过将对应于最后一个变换器块最后一个标记的文本嵌入 $\mathcal{L}_K$ 投影到一个共同的 V-L 潜在嵌入空间，使用 TextProj 获得的，

$$z = \text{Text Proj}\left(w_K^N\right)\ z \in \mathbb{R}^{d_{vl}}.$$

Zero-shot Classification: For zero-shot classification, text prompts are hand-crafted with class labels $y \in \{1, 2, \ldots C\}$ (e.g.,' a photo of a <category>') having C classes. Prediction $\widehat{y}$ corresponding to the

image I having the highest cosine similarity score ($\text{sim}(\cdot)$) is calculated with a temperature parameter τ ,

零样本分类: 对于零样本分类，文本提示是手工制作的，包含类别标签 $y \in \{1, 2, \ldots C\}$ (例如，“一张 < 类别 > 的照片”) 具有 C 类别。预测 $\widehat{y}$ 对应于具有最高余弦相似度分数的图像 I ，使用温度参数 τ 进行计算，

$$p(\widehat{y} \mid x) = \frac{\exp(\text{sim}(x, z_{\widehat{y}})/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(x, z_i))}.$$

3.2. MaPLe: Multi-modal Prompt Learning

3.2. MaPLe: 多模态提示学习

To efficiently fine-tune CLIP for downstream image recognition tasks, we explore the potential of multi-modal prompt tuning. We reason that prior works that have predominantly explored uni-modal approaches are less suitable as they do not offer the flexibility to dynamically adapt both language and vision representation spaces. Thus to achieve completeness in prompting, we underline the importance of multimodal prompting approach. In Fig. 3, we visualize and compare the image embeddings of MaPLe with recent state-of-the-art work, Co-CoOp. Note that the image embeddings of CLIP, CoOp and Co-CoOp will be identical as they do not learn prompts in the vision branch. The visualization shows that image embeddings of MaPLe are more separable indicating that learning vision prompts in addition to language prompts leads to better adaptation of CLIP.

为了高效地微调 CLIP 以用于下游图像识别任务，我们探索了多模态提示调优的潜力。我们认为，先前主要探索单模态方法的工作不太适合，因为它们无法提供动态适应语言和视觉表示空间的灵活性。因此，为了实现提示的完整性，我们强调多模态提示方法的重要性。在图 3 中，我们可视化并比较了 MaPLe 的图像嵌入与最近的最先进工作 Co-CoOp。请注意，CLIP、CoOp 和 Co-CoOp 的图像嵌入将是相同的，因为它们在视觉分支中不学习提示。可视化结果显示，MaPLe 的图像嵌入更具可分性，表明除了语言提示之外，学习视觉提示有助于更好地适应 CLIP。

In addition to multi-modal prompting, we find that it is essential to learn prompts in the deeper transformer layers to progressively model stage-wise feature representations. To this end, we propose to introduce learnable tokens in the first J (where $J < K$) layers of both vision and language branches. These multi-modal hierarchical prompts utilize the knowledge embedded in CLIP model to effectively learn task relevant contextual representations (see Fig. 4).

除了多模态提示，我们发现学习深层变换器层中的提示对于逐步建模阶段特征表示是至关重要的。为此，我们建议在视觉和语言分支的前 J 层 (其中 $J < K$) 引入可学习的标记。这些多模态层次提示利用嵌入在 CLIP 模型中的知识，有效地学习与任务相关的上下文表示 (见图 4)。

3.2.1 Deep Language Prompting

3.2.1 深度语言提示

To learn the language context prompts, we introduce b learnable tokens $\left\{P^i \in \mathbb{R}^{d_l}\right\}_{i=1}^{b}$, in the language branch of CLIP. The input embeddings now follow the form $\left[P^1, P^2, \cdots, P^b, W_0\right]$, where $W_0 = \left[w^1, w^2, \cdots, w^N\right]$ corresponds to fixed input tokens. New learnable tokens are further introduced in each transformer block of the language encoder ($\mathcal{L}_i$) up to a specific depth J ,

为了学习语言上下文提示，我们在 CLIP 的语言分支中引入 b 可学习的标记 $\left\{P^i \in \mathbb{R}^{d_l}\right\}_{i=1}^{b}$ 。输入嵌入现在遵循形式 $\left[P^1, P^2, \cdots, P^b, W_0\right]$ ，其中 $W_0 = \left[w^1, w^2, \cdots, w^N\right]$ 对应于固定输入标记。在语言编码器的每个变换器块中进一步引入新的可学习标记 ($\mathcal{L}_i$) ，直到特定深度 J 。

$$[, W_i] = \mathcal{L}_i([P_{i-1}, W_{i-1}]) \;\; i = 1, 2, \cdots, J. \tag{1}$$

Here $[\cdot, \cdot]$ refers to the concatenation operation. After J^{th} transformer layer, the subsequent layers process previous layer prompts and final text representation z is computed,

这里 $[\cdot, \cdot]$ 指的是连接操作。在 J^{th} 变换器层之后，后续层处理前一层的提示，并计算最终文本表示 z 。

$$[P_j, W_j] = \mathcal{L}_j\left([P_{j-1}, W_{j-1}]\right)\ j = J+1, \cdots, K, \tag{2}$$

$$z = \text{TextProj}\left(w_K^N\right)\ . \tag{3}$$

When $J = 1$, the learnable tokens P are only applied at the input of first transformer layer, and this deep language prompting technique degenerates to CoOp [49].

当 $J = 1$ 时，可学习的标记 P 仅在第一个变换器层的输入中应用，这种深度语言提示技术退化为 CoOp [49]。

3.2.2 Deep Vision Prompting

3.2.2 深度视觉提示

Similar to deep language prompting, we introduce b learnable tokens $\left\{\widetilde{P}^i \in \mathbb{R}^{d_v}\right\}_{i=1}^{b}$, in the vision branch of CLIP alongside the input image tokens. New learnable tokens are further introduced in deeper transformer layers of the image encoder(V)up to depth J .

类似于深度语言提示，我们在 CLIP 的视觉分支中引入 b 可学习的标记 $\left\{\widetilde{P}^i \in \mathbb{R}^{d_v}\right\}_{i=1}^{b}$ ，与输入图像标记一起使用。在图像编码器 (V) 的更深变换器层中进一步引入新的可学习标记，深度达到 J 。

$$[c_i, E_i, \_] = \mathcal{V}_i\left(\left[c_{i-1}, E_{i-1}, \widetilde{P}_{i-1}\right]\right)\ i = 1, 2, \cdots, J,$$

$$\left[c_j, E_j, \widetilde{P}_j\right] = \mathcal{V}_j\left(\left[c_{j-1}, E_{j-1}, \widetilde{P}_{j-1}\right]\right)\ j = J+1, \cdots, K,$$

$$x = \text{ImageProj}\left(c_K\right)\ .$$

Our deep prompting provides the flexibility to learn prompts across different feature hierarchies within the ViT architecture. We find that sharing prompts across stages is better compared to independent prompts as features are more correlated due to successive transformer block processing. Thus, the later stages do not provide independently-learned complimentary prompts as compared to the early stages.

我们的深度提示提供了在 ViT 架构中跨不同特征层次学习提示的灵活性。我们发现，与独立提示相比，在各个阶段共享提示更好，因为由于连续的变换器块处理，特征之间的相关性更高。因此，后期阶段提供的提示与早期阶段相比并不是独立学习的互补提示。

3.2.3 Vision Language Prompt Coupling

3.2.3 视觉语言提示耦合

We reason that in prompt tuning it is essential to take a multi-modal approach and simultaneously adapt both the vision and language branch of CLIP in order to achieve completeness in context optimization. A simple approach would be to naively combine deep vision and language prompting, where both the language prompts P , and the vision prompts $\widetilde{P}$, will be learned during the same training schedule. We name this design as 'Independent V-L Prompting'. Although this approach satisfies the requirement of completeness in prompting, this design lacks synergy between vision and language branch as both branches do not interact while learning the task relevant context prompts.

我们认为，在提示调优中，采取多模态方法并同时适应 CLIP 的视觉和语言分支对于实现上下文优化的完整性至关重要。一种简单的方法是天真地结合深度视觉和语言提示，其中语言提示 P 和视觉提示 $\widetilde{P}$ 将在相同的训练计划中学习。我们将这种设计称为 “独立 V-L 提示”。尽管这种方法满足了提示完整性的要求，但由于两个分支在学习任务相关上下文提示时没有相互作用，因此这种设计缺乏视觉和语言分支之间的协同作用。

To this end, we propose a branch-aware multi-modal prompting which tunes vision and language branch of CLIP together by sharing prompts across both modalities. Language prompt tokens are introduced in the language branch up to $J^{\text{ th}}$ transformer block similar to deep language prompting as illustrated in Eqs. 1-3. To ensure mutual synergy between V-L prompts, vision prompts $\widetilde{P}$, are obtained by projecting language prompts P via vision-to-language projection which we refer to as V -L coupling

function $\mathcal{F}(\cdot)$, such that $\widetilde{P}_k = \mathcal{F}_k(P_k)$. The coupling function is implemented as a linear layer which maps d_l dimensional inputs to d_v . This acts as a bridge between the two modalities, thus encouraging mutual propagation of gradients.

为此，我们提出了一种分支感知的多模态提示，通过在两个模态之间共享提示来共同调整 CLIP 的视觉和语言分支。语言提示令牌在语言分支中引入，直到 J^{th} 变换器块，类似于深度语言提示，如方程 1-3 所示。为了确保 V-L 提示之间的相互协同，视觉提示 $\widetilde{P}$ 是通过我们称之为 V 的视觉到语言投影将语言提示 P 投影而获得的，满足 $\widetilde{P}_k = \mathcal{F}_k(P_k)$ 。耦合函数实现为一个线性层，将 d_l 维输入映射到 d_v 。这充当了两个模态之间的桥梁，从而促进梯度的相互传播。

$$[c_i, E_i, \_] = \mathcal{V}_i([c_{i-1}, E_{i-1}, \mathcal{F}_{i-1}(P_{i-1})]) \, i = 1, \cdots, J$$

$$\left[c_j, E_j, \widetilde{P}_j\right] = \mathcal{V}_j\left(\left[c_{j-1}, E_{j-1}, \widetilde{P}_{j-1}\right]\right) \; j = J+1, \cdots, K$$

$$x = \text{ImageProj}(c_K)$$

Unlike independent V-L prompting, explicit conditioning of $\widetilde{P}$ on P helps learn prompts in a shared embedding space between the two branches, thus improving mutual synergy.

与独立的 V-L 提示不同，明确地将 $\widetilde{P}$ 条件化于 P 有助于在两个分支之间的共享嵌入空间中学习提示，从而改善相互协同。

4. Experiments

4. 实验

4.1. Benchmark setting

4.1. 基准设置

Generalization from Base-to-Novel Classes: We evaluate the generalizability of MaPLe, and follow a zero-shot setting where the datasets are split into base and novel classes. The model is trained only on the base classes in a few-shot setting and evaluated on base and novel categories.

从基础类到新类的泛化: 我们评估 MaPLe 的泛化能力，并遵循零-shot 设置，其中数据集被分为基础类和新类。模型仅在少量样本的设置下对基础类进行训练，并在基础类和新类上进行评估。

Cross-dataset Evaluation: To validate the potential of our approach in cross-dataset transfer, we evaluate our Ima-geNet trained model directly on other datasets. Consistent with Co-CoOp, our model is trained on all 1000 ImageNet classes in a few-shot manner.

跨数据集评估: 为了验证我们的方法在跨数据集迁移中的潜力，我们直接在其他数据集上评估我们在 ImageNet 上训练的模型。与 Co-CoOp 一致，我们的模型以少量样本的方式在所有 1000 个 ImageNet 类上进行训练。

Domain Generalization: We evaluate the robustness of our method on out-of-distribution datasets. Similar to cross-dataset evaluation, we test our ImageNet trained model directly on four other ImageNet datasets that contain various types of domain shifts.

领域泛化: 我们评估我们的方法在分布外数据集上的鲁棒性。与跨数据集评估类似，我们直接在包含各种领域转移类型的四个其他 ImageNet 数据集上测试我们在 ImageNet 上训练的模型。

Datasets: For generalization from base-to-novel classes and cross-dataset evaluation, we follow [48, 49] and evaluate the performance of our method on 11 image classification datasets which covers a wide range of recognition tasks. This includes two generic-objects datasets, Im-ageNet [4] and Caltech101 [7]; five fine-grained datasets, OxfordPets [31], StanfordCars [20], Flowers102 [30], Food101 [2], and FGVCAircraft [28]; a scene recognition dataset SUN397 [40]; an action recognition dataset UCF101 [36]; a texture dataset DTD [3] and a satellite-image dataset EuroSAT [11]. For domain generalization, we use ImageNet as source dataset and its four variants as target datasets including ImageNetV2 [35], ImageNet-Sketch [37], ImageNet-A [13] and ImageNet-R [12].

数据集: 为了从基础类到新颖类的泛化以及跨数据集评估，我们遵循 [48, 49] 并在 11 个图像分类数据集上评估我们方法的性能，这些数据集涵盖了广泛的识别任务。这包括两个通用对象数据集，ImageNet [4] 和 Caltech101 [7]；五个细粒度数据集，OxfordPets [31]、StanfordCars [20]、Flowers102 [30]、Food101 [2] 和 FGVCAircraft [28]；一个场景识别数据集 SUN397 [40]；一个动作识别数据集 UCF101 [36]；一个

纹理数据集 DTD [3] 和一个卫星图像数据集 EuroSAT [11]。对于领域泛化，我们使用 ImageNet 作为源数据集及其四个变体作为目标数据集，包括 ImageNetV2 [35]、ImageNet-Sketch [37]、ImageNet-A [13] 和 ImageNet-R [12]。

Implementation Details We use a few-shot training strategy in all experiments at 16 shots which are randomly sampled for each class. We apply prompt tuning on a pre-trained ViT-B/16 CLIP model where $d_l = 512, d_v = 768$ and $d_{vl} = 512$. For MaPLe, we set prompt depth J to 9 and the language and vision prompt lengths to 2 . All models are trained for 5 epochs with a batch-size of 4 and a learning rate of 0.0035 via SGD optimizer on a single NVIDIA A100 GPU. We report base and novel class accuracies and their harmonic mean (HM) averaged over 3 runs. We initialize the language prompts of the first layer P_0 with the pre-trained CLIP word embeddings of the template 'a photo of a <category>', while for the subsequent layers they are randomly initialized from a normal distribution. For training MaPLe on all 1000 classes of ImageNet as a source model, prompt depth J is set to 3 and the model trained for 2 epochs with learning rate of 0.0026 . Hyper-parameters for deep language prompting, deep vision prompting, and independent V-L prompting are detailed in Appendix A. The hyper-parameters are fixed across all datasets.

实施细节我们在所有实验中使用少样本训练策略，每个类别随机抽取 16 个样本。我们在预训练的 ViT-B/16 CLIP 模型上应用提示调优，其中 $d_l = 512, d_v = 768$ 和 $d_{vl} = 512$ 。对于 MaPLe，我们将提示深度 J 设置为 9，语言和视觉提示的长度设置为 2。所有模型在单个 NVIDIA A100 GPU 上训练 5 个周期，批量大小为 4，学习率为 0.0035，使用 SGD 优化器。我们报告基础类和新颖类的准确率及其调和平均数 (HM)，结果是基于 3 次运行的平均值。我们使用预训练的 CLIP 词嵌入初始化第一层的语言提示 P_0 ，模板为 “一个 < 类别 > 的照片”，而后续层则从正态分布随机初始化。为了在 ImageNet 的所有 1000 个类别上训练 MaPLe 作为源模型，提示深度 J 设置为 3，模型训练 2 个周期，学习率为 0.0026。深度语言提示、深度视觉提示和独立 V-L 提示的超参数详见附录 A。超参数在所有数据集上保持不变。

Method	Base Acc.	Novel Acc.	HM	GFLOPS
1: MaPLe shallow ($J = 1$)	80.10	73.52	76.67	167.1
2: Deep vision prompting	80.24	73.43	76.68	18.0
3: Deep language prompting	81.72	73.81	77.56	166.8
4: Independent V-L prompting	82.15	74.07	77.90	167.0
5: MaPLe (Ours)	82.28	75.14	78.55	167.0

方法	基础准确率	新颖准确率	HM	GFLOPS
1: MaPLe 浅层 ($J = 1$)	80.10	73.52	76.67	167.1
2: 深度视觉提示	80.24	73.43	76.68	18.0
3: 深度语言提示	81.72	73.81	77.56	166.8
4: 独立的视觉-语言提示	82.15	74.07	77.90	167.0
5: MaPLe(我们的)	82.28	75.14	78.55	167.0

Table 1. Comparison of MaPLe with different prompting designs in base-to-novel generalization. Results are averaged over 11 datasets. HM refers to harmonic mean.

表 1. MaPLe 与不同提示设计在基础到新颖泛化中的比较。结果是基于 11 个数据集的平均值。HM 指调和平均数。

4.2. Prompting CLIP via Vision-Language Prompts

4.2. 通过视觉-语言提示对 CLIP 进行提示

Prompting Variants: We first evaluate the performance of different possible prompting design choices as an ablation for our proposed branch-aware multi-modal prompting, MaPLe. These variants include shallow MaPLe, deep language prompting, deep vision prompting and independent V-L prompting. In Table 1, we present the results averaged over the 11 image recognition datasets. Shallow MaPLe (row-1) provides consistant improvements over CoOp and Co − CoOp in terms of generalization. Deep language prompting (row-3) shows improvements over deep vision prompting (row-2), indicating that prompts learned at the language branch provide better adaptation of CLIP. Although separately combining the above two approaches (row-4) further improves the performance, it struggles to achieve comprehensive benefits from the language and vision branches. We hypothesize that this is due to the lack of synergy between the learned vision and language prompts as they do not interact with each other during training. Meanwhile, MaPLe tied with deep prompting (row-4) combines the benefits of prompting in both branches

by enforcing interactions through explicit conditioning of vision prompts on the language prompts. It provides improvements on novel and base class accuracies which leads to the best HM of 78.55% . We explore other possible design choices and present the ablations in Appendix B.

提示变体: 我们首先评估不同可能的提示设计选择的性能，作为我们提出的基于分支的多模态提示 MaPLe 的消融实验。这些变体包括浅层 MaPLe、深层语言提示、深层视觉提示和独立的 V-L 提示。在表 1 中，我们展示了在 11 个图像识别数据集上平均得到的结果。浅层 MaPLe(第 1 行) 在泛化方面相较于 CoOp 和 Co − CoOp 提供了一致的改进。深层语言提示 (第 3 行) 相较于深层视觉提示 (第 2 行) 显示出改进，表明在语言分支学习的提示提供了更好的 CLIP 适应性。尽管将上述两种方法单独结合 (第 4 行) 进一步提高了性能，但它在语言和视觉分支之间难以实现全面的收益。我们假设这是由于学习的视觉和语言提示之间缺乏协同作用，因为它们在训练期间并未相互作用。同时，MaPLe 与深层提示 (第 4 行) 结合，通过对语言提示的显式条件化强制视觉提示之间的交互，从而结合了两个分支提示的优势。它在新颖类和基础类准确性上提供了改进，从而导致了最佳的 78.55% HM。我们探索了其他可能的设计选择，并在附录 B 中展示了消融实验。

4.3. Base-to-Novel Generalization

4.3. 基础到新颖的泛化

Generalization to Unseen Classes: Table 3 presents the performance of MaPLe in base-to-novel generalization setting on 11 recognition datasets. We compare its performance with CLIP zero-shot, and recent prompt learning works including CoOp [49] and Co-CoOp [48]. In case of CLIP, we use hand-crafted prompts that are specifically designed for each dataset.

对未见类别的泛化: 表 3 展示了 MaPLe 在 11 个识别数据集上基础到新颖泛化设置中的性能。我们将其性能与 CLIP 的零-shot 以及最近的提示学习工作进行比较，包括 CoOp [49] 和 Co-CoOp [48]。在 CLIP 的情况下，我们使用专门为每个数据集设计的手工制作提示。

In comparison with the state-of-the-art Co-CoOp, MaPLe shows improved performance on both base and novel categories on all 11 datasets with an exception of marginal reduction on only the base class performance of Caltech101. With mutual synergy from the branch-aware multi-modal prompting, MaPLe better generalizes to novel categories on all 11 datasets in comparison with Co − CoOp , and obtains an overall gain from 71.69% to 75.14% . When taking into account both the base and novel classes, MaPLe shows an absolute average gain of 2.72% over Co-CoOp.

与最先进的 Co-CoOp 相比，MaPLe 在所有 11 个数据集的基础类和新类别上表现出更好的性能，唯一的例外是 Caltech101 的基础类性能仅有微小下降。通过分支感知的多模态提示的相互协同，MaPLe 在所有 11 个数据集上相较于 Co − CoOp 更好地泛化到新类别，并从 71.69% 获得了整体增益至 75.14% 。考虑到基础类和新类别，MaPLe 相较于 Co-CoOp 显示出绝对平均增益 2.72% 。

In comparison with CLIP on novel classes, Co-CoOp improves only on 4/11 datasets dropping the average novel accuracy from 74.22% to 71.69% . MaPLe is a strong competitor which improves accuracy over CLIP on novel classes on 6/11 datasets, with an average gain from 74.22% to 75.14%.

与 CLIP 在新类别上的表现相比，Co-CoOp 仅在 4/11 数据集上有所改善，导致平均新准确率从 74.22% 降至 71.69% 。MaPLe 是一个强有力的竞争者，在 6/11 个数据集上提高了新类别的准确率，相较于 CLIP 平均增益从 74.22% 提升至 75.14%。

Generalization and Performance on Base Classes: Co- CoOp solves the poor generalization problem in CoOp by conditioning prompts on image instances and shows significant gains in novel categories. However on base classes, it improves over CoOp only on 3/11 datasets with an average drop in performance from 82.69% to 80.47% . Meanwhile, the completeness in prompting helps MaPLe improve over CoOp on base classes in 6/11 datasets maintaining the average base accuracy to around 82.28% , in addition to its improvement in generalization to novel classes.

基础类的泛化与性能:Co- CoOp 通过对图像实例进行条件提示，解决了 CoOp 中的较差泛化问题，并在新类别上显示出显著增益。然而，在基础类上，它仅在 3/11 数据集上相较于 CoOp 有所改善，性能平均下降从 82.69% 降至 80.47% 。与此同时，提示的完整性帮助 MaPLe 在 6/11 数据集上相较于 CoOp 提高了基础类的表现，保持基础准确率在约 82.28% 的水平，此外还改善了对新类别的泛化。

We find that the training strategies of Co-CoOp can be used to substantially boost the generalization performance of vanilla CoOp (6.8% gain in novel classes). We therefore compare our method with CoOp† , which trains CoOp in Co - CoOp setting (refer to Appendix A for more details).

我们发现 Co-CoOp 的训练策略可以显著提升普通 CoOp 的泛化性能 (新类别增益 6.8%)。因此，我们将我们的方法与 CoOp† 进行比较，后者在 Co - CoOp 设置中训练 CoOp (更多细节请参见附录 A)。

	Base	Novel	HM
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
CoOp+	80.85	70.02	75.04
MaPLe	82.28	75.14	78.55

	基础	新颖	HM
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
CoOp+	80.85	70.02	75.04
MaPLe	82.28	75.14	78.55

Table 2. Generalization comparison of MaPLe with CoOp † .

表 2. MaPLe 与 CoOp 的泛化比较 † 。

Compare to CoOp† , the vanilla CoOp model seems to overfit on base classes. When compared to CoOp† which attains an average base accuracy of 80.85% , MaPLe shows an improvement of 1.43% with the average base accuracy of 82.28% (Table 2).

与 CoOp† 相比，基础模型 CoOp 似乎在基础类别上过拟合。当与 CoOp† 相比，该模型的平均基础准确率为 80.85% ，MaPLe 显示出 1.43% 的改进，平均基础准确率为 82.28%(表 2)。

4.4. Cross-Dataset Evaluation

4.4. 跨数据集评估

We test the cross-dataset generalization ability of MaPLe by learning multi-modal prompts on all the 1000 ImageNet classes and then transferring it directly on the remaining 10 datasets. Table 4 shows the performance comparison between MaPLe, CoOp and Co-CoOp. On the ImageNet source dataset, MaPLe achieves performance comparable (k) EuroSAT

我们通过在所有 1000 个 ImageNet 类别上学习多模态提示，然后直接在剩余的 10 个数据集上进行迁移，来测试 MaPLe 的跨数据集泛化能力。表 4 显示了 MaPLe、CoOp 和 Co-CoOp 之间的性能比较。在 ImageNet 源数据集上，MaPLe 的性能可与 (k) EuroSAT 相媲美。

(a) Average over 11 datasets

(a) 在 11 个数据集上的平均值

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
	+1.81	+3.45	+2.72

	基础	新颖	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
	+1.81	+3.45	+2.72

	(b) ImageNet.		
	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
Co-CoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
	+0.68	+0.11	+0.37

	(b) ImageNet.		
	基础	新颖	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
Co-CoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
	+0.68	+0.11	+0.37

(c) Caltech101			
	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
Co-CoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
	-0.22	+0.55	+0.18

(c) Caltech101			
	基础	新颖	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
Co-CoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
	-0.22	+0.55	+0.18

(d) OxfordPets
(d) 牛津宠物

	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
Co-CoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
	+0.23	+0.07	+0.15

	基础	新颖	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
Co-CoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
	+0.23	+0.07	+0.15

(e) StanfordCars
(e) 斯坦福汽车

	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
Co-CoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
	+2.45	+0.41	+1.46

	基础	新颖	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
Co-CoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
	+2.45	+0.41	+1.46

(f) Flowers102
(f) 花卉 102

	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
Co-CoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
	+1.05	+0.71	+0.85

	基础	新颖	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
Co-CoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
	+1.05	+0.71	+0.85

(g) Food101
(g) 食物 101

	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
Co-CoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
	+0.01	+0.76	+0.39

	基础	新颖	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
Co-CoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
	+0.01	+0.76	+0.39

(h) FGVCAircraft
(h) FGVCAircraft

	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
Co-CoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
	+4.03	+11.90	+8.76

	基础	新颖	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
Co-CoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
	+4.03	+11.90	+8.76

(i) SUN397
(i) SUN397

	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
Co-CoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
	+1.08	+1.84	+1.48

	基础	新颖	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
Co-CoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
	+1.08	+1.84	+1.48

(j) DTD

(j) DTD

	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
Co-CoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
	+3.35	+3.18	+3.31

	基础	新颖	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
Co-CoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
	+3.35	+3.18	+3.31

(l) UCF101

(l) UCF101

	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
Co-CoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
	+0.67	+5.21	+3.13

	基础	新颖	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
Co-CoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
	+0.67	+5.21	+3.13

	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
Co-CoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
	+6.58	+13.19	+11.14

	基础	新颖	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
Co-CoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
	+6.58	+13.19	+11.14

Table 3. Comparison with state-of-the-art methods on base-to-novel generalization. MaPLe learns multi-modal prompts and demonstrates strong generalization results over existing methods on 11 recognition datasets. Absolute gains over Co-CoOp are indicated in blue.

表 3. 与最先进方法在基础到新颖泛化方面的比较。MaPLe 学习多模态提示，并在 11 个识别数据集上展示出强大的泛化结果。相对于 Co-CoOp 的绝对增益以蓝色标示。

	Source	Target									
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
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
Co-CoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
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**

Table 4. Comparison of MaPLe with existing approaches on cross-dataset evaluation. Overall, MaPLe achieves competitive performance providing highest average accuracy, indicating better generalization. to competing approaches but demonstrates a much stronger generalization performance by surpassing CoOp in 9/10 and Co-CoOp in 8/10 datasets. Overall, MaPLe shows competitive performance leading to the highest averaged accuracy of 66.30% . This suggests that the use of branch-aware V-L prompting in MaPLe facilitates better generalization.

表 4. MaPLe 与现有方法在跨数据集评估中的比较。总体而言，MaPLe 实现了竞争性的表现，提供了最高的平均准确率，表明其泛化能力优于竞争方法，但在 CoOp 和 Co-CoOp 的 8/10 个数据集中表现出更强的泛化性能。总体而言，MaPLe 显示出竞争力表现，导致最高的平均准确率为 66.30% 。这表明在 MaPLe 中使用分支感知的 V-L 提示有助于更好的泛化。

4.5. Domain Generalization

4.5. 域泛化

We show that MaPLe generalizes favourably on out-of-distribution datasets as compared to CoOp and Co − CoOp . We evaluate the direct transferability of ImageNet trained model to various out-of-domain datasets, and observe that it consistently improves against all the existing approaches as indicated in Table 5. This indicates that utilizing multimodal branch-aware prompting helps MaPLe in enhancing the generalization and robustness of V-L models like CLIP.

我们展示了 MaPLe 在分布外数据集上的良好泛化，相较于 CoOp 和 Co − CoOp 。我们评估了在各种域外数据集上直接转移 ImageNet 训练模型的能力，并观察到它在所有现有方法中始终有所改善，如表 5 所示。这表明利用多模态分支感知提示有助于 MaPLe 提升 V-L 模型 (如 CLIP) 的泛化和鲁棒性。

	Source	Target			
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	71.51	64.20	47.99	49.71	75.21
Co-CoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	49.15	50.90	76.98

	源	目标			
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	71.51	64.20	47.99	49.71	75.21
Co-CoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	49.15	50.90	76.98

Table 5. Comparison of MaPLe with existing approaches in domain generalization setting. MaPLe shows consistant improvements on all target datasets.

表 5. MaPLe 与现有方法在域泛化设置中的比较。MaPLe 在所有目标数据集上显示出一致的改善。

4.6. Ablation Experiments

4.6. 消融实验

Prompt Depth: In Fig. 4 (left), we illustrate the effect of prompt depth J for MaPLe and ablate on the depth of language and vision branch individually. In general, the performance improves as prompt depth

increases. We note that performance sensitivity increases when randomly initialized prompts are inserted in the deeper layers of a frozen model where the model feature space is already mature. Similar trend is also reported by [16]. As earlier methods utilize shallow language prompting ($J = 1$) , we compare our method with deep language prompting. Overall, MaPLe achieves better performance than deep language prompting and achieves maximum performance at a depth of 9 .

提示深度: 在图 4(左侧)，我们展示了提示深度 J 对 MaPLe 和 ablate 在语言和视觉分支深度上的影响。一般来说，随着提示深度的增加，性能会有所提升。我们注意到，当随机初始化的提示被插入到冻结模型的更深层时，性能的敏感性会增加，因为此时模型特征空间已经成熟。类似的趋势也在 [16] 中被报告。由于早期方法使用了浅层语言提示 ($J = 1$) ，我们将我们的方法与深层语言提示进行了比较。总体而言，MaPLe 的性能优于深层语言提示，并在深度为 9 时达到了最佳性能。

Prompt Length: Fig. 4 (right) shows the effect of prompt length for MaPLe. As the prompt length increases, the performance on base classes is generally maintained, while the novel class accuracy decreases. This indicates over-fitting which inherently hurts the generalization to novel classes.

提示长度: 图 4(右侧) 显示了 MaPLe 的提示长度的影响。随着提示长度的增加，基础类别的性能通常保持不变，而新类别的准确性则下降。这表明过拟合，这本质上会损害对新类别的泛化能力。

Effectiveness of Multi-modal Prompting: Fig. 5 shows the analysis of per class accuracy for selected datasets in the order of increasing domain shift. It indicates that the performance gains of MaPLe in comparison to Co-CoOp varies across different datasets. MaPLe provides significant gains over Co-CoOp for datasets that have large distribution shifts from the pretraining dataset of CLIP, and vision concepts that are usually rare and less generic. Further detailed analysis is provided in Appendix C.

多模态提示的有效性: 图 5 显示了选定数据集的每类准确性的分析，按领域转移的增加顺序排列。它表明，与 Co-CoOp 相比，MaPLe 的性能提升在不同数据集之间有所不同。对于与 CLIP 预训练数据集存在较大分布转移的数据集，以及通常较少且不太通用的视觉概念，MaPLe 提供了显著的提升。更多详细分析见附录 C。

Prompting complexity: Table 6 shows the computational complexity of MaPLe in comparison with other approaches.

提示复杂性: 表 6 显示了 MaPLe 与其他方法相比的计算复杂性。

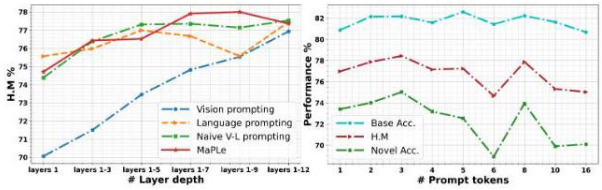
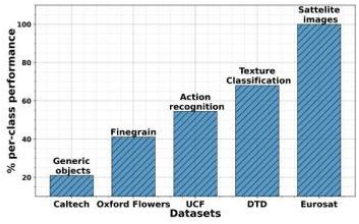


Figure 4. Ablation on prompt depth (left) and prompt length (right) in MaPLe. We report average results on the held-out validation sets of all datasets.

图 4. MaPLe 中提示深度 (左侧) 和提示长度 (右侧) 的消融实验。我们报告了所有数据集的保留验证集上的平均结果。

Figure 5. Percentage classes where MaPLe shows improved performance over Co-CoOp, which increases as dataset domain shift from generic categories increases (→) .

图 5. MaPLe 在 Co-CoOp 上表现出改进性能的类别百分比，随着数据集领域转移从通用类别增加而增加 (→) 。



Although MaPLe utilizes multi-modal prompts, its overall FLOPS (Floating Point Operations) exceeds only by 0.1% over CoOp and Co−CoOp . The independent V-L prompting also provides comparable FLOP count. In terms of inference speed, Co − CoOp is significantly slower and the FPS (Frames Per Second) remains constant as the batch size increases. In contrast, MaPLe has no such overhead and provides much better inference and training speeds. Further, MaPLe provides better convergence as it

requires only half training epochs as compared to Co-CoOp (5 vs 10 epochs). MaPLe adds about 2.85% training parameters on top of CLIP. To study if the performance gain is mainly attributed to more parameters, we experiment with MaPLe†, which uses a unified V-L coupling function for all layer prompts. MaPLe† with about 9x lesser parameters than MaPLe also improves over existing methods. We also ablate by comparing MaPLe with heavier CoCoOp in Appendix D.

尽管 MaPLe 利用多模态提示，但其整体 FLOPS(浮点运算) 仅比 0.1% 超过 CoOp 和 Co − CoOp 。独立的 V-L 提示也提供了可比的 FLOP 计数。在推理速度方面，Co − CoOp 明显较慢，且随着批量大小的增加，FPS(每秒帧数) 保持不变。相比之下，MaPLe 没有这样的开销，并且提供了更好的推理和训练速度。此外，MaPLe 提供了更好的收敛性，因为与 Co-CoOp(5 轮与 10 轮) 相比，它只需要一半的训练轮数。MaPLe 在 CLIP 的基础上增加了大约 2.85% 的训练参数。为了研究性能提升是否主要归因于更多的参数，我们实验了 MaPLe†，它对所有层提示使用统一的 V-L 结合函数。MaPLe† 的参数比 MaPLe 少约 9x ，也优于现有方法。我们还通过在附录 D 中将 MaPLe 与更重的 CoCoOp 进行比较来进行消融实验。

Method	Params	Params % CLIP	FPS (with BS)			HM
			1	4	100	
CoOp	2048	0.002	13.8	55.3	1353.0	71.66
CoCoOp	35360	0.03	64.6	114.7	15.1	75.83
Independent V-L	31488	0.02	62.5	239.4	1383.8	77.90
MaPLe	3.55 M	2.85	60.2	239.0	1365.1	78.55
MaPLe*	0.41 M	0.33	60.2	238.0	1365.0	78.11

方法	参数	参数 % CLIP	FPS(带 BS)			HM
			1	4	100	
CoOp	2048	0.002	13.8	55.3	1353.0	71.66
CoCoOp	35360	0.03	64.6	114.7	15.1	75.83
独立 V-L	31488	0.02	62.5	239.4	1383.8	77.90
MaPLe	3.55 M	2.85	60.2	239.0	1365.1	78.55
MaPLe*	0.41 M	0.33	60.2	238.0	1365.0	78.11

Table 6. Comparison of computational complexity among different prompting methods. MaPLe† is a MaPLe version which utilizes a common V-L coupling function for all layers.

表 6. 不同提示方法之间计算复杂度的比较。MaPLe† 是一个 MaPLe 版本，它对所有层使用共同的 V-L 结合函数。

5. Conclusion

5. 结论

Adaptation of large-scale V-L models, e.g., CLIP [32] to downstream tasks is a challenging problem due to the large number of tunable parameters and limited size of downstream datasets. Prompt learning is an efficient and scalable technique to tailor V-L models to novel downstream tasks. To this end, the current prompt learning approaches either consider only the vision or language side prompting. Our work shows that it is critical to perform prompting for both vision and language branches to appropriately adapt V-L models to downstream tasks. Further, we propose a strategy to ensure synergy between vision-language modalities by explicitly conditioning the vision prompts on textual prompt across different transformer stages. Our approach improves the generalization towards novel categories, cross-dataset transfer and datasets with domain shifts.

大规模视觉-语言模型 (V-L 模型) 的适应，例如 CLIP [32]，在下游任务中是一个具有挑战性的问题，因为可调参数数量庞大且下游数据集的规模有限。提示学习是一种高效且可扩展的技术，可以将 V-L 模型调整到新的下游任务。为此，目前的提示学习方法要么仅考虑视觉方面的提示，要么仅考虑语言方面的提示。我们的工作表明，同时对视觉和语言分支进行提示对于适当地调整 V-L 模型以适应下游任务至关重要。此外，我们提出了一种策略，通过在不同的变换器阶段显式地将视觉提示与文本提示进行条件关联，以确保视觉-语言模态之间的协同作用。我们的方法改善了对新类别的泛化能力、跨数据集迁移以及具有领域转移的数据集的适应能力。

References

参考文献

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274, 2022. 3, 11, 12

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101-mining discriminative components with random forests. In The European Conference on Computer Vision, pages 446-461. Springer, 2014. 5

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3606-3613, 2014. 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 248-255. Ieee, 2009. 5

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11583-11592, 2022. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021. 3

[7] 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 2004 conference on computer vision and pattern recognition workshop, pages 178-178. IEEE, 2004. 5

[8] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-det: Towards open-vocabulary detection using uncurated images. In The European Conference on Computer Vision, 2022. 2

[9] 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. arXiv preprint arXiv:2110.04544, 2021. 2, 3

[10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921, 2021.2

[11] 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. 5

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kada-vath, 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8340-8349, 2021. 5

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Stein-hardt, and Dawn Song. Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15262-15271, 2021. 5

[14] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649, 2022. 1

[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 International Conference on Machine Learning, pages 4904-4916. PMLR, 2021.2

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In The European Conference on Computer Vision, 2022. 2, 3, 8

[17] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. arXiv preprint arXiv:2110.08484, 2021.

[18] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In The European Conference on Computer Vision, 2021. 3

[19] Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. How to adapt your large-scale vision-and-language model, 2022. 2

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3 d object representations for fine-grained categorization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 554-561, 2013. 5

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Conference on Empirical Methods in Natural Language Processing, 2021.2

[22] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In International Conference on Learning Representations, 2022. 2

[23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 2

[24] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.2

[25] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5206-5215, 2022. 1, 3

[26] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7086-7096, 2022. 2

[27] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fa-had Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multimodal transformer. In The European Conference on Computer Vision. Springer, 2022. 2

[28] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151, 2013. 5

[29] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. 2022. 1

[30] 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. 5

[31] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In 2012 IEEE conference on computer vision and pattern recognition, pages 3498-3505. IEEE, 2012. 5

[32] 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 International Conference on Machine Learning, pages 8748-8763. PMLR, 2021. 1, 2, 8

[33] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18082-18091, 2022. 2

[34] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In Advances in Neural Information Processing Systems, 2022. 2

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im-agenet? In International Conference on Machine Learning, pages 5389-5400. PMLR, 2019. 5

[36] 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. 5

[37] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In Advances in Neural Information Processing Systems, volume 32, 2019. 5

[38] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary

prompting for rehearsal-free continual learning. In The European Conference on Computer Vision, 2022. 2

[39] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 139-149, 2022.2

[40] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE computer society conference on computer vision and pattern recognition, pages 3485-3492. IEEE, 2010. 5

[41] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783, 2021. 2

[42] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797, 2021. 1

[43] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021. 2

[44] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In The European Conference on Computer Vision, 2022.2

[45] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18123-18133, 2022. 2

[46] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In The European Conference on Computer Vision, 2022. 2

[47] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. arXiv preprint arXiv:2206.04673, 2022. 2

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816-16825, 2022. 1, 2, 3, 5, 6

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, pages 1-12, 2022. 1, 2,3,5,6

[50] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In The European Conference on Computer Vision, 2022. 2

[51] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. arXiv preprint arXiv:2205.14865, 2022. 2