Improving Zero-shot Generalization of Learned Prompts via Unsupervised Knowledge Distillation

*Marco Mistretta[1], *Alberto Baldrati[1,2], Marco Bertini[1], and Andrew D. Bagdanov[1]

[1] University of Florence - Media Integration and Communication Center (MICC)
[2] University of Pisa
Florence, Italy - Pisa, Italy
{name.surname}@unifi.it

Abstract. Vision-Language Models (VLMs) demonstrate remarkable zero-shot generalization to unseen tasks, but fall short of the performance of supervised methods in generalizing to downstream tasks with limited data. Prompt learning is emerging as a parameter-efficient method for adapting VLMs, but state-of-the-art approaches require annotated samples. In this paper we propose a novel approach to prompt learning based on unsupervised knowledge distillation from more powerful models. Our approach, which we call Knowledge Distillation Prompt Learning (KDPL), can be integrated into existing prompt learning techniques and eliminates the need for labeled examples during adaptation. Our experiments on more than ten standard benchmark datasets demonstrate that KDPL is very effective at improving generalization of learned prompts for zero-shot domain generalization, zero-shot cross-dataset generalization, and zero-shot base-to-novel class generalization problems. KDPL requires no ground-truth labels for adaptation, and moreover we show that even in the absence of *any* knowledge of training class names (CA-KDPL) can be used to effectively transfer knowledge. The code is publicly available at https://github.com/miccunifi/KDPL.

Keywords: Prompt Learning · Unsupervised Knowledge Distillation · Vision-Language Models · Few-Shot Learning · Zero-shot Transfer.

1 Introduction

Vision-Language Models (VLMs) like Contrastive Language-Image Pre-training (CLIP) are remarkably effective at zero-shot generalization to downstream tasks [3, 42, 44, 65]. These models leverage a dual encoder architecture and are trained to align image and text features in a shared embedding space. CLIP and similar models are able to perform zero-shot classification by predicting the output class based on the similarity between the test image embedding and text embeddings of words from a fixed vocabulary.

* These authors contributed equally to this work.

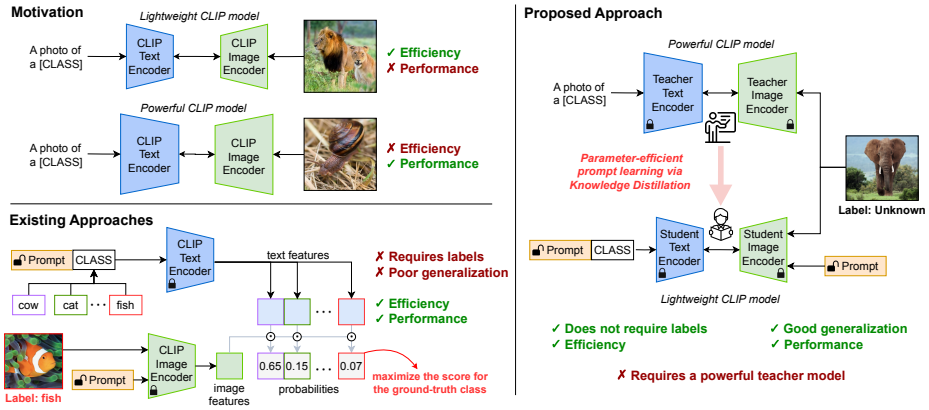


Fig. 1: Motivation and overview. **(Top left)** Lightweight VLMs like CLIP achieve impressive zero-shot performance but lag behind supervised approaches; large VLMs incur a high computational burden. **(Bottom left)** Parameter-efficient prompt learning offers a non-destructive approach to adapting VLMs to downstream tasks; however, existing methods require annotated samples and struggle to generalize to unseen classes. **(Right)** Our approach does not require labeled samples and learns by distilling knowledge from a more powerful VLM. It can be seamlessly integrated into existing prompt learning techniques and generalizes better to unseen classes on downstream tasks.

While CLIP exhibits remarkable zero-shot performance, it can fall short of supervised adaptation. Narrowing this gap through fine-tuning is challenging when large training datasets are not available for downstream tasks. Moreover, fine-tuning can also lead to overfitting and is destructive in that it can cause the model to forget knowledge acquired during large-scale pre-training [14, 67].

An intuitive way to achieve better performance is to use larger and more powerful VLMs [10] with better zero-shot generalization. However, though performance improves with increasing size, their practical applicability in real-world scenarios decreases. Lightweight CLIP models based on ResNet-50 and ViT-B/32 require about 18 and 14 GFLOPs for inference, respectively. In contrast, more powerful models based on ViT-H/14 [10, 18] require nearly 30× the computation at 381 GFLOPs [18] for inference.

To address these challenges, parameter-efficient prompt learning is emerging as a promising and non-destructive approach to adapting VLMs to downstream tasks [34, 35, 64]. Zhou et al. [67] proposed a text-only prompt learning approach called CoOp to overcome the limitations of manually crafted prompts. Subsequently, Jia et al. [29] extended prompt learning to the visual domain with Visual Prompt Tuning (VPT), and Khattak et al. [30, 31] proposed a multi-modal prompt learning approach called MaPLe [30] and a self-regulated approach called PromptSRC [31]. Although significantly improving over carefully tuned, handcrafted prompts, these state-of-the-art techniques all require annotated samples from the dataset and have problems generalizing to other datasets or classes unseen during training [30, 66].

To eliminate the need for labeled training examples and improve the generalization of learned prompts, we propose a novel approach to prompt learning which we call Knowledge Distillation Prompt Learning (KDPL). KDPL adapts lightweight VLMs and improves performance on downstream tasks by distilling knowledge from a more powerful VLM without the need for annotated examples. KDPL is a flexible method that can be seamlessly integrated into existing prompt learning framework such as CoOp [67], CoCoOp [66], VPT [29], MaPLe [30], and PromptSRC [31]. Figure 1 summarizes the motivations and illustrates the workflow of the proposed approach. We validate KDPL on two increasingly challenging scenarios: in the first, we assume knowledge of training class names, while in the second we assume no knowledge of training classes. Importantly, in both scenarios *no annotated training samples are used.*

The contributions of this work are:

- we propose a novel parameter-efficient prompt-learning approach that eliminates the need for labeled training examples by distilling knowledge from a large Vision-Language Model;
- we show that our approach can be integrated into existing prompt learning techniques to learn visual, textual, and multimodal prompts which generalize significantly better than state-of-the-art baselines to unseen classes in downstream tasks;
- we demonstrate the effectiveness of our approach through extensive experiments on more than ten standard benchmark datasets covering a broad range of downstream problems, including domain generalization, cross-dataset adaptation, and base-to-novel generalization; and
- we introduce a new, class agnostic evaluation setting in which training class names are unknown and show the superiority of our approach in this scenario.

To the best of our knowledge ours is the first approach to parameter-efficient VLM adaptation which is applicable in scenarios that are *label agnostic* (i.e. no label supervision is required during adaptation) and also in scenarios that are additionally *class agnostic* (i.e. no knowledge of training class names is assumed).

2 Related Work

In this section we review the recent literature most relevant to our contribution.

Vision-Language Models. Vision-Language Models (VLMs) learn robust visual representations thanks to their extensive pre-training on image-caption datasets [47, 48]. These representations are very effective at generalizing to a variety of downstream tasks [13, 28, 44]. In contrast to vision models trained only with image supervision, vision-language models can interpret both visual and textual data, showing improvement in tasks requiring cross-model understanding [3, 4, 52]. We focus on contrastively-trained VLMs such as CLIP [44], ALIGN [28], and LiT [60]. These models use a dual encoder architecture and contrastive learning to align images and text in a common embedding space. Thanks to this shared embedding space and massive pre-training datasets these

models achieve remarkable performance. However, there is still a gap between the zero-shot capabilities of these VLMs and the performance of tailored state-of-the-art models [15,44], and thus there is active research on efficient and effective methods to adapt VLMs to downstream tasks [21,38,55,61].

Prompt Learning. The performance of VLMs is highly dependent on textual prompts used to characterize downstream tasks. Given the significant impact of small changes in wording on performance, the manual generation of optimal textual prompts is a challenging task. Inspired by developments in NLP [34,35, 64], CoOp [67] proposes to learn continuous vectors in the word embedding space instead of using tailored hand-crafted prompts.

The promising results of CoOp have attracted considerable interest in prompt learning for VLMs [1, 2, 7, 8, 30, 31, 37, 49, 50, 57, 68]. CoCoOp, in particular, highlights the poor performance of CoOp on unseen classes and uses image-conditional dynamic prompts to improve generalization [66]. PLOT [8] instead is based on learning multiple prompts by exploiting local visual features. KgCoOp [57] adds a regularization loss to minimize the discrepancy between learned and hand-crafted prompts, while VPT [29] adapts prompt learning to the visual domain by learning continuous vectors in the input space of a ViT [16]. Finally, MaPLe [30] introduces a multimodal prompt learning strategy for VLMs. PromptSRC [31] introduces a self-regularization framework for prompting, addressing overfitting through agreement with a frozen model, a self-ensemble of prompts, and textual diversity.

Similar to the approach we propose in this paper, UPL [27] performs prompt learning without relying on annotated samples. It leverages pseudolabels derived from a larger CLIP model and selects the top-K most confident samples per class to construct a balanced set of pseudo-labels. UPL requires a substantial collection of unlabeled images and is not directly applicable when only a few unlabeled samples are available, like in the few-shot adaptation scenarios considered here. In contrast to UPL, KDPL does not use pseudolabeling; instead, we directly learn from the logits of a CLIP teacher model via knowledge distillation, thus eliminating the need for any selection strategy of the training samples.

Knowledge Distillation. Knowledge distillation (KD) is a machine learning technique in which a simple model (*student*) is trained to mimic the behavior of a larger model (*teacher*) by learning from its output or intermediate activation [26]. This technique has found success in many contexts, including image classification [5,9,26], self-supervised learning [19,56], and image/video segmentation [17, 22, 40, 51], leading to improvements in model compression, computational efficiency, and performance. DML [62] introduces a mutual learning approach to train students and teachers simultaneously. DKD [63] reformulates the classical KD loss into a term related to the target class and a term related to the non-target classes. In contrast to the majority of knowledge distillation approaches [58,59,62,63], we do not utilize any labels during training. Instead, we solely employ the distillation loss. In our work, we apply knowledge distillation in a parameter-efficient prompt learning scenario. Specifically, we distill

knowledge from a large and powerful VLM (teacher) into a lightweight VLM (student) by updating only a reduced number of parameters (the prompt).

Techniques Using Teacher-Student Distillation. Recently Li et al. [36] introduced PromptKD which also uses teacher-student distillation in prompt learning. Although it too leverages a larger teacher to guide a smaller student, it differs from our contribution in several key aspects. PromptKD uses PromptSRC [31] to pre-train the teacher model with labeled examples, whereas KDPL does not pre-train the teacher and requires no labeled examples. PromptKD uses an MLP to project the student image encoder output, which adds extra parameters. KDPL is instead a general framework with no additional parameters, adaptable to any prompt-tuning technique to render them completely unsupervised. Moreover, PromptKD uses the entire training set, including both base and novel images, for distillation. In contrast, KDPL follows the standard 16-shot setting.

3 Knowledge Distillation Prompt Learning (KDPL)

We first introduce preliminary concepts related to prompt learning and then describe our approach to applying knowledge distillation to the problem.

3.1 Preliminaries

Our approach is based on knowledge distillation applied to prompt learning. Here we discuss the base teacher and student models (CLIPs) and five state-of-the-art prompt learning approaches into which we will incorporate KDPL.

CLIP. Contrastive Language-Image Pre-training (CLIP) is a vision-language model trained to align images and textual captions in shared semantic space [44]. CLIP consists of an image encoder f_θ and a text encoder g_ϕ. Given an image I, the image encoder computes its feature representation $f_\theta(I) \in \mathbb{R}^d$, where d is the size of the semantic embedding space. Similarly, for a given textual caption Y, a word embedding layer E_L maps each tokenized word to the token embedding space $\mathcal{W}$. Then, the text encoder g_ϕ generates the textual feature representation $g_\phi(E_L(Y)) \in \mathbb{R}^d$. The main goal of CLIP training is to learn θ and ϕ such that $f_\theta(I) \approx g_\phi(E_L(Y))$ for associated image/text pairs (I, Y).

When using a Vision Transformer (ViT) [16] as the visual encoder f_θ, the encoding process begins by splitting the image into U fixed-size patches. These patches are then projected into patch embeddings $\{w_1, w_2, \ldots, w_U\}$, where each w_i belongs to the patch embedding space $\mathcal{V}$. A learnable class (CLS) token c_i is concatenated with the patch embeddings, resulting in the input to the vision transformer being $\{c_i, w_1, w_2, \ldots, w_U\}$. Finally, the CLS token of the final transformer layer is projected to the shared embedding space via a linear projection to obtain the final representation.

To perform zero-shot classification using CLIP, we start with an image I and build a set of textual prompts $\{Y_i\}_{i=1}^C$, where C denotes the number of classes.

Each handcrafted text prompt Y_i takes the format "*a photo of a [CLASS$_i$]*", where CLASS$_i$ represents a specific class name, such as *airplane*, *bird*, etc.

Then the feature representations $\psi_I = f_\theta(I)$ and $\psi_T^i = g_\phi(E_L(Y_i))$ are extracted using the CLIP encoders. The predicted probability for each class is:

$$p(y=i|I) = \frac{\exp(\cos(\psi_T^i, \psi_I)/\tau)}{\sum_{j=1}^{C} \exp(\cos(\psi_T^j, \psi_I)/\tau)}, \tag{1}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity and τ is a temperature hyperparameter.

Prompt-Learning Techniques. Here we summarize how the state-of-the-art techniques for textual, visual, and multimodal prompt learning work.

- **CoOp** [67] is a textual prompt learning technique that learns continuous context tokens (i.e. the learnable prompt) in the CLIP token embedding space. Specifically, CoOp introduces M learnable vectors, $\{v_1, v_2, \ldots, v_M\}$ where each context vector $v_i \in \mathcal{W}$. For each of the k classes of a dataset, the input to the text encoder is $\{v_1, v_2, \ldots, v_M, c_k\}$, where $c_k = E_L([CLASS_k])$.
- **CoCoOp** [66] extends CoOp by incorporating a lightweight network h_θ. Each context token is obtained as $v_i(I) = v_i + \pi$, where $\pi = h_\theta(I)$. This method ensures that the textual prompts are conditioned on the input image.
- **VPT** [29] is a visual prompt learning method that can be viewed as the counterpart of CoOp in the visual domain. Unlike CoOp, which operates entirely in the textual token embedding space $\mathcal{W}$, VPT performs prompt learning in the visual patch embedding space $\mathcal{V}$. Specifically, VPT learns P visual tokens, leading to the input to the ViT being $\{z_1, \ldots, z_P, c_i, w_1, w_2, \ldots, w_U\}$. VPT offers two prompting variants: deep and shallow. The deep variant learns a distinct prompt for each ViT layer, while the shallow variant incorporates the prompt parameters only into the input of the first layer.
- **MaPLe** [30] is a deep multi-modal prompt learning technique that promotes strong coupling between vision and language prompts. In practice, MaPLe learns different textual context tokens for each layer in g_ϕ. The visual prompts, on the other hand, are not directly learned but are obtained through a linear mapping from the textual ones.
- **PromptSRC** [31] is another multi-modal prompt learning technique that optimizes prompts through mutual agreement maximization between prompted and frozen model features using self-distillation, Gaussian-weighted prompt aggregation over the training session, and promoting textual diversity to address sample diversity imbalance.

Note that methods involving visual prompt learnable tokens like VPT [29], MaPLe [30], and PromptSRC [31] can only be applied to VLMs equipped with a ViT-based image encoder.

3.2 Label Agnostic Prompt Learning

The methods described above all rely on ground-truth labels during adaptation. Here we show how unsupervised knowledge distillation can be used instead to replace the need for annotated training examples.

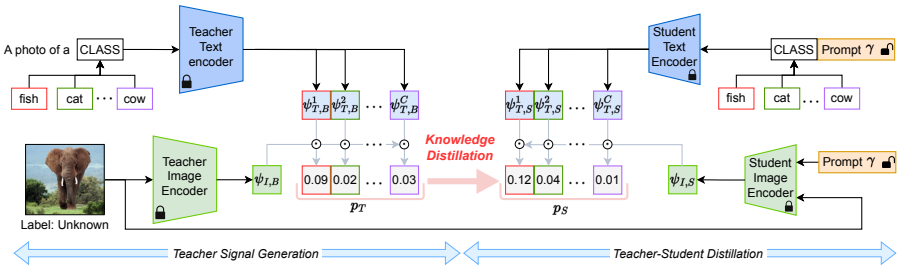


Fig. 2: Knowledge Distillation Prompt Learning (KDPL) overview. Given a lightweight VLM *student* and a larger, more powerful VLM *teacher*, KDPL updates the student prompt parameters by distilling knowledge from the teacher. KDPL first performs zero-shot classification with the teacher to obtain teacher probabilities p_T. It then computes the student probabilities p_S and performs knowledge distillation to update the student prompt parameters γ.

Overview. Our proposed approach, which we call Knowledge Distillation Prompt Learning (KDPL), is a general method designed to enhance the performance of the CLIP model on downstream tasks through parameter-efficient prompt learning. Unlike previous approaches [30, 66, 67], which rely on labeled examples for training, KDPL eliminates the need for manually-labeled samples by learning only through knowledge distillation from a larger and more powerful VLM. Note that KDPL is a method that can be seamlessly integrated with any existing prompt learning approach in scenarios where no information about class names or labels is available.

We validate KDPL in two progressively challenging supervision regimes. In the **label agnostic** scenario we do not use ground-truth labels, but we assume knowledge of the class names in the training dataset. In the **class agnostic** scenario (see Section 3.3) we go one step further and assume that even the training class names are unknown. For this class agnostic scenario we propose an effective and efficient online strategy for automatically filtering the classes from a large dictionary of approximately 20K class names [33].

Prompt Learning via Unsupervised Knowledge Distillation (KDPL). Given a lightweight CLIP model (the *student*) and a larger, more powerful CLIP model (the *teacher*), we aim to improve the downstream performance of the student model by distilling knowledge from teacher to student. For an image I and a set of classes $\mathcal{C} = \{\text{CLASS}_i\}_{i=1}^{C}$, we start by performing zero-shot classification using the frozen teacher model. Specifically, we use the teacher image encoder f_θ^B and text encoder g_ϕ^B to compute the teacher image features $\psi_{I,B} = f_\theta^B(I)$ and text features $\psi_{T,B}^i = g_\phi^S(E_L([\text{CLASS}_i]))$. For the teacher model we use the fixed text prompt "a photo of [CLASS]". We then apply Eq. (1) to produce the probabilities $p_T(I, \mathcal{C})$ predicted by the teacher on image I for classes $\mathcal{C}$.

The teacher model does not rely on a learnable prompt and its predictions remain fixed during training. Our aim is to learn text and image prompts for the student model that enhance its generalization to downstream tasks.

We denote with f_θ^S and g_ϕ^S the student image and text encoders, respectively, and with γ the parameters associated with the learnable student prompts (see Fig. 2). Given the same image I processed by the teacher and the same set of classes $\mathcal{C}$, the student extracts image features $\psi_{I,S} = f_{\theta,\gamma}^S(I)$ and text features $\psi_{T,S}^i = g_{\phi,\gamma}^S(E_L([\text{CLASS}_i]))$. Note that the text and image encoders can both depend on the prompt parameters γ. According to the prompt learning technique used, the γ parameters can be used only by the text encoder (CoOp, CoCoOp), the visual encoder (VPT), or both (MaPLe, PromptSRC). Finally, using Eq. (1) we produce student class probabilities $p_S(I, \mathcal{C})$ predicted on image I for classes $\mathcal{C}$. Note that all encoder parameters except for the learnable prompt γ are frozen.

We use the symmetric KL-divergence between the teacher ($p_T(I, \mathcal{C})$) and the student ($p_S(I, \mathcal{C})$) probabilities in a distillation loss:

$$\mathcal{L}(\mathcal{I}, \mathcal{C}) = \left[D_{\text{KL}}(p_T(I, \mathcal{C}) \,||\, p_S(I, \mathcal{C})) + D_{\text{KL}}(p_S(I, \mathcal{C}) \,||\, p_T(I, \mathcal{C}))\right], \tag{2}$$

where $D(q \,||\, p)$ is the asymmetric KL-divergence between the discrete probability distributions p and q:

$$D_{\text{KL}}(p \,||\, q) = \sum_{i \in \mathcal{C}} p_i \log\left(\frac{p_i}{q_i}\right) \tag{3}$$

This distillation loss depends only on the fixed predictions of the teacher, the prompt-conditioned predictions of the students, and the set of classes $\mathcal{C}$.

Importantly, *our distillation loss does not assume any knowledge of class labels for image I*. Nor does it require that $\mathcal{C}$ be the classes of the downstream task – that is, KDPL can be used for Label Agnostic and Class Agnostic adaptation scenarios. We found in early experiments that the symmetric KL-divergence works slightly better than either asymmetric option (see Section B.3 in the Supplementary Material for an ablation study on this choice).

3.3 Class Agnostic Prompt Learning

To further evaluate the generalization capabilities of KDPL, we introduce a scenario where not only do we not know the *labels* of training images (label agnostic, Section 3.2) but where we also do not even know the *class names* associated with the dataset (class agnostic). This scenario is considerably more challenging as we make no assumptions about the few-shot training data.

To address the unavailability of class names, we propose a strategy for automatically selecting a set of class names for each batch. We start with a large vocabulary of class names from which to select. Specifically, we use the Open Images V7 dataset [33], which contains ∼20K classes. The most straightforward method would simply use all 20K classes during training. However, this is impractical as the memory required by prompt learning methods increases linearly with the number of classes. According to Ren et al. [46], CoOp requires nearly 15MB of VRAM *per class*, resulting in approximately 300GB of memory when

multiplied by 20K classes. Therefore, we propose a method to automatically select which classes to use for each batch by retaining only the ones most relevant to the calculation of the loss in Eq. (2).

Given a batch of images $X = \{I_i\}_{i=1}^{N}$ and all the class names in the vocabulary $\mathcal{C} = \{\text{CLASS}_i\}_{i=1}^{C}$, where N represents the number of images in a batch and C the size of the dictionary, we let the teacher model select the most relevant classes. Our objective is to identify the most useful K classes in each batch for student prompt learning. After extracting the teacher image and text features, for each image $I_i \in X$, we apply Eq. (1) to obtain the teacher probabilities $p_T(I_i, \mathcal{C})$. By stacking the probabilities along the batch dimension, we obtain the matrix $P_T = [p_T(I_1, \mathcal{C}); \cdots ; p_T(I_N, \mathcal{C})]^T \in \mathbb{R}^{N \times C}$, where the i-th row corresponds to the probabilities associated with image I_i. We then compute the average probabilities along the batch axis, resulting in $\overline{P_T} \in \mathbb{R}^C$. Finally, we select the classes corresponding to the K highest values in $\overline{P_T}$.

Using the teacher model to perform this class filtering for each batch is feasible and does not incur excessive memory costs since the teacher requires no gradient computation. Therefore, the memory consumption does not depend on the number of classes in $\mathcal{C}$. Conversely, although the student model remains frozen, gradients must still be propagated to update the prompt parameters γ. Once the classes are selected, the training strategy remains the same as described above, with the only difference being that the class names observed by the student vary in each batch based on teacher predictions.

4 Experimental Results

In this section we report on experiments validating our proposed approach.

4.1 Evaluated Scenarios and Implementation Details

Following previous works [30,66,67], we validate KDPL in three distinct settings: 1) domain generalization; 2) cross-dataset transfer; 3) generalization to unseen classes. Additionally, to evaluate the scenario where class names are also unknown, we introduce a new evaluation setting we call *class agnostic adaptation*. We use the train/val/test splits and seeds provided by Zhou et al. [67] for all datasets.[3] All reported results are averages over three independent runs.

Evaluated Scenarios. We evaluate KDPL on the following scenarios:

- **Domain Generalization**. To assess the ability of the learned prompts to generalize to out-of-distribution datasets, we apply the prompt learned from ImageNet [12] to four different versions of ImageNet datasets exhibiting various types of domain shift. The target datasets are ImageNetV2 [45], ImageNet-Sketch [54], ImageNet-A [25], and ImageNet-R [24].

[3] `https://github.com/KaiyangZhou/CoOp/blob/main/DATASETS.md`

- **Cross-dataset Transfer**. To evaluate the ability of learned prompts to generalize to unseen classes, we evaluate the same prompt trained on ImageNet on a broad range of downstream recognition tasks. We validate on ten datasets with varying characteristics: Caltech101 [20] for general object classification; OxfordPets [43], StanfordCars [32], Flowers102 [41], Food101 [6], and FGVCAircraft [39] for fine-grained classification; the Describable Textures Dataset (DTD) [11] for texture classification; EuroSAT [23] for satellite-image recognition; and UCF101 [53] for action recognition.
- **Generalization to Unseen Classes**. To evaluate how learned prompts generalize to unseen classes from the same dataset, we divide classes into two subsets: base and novel classes. The prompt is trained exclusively on the base classes and then tested on the novel ones. We evaluate on ImageNet as well as the ten benchmark datasets used for cross-dataset evaluation.
- **Class Agnostic Adaptation**. In addition we introduce a novel evaluation setting in which the training class names are unknown. The prompt is trained on ImageNet and tested on all the benchmark datasets used in the domain generalization and cross-dataset evaluations.

Implementation Details. We use a CLIP model with a ViT-H-14 visual backbone as the teacher model.[4] For student models, we evaluate both on a CLIP model based on ResNet-50 and a model based on ViT-B/32. By experimenting with both ResNet-based and ViT-based CLIP models we show the architecture independence of our approach. To assess how KDPL performs when integrated into different prompt learning techniques, we selected five distinct textual, visual, and multimodal approaches. With the ResNet-50 student model, we experiment with CoOp [67] and CoCoOp [66]. For the ViT-B/32 student, since it supports both visual and textual prompting, we experiment with CoOp [67], VPT [29], MaPLe [30] and PromptSRC [31].

We denote the integration of our unsupervised knowledge distillation strategy with a specific prompt learning technique by adding the suffix +KDPL to the name of the corresponding technique. For example, CoOp+KDPL refers to the application of KDPL to CoOp. For class agnostic settings we instead use the suffix +CA-KDPL to indicate that no training class names are used.

Unless otherwise stated, all experiments are conducted in the few-shot setting with 16 examples per class randomly sampled from the training set. We set the temperature hyperparameter τ in Eq. (1) to 0.01. In the class agnostic experiments in which we assume no knowledge of class names in the training set, we set the number of class names selected in each iteration to $K = 1000$.

For each prompt learning method we use the original implementation and hyperparameters reported in the respective papers. Since we use the compact ResNet-50 and ViT-B/32 backbones, we cannot directly compare to the originally published results. Thus, all numbers we report here were computed by us and are averages over three independent runs. See Section A in the Supplementary Material for additional implementation details.

[4] `https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14`

Table 1: Domain Generalization. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). In this setting, the prompt is trained on ImageNet (source) and then tested on four different versions of ImageNet (targets) that exhibit some kind of domain shift. Average performance improvements over the baselines are shown in **green**, and average performance deterioration in **red**.

		Source	Target				
Backbone	Method	ImageNet	-V2	-S	-A	-R	*Average*
RN50	CLIP (student)	58.20	51.50	33.30	21.70	56.00	40.63
	CoOp	62.40	55.17	33.70	23.13	56.20	42.05
	CoOp + KDPL	**62.73**	**55.37**	**35.20**	**23.27**	**57.77**	**42.90**
	CoCoOp	**63.07**	55.53	34.77	**23.73**	**59.47**	43.38
	CoCoOp + KDPL	62.70	**55.60**	**35.30**	23.43	57.90	**43.06**
ViT-B/32	CLIP (student)	62.00	54.70	40.80	29.60	66.00	47.78
	CoOp	**66.33**	**58.30**	41.40	31.47	65.87	49.26
	CoOp + KDPL	65.97	58.10	**42.50**	**31.63**	**67.37**	**49.90**
	VPT	64.97	56.73	41.27	27.00	66.50	47.88
	VPT + KDPL	**65.10**	**57.37**	**41.67**	**27.77**	**67.47**	**48.57**
	MaPLe	**66.80**	**58.53**	42.23	**30.13**	66.40	49.32
	MaPLe + KDPL	66.50	58.47	**42.77**	29.87	**67.70**	**49.70**
	PromptSRC	**66.33**	**58.70**	42.97	**32.07**	**68.93**	50.67
	PromptSRC + KDPL	66.27	58.60	**43.07**	31.70	68.83	**50.55**
ViT-H/14	CLIP (teacher)	82.80	76.60	71.10	71.10	91.30	77.53

4.2 Results on Domain Generalization

Table 1 outlines the performance in the domain generalization setting. We report the performance of each baseline method alongside their unsupervised KDPL variants. Additionally, the performance of the zero-shot CLIP student and teacher models using a handcrafted prompt is included for comparison. We observe that applying our unsupervised approach to each baseline does not result in a significant decrease in performance on the source dataset. Notably, when transferring the learned prompts to a different domain, incorporating KDPL in each baseline can lead to a slight improvement. This suggests that our unsupervised teacher-student distillation learns prompts that generalize better than those trained using ground-truth labels. The only exceptions where we observe an average decrease compared to the baseline are when using CoCoOp or PromptSRC. CoCoOp-KDPL achieves an average accuracy significantly lower than CoCoOp only on the ImageNet-R dataset. Finally, note that all our unsupervised KDPL variants significantly improve the performance over the zero-shot CLIP student model in both the source and the target datasets.

Table 2: Cross-dataset Transfer. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). The prompt is learned on ImageNet and then tested on the ten different target datasets. Average performance improvements over the baselines are indicated in **green**.

		Target										
Backbone	Method	*OxfordPets*	*Flowers102*	*FGVCAircraft*	*DTD*	*EuroSAT*	*StanfordCars*	*Food101*	*SUN397*	*Caltech101*	*UCF101*	*Average*
RN50	CLIP (student)	83.70	61.00	15.60	40.00	24.20	55.60	75.20	58.50	86.00	58.30	55.81
	CoOp	84.60	61.63	13.77	36.83	22.33	54.20	75.03	59.00	87.67	57.73	55.28
	CoOp + KDPL	**87.33**	**62.83**	**14.97**	**39.27**	**30.17**	**56.60**	**77.27**	**60.30**	**88.57**	**58.63**	**57.59**
	CoCoOp	86.93	63.47	**16.27**	40.53	27.20	55.47	77.73	61.40	88.07	**60.43**	57.75
	CoCoOp + KDPL	**87.33**	**64.97**	15.97	**41.27**	**29.17**	**56.83**	**78.00**	**61.50**	**88.50**	59.53	**58.31**
ViT-B/32	CLIP (student)	85.00	64.30	18.20	42.80	38.10	60.40	79.10	62.00	91.10	60.70	60.17
	CoOp	**87.13**	62.20	12.13	**40.57**	36.97	57.83	80.13	62.83	91.13	**61.80**	59.27
	CoOp + KDPL	**87.13**	**62.23**	**15.57**	40.40	**40.93**	**58.83**	**80.57**	**64.23**	**92.77**	61.63	**60.43**
	VPT	**87.43**	**64.77**	**19.53**	**44.20**	28.47	57.60	78.27	**63.73**	91.40	62.50	59.79
	VPT + KDPL	**87.43**	64.43	17.77	42.93	**30.80**	**57.97**	**78.47**	63.63	**92.00**	**62.77**	**59.82**
	MaPLe	**88.37**	**66.43**	**18.47**	42.03	**37.77**	**60.00**	80.10	64.43	91.33	62.50	61.14
	MaPLe + KDPL	88.27	64.77	17.33	**43.13**	37.13	59.93	**80.47**	**65.43**	**93.43**	**62.70**	**61.26**
	PromptSRC	88.13	**65.33**	17.17	**43.90**	40.50	60.17	**81.23**	65.40	92.37	**63.63**	61.78
	PromptSRC + KDPL	**88.17**	65.10	**18.23**	43.27	**44.30**	**60.43**	80.90	**65.67**	**93.63**	63.40	**62.31**
ViT-H/14	CLIP (teacher)	94.80	89.40	63.20	66.80	63.30	95.60	93.60	76.40	97.90	76.40	81.74

4.3 Results on Cross-dataset Transfer

In Table 2 we present the results of the cross-dataset evaluation setting in which the prompt is trained on ImageNet and tested on ten different target datasets. For all datasets our KDPL-based variants consistently outperform the corresponding baselines and demonstrate superior generalization performance. The greater generalization capabilities of KDPL are evident even for fine-grained datasets like EuroSAT, on which the CoOp+KDPL achieves an 8% improvement over the baseline CoOp when using ResNet-50 as the backbone. Although adding KDPL yields only minor improvement to the VPT and MaPLe prompt learning techniques, we emphasize that our VPT+KDPL and MaPLe+KDPL do not have access to ground-truth labels. Notably, PromptSRC+KDPL achieves the highest average performance.

4.4 Results on Generalization to Unseen Classes

In Table 3 we give results for the unseen class generalization task. In these experiments each dataset is split into 50% of the classes as a base for training few-shot adaptation, and the remaining 50% as new classes on which zero-shot performance is evaluated. KDPL consistently outperforms the corresponding baseline

Table 3: Generalization to Unseen Classes. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). In this setting the prompt is tested on classes never observed during training. Average performance improvements over the baselines are indicated in **green**.

		Unseen Classes											
Backbone	Method	*ImageNet*	*OxfordPets*	*Flowers102*	*FGVCAircraft*	*DTD*	*EuroSAT*	*StanfordCars*	*Food101*	*SUN397*	*Caltech101*	*UCF101*	*Average*
RN50	CLIP (student)	60.10	93.70	71.30	24.80	53.90	43.70	66.60	82.20	70.10	90.50	67.80	65.88
	CoOp	60.00	91.63	58.30	**22.17**	**40.93**	42.13	**59.87**	82.03	66.77	87.53	57.63	60.82
	CoOp + KDPL	**60.53**	**94.97**	**63.40**	21.27	38.70	**62.07**	56.63	**83.07**	**69.83**	**88.75**	**62.63**	**63.80**
	CoCoOp	**63.07**	94.93	67.47	**24.87**	45.40	35.10	**64.73**	84.50	72.87	**90.73**	64.67	64.39
	CoCoOp + KDPL	62.43	**95.10**	**69.17**	24.57	**48.00**	**55.93**	63.70	**85.67**	**73.33**	90.53	**66.40**	**66.80**
ViT-B/32	CLIP (student)	64.00	96.30	72.30	28.30	53.90	61.50	69.80	85.70	73.10	94.00	71.60	70.05
	CoOp	64.53	93.37	**59.77**	**23.03**	45.57	49.13	61.73	85.50	69.37	92.13	65.60	64.52
	CoOp + KDPL	**64.90**	**94.00**	56.17	22.00	**47.33**	**62.07**	**62.37**	**87.23**	**74.17**	**93.43**	**66.20**	**66.35**
	VPT	64.60	94.83	66.03	29.70	**50.17**	50.03	**69.90**	85.83	75.13	92.57	69.47	68.02
	VPT + KDPL	**65.20**	**95.20**	**67.87**	**30.40**	47.53	**54.97**	69.80	**86.33**	**75.17**	**92.80**	**71.07**	**68.76**
	MaPLe	66.70	**96.87**	**69.77**	20.43	**53.50**	66.70	68.03	87.47	76.57	92.27	69.40	69.79
	MaPLe + KDPL	**66.73**	**96.87**	68.23	**27.37**	50.33	**75.70**	**68.40**	**87.83**	**76.83**	**93.67**	**73.53**	**71.41**
	PromptSRC	65.93	96.30	**71.37**	23.67	**54.43**	**61.97**	**70.00**	87.13	76.77	**94.70**	72.63	70.45
	PromptSRC + KDPL	**66.33**	**96.47**	68.83	**28.43**	50.87	74.17	68.50	**87.63**	**77.33**	94.47	**74.73**	**71.61**
ViT-H/14	CLIP (teacher)	84.10	99.20	85.60	64.30	75.00	82.60	98.20	96.30	85.10	97.30	85.20	86.63

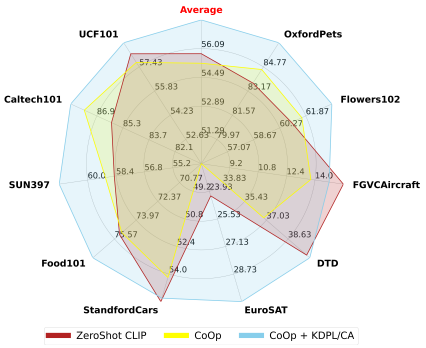
methods for both backbones, demonstrating improvement in all scenarios for the majority of the datasets. On average, the performance improvement over the supervised baseline methods ranges from about 1% for VPT to about 3% for CoOp with the ResNet-50 backbone. See Section C.1 in the Supplementary Material for further analysis of base and unseen performance.

4.5 Results on Class Agnostic Adaptation

Figure 3(a-b) summarizes the main results in the proposed Class Agnostic (CA) scenario in which even the training class names are unknown at training time. We report the accuracy on the ImageNet dataset (source), as well as the average accuracy in domain generalization and cross-dataset settings. Note that, even without knowing the class names, the performance on the source dataset steadily improves compared to the zero-shot CLIP model. Moreover, the prompts learned via the proposed unsupervised class agnostic knowledge distillation also exhibit improved average domain generalization (AVG^1) and cross-dataset capabilities (AVG^2). Figure 3(b) visually depicts how the ResNet-50-based CoOp+CA-KDPL compares with supervised CoOp and zero-shot CLIP student performance in the cross-dataset transfer setting. Notably, CA-KDPL outperforms both baselines despite being unsupervised and class agnostic during training.

Method	Source	AVG^1	AVG^2
ResNet-50 (student)	58.20	40.63	55.81
CoOp + CA-KDPL	61.83	**43.02**	**57.69**
CoCoOp + CA-KDPL	61.43	**43.18**	**58.03**
ViT-B/32 (student)	62.00	47.78	60.17
CoOp + CA-KDPL	64.73	**49.63**	**60.74**
VPT + CA-KDPL	63.73	**47.89**	**59.80**
MaPLe + CA-KDPL	65.23	**49.28**	**61.23**
PromptSRC + CA-KDPL	65.00	**50.37**	**61.93**

(a) Average class agnostic adaptation



(b) Per-dataset class agnostic adaptation

Fig. 3: Class agnostic adaptation. **(a)** Comparison between the zero-shot baselines and our class agnostic CA-KDPL variants (highlighted in cyan). The prompt is learned in an unsupervised and class agnostic setting on ImageNet and evaluated on the benchmark datasets for domain generalization (AVG^1) and cross-dataset (AVG^2) evaluations. Average performance improvements are indicated in **green**, and deterioration in **red**. **(b)** Per-dataset accuracy comparison between our unsupervised and class agnostic method (CoOp+CA-KDPL), the supervised baseline CoOp, and the zero-shot student on the cross-dataset benchmark datasets.

5 Conclusion

In this paper we proposed an approach to Knowledge Distillation Prompt Learning (KDPL) which is easily integrated into existing supervised prompt learning methods. Our experiments show that for CoOp, CoCoOp, VPT, MaPLe and PromptSRC adding KDPL: (1) renders them *label agnostic* by eliminating the need for ground-truth labels for few-shot adaptation; (2) can also render them *class agnostic* in cases where no knowledge of training class labels is available; and (3) remarkably improves generalization to downstream tasks.

The additional computation cost incurred by distilling from a large VLM is a limitation of KDPL, though all encoder parameters are fixed and the predictions of the teacher can be precomputed, which helps mitigate the extra cost. This extra computation only amounts to about a 15% increase for MaPLe.

Note that we did not tune any hyperparameters when training the KDPL variants. We use the same settings as in the original papers, which are likely suboptimal for distillation-based adaptation. With careful tuning, there is potential for improvement. Our experiments indicate that distillation-learned prompts are more transferable, and we think it would be interesting to see if this idea can generalize to very different downstream tasks and scale to even bigger teacher models or – more interestingly – to even smaller student models.

Acknowledgements

This work was supported by funding from the European Commission Horizon 2020 grant #951911 (AI4Media).

References

1. Abdul Samadh, J., Gani, M.H., Hussein, N., Khattak, M.U., Naseer, M.M., Shahbaz Khan, F., Khan, S.H.: Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. Advances in Neural Information Processing Systems **36** (2024)
2. Agnolucci, L., Baldrati, A., Todino, F., Becattini, F., Bertini, M., Del Bimbo, A.: Eco: Ensembling context optimization for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2811–2815 (2023)
3. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15338–15347 (October 2023)
4. Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4662–4670 (2022)
5. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10925–10934 (2022)
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014)
7. Bulat, A., Tzimiropoulos, G.: Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23232–23241 (2023)
8. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Plot: Prompt learning with optimal transport for vision-language models. In: The Eleventh International Conference on Learning Representations (2022)
9. Chen, W.C., Chang, C.C., Lee, C.R.: Knowledge distillation with feature maps for image classification. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 200–215. Springer (2019)
10. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
11. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

13. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11162–11173 (2021)
14. Ding, Y., Liu, L., Tian, C., Yang, J., Ding, H.: Don't stop learning: Towards continual learning for the clip model. arXiv preprint arXiv:2207.09248 (2022)
15. Dong, X., Bao, J., Zhang, T., Chen, D., Gu, S., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. arXiv preprint arXiv:2212.06138 (2022)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
17. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging **39**(7), 2415–2425 (2020)
18. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A.T., Shankar, V.: Data filtering networks. In: The Twelfth International Conference on Learning Representations (2023)
19. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: Self-supervised distillation for visual representation. In: International Conference on Learning Representations (2020)
20. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178. IEEE (2004)
21. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2021)
22. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 578–587 (2019)
23. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)
24. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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. pp. 8340–8349 (2021)
25. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat **1050**, 9 (2015)
27. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649 (2022)
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

29. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (2022)
30. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
31. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15190–15200 (2023)
32. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
33. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision (IJCV) **128**(7), 1956–1981 (2020)
34. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021)
35. Li, X.L., Liang, P.: 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 (Volume 1: Long Papers). pp. 4582–4597 (2021)
36. Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26617–26626 (2024)
37. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5206–5215 (2022)
38. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096 (2022)
39. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
40. Mullapudi, R.T., Chen, S., Zhang, K., Ramanan, D., Fatahalian, K.: Online model distillation for efficient video inference. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3573–3582 (2019)
41. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
42. Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T.: Clip-guided vision-language pre-training for question answering in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5606–5611 (2023)
43. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
45. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
46. Ren, S., Zhang, A., Zhu, Y., Zhang, S., Zheng, S., Li, M., Smola, A.J., Sun, X.: Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. Advances in Neural Information Processing Systems **36** (2024)
47. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
48. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
49. Shi, C., Yang, S.: Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2932–2941 (2023)
50. Shi, Z., Lipani, A.: DePT: Decomposed prompt tuning for parameter-efficient fine-tuning. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=KjegfPGRde`
51. Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., Jagersand, M.: Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 50–56. IEEE (2019)
52. Song, H., Dong, L., Zhang, W., Liu, T., Wei, F.: Clip models are few-shot learners: Empirical studies on vqa and visual entailment. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6088–6100 (2022)
53. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
54. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in Neural Information Processing Systems **32** (2019)
55. Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., Tan, T.: Improving zero-shot generalization for clip with synthesized prompts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3032–3042 (2023)
56. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European Conference on Computer Vision. pp. 588–604. Springer (2020)
57. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6757–6767 (2023)
58. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3903–3911 (2020)
59. Yun, S., Park, J., Lee, K., Shin, J.: Regularizing class-wise predictions via self-knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13876–13885 (2020)
60. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022)
61. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
62. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
63. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022)
64. Zhong, Z., Friedman, D., Chen, D.: Factual probing is [mask]: Learning vs. learning to recall. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5017–5033 (2021)
65. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
66. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
67. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
68. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15613–15623. IEEE Computer Society (2023)