PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

Zheng Li[1], Xiang Li[2,1]*, Xinyi Fu[3], Xin Zhang[1], Weiqiang Wang[3], Shuo Chen[4], Jian Yang[1]*
[1] PCA Lab, VCIP, College of Computer Science, Nankai University
[2] NKIARI, Shenzhen Futian, [3] Tiansuan Lab, Ant Group, [4] RIKEN
{zhengli97, zhasion}@mail.nankai.edu.cn, {xiang.li.implus, csjyang}@nankai.edu.cn
{fxy122992, weiqiang.wwq}@antgroup.com, shuo.chen.ya@riken.jp

Abstract

Prompt learning has emerged as a valuable technique in enhancing vision-language models (VLMs) such as CLIP for downstream tasks in specific domains. Existing work mainly focuses on designing various learning forms of prompts, neglecting the potential of prompts as effective distillers for learning from larger teacher models. In this paper, we introduce an unsupervised domain prompt distillation framework, which aims to transfer the knowledge of a larger teacher model to a lightweight target model through prompt-driven imitation using unlabeled domain images. Specifically, our framework consists of two distinct stages. In the initial stage, we pre-train a large CLIP teacher model using domain (few-shot) labels. After pre-training, we leverage the unique decoupled-modality characteristics of CLIP by pre-computing and storing the text features as class vectors only once through the teacher text encoder. In the subsequent stage, the stored class vectors are shared across teacher and student image encoders for calculating the predicted logits. Further, we align the logits of both the teacher and student models via KL divergence, encouraging the student image encoder to generate similar probability distributions to the teacher through the learnable prompts. The proposed prompt distillation process eliminates the reliance on labeled data, enabling the algorithm to leverage a vast amount of unlabeled images within the domain. Finally, the well-trained student image encoders and pre-stored text features (class vectors) are utilized for inference. To our best knowledge, we are the first to (1) perform unsupervised domain-specific prompt-driven knowledge distillation for CLIP, and (2) establish a practical pre-storing mechanism of text features as shared class vectors between teacher and student. Extensive experiments on 11 datasets demonstrate the effectiveness of our method. Code is publicly available at `https://github.com/zhengli97/PromptKD`.

*Corresponding author.

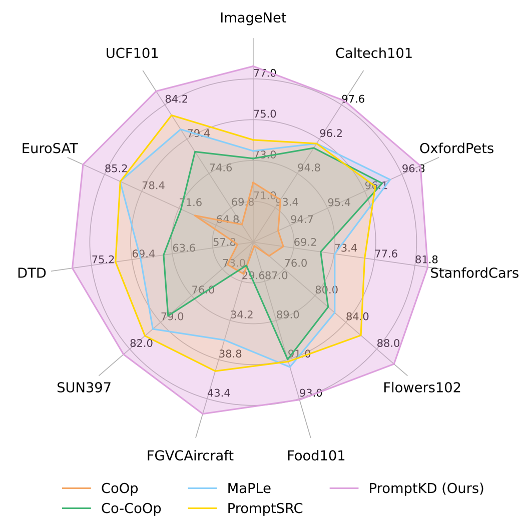

Figure 1. Harmonic mean (HM) comparison on base-to-novel generalization. All methods adopt the **ViT-B/16 image encoder** from the pre-trained CLIP model. PromptKD achieves state-of-the-art performance on 11 diverse recognition datasets.

1. Introduction

Recently large pretrained vision-language models (VLMs), such as CLIP [30, 49] and ALIGN [10], have demonstrated superior generalization ability for domain-specific downstream tasks. Unlike conventional visual frameworks, the vision-language model, like CLIP, usually employs a two-tower architecture that includes an image encoder and a text encoder. These models are trained using a contrastive loss to learn a unified embedding space that aligns the representations of multi-modal signals.

To better optimize the models for domain-specific downstream tasks, various methods [6, 14, 46, 52, 53] have been proposed to adapt the representation while keeping the original CLIP model fixed. Inspired by the success of Nature Language Processing (NLP) [18, 20] area, prompt learn-

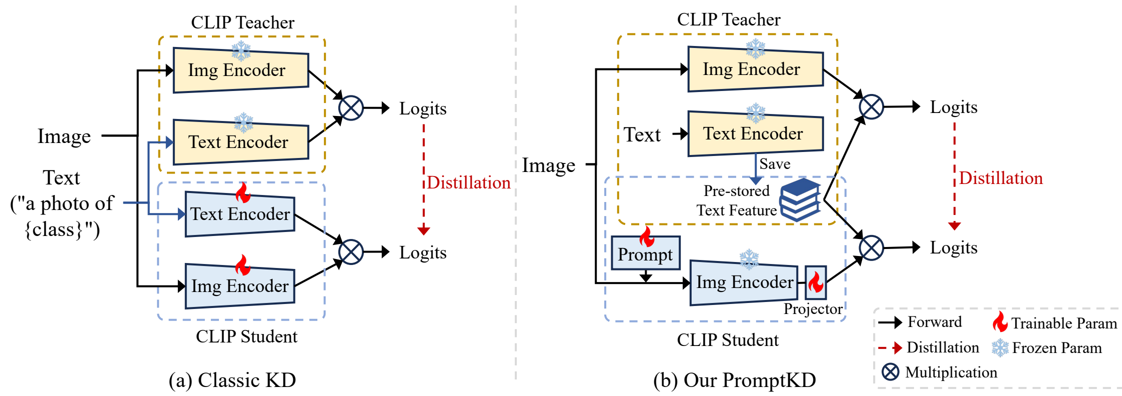

Figure 2. Architecture comparison between classic KD paradigm for CLIP (likewise CLIP-KD [44]) and our prompt distillation framework. (a) Classic KD methods perform distillation between independent teacher and student models. Students are typically fully fine-tuned by teachers' soft labels. (b) PromptKD breaks the rules of teacher-student independence. We propose to reuse the previously well-trained text features from the teacher pre-training stage and incorporate them into the student image encoder for both distillation and inference.

ing [11, 52, 53] has been proposed to acquire continuous prompt representations as a replacement for meticulously designed hard prompts. Based on the type of information learned by prompt, existing methods can be roughly divided into three types: text-based, visual-based, and both. Text-based methods [52, 53] propose to adaptively learn appropriate text prompts for downstream tasks, rather than fixed forms. Visual-based methods [3, 11] follow similar principles and further apply them to visual modalities. Text-visual-based prompt methods [14, 15, 17, 35] suggest a simultaneous learning strategy for prompts in both image and text branches, instead of treating them separately.

Prior research has primarily concentrated on acquiring effective *formats of prompts* using scarce labeled data while preserving the outstanding generalization capabilities. In this paper, we introduce a novel unsupervised framework (termed "PromptKD") where the prompt acts as a domain knowledge distiller, allowing the CLIP student model to absorb knowledge from a vast CLIP teacher model on extensive unlabeled domain data. Specifically, our framework consists of two distinct stages: the teacher pre-training stage and the student distillation stage.

In the initial stage, we first pre-train a large CLIP teacher model using existing advanced approaches [14, 15] on domain few-shot labeled data. After pre-training, we propose to leverage the unique decoupled-modality characteristics of CLIP by pre-computing and storing the text features as class vectors *only once* through the teacher text encoder.

In the subsequent stage, the stored class vectors are shared across the teacher and student image encoder to calculate the predicted logits without any extra computation costs from text branches. Different from the traditional knowledge distillation scheme where the weights of a student are usually fully tuned to mimic the teachers' statistical behavior as shown in Fig. 2(a), we propose to utilize the student's learnable visual prompts to align the logits of both teacher and student models via KL divergence, encouraging the student image encoder to generate similar probability distributions to the teacher through prompt distillation. Due to the dimensional differences between the features of teacher and student, an extra projector is implemented to adjust the features to account for the dimension disparity.

With the benefits of the teacher-student paradigm, we can leverage the pre-trained teacher to generate soft labels for unlabeled images from the target domain, thus enabling the training of students without the need for labeled images. Finally, the well-trained student image encoder, along with the pre-stored teacher text features (class vectors), are employed for inference purposes. An architectural comparison of the classic distillation paradigm for CLIP and our proposed prompt distillation framework is illustrated in Fig. 2.

Experimental results in Fig. 1 show that our PromptKD outperforms previous methods and achieves state-of-the-art performance on 11 diverse recognition datasets with the ViT-B/16 image encoder CLIP model. Specifically, our method achieves average improvements of 2.70% and 4.63% on the base and new classes on 11 diverse datasets.

Our contributions can be summarized as follows:

- To our best knowledge, we are the first method to perform domain-specific prompt-based knowledge distillation for CLIP using unlabeled domain data.
- We leverage CLIP's unique decoupled-modality property to reuse pre-stored text features without incurring any additional computation costs from the text branch, thereby facilitating the distillation and inference processes.
- With the benefits of the teacher-student paradigm, we can utilize the teacher to generate soft labels on extensive unlabeled domain data, enabling the training of students without the need for labeled images.
- Extensive experiments on 11 datasets demonstrate the effectiveness of our method.

2. Related Work

Prompt Learning in Vision-Language Models. Prompt learning is a technique that can transfer the large pre-trained model, like CLIP [30], towards downstream tasks [7, 31, 47] without the need for completely re-training the original model. It proposes to adapt the representations for specific tasks through learnable text or visual soft prompts instead of manually crafted hard prompts (e.g., "a photo of a {classname}"). Soft prompts [11, 17, 32, 52, 53] can be optimized by back-propagating through the frozen pre-trained model, resulting in better performance. Existing works mainly focus on designing various efficient forms of prompts using scarce labeled domain data. MaPLe [14] proposes to learn prompts for the image and text branches simultaneously, rather than a separate side. PromptSRC [15] utilizes its original features to regularize the learning of prompts for each branch. Previous works necessitated forward and backward computations for each input in both image [5, 38] and text branches. In our work, we leverage the unique decoupled-modality characteristic of CLIP, saving well-trained teacher text features as class vectors for student distillation. In this way, the training of student CLIP is simplified to solely include forward and backward calculations of the image branch, without requiring the text branch.

Zero-shot Learning. Given the labeled training set of the seen classes, zero-shot learning (ZSL) [24, 37, 40] aims to learn a classifier that can classify testing samples of unseen classes. Existing methods can be roughly divided into two types based on whether test images are available: Inductive [41, 48] and Transductive [33, 34] ZSL. Previous works on prompt learning, such as MaPLe and PromptSRC, have mainly focused on the instance inductive settings where only labeled training instances are available. In our paper, we explore the transductive ZSL setting where both seen and unseen class images are all utilized in model learning. Specifically, our teacher model follows the same training scheme as PromptSRC, which is trained on samples from seen classes with ground truth labels. The difference is that the target student model is trained on the full unlabeled dataset, which contains all samples of both seen and unseen classes, without using any ground truth labels.

Knowledge Distillation. Knowledge distillation [8] aims to train a lightweight student model under the supervision of a large pretrained teacher model. In recent years, various distillation forms have emerged for effective knowledge transfer from teachers to students, such as logits alignment [21, 23, 50, 51], feature imitation [2, 19, 45] and sample relationship matching [28, 42]. In addition to traditional image classification topics, knowledge distillation has achieved great success in many vision tasks, including object detection [1, 12, 36], image segmentation [25, 43], and pose estimation [22]. Recently, many works [16, 29, 39, 44] have turned their attention to the CLIP model. These works propose leveraging the CLIP model's exceptional generalization capabilities to enhance the learning of existing models. CLIP-KD [44] find that in distilling pre-trained CLIP models, the simplest feature mimicry with the MSE loss approach yields the best results. TinyCLIP [39] performs cross-modal feature alignment in affinity space between teacher and student. Our approach differs from previous distillation methods that train the *entire student* model by leveraging a pre-trained large CLIP teacher. In our work, we employ a more efficient approach by utilizing *student prompts* for distillation while keeping the student's original CLIP weights frozen. This allows us to achieve the desired knowledge transfer without the need for extensive re-training of the student model.

3. Method

Prompt learning [11, 53] aims to enhance the performance of existing VLMs like CLIP to downstream tasks by incorporating learnable prompts. Existing works mainly focus on devising effective learning formats of prompts using scarce labeled domain data while ensuring strong generalization capabilities to unseen images. In this paper, we first explore prompts as an effective knowledge distiller, allowing the CLIP student model to learn from the large CLIP teacher model by aligning their predictions on extensive unlabeled domain images. An overview of our proposed prompt distillation method is illustrated in Fig. 3. Specifically, our method comprises two main stages: the teacher pre-training stage and the student prompt distillation stage. In the initial stage, we first pre-train a large CLIP teacher model using existing advanced approaches on few-shot labeled data, as depicted in Fig. 3(a). After pre-training, we extract and preserve the highly proficient text features obtained from the teacher text encoder as class vectors. In the subsequent stage, the pre-stored class vectors are effectively reused by multiplying them with the outputs of both the teacher and student image encoders, resulting in predictions for each model. Then we initiate the distillation process by promoting prompt imitation, encouraging the student model to generate similar predictions to the teacher model, as illustrated in Fig. 3(b). An additional projector is introduced to align the dimensions of teacher text features and student image features. Finally, the well-trained student image encoder branch and pre-stored teacher text features (class vectors) are utilized for inference (see Fig. 3(c)).

Below we first introduce the background knowledge of VLMs and the knowledge distillation method in Sec. 3.1. Then we introduce our method in detail in Sec. 3.2.

3.1. Background

Vision-Language Models. Existing VLMs like CLIP [30] and ALIGN [10] are designed to align images and texts in order to learn a joint embedding space. Following [14, 15,

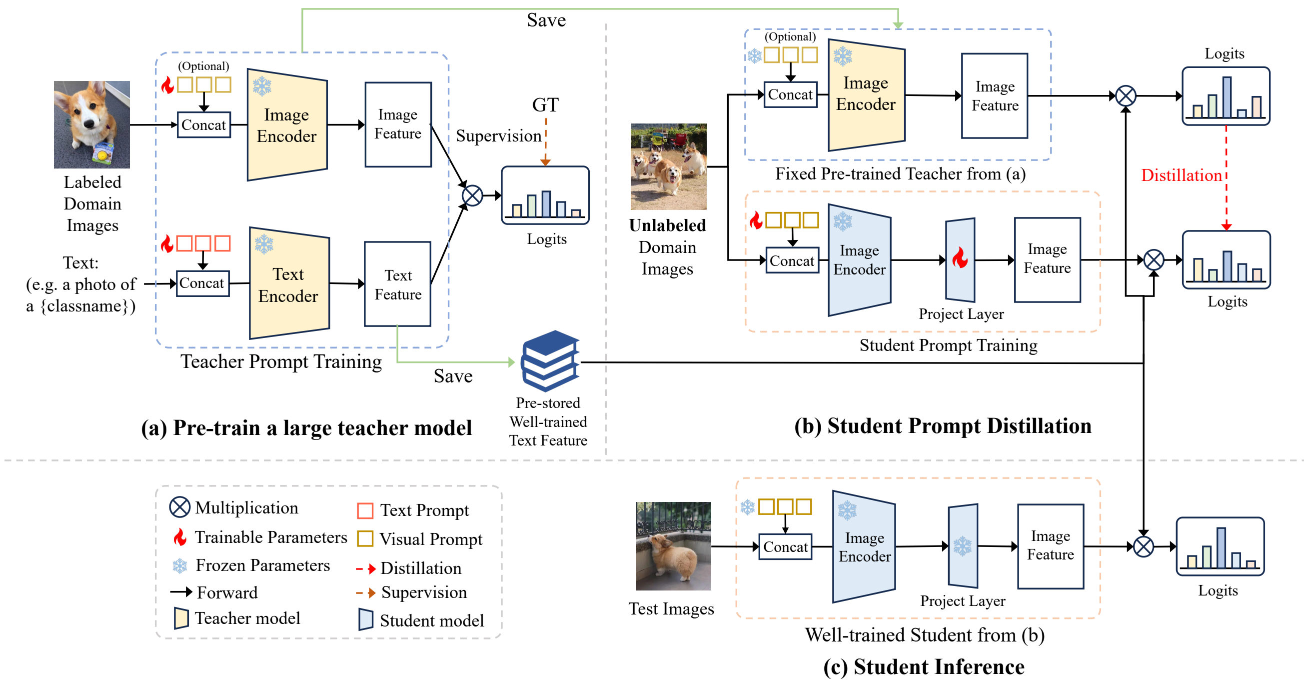

Figure 3. An overview of our proposed prompt distillation (PromptKD) framework. (a) We first pre-train a large CLIP teacher model using existing state-of-the-art prompt learning methods with labeled training images. Then we save the well-trained text features of all possible classes for the next stages. (b) During the distillation stage, the training is focused on student image prompts and the project layer, and there are no extra computational expenses associated with the text encoding process when utilizing the pre-saved text features as class vectors. (c) Finally, the well-trained student and pre-stored class vectors are utilized for inference.

52], we consider CLIP as our foundation model. Specifically, CLIP consists of two encoders, one for image and the other for text. Given a labeled visual recognition dataset $D = \{x_j, y_j\}_{j=1}^{M}$ that includes a set of N class names $\mathrm{c} = \{c_i\}_{i=1}^{N}$, CLIP generates textual descriptions t_i using the template "a photo of a $\{c_i\}$" for each class name. Then each text description t_i is fed into the text encoder f_T to obtain the normalized text feature $w_i = f_T(t_i)/||f_T(t_i)||_2 \in \mathbb{R}^d$, where d represents the feature dimension. The complete text features $\mathrm{W} = [w_1, w_2, ..., w_N] \in \mathbb{R}^{N\times d}$ of all classes can be considered as the classification weight vector for classifying an image. Given an input image x from the dataset D, the image encoder f_I takes as input and generates the normalized image feature $u = f_I(x)/||f_I(x)||_2 \in \mathbb{R}^d$. The output probability is calculated as follows:

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

where uw^T represent the output logit and τ is the temperature parameter.

Instead of manually crafted hard prompts, recent works like CoOp [53] propose to adaptively learn appropriate soft *textual* prompts for downstream tasks. Concretely, M learnable textual vectors $\{v_1, v_2, ..., v_M\}$, i.e., prefix, are added before the CLASS token to create a contextualized representation. Then the prompt t_i for class c_i becomes $t_i = \{v_1, v_2, ..., v_M, c_i\}$, where each vector v_i $(i \in 1, 2, ..., M)$ have the same dimension with the word embeddings and M is a hyperparameter that determines the length of the prefix. In addition to text prompt tuning methods, visual prompts have also been extensively explored. Some works [11, 14, 15] follow the same idea as the text prompt method, adding multiple learnable visual prefixes to the image patch as input to the image encoder. These visual prompts aim to guide the image encoder to extract more meaningful and task-relevant visual features. By incorporating these learnable visual prefixes, the model can leverage additional context and prior knowledge to improve its performance on image understanding tasks.

Knowledge Distillation. Originally proposed by Hinton et al. [8], knowledge distillation aims to transfer the knowledge of a pretrained heavy teacher model to a lightweight student model. After the distillation, the student can master the expertise of the teacher and be used for final deployment. Specifically, the Kullback-Leibler (KL) divergence loss is utilized to match the output distribution of two models, which can be formulated as follows:

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

where q^t and q^s denote the logits predicted by the teacher and student. $\sigma(\cdot)$ is the softmax function and τ is the temperature [8, 23] which controls the softness of distribution.

3.2. PromptKD: Prompt Distillation for VLMs

Our proposed prompt distillation framework comprises two stages: teacher pre-training and student prompt distillation, as illustrated in Fig. 3. In this section, we provide a comprehensive explanation of each stage.

Stage I: Teacher Pretraining. In the initial stage, we begin by pre-training a large CLIP teacher model using labeled domain data, as illustrated in Fig. 3(a). To accomplish this, we can employ existing prompt learning methods such as MaPLe [14] and PromptSRC [15], or alternatively, utilize a publicly available pretrained CLIP model for simplicity. Given a labeled domain dataset $D_{labeled} = \{x_i, y_i\}_{i=1}^{M}$ with a set class name, the teacher CLIP model takes training images and text descriptions with category names as input, and passes through the image encoder f_I^t and text encoder f_T^t to obtain the corresponding normalized image features $u \in \mathbb{R}^d$ and text features $w \in \mathbb{R}^d$. The final output result p^t is calculated by Eqn. (1). Typically, the parameters of teacher soft prompts are updated by minimizing the cross-entropy loss between predicted probabilities p and ground truth labels y.

Once the training of the text encoder is completed, the output features remain fixed and do not require further updates. In this case, we save the well-trained teacher text features of all N classes $W = [w_1, w_2, ..., w_N] \in \mathbb{R}^{N\times d}$ as *shared* class vectors that will be utilized in the subsequent stages of the process. This operation eliminates the necessity of having the student CLIP text branch, resulting in substantial computational cost savings during the training process. In addition, through our PromptKD method, we can replace the large teacher's heavy image encoder with a student lightweight image encoder, reducing the computational cost during deployment while maintaining competitive performance.

Stage II: Student Prompt Distillation. At this stage, we aim to train a student model by encouraging the student to align with the teacher's output results through prompt imitation, as shown in Fig. 3(b). Thanks to the strategy of reusing teacher text features, we only need to train the student image encoder branch f_I^s with learnable visual prompts and the feature projector. In the context of an unlabeled domain dataset $D_{unlabeled}$, by inputting the image x into both the pre-trained teacher's and the untrained student's image branches, we can acquire the normalized teacher image features $u^t = f_I^t(x)/||f_I^t(x)||_2 \in \mathbb{R}^d$ and student image features $u^s = P(f_I^s(x))/||P(f_I^s(x))||_2 \in \mathbb{R}^d$. The learnable projector $P(\cdot)$ in the student image encoder branch is introduced to match the feature dimensions at a relatively small cost while being effective enough to ensure accurate alignment. Then we multiply the pre-stored teacher text features $W \in \mathbb{R}^{N\times d}$ with the teacher and student image features to obtain the output logits $q^t = u^t W^{\mathsf{T}} \in \mathbb{R}^N$ and $q^s = u^s W^{\mathsf{T}} \in \mathbb{R}^N$, respectively. We optimize the student model to produce similar output to the teacher model on the unlabeled domain dataset $D_{unlabeled}$, which can be formulated as follows:

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

Algorithm 1 Pseudocode of PromptKD in PyTorch.

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

# init
f_txt_t = tea_t(txt_of_all_classes)

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

    f_img_s = Proj(f_img_s)

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

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

Algorithm 1 provides PromptKD's PyTorch-style pseudocode.

Inference. Finally, the well-trained student image encoder f_I^s, along with the pre-stored teacher text features W (class vectors), are employed for inference purposes.

4. Experiments

4.1. Settings

Base-to-novel Generalization. Following [14, 15, 52], we split the training and testing datasets into base and novel classes. The teacher is pre-trained using the PrompSRC [15] method, following the same training setting as PromptSRC. During distillation, we use the entire unlabeled training set to train our students. After distillation, the student's performance on the base and the novel class is evaluated on the testing set.

Cross-dataset Evaluation. Same as PromptSRC [15], our teacher model is pre-trained on the source dataset (i.e., ImageNet) with a 16-shot training data configuration. Then we use the training set of unlabeled target datasets to train students and evaluate their performance on the test set after training. In PromptKD, we use unlabeled images of unseen classes for student training which belongs to the *transductive* zero-shot learning method. For previous methods such as CoOp, MaPLe, and PromptSRC, their training is based on seen class data and belongs to the *inductive* paradigm.

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

(a) Average over 11 datasets.

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

(b) ImageNet

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

(c) Caltech101

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

(d) OxfordPets

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

(e) StanfordCars

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

(f) Flowers102

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

(g) Food101

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

(h) FGVCAircraft

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

(i) SUN397

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

(j) DTD

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

(k) EuroSAT

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

(l) UCF101

Table 1. Comparison with existing state-of-the-art methods on base-to-novel generalization. Our proposed PromptKD demonstrates strong generalization ability and achieves significant improvements on 11 recognition datasets given the **ViT-B/16 image encoder** of the CLIP model. In our approach, the default teacher model is the ViT-L/14 CLIP model. The symbol Δ denotes the performance improvement compared to the previous SOTA method PromptSRC. Our PromptKD outperforms previous methods on all datasets.

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

Table 2. Comparison of PromptKD with existing advanced approaches on cross-dataset benchmark evaluation. Based on our pipeline, we perform unsupervised prompt distillation using the unlabeled domain data respectively (i.e., the transductive setting). The source model is trained on ImageNet [4]. "ZSL" denotes the setting type for Zero-Shot Learning. PromptKD achieves better results on 9 of 10 datasets.

Datasets. We evaluate the model performance on 11 popular recognition datasets. The details of each dataset are attached in the Appendix.

Implementation Details. We use the ViT-L/14 CLIP model as our teacher model and the ViT-B/16 CLIP model as our target student model. Unless otherwise stated, the PromptSRC [15] is leveraged as our default method to pre-train our teacher model. We report base and novel class accuracy and their harmonic mean (HM) averaged over 3 runs. Due to page limitations, please refer to the Appendix for more implementation details and experimental results.

4.2. Base-to-novel Generalization

As shown in Table 1, based on the same ViT-B/16 image encoder of the pre-trained CLIP, we compare the performance of our proposed PromptKD with recent state-of-the-art prompt learning methods including CoOp, CoCoOp, MaPLe and PromptSRC on 11 recognition datasets. In comparison with previous works, PromptKD shows superior performance on all 11 datasets. The accuracy of our pre-trained teacher model with ViT-L/14 image encoder on each dataset is provided in the Appendix.

4.3. Cross-dataset Evaluation

Table 2 shows the performance comparison between CoOp, CoCoOp, MaPLe, PromoptSRC, and PromptKD. In comparison with previous methods, our method demonstrates better performance on 9 of 10 datasets. leading to an average improvement of 5.52% over the previous method.

4.4. Comparison with Other Methods

In PromptKD, we utilize unlabeled images to train the target student model. Table 3 presents a comprehensive comparison between our method and other recent methods that also leverage unlabeled data for model training. While many methods resort to pseudo-labeling of unlabeled data for training, our approach adopts a teacher-student paradigm. In this paradigm, the teacher model plays a pivotal role by furnishing soft labels to train the student models on the unlabeled data. For fair comparisons, the methods using few-shot labels ([26] and PromptKD) are all implemented based on PromptSRC framework. All experiments utilize **ViT-B/16** CLIP with the few-shot number being 16. The results on Flowers102 underscore the clear performance advantages of our approach over previous methods.

4.5. Ablation Study

By default, the distillation experiments are conducted on the entire ImageNet. For experimental efficiency, we use 64 images per class by default, that is, a total of 64,000 images for 1,000 classes, as an unlabeled training set for distillation, unless we state otherwise. The accuracy of the base and new classes is evaluated on the test set.

Method	Domain Data	Base	Novel	HM
CLIP	Zero-shot	72.08	77.80	74.83
PromptSRC	Few-shot	98.07	76.50	85.95
CLIP-PR [13]	Unlabeled	65.05	71.13	67.96
UPL [9]		74.83	78.04	76.40
LaFTer [27]		79.49	82.91	81.16
FPL [26]	Few-shot + Unlabeled	97.60	78.27	86.87
IFPL [26]		97.73	80.27	88.14
GRIP [26]		97.83	80.87	88.54
PromptKD		99.42	82.62	90.24
Δ		+1.59	+1.75	+1.70

Table 3. Comparison with existing works using unlabeled data on Flowers102. Our method performs better than previous methods.

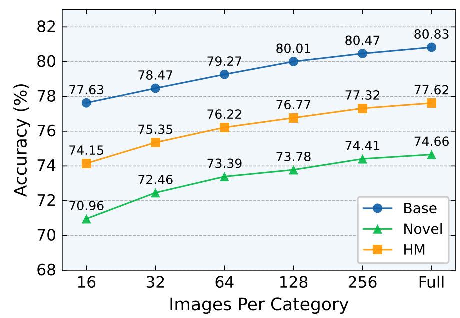

Figure 4. Improved ImageNet classification accuracy of the student model with increasing numbers of unlabeled images per category used for distillation.

Number of Images Used for Training. In this section, our objective is to assess the influence of training data volume on distillation performance, as depicted in Fig. 4. The figure illustrates that as the number of training images rises towards the complete training dataset, the accuracy consistently improves. It is noteworthy that with a further increase in the number of training images, the rate of performance improvement starts to plateau.

KD Form	Loss	Base	Novel	HM
Feature	L1	73.09	65.98	69.35
	MSE	71.89	66.17	68.91
Logit	KL	**79.27**	**73.39**	**76.22**

Table 4. Comparison of different distillation forms. The logit-based form works best.

Distillation Form. In Table 4, we compare the performance of feature- and logit-based distillation. In feature distillation, we align the features extracted by the teacher and student image encoders. Through careful hyperparameter tuning, we find that logit distillation yields significantly better results than the feature method. One possible reason is that the image feature space alignment is more difficult than the logit space alignment due to differences in the structure of the teacher and student models.

Distillation Method. In Table 5, we compare the performance of different distillation methods. "Projector Only" represents that there is only the projector module in the student image encoder and no learnable prompts. "Full Fine-tune" means that we fine-tune all parameters of the student CLIP model like CLIP-KD [44]. "w/o Shared Text Feature" means that we train the student model using its own text encoder along with learnable prompts to generate text features. The results indicate that the foundational designs of PromptKD, encompassing prompt-based distillation and shared class vectors from the teacher, play a crucial role in determining the ultimate performance.

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

Table 5. Ablation study of different distillation ways.

Teacher Pre-training Method. In Table 6, we conduct experiments employing various methods for teacher pre-training, such as vanilla CLIP and MaPLe. The table illustrates that a higher accuracy attained by the teacher model through pre-training aligns with the improved distillation performance of the student model. Notably, any type of teacher model can enhance the student model with a non-trivial improvement.

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

Table 6. Comparison of different pre-training methods. Teacher pre-training with PromptSRC brings the best student performance.

Distillation with Different Pre-trained Teachers. In this part, we investigate the impact of using teacher models with different capacities on the performance of the student models, as shown in Fig. 5. We adopt the ViT-B/16 and ViT-B/32 CLIP models using the official PromptSRC code and employ them as pre-trained teacher models. The results indicate that stronger teacher models lead to better performance in distillation.

Inference Cost Analysis. In Table 7, we show the inference cost analysis and compare it with other prompt learning methods including CoOp, CoCoOp, and PromptSRC. The inference cost for all methods is calculated on a single A100 GPU on SUN397. The results indicate that our method is more efficient than previous methods during inference, affirming its practicality in real-world applications.

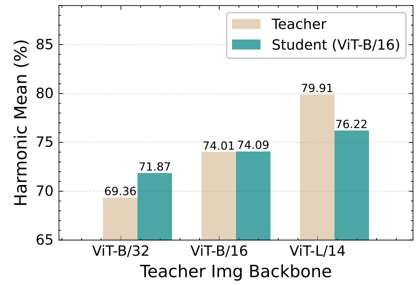

Figure 5. Comparison of distillation results for teachers with different capacities. Better teachers lead to better performance.

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

Table 7. Comparison of computation costs among existing methods on the SUN397 dataset. Our PromptKD is more efficient than previous methods during testing.

5. Conclusion

In this paper, we introduce a two-stage unsupervised prompt distillation framework for Vision-Language Models, which aims to transfer the knowledge of a large CLIP teacher model to a lightweight CLIP student model through prompt imitation using unlabeled domain data. Our method first pre-trains a large teacher model on domain few-shot labeled data and then performs student prompt distillation on extensive unlabeled domain data. By leveraging CLIP's unique decoupled-modality property, we propose to reuse pre-stored teacher text features and incorporate them into the student image encoder for both distillation and inference purposes. Extensive experiments on 11 recognition datasets demonstrate the effectiveness of our method.

Limitations and future work. The effectiveness of the distillation method is intricately tied to the knowledge transferred through unlabeled domain samples. When the distillation data lacks representation from the target domain, the generalization capability of the distilled student model towards that specific domain may be biased or weakened. In the future, we plan to explore potential regularization methods to mitigate these issues.

Acknowledgement. This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No.62206134), the National Science Fund of China (No. 62361166670), the Fundamental Research Funds for the Central Universities 070-63233084 and the Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP). Computation is supported by the Supercomputing Center of Nankai University (NKSC).

References

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

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

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

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

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

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

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

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

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

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

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 2, 3, 4

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

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

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1, 2, 3, 4, 5

[15] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 2, 3, 4, 5, 7

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

[17] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, pages 1401–1411, 2023. 2, 3

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

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

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

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

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

[23] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023. 3, 4

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

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

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

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

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

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

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3

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

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

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

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

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

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

[37] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019. 3

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

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

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

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

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

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

[44] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, and Yongjun Xu. Clip-kd: An empirical study of distilling clip models. *arXiv preprint arXiv:2307.12732*, 2023. 2, 3, 8

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

[46] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *NeurIPS*, 36, 2024. 1

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

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

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

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

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

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

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