Supplementary Material for: 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

Overview

This document contains Supplementary Material that provides additional details and further experimental analysis. The contents are organized as follows:

- **A. Additional Implementation Details**: This section provides additional implementation details, including hyperparameters and configuration settings for reproducing all reported results and further details about the datasets used.
- **B. Ablation Study**
  - B.1. **Pseudolabelling versus KDPL**: This subsection compares KDPL performances with a famous pseudolabeling prompt learning strategy demonstrating the superiority of KDPL over the baseline approach.
  - B.2. **Sampling Strategies for Class Agnostic Adaptation**: This part evaluates different sampling strategies for class agnostic adaptation and their performance varying the number of sampled class names.
  - B.3. **Symmetric versus Asymmetric KL**: This section explores the impact of using symmetric versus asymmetric KL-divergence in the loss function explaining why is more effective for class agnostic scenarios.
  - B.4. **Is a smaller Teacher enough?** Here, we investigate whether smaller teachers model can achieve competitive performance compared to the larger ViT-H-14 teacher model used in the main paper.
- **C. Additional Results**
  - C.1. **Generalization to Unseen Classes**: Here we present additional experimental results for the generalization to unseen classes within the same dataset scenario, including the downstream performance of few-shot base as well as the harmonic mean of base and unseen classes.
  - C.2. **Class Agnostic Scenario**: This section provides additional results for the class agnostic setting, our proposed new setting in which we know neither the labels nor the names of the training classes.

* These authors contributed equally to this work.

Table 1: Configuration of hyperparameters for the re-computed baselines. Note that CoOp and CoCoOp, being textual prompt learning methods, do not learn visual tokens.

Dataset	CoOp	CoCoOp	VPT	MaPLe	PromptSRC
Batch Size	32	1	4	4	4
Optimizer	SGD	SGD	SGD	SGD	SGD
LR	0.02	0.02	0.0025	0.0035	0.0025
Epochs	50	10	5	5	20
Text-tokens	4	4	4	4	4
Visual-tokens	—	—	8	2	4
Vision-depth	—	—	12	9	9

Table 2: Datasets statistics.

Dataset	Classes	Train	Val	Test
ImageNet [3]	1,000	1.28M	N/A	50,000
Caltech101 [4]	100	4,128	1,649	2,465
OxfordPets [15]	37	2,944	736	3,669
StanfordCars [12]	196	6,509	1,635	8,041
Flowers102 [14]	102	4,093	1,633	2,463
Food101 [1]	101	50,500	20,300	30,300
FGVCAircraft [13]	100	3,334	3,333	3,333
SUN397 [20]	397	15,880	3,970	19,850
DTD [2]	47	2,820	1,128	1,692
EuroSAT [5]	10	13,500	5,400	8,100
UCF101 [18]	101	7,639	1,898	3,783
ImageNetV2 [16]	1,000	N/A	N/A	10,000
ImageNet-S [19]	1,000	N/A	N/A	50,889
ImageNet-A [7]	200	N/A	N/A	7,500
ImageNet-R [6]	200	N/A	N/A	30,000

A. Additional Implementation Details

In this section we provide further details on hyperparameters of the proposed approaches presented in the main paper and additional details on the benchmark datasets evaluated.

Hyperparameter settings. For all evaluated scenarios we recalculated all baseline results using the parameter and hyperparameter settings reported in the original publications [9–11, 21, 22].

In Table 1 we summarize all configuration settings. Notably, all the evaluated baselines use a cosine scheduler with one warm-up epoch with a constant learning rate of 1e-5, and all authors suggest initializing the textual context with "a photo of a" as a better starting point. Note that our variants augmented with Knowledge Distillation Prompt Learning (KDPL) were trained using exactly the same parameters and configuration settings as the baseline methods. Therefore, it is reasonable to expect that further hyperparameter tuning could potentially improve our performance even further.

In addition, all the baselines – and so also our augmented versions – adhere to the same preprocessing pipeline in which input images are resized to 224×224 pixels using bicubic interpolation. Data augmentations including random resized crop, random flip, and normalization are used to ensure robustness.

To ensure consistency with baselines and to facilitate easy comparison and evaluation, our implementation is built upon the Dassl framework[3], which serves as the foundation for all these methods. We use ResNet-50 and ViT-B/32 as the backbones for all experiments. These backbones are the original publicly available models released by OpenAI.[4] As mentioned in the main paper, the code is publicly available at `https://github.com/miccunifi/KDPL`.

[3] `https://github.com/KaiyangZhou/Dassl.pytorch`
[4] `https://github.com/openai/CLIP`

Table 3: Ablations on labeling, sampling, and KL-divergence. **(a)** Comparison of KDPL and a pseudolabeling strategy adapted from UPL [8]. **(b)** Ablation on a number of classes sampled for class agnostic adaptation. **(c)** Comparison of forward, reverse, and symmetric KL-divergence used in the loss (see Eq. 2 of the main paper).

	Source	AVG^1	AVG^2
CLIP (student)	58.20	40.63	55.81
CoOp + UPL*	62.37	42.72	56.24
CoOp + KDPL	**62.73**	**42.90**	**57.59**
CoOp + CA-UPL*	60.13	42.48	55.73
CoOp + CA-KDPL	**61.83**	**43.02**	**57.69**

(a) KDPL vs. pseudolabeling

	Source	AVG^1	AVG^2
POMP*	60.80	42.37	57.35
CA-KDPL (K=100)	61.50	42.45	57.06
CA-KDPL (K=500)	61.73	42.68	57.38
CA-KDPL (K=1K)	**61.83**	43.02	**57.69**
CA-KDPL (K=2K)	61.60	**43.07**	57.51

(b) Sampling strategy

	Source	AVG^1	AVG^1
CoOp + KDPL F	**63.13**	**43.16**	57.33
CoOp + KDPL R	62.47	42.44	57.24
CoOp + KDPL	62.73	42.90	**57.59**
CoOp + CA-KDPL F	60.77	42.68	56.39
CoOp + CA-KDPL R	61.37	42.71	57.12
CoOp + CA-KDPL	**61.83**	**43.02**	**57.69**

(c) Symmetric/asymmetric KL

Dataset Details.. The detailed statistics of the 11 datasets used in the Cross-dataset transfer setting, as well as the four variants of ImageNet used in the Cross-domain generalization setting, are given in Table 2. Note that according to the authors of CoOp, for Caltech101, the "BACKGROUND Google" and "Faces easy" classes are discarded, and for the video dataset UCF101 the middle frame of each video is used as input to the image encoder [22].

B. Ablation Study

We performed a range of ablation studies to evaluate all aspects of KDPL. All ablations reported here are averages over three independent runs using CoOp as the baseline approach and ResNet-50 as the student backbone. The prompt is learned on ImageNet and evaluated on the benchmark datasets used in the domain generalization (AVG^1) and cross-dataset (AVG^2) evaluations.

B.1. Pseudolabeling versus KDPL

We compare KDPL and CA-KDPL with a pseudolabeling strategy for exploiting unlabeled samples. Similar to the approach outlined in UPL [8], we generate teacher-derived labels for the training samples. Instead of selecting the top-K confident pseudolabels over the entire training set, for a fair comparison we pseudolabel the fixed, few shot samples in each run. We call this baseline UPL*. Table 3(a) shows that KDPL consistently outperforms UPL* on all the evaluated scenarios.

B.2. Sampling Strategies for Class Agnostic Adaptation

An important detail in the class agnostic adaptation scenario is how class names are selected for computing the loss. Existing supervised approaches, such as POMP [17], to handling large numbers of classes suggest simply using the ground truth labels of batch samples and supplementing them with randomly selected ones to reach a predetermined number K. We adapted this technique and compare it with ours in Table 3(b). These results show that KDPL outperforms the random selection strategy denoted as POMP*. Notably, the optimal number of classes for domain generalization and cross-dataset adaptation is 1000, suggesting that the optimum closely aligns with the actual number of classes in the training dataset.

Table 4: Average accuracy comparison on the three evaluated scenarios between the supervised baseline CoOp and our unsupervised approach (CoOp + KDPL) using different and smaller teachers.

		Target downstream *Average*		
Method	Source	Domain Generalization	Cross-Dataset Transfer	Generalization to Unseen Classes
CoOp	62.40	42.05	55.28	60.82
CoOp + KDPL (B/16)	60.73	42.27	56.52	**67.57**
CoOp + KDPL (L/14)	61.43	42.61	57.10	66.90
CoOp + KDPL (H/14)	**62.73**	**42.90**	**57.59**	63.80

B.3. Symmetric versus Asymmetric KL

In Eq. 2 of the main paper we propose to use the symmetric KL-divergence, which is a sum of the forward (from teacher to student) and reverse (from student to teacher) KL-divergences. An asymmetric forward loss may yield benefits when the target distribution is similar to the source distribution (see Table 3(c)). Conversely, a reverse loss may prove better in the opposite scenario. If the Teacher outputs probability zero for a class on a training sample, the *forward KL-divergence* is zero for that sample even if the Student outputs a very high probability (see Eq. (3), main paper). Adding the *reverse KL-divergence* prevents the distillation loss from going to zero in such cases (Eq. (2), main paper). Indeed, in the class agnostic scenario the optimal strategy is the symmetric KL divergence. This enables learning from both in- and out-of-domain training samples, as demonstrated by the results in Figure 3 of the main paper.

B.4. Is a smaller Teacher enough?

All the results presented so far involve ViT-H-14 as the teacher. To evaluate if a smaller teacher would be enough we ran additional experiments with ViT-L/14 and ViT-B/16 as Teacher and ResNet-50 as Student (see Table 4). Larger Teachers yield better results, but even more modest ones still maintain excellent performance on domain generalization and cross-dataset transfer. Interestingly, smaller Teachers work *better* on generalization to unseen classes, which we believe is due to the weaker supervision causing less adaptation and thus maintaining better zero-shot transfer.

C. Additional Results

In the following sections we report on additional experiments validating our proposed approach.

C.1. Generalization to Unseen Classes

In the main paper, due to brevity and to maintain focus on the core aspects of our method, we reported only the performance metrics on unseen classes. Here we provide the performance metrics on base classes as well as the harmonic mean of both base and unseen classes to provide a more comprehensive perspective.

Table 5: Base-to-new generalization. Prompts are learned from the base classes (16-shots) and evaluated on the same classes on the test set (Base), zero-shot on the unseen classes of the test set (New), and in terms of harmonic mean (H) between the base and the new accuracy. Average *New* performance improvements over the baselines are indicated in **green**. Here we see how KDPL generalizes better to the new classes, limiting deterioration in performance caused by the unbalanced few-shot training on half the classes of the source domain. Remarkably, we are competitive even on the base classes without requiring training labels.

		Average			ImageNet			OxfordPets			Flowers102		
		Base	**New**	H	Base	New	H	Base	New	H	Base	New	H
RN50	CLIP (student)	61.61	65.88	63.67	64.40	60.10	58.20	85.80	93.70	83.70	63.80	71.30	61.00
	CoOp	**77.13**	60.82	68.01	**68.40**	60.00	60.30	90.80	91.63	85.03	**95.17**	58.30	67.03
	CoOp + KDPL	73.66	**63.80**	**68.38**	68.37	**60.53**	**60.63**	**93.13**	**94.97**	**88.63**	93.93	**63.40**	**69.77**
	CoCoOp	**75.47**	64.39	**69.49**	**68.30**	**63.07**	**61.90**	92.13	94.93	87.73	**91.10**	67.47	68.37
	CoCoOp + KDPL	71.69	**65.90**	68.67	68.13	62.87	61.63	**93.40**	**95.33**	**89.07**	87.77	**70.63**	**70.43**
ViT-B/32	CLIP (student)	64.81	70.05	67.33	67.40	64.00	62.00	86.90	96.30	85.00	68.70	72.30	64.30
	CoOp	**79.02**	64.52	**71.04**	**70.90**	64.53	**63.90**	91.90	93.37	84.77	**95.13**	**59.77**	**67.30**
	CoOp + KDPL	75.20	**66.35**	70.50	70.57	**64.90**	63.83	**93.67**	**94.00**	**87.57**	93.47	56.17	65.23
	VPT	**75.15**	68.02	**71.41**	70.30	64.60	63.85	**94.03**	94.83	**85.87**	**88.10**	66.03	**67.73**
	VPT + KDPL	72.12	**68.76**	70.40	**70.70**	**65.20**	**64.30**	93.80	**95.20**	85.53	82.73	**67.87**	67.30
	MaPLe	**78.26**	69.79	**73.79**	71.40	66.70	65.40	93.07	**96.87**	90.03	**93.37**	**69.77**	**71.83**
	MaPLe + KDPL	74.74	**71.41**	73.03	**71.60**	**66.73**	**65.53**	**93.90**	**96.87**	**90.43**	89.73	68.23	69.70
	PromptSRC	**81.17**	70.45	**75.43**	72.50	65.93	65.63	93.40	96.30	89.30	**96.20**	**71.37**	**74.90**
	PromptSRC + KDPL	77.11	**71.61**	74.26	**72.70**	**66.33**	**65.93**	**94.37**	**96.47**	**90.17**	94.93	68.83	73.10
ViT-H/14	CLIP (teacher)	86.11	86.63	86.37	86.50	84.10	82.80	94.40	99.20	94.80	96.60	85.60	89.40

		FGVCAircraft			DTD			EuroSAT			StanfordCars		
		Base	New	H	Base	New	H	Base	New	H	Base	New	H
RN50	CLIP (student)	17.20	24.80	15.50	49.00	53.90	40.00	39.20	43.70	24.20	55.40	66.60	55.60
	CoOp	**27.93**	**22.17**	**18.90**	**75.03**	**40.93**	**47.13**	**89.30**	42.13	46.07	67.77	**59.87**	**58.30**
	CoOp + KDPL	24.33	21.27	17.03	66.73	38.70	42.70	71.17	**62.07**	**47.23**	**67.93**	56.63	56.70
	CoCoOp	**24.27**	**24.87**	**18.30**	**72.37**	45.40	**45.73**	**87.73**	35.10	**42.00**	**63.43**	**64.73**	**58.87**
	CoCoOp + KDPL	21.77	22.67	16.93	62.70	**49.07**	44.07	69.73	**44.27**	37.67	62.40	63.67	57.53
ViT-B/32	CLIP student	20.30	28.30	18.20	53.20	53.90	42.80	43.40	61.50	38.10	61.00	69.80	60.40
	CoOp	**30.80**	**23.03**	**20.27**	**77.17**	45.57	**49.50**	**88.37**	49.13	**51.23**	**71.13**	61.73	**61.17**
	CoOp + KDPL	26.67	22.00	18.23	67.13	**47.33**	44.10	71.40	**62.07**	46.90	69.37	**62.37**	60.47
	VPT	**25.23**	29.70	**20.77**	**72.93**	**50.17**	**46.27**	**76.27**	50.03	**43.87**	**64.07**	**69.90**	**61.27**
	VPT + KDPL	24.10	**30.40**	20.03	63.40	47.53	42.07	65.43	**54.97**	42.40	63.77	69.80	61.23
	MaPLe	23.57	20.43	16.50	**77.27**	**53.50**	**53.90**	**91.17**	66.70	**63.93**	**67.50**	68.03	62.50
	MaPLe + KDPL	**25.43**	**27.37**	**19.93**	66.27	50.33	46.53	74.03	**75.70**	62.30	66.97	**68.40**	**62.40**
	PromptSRC	**33.77**	23.67	22.50	**80.57**	**54.43**	**54.67**	**93.77**	61.97	**61.30**	**72.90**	**70.00**	**66.37**
	PromptSRC + KDPL	30.37	**28.43**	**23.20**	71.93	50.87	47.97	73.93	**74.17**	58.53	72.53	68.50	65.43
ViT-H/14	CLIP (teacher)	71.40	64.30	63.20	78.20	75.00	66.80	70.70	82.60	63.30	94.60	98.20	95.60

		Food101			SUN397			Caltech101			UCF101		
		Base	New	H	Base	New	H	Base	New	H	Base	New	H
RN50	CLIP (student)	81.60	82.20	75.20	66.50	70.10	58.50	91.10	90.50	86.00	63.70	67.80	58.30
	CoOp	82.63	82.03	75.37	**76.37**	66.77	61.50	**95.50**	87.53	88.53	**79.50**	57.63	**61.53**
	CoOp + KDPL	**83.73**	**83.07**	**76.77**	73.43	**69.83**	**61.93**	95.10	**88.75**	**88.90**	72.43	**62.63**	59.97
	CoCoOp	84.43	84.50	78.00	**74.53**	72.87	**63.93**	**94.97**	90.73	89.33	**76.90**	64.67	**63.77**
	CoCoOp + KDPL	**84.63**	**85.90**	**79.00**	71.83	**73.23**	62.80	94.87	**91.00**	**89.37**	71.33	**66.23**	61.70
ViT-B/32	CLIP student	84.50	85.70	79.10	69.80	73.10	62.00	93.70	94.00	91.10	64.00	71.60	60.70
	CoOp	85.17	85.50	79.27	**79.17**	69.37	64.93	**97.40**	92.13	92.37	**82.13**	65.60	**66.93**
	CoOp + KDPL	**86.40**	**87.23**	**81.10**	75.87	**74.17**	**65.70**	96.47	**93.43**	**92.57**	76.17	**66.20**	64.20
	VPT	85.47	85.83	79.60	**75.70**	75.13	**65.47**	**97.10**	92.57	**92.80**	**77.43**	69.47	**65.10**
	VPT + KDPL	**85.63**	**86.33**	**79.97**	74.07	**75.17**	65.07	96.33	**92.80**	92.53	73.33	**71.07**	63.70
	MaPLe	86.40	87.47	81.27	**78.90**	76.57	**68.60**	**97.03**	92.27	93.00	**81.23**	69.40	**68.17**
	MaPLe + KDPL	**86.50**	**87.83**	**81.73**	75.57	**76.83**	67.07	97.00	**93.67**	**93.33**	75.10	**73.53**	67.07
	PromptSRC	86.33	87.13	80.93	**80.80**	76.77	**69.90**	**97.53**	**94.70**	**94.30**	**85.07**	72.63	**72.67**
	PromptSRC + KDPL	**86.53**	**87.63**	**81.27**	76.53	**77.33**	68.00	96.90	94.47	93.93	77.50	**74.73**	68.53
ViT-H/14	CLIP (teacher)	95.50	96.30	93.60	82.00	85.10	76.40	99.20	97.30	97.90	78.10	85.20	76.40

Table 6: Class Agnostic adaptation. Comparison between baselines and the proposed unsupervised 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 generalization and on the cross-dataset benchmark datasets. Average performance improvements are indicated in **green**, and deterioration in **red**. Notably, our proposed unsupervised and class-agnostic approach is competitive against the reference supervised and class-aware baselines. Overall, on the cross-dataset benchmark we achieve better generalization, demonstrating that using CA-KDPL is a valid option in such settings.

		Source	Cross-Domain Target					Cross-Dataset Target										
Backbone	Method	*ImageNet*	*ImageNet-V2*	*ImageNet-S*	*ImageNet-A*	*ImageNet-R*	*Average*	*OxfordPets*	*Flowers102*	*FGVCAircraft*	*DTD*	*EuroSAT*	*StanfordCars*	*Food101*	*SUN397*	*Caltech101*	*UCF101*	*Average*
RN50	CLIP (student)	58.20	51.50	33.30	21.70	56.00	40.63	83.70	61.00	15.60	40.00	24.20	55.60	75.20	58.50	86.00	58.30	55.81
	CoOp	**62.40**	**55.17**	33.70	23.13	56.20	42.05	84.60	61.63	13.77	36.83	22.33	54.20	75.03	59.00	87.67	57.73	55.28
	CoOp + CA-KDPL	61.83	54.27	**35.20**	**23.50**	**59.10**	**43.02**	**86.37**	**63.47**	**14.80**	**40.23**	**30.33**	**55.40**	**77.17**	**61.60**	**88.50**	**59.03**	**57.69**
	CoCoOp	**63.07**	**55.53**	34.77	23.73	**59.47**	43.38	86.93	63.47	**16.27**	**40.53**	27.20	**55.47**	**77.73**	61.40	**88.07**	**60.43**	57.75
	CoCoOp + CA-KDPL	61.43	54.53	**35.33**	**23.77**	59.10	**43.18**	**87.47**	**67.30**	15.30	39.50	**28.57**	54.80	77.53	**61.63**	88.03	60.13	**58.03**
ViT-B/32	CLIP (student)	62.00	54.70	40.80	29.60	66.00	47.78	85.00	64.30	18.20	42.80	38.10	60.40	79.10	62.00	91.10	60.70	60.17
	CoOp	**66.33**	**58.30**	41.40	**31.47**	65.87	49.26	**87.13**	62.20	12.13	40.57	36.97	57.83	**80.13**	62.83	91.13	**61.80**	59.27
	CoOp + CA-KDPL	64.73	56.93	**42.27**	**31.47**	**67.83**	**49.63**	86.57	**62.73**	**15.00**	**42.23**	**43.93**	**58.57**	79.47	**65.03**	**92.30**	61.53	**60.74**
	VPT	**64.97**	**56.73**	**41.27**	27.00	66.50	47.88	**87.43**	**64.77**	**19.53**	**44.20**	28.47	**57.60**	**78.27**	**63.73**	91.40	62.50	59.79
	VPT + CA-KDPL	63.73	55.83	41.10	**27.33**	**67.30**	**47.89**	87.20	63.17	18.43	44.00	**33.53**	55.47	77.70	63.63	**91.93**	**62.93**	**59.80**
	MaPLe	**66.80**	**58.53**	**42.23**	**30.13**	66.40	49.32	**88.37**	**66.43**	**18.47**	42.03	37.77	**60.00**	**80.10**	64.43	91.33	**62.50**	61.14
	MaPLe + CA-KDPL	65.23	57.43	42.07	30.00	**67.63**	**49.28**	88.30	64.97	16.07	**44.03**	**41.53**	57.90	79.80	**64.97**	**92.77**	61.93	**61.23**
	PromptSRC	**66.33**	**58.70**	**42.97**	32.07	**68.93**	50.67	**88.13**	65.33	17.17	**43.90**	40.50	**60.17**	**81.23**	65.40	92.37	**63.63**	61.78
	PromptSRC + CA-KDPL	65.00	57.50	42.90	**32.30**	68.77	**50.37**	87.53	**65.47**	**17.87**	42.17	**44.53**	60.07	80.50	**65.67**	**93.10**	62.37	**61.93**
ViT-H/14	CLIP (teacher)	82.80	76.60	71.10	71.10	91.30	77.53	94.80	89.40	63.20	66.80	63.30	95.60	93.60	76.40	97.90	76.40	81.74

In Table 5, we observe that the baseline methods without KDPL outperform our approach in few-shot transfer learning on base classes. Thanks to the supervision label signal, the baselines can exploit the ground-truth source training classes. Without using KDPL, however, all the baseline methods suffer from forgetting the unseen classes. In particular, we see that training on base-shots causes a decrease in performance on the unseen classes. Certain methods in particular suffer from this phenomenon, typical of an Incremental Learning setting (e.g. in Figure 1 we see that the average performance of MaPLe on unseen classes is consistently *below* the average zero-shot unseen performance). Applying our method instead exhibits a more favorable trend in zero-shot transfer learning on unseen classes, alleviating the forgetting of unseen classes. Our method is preferable in scenarios where labeled data for base classes are *not* available, and improvement in base class performance is desired without compromising zero-shot transfer performance on unseen classes.

C.2. Class Agnostic Scenario

In the main paper we did not include all results for the class agnostic setting due to space limitations, but only reported the average domain generalization performance and the average cross-dataset performance. Here, in Table 6 we provide all class agnostic results comparing our proposed method (CA-KDPL) against the baselines. As mentioned in the main paper, even without knowing either the labels or the training class names, our proposed unsupervised method is competitive, and sometimes even outperforms the baselines.

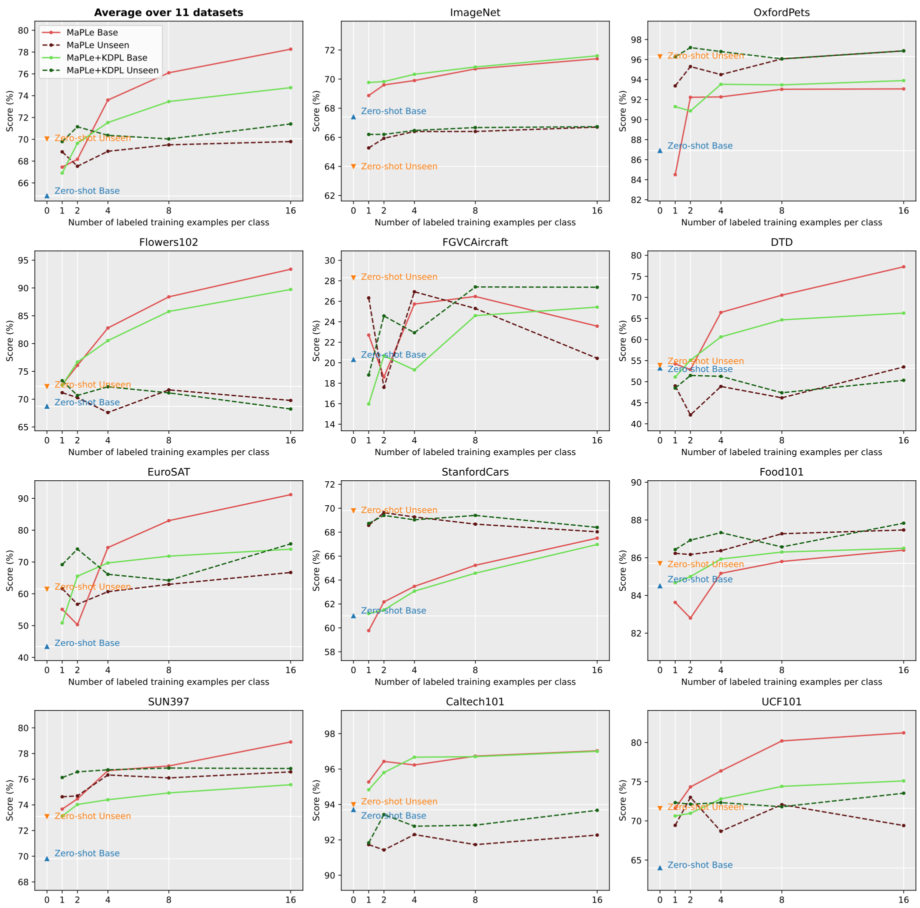

Fig. 1: Generalization to Unseen Classes. The base-to-unseen generalization results of MaPLe+KDPL versus MaPLe are reported. Prompts are learned from the base classes and evaluated on the same base classes and on the unseen classes of the test set. Instead of reporting only the 16-shot performance, we include results for 1-, 2-, 4-, 8-, and 16-shots. Here we see that increasing the number of shots increases the performance of the base class but harms the performance on unseen classes. However, with our approach we can alleviate this problem, reaching an average 16-shots performance on unseen classes greater than the zero-shot reference, while the average MaPLe performance remains below the average zero-shot reference for all shots.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649 (2022)
9. 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)
10. 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)
11. 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)
12. 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)
13. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
14. 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)
15. 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)

16. 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)
17. 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)
18. 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)
19. 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)
20. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
21. 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)
22. 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)