Visual-Language Prompt Tuning with Knowledge-guided Context Optimization

Hantao Yao[1],Rui Zhang[2], Changsheng Xu[1,3]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS
[2] State Key Lab of Processors, Institute of Computing Technology, CAS; [3] University of Chinese Academy of Sciences(CAS),
{hantao.yao,csxu}@nlpr.ia.ac.cn;zhangrui@ict.ac.cn

Abstract

Prompt tuning is an effective way to adapt the pretrained visual-language model (VLM) to the downstream task using task-related textual tokens. Representative CoOp-based work combines the learnable textual tokens with the class tokens to obtain specific textual knowledge. However, the specific textual knowledge is worse generalization to the unseen classes because it forgets the essential general textual knowledge having a strong generalization ability. To tackle this issue, we introduce a novel Knowledge-guided Context Optimization (KgCoOp) to enhance the generalization ability of the learnable prompt for unseen classes. The key insight of KgCoOp is that the forgetting about essential knowledge can be alleviated by reducing the discrepancy between the learnable prompt and the hand-crafted prompt. Especially, KgCoOp minimizes the discrepancy between the textual embeddings generated by learned prompts and the hand-crafted prompts. Finally, adding the KgCoOp upon the contrastive loss can make a discriminative prompt for both seen and unseen tasks. Extensive evaluation of several benchmarks demonstrates that the proposed Knowledge-guided Context Optimization is an efficient method for prompt tuning, i.e., *achieves better performance with less training time. code.*

1. Introduction

With the help of the large scale of the image-text association pairs, the trained visual-language model (VLM) contains essential general knowledge, which has a better generalization ability for the other tasks. Recently, many visual-language models have been proposed, such as Contrastive Language-Image Pretraining (CLIP) [29], Flamingo [1], ALIGN [18], *etc*. Although VLM is an effective model for extracting the visual and text description, training VLM needs a large scale of high-quality datasets. However, collecting a large amount of data for training a task-related model in real visual-language tasks is difficult. To address the above problem, the prompt tuning [4] [10] [19] [22] [28] [30] [33] [38] has been proposed to adapt the pretrained VLM to downstream tasks, achieving a fantastic performance on various few-shot or zero-shot visual recognization tasks.

Table 1. Compared to existing methods, the proposed KgCoOp is an efficient method, obtaining **a higher performance with less training time**.

Methods	Prompts	Accuracy			Training-time
		Base	New	H	
CLIP	hand-crafted	69.34	74.22	71.70	-
CoOp	textual	**82.63**	67.99	74.60	6ms/image
ProGrad	textual	82.48	70.75	76.16	22ms/image
CoCoOp	textual+visual	80.47	71.69	75.83	160ms/image
KgCoOp	textual	80.73	**73.6**	**77.0**	**6ms**/image

The prompt tuning[1] usually applies task-related textual tokens to embed task-specific textual knowledge for prediction. The hand-crafted template "a photo of a [Class]" in CLIP [29] is used to model the textual-based class embedding for zero-shot prediction. By defining the knowledge captured by the fixed (hand-crafted) prompts as the *general textual knowledge*[2], it has a high generalization capability on unseen tasks. However, the general textual knowledge is less able to describe the downstream tasks due to not consider the specific knowledge of each task. To obtain discriminative task-specific knowledge, Context Optimization(CoOp) [41], Conditional Context Optimization(CoCoOp) [40], and ProGrad [42] replace the hand-crafted prompts with a set of learnable prompts inferred by the labeled few-shot samples. Formally, the discriminative knowledge generated by the learned prompts is defined as the *specific textual knowledge*. However, CoOp-based methods have a worse generalization to the unseen classes with the same task, *e.g.,* obtaining a worse performance than CLIP for the unseen classes (*New*), shown in Table 1.

As the specific textual knowledge is inferred from the

[1] In this work, we only consider the textual prompt tuning and do not involve the visual prompt tuning.

[2] Inspired from [17], 'knowledge'in this work denotes the information contained in the trained model.

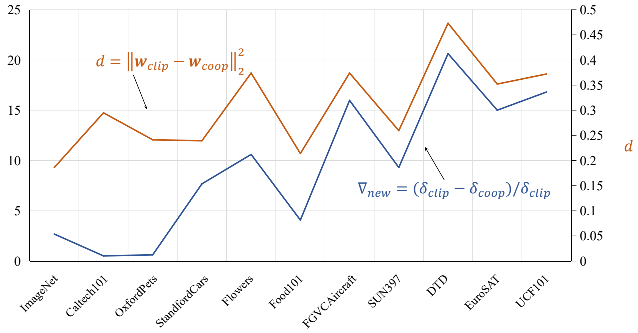

Figure 1. For the CoOp-based prompt tuning, the degree of performance degradation $\triangledown_{new}$ on the *New* classes is consistent with the distance between the learnable textual embedding $\mathbf{w}_{coop}$ and the hand-crafted textual embedding $\mathbf{w}_{clip}$. The larger distance, the more severe the performance degradation. σ_{clip} and σ_{coop} are the accuracy of *New* classes for CLIP and CoOp, respectively.

labeled few-shot samples, it is discriminative for the seen classes and biased away from the unseen classes, leading to worse performance on the unseen domain. For example, non-training CLIP obtains a higher *New* accuracy on the unseen classes than CoOp-based methods, *e.g.,* 74.22%/63.22%/71.69% for CLIP/CoOP/CoCoOp. The superior performance of CLIP on unseen classes verifies that its general textual knowledge has a better generalization for unseen classes. However, the specific textual knowledge inferred by the CoOp-based methods always forgets the essential general textual knowledge, called catastrophic knowledge forgetting, *i.e.,* the more serve catastrophic forgetting, the larger performance degradation.

To address this issue, we introduce a novel prompt tuning method Knowledge-guided Context Optimization (KgCoOp) to boost the generality of the unseen class by reducing the forgetting of the general textual knowledge. The key insight of KgCoOp is that the forgetting about general textual knowledge can be alleviated by reducing the discrepancy between the learnable prompt and the hand-crafted prompt. The observation of relationship between the discrepancy of two prompts and the performance drop also verify the insight. As shown in Figure 1, the larger the distance between textual embeddings generated by the learnable prompt and the hand-crafted prompt, the more severe the performance degradation. Formally, the hand-crafted prompts "a photo of a [Class]" are fed into the *text encoder* of CLIP to generate the general textual embedding, regarded as the general textual knowledge. Otherwise, a set of learnable prompts is optimized to generate task-specific textual embedding. Furthermore, Knowledge-guided Context Optimization(KgCoOp) minimizes the euclidean distance between general textual embeddings and specific textual embeddings for remembering the essential general textual knowledge. Similar to the CoOp and CoCoOp, the contrastive loss between the task-specific textual and visual embeddings is used to optimize the learnable prompts.

We conduct comprehensive experiments under base-to-new generalization setting, few-shot classification, and domain generalization over 11 image classification datasets and four types of ImageNets. The evaluation shows that the proposed KgCoOp is an efficient method: **using the less training time obtains a higher performance,** shown in Table 1. In summary, the proposed KgCoOp obtains: 1) higher performance: KgCoOp obtains a higher final performance than existing methods. Especially, KgCoOp obtains a clear improvement on the *New* class upon the CoOp, CoCoOp, and ProGrad, demonstrating the rationality and necessity of considering the general textual knowledge. 2) less training time: the training time of KgCoOp is the same as CoOp, which is faster than CoCoOp and ProGrad.

2. Related Work

Visual-Language Models: Recently, research has shown that using image-text association pairs can model a powerful visual-language model rather than merely considering the images. The model inferred based on the image-text association pairs is defined as Visual-Language Model (VLM). Recently, the visual-language models can be improved from the following aspects: 1) using a stronger text encoder or visual encoder, *e.g.,* Transformers [34]; 2) contrastive representation learning [3]; 3) using more images [29] [18]. As training VLM needs a large-scale annotated dataset, unsupervised learning or weakly supervised learning [36] are used to train the visual-language model with the unannotated images. Specially, Masked Language Modeling (MLM) [20] [23] improves the robustness of visual and text embedding by randomly erasing the words in the text, and Masked autoencoders [13] is a scalable self-supervised learner by masking random patches of the input image. As representive work is CLIP, which trains the visual encoder and visual encoder using the contrastive loss based on 400 millions image-text association pairs, which demonstrates a good generability for the unseen classes. Similar to the previous work CoOp and CoCoOp, we apply the pretrained CLIP for knowledge transfer.

Prompt Tuning: To adapt the pretrained VLM to the downstream tasks, the prompt tuning [10] always applies task-related textual tokens to infer the task-specific textual knowledge [29,39]. For example, the hand-crafted template "a photo of a [CLASS]" in CLIP [29] is used to model the textual embedding for zero-shot prediction. However, the hand-crafted prompts have less ability to describe the downstream task because they do not consider the specific knowledge of the current task. To address the above problem, Context Optimization(CoOp) [41] replaces the hand-crafted prompts with the learnable soft prompts inferred by the labeled few-shot samples. The disadvantage of CoOp is that the learnable prompts are unique and fixed for each task's images. That is to say, CoOP infers task-related prompts and ignores the characteristics of different images. Furthermore, Conditional Context Optimization(CoCoOp) [40] is proposed to generate an image-conditional context for each image and combine the textual-conditional context for prompt tuning. Specialy, it uses a lightweight neural network to generate a vector, which is learnable text prompts. To obtain high-quality task-related tokens, ProDA [24] considers the prompt's prior distribution learning. Furthermore, ProGrad [42] only updates the prompts whose gradient is aligned to the "general knowledge" generated by the original prompts. DenseCLIP [30] uses the context-aware prompt strategy to generate dense prediction tasks, and CLIP-Adapter [12] applies an adapter to adjust the visual or text embeddings.

Among existing methods, the most related to ours are the CoOp and ProGrad. The CoOp can be treated as the baseline model for the proposed KgCoOp. Compared with CoOp, the proposed KgCoOp considers an additional term to ensure learnable prompts have a low discrepancy with the original prompts, leading the proposed KgCoOp to obtain a higher performance on the terms of unseen classes than CoOp. ProGrad has the same idea as the proposed KgCoOp, ensuring that the learnable specific knowledge is aligned with the general knowledge. However, ProGrad only optimizes the prompts with the aligned direction and discards a conflicting update. That is to say, ProGrad discards a lot of the conflict knowledge during prompt tuning. Unlike ProGrad, the proposed KgCoOp will not discard any knowledge and only ensures that the learnable specific knowledge is close to the general knowledge. Furthermore, KgCoOp is more efficient than ProGrad because it does not need additional computation. The comprehensive evaluation shows that the proposed KgCoOp is an efficient method: using less training time obtains a higher performance.

3. Methodolgy

As Knowledge-guided Context Optimization(KgCoOp) is proposed based on Context Optimization (CoOp), we first give a brief review of Context Optimization (CoOp) for visual-language prompt tuning. Then, we give a detailed introduction to the proposed KgCoOp.

3.1. Preliminaries

Among the existing visual-language models, Contrastive Language-Image Pre-training(CLIP) is a representative model trained with 400 million image-text association pairs, having a powerful generability for zero-shot image recognition. Since CLIP is trained based on the image-text association pairs, it contains two types of encoders: visual encoder, and textual encoder, where the visual encoder is used to map the given image into the visual embedding, and the textual encoder is applied to embedding the corresponding textual information. By fixing the pretrained visual and textual encoders in CLIP, the prompt tuning uses the hand-crafted prompts or the learnable prompts for adapting the pre-trained CLIP to downstream tasks.

Formally, we define the visual encoder and textual encoder as ϕ and θ, respectively. For a downstream task consisting of N_c categories, CLIP employs a hand-crafted prompt to generate the textual class embeddings, *i.e.,* $\mathbf{W}^{clip} = \{\mathbf{w}_i^{clip}\}_{i=1}^{N_c}$ denotes the textual embedding of all categories, where $\mathbf{w}_i^{clip}$ denotes the textual embedding of the i-th class. Specifically, assuming the name of the i-th class as "class-name", the corresponding textual embedding $\mathbf{w}_i^{clip}$ is generated from a hand-crafted prompt: "a photo of a [class-name]" with the textual encoder $\theta(\cdot)$ and a transformer-based encoder $e(\cdot)$, where $e(\cdot)$ takes a sequence of words as input and outputs a vectorized textual tokens. Formally, the vectorized textual tokens of the i-th class template "a photo of a [class-name]" is defined as: $\mathbf{t}_i^{clip} = e($"a photo of a [class-name]"$)$. $\mathbf{t}_i^{clip}$ is further project to the textual class embedding $\mathbf{w}_i^{clip}$ with the textual encoder θ: $\mathbf{w}_i^{clip} = \theta(\mathbf{t}_i^{clip})$.

Given an image I along with its label y, the visual embedding is extracted with the visual encoder $\phi(\cdot)$: $\mathbf{x} = \phi(I)$. After that, the prediction probability between the visual embedding $\mathbf{x}$ and textual embedding $\mathbf{W}^{clip}$ is computed for prediction:

$$p(y|\mathbf{x}) = \frac{\exp(d(\mathbf{x}, \mathbf{w}_y^{clip})/\tau)}{\sum_{i=1}^{N_c} \exp(d(\mathbf{x}, \mathbf{w}_i^{clip})/\tau)}, \tag{1}$$

where $d(\cdot)$ denotes the cosine similarity, and τ is a learnable temperature parameter.

Although Eq.(1) can be easily applied for zero-shot prediction, CLIP employs a fixed hand-crafted prompt("a photo of a []") to generate the textual embedding, leading to weak generability to the downstream tasks. To address the above problem, Context Optimization (CoOp) automatically learns a set of continuous context vectors for generating task-related textual embeddings. Specifically, CoOp introduces M context vectors $\mathbb{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M\}$ as the learnable prompt. Finally, the corresponding class token embedding $\mathbf{c}_i$ of the i-th class is concatenated with the learnable context vector $\mathbb{V}$ for generating the prompts $\mathbf{t}_i^{coop} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M, \mathbf{c}_i\}$. After that, the textual class embedding $\mathbf{w}_i^{coop}$ is obtained by fedding the learnable prompts $\mathbf{t}_i^{coop}$ into the textual encoder θ, *i.e.*, $\mathbf{w}_i^{coop} = \theta(\mathbf{t}_i^{coop})$. Therefore, the final textual class embedding for all class is defined as: $\mathbf{W}^{coop} = \{\mathbf{w}_i^{coop}\}_{i=1}^{N_c}$.

With the given few-shot samples, CoOp optimizes the learnable context tokens $\mathbb{V}$ by minimizing the negative log-likelihood between the image feature $\mathbf{x}$ and its class textual embeding $\mathbf{w}_y^{coop}$:

$$p_{coop}(y|\mathbf{x}) = \frac{\exp(d(\mathbf{x}, \mathbf{w}_y^{coop})/\tau)}{\sum_{i=1}^{N_c} \exp(d(\mathbf{x}, \mathbf{w}_i^{coop})/\tau)}. \tag{2}$$

Note that the visual encoder ϕ and the pretrained textual encoder θ are frozen during training for CLIP and CoOp. Different from CLIP using the fixed prompts, CoOp only infers the suitable task-related prompts $\mathbf{t}_i^{coop}$ to boost its generability and discrimination.

3.2. Knowledge-guided Context Optimization

Although existing CoOp-based prompt tuning methods can effectively adapt the pretrained CLIP to the downstream tasks, it might easily overfit the seen classes because only a few labeled images are used for training. For example, CoOp obtains a noticeable improvement for the *Base* accuracy upon CLIP, *e.g.,* 69.34%(CLIP) vs 82.89%(CoOp). However, CoOp obtains a worse *New* accuracy than CLIP on the unseen classes, *e.g.,* 74.22%(CLIP) vs 63.22%(CoOp). By further analyzing the *New* accuracy between CLIP and CoOp on all 11 datasets, an interesting phenomenon is that the performance degradation on the unseen classes is consistent with the distance between the learnable prompts and fixed prompts. In this work, the relative ratio of performance drop ∇_{new} between CLIP and CoOp indicates the degree of performance degradation. Moreover, the distance between learnable textual embedding (CoOp) and fixed textual embedding(CLIP) is used to measure the similarity between the two types of prompts. As shown in Figure 1, the larger distance, the more severe the performance drop. For example, among all 11 datasets, CoOp obtains the largest drop ratio of 20.63% on the DTD dataset, while its special class embeddings also have the largest distance compared to CLIP ones. Based on the above results, we can conclude that enhancing the similarity between the learnable prompt and fixed prompts can alleviate the forgetting of general textual knowledge for boosting the generability of the unseen domain, which is the core motivation of our work. Formally, we propose a novel prompt tuning method named Knowledge-guided Context Optimization (KgCoOp) to infer learnable prompts which have a high discriminative on the seen classes and high generability on the unseen classes, shown in Figure 2.

For CLIP, given an image I along with its embedding $\mathbf{x}$, the predictions are obtained by computing the visual-textual similarity between the visual embedding and textual class embeddings. Since CLIP and KgCoOp apply different textual embeddings to match the visual embeddings, the general textual knowledge and special textual knowledge are majorly controlled by the textual embeddings of CLIP and KgCoOp. Furthermore, the discrepancy between general textual knowledge and special textual knowledge can be measured by the distance between the corresponding textual embeddings.

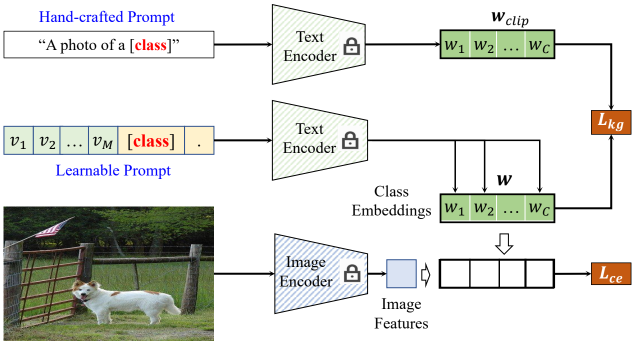

Figure 2. The framework of the Knowledge-guided Context Optimization for prompt tuning. $\mathcal{L}_{ce}$ is the standard cross-entropy loss, and $\mathcal{L}_{kg}$ is the proposed Knowledge-guided Context Optimization contraint to minimize the discrepancy between the special knowledge (learnable textual embeddings) and the general knowledge(the textual embeddings generated by the hand-crafted prompt).

Formally, we define the textual embedding generated by the CLIP and KgCoOp as $\mathbf{w}_i^{clip} = \theta(\mathbf{t}_i^{clip})$ and $\mathbf{w}_i = \theta(\mathbf{t}_i)$, where $\mathbf{t}_i^{clip}$ is the vectorized textual tokens in CLIP, and $\mathbf{t}_i = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M, \mathbf{c}_i\}$ denotes the learnable prompt of the i-th class. The discrepancy between the special knowledge and general knowledge is to compute the euclidean distance between $\mathbf{w}_i$ and $\mathbf{w}_i^{clip}$, As shown in Figure 1, the distance is positively related to the performance degradation, and a lower distance produces a lower performance degradation. Therefore, we can minimize the distance between $\mathbf{w}_i$ and $\mathbf{w}_i^{clip}$ for boosting the generability of the unseen classes,

$$\mathcal{L}_{kg} = \frac{1}{N_c} \sum_{i=1}^{N_c} ||\mathbf{w}_i - \mathbf{w}_i^{clip}||_2^2, \tag{3}$$

where $||\cdot||$ is the euclidean distance, N_c is the number of seen classes. Meanwhile, the standard contrastive loss is:

$$\mathcal{L}_{ce} = -\sum_{\mathbf{x} \in \mathbf{X}} \log \frac{\exp(d(\mathbf{x}, \mathbf{w}_y)/\tau)}{\sum_{i=1}^{N_c} \exp(d(\mathbf{x}, \mathbf{w}_i)/\tau)}, \tag{4}$$

where y is the corresponding label of the image embedding.

By combining the standard cross-entropy loss $\mathcal{L}_{ce}$, the final objective is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{kg}, \tag{5}$$

where λ is used balance the effect of $\mathcal{L}_{kg}$.

4. Experiments

Similar to CoCoOp [40] and ProGrad [42], we evaluate the proposed method based on the following settings: 1)

Table 2. Comparison in the base-to-new setting with different K-shot samples in terms of the average performance among all 11 datasets and backbones(ViT-B/16 and ResNet-50).

Backbones	Methods	K=4			K=8			K=16		
		Base	New	H	Base	New	H	Base	New	H
ViT-B/16	CoOp	78.43	68.03	72.44	80.73	68.39	73.5	82.63	67.99	74.60
	CoCoOp	76.72	**73.34**	74.85	78.56	72.0	74.9	80.47	71.69	75.83
	ProGrad	79.18	71.14	74.62	**80.62**	71.02	75.2	**82.48**	70.75	76.16
	KgCoOp	**79.92**	73.11	**75.90**	78.36	**73.89**	**76.06**	80.73	**73.6**	**77.0**
ResNet-50	CoOp	72.06	59.69	65.29	74.72	58.05	65.34	77.24	57.4	65.86
	CoCoOp	71.39	65.74	68.45	73.4	66.42	69.29	75.2	63.64	68.9
	ProGrad	**73.88**	64.95	69.13	**76.25**	64.74	70.03	**77.98**	64.41	69.94
	KgCoOp	72.42	**68.00**	**70.14**	74.08	**67.86**	**70.84**	75.51	**67.53**	**71.30**

generalization from base-to-new classes within a dataset; 2) few-shot image classification; 3) domain generalization. All experiments are conducted based on the pretrained CLIP [29] model. More detailed results will be given in the Supplementary materials.

Dataset: Following CLIP [29], CoOp [41], CoCoOp [40], and ProGrad [42], the base-to-new generaliation is conducted on 11 image classification datasets, *i.e.,* ImageNet [6] and Caltech [9] for generic object classification; OxfordPets [27], StanfordCars [21], Flowers [26], Food101 [2], and FGVCAircraft [25] for fine-grained visual categorization, EuroSAT [15] for satellite image classification, UCF101 [32] for action recognization, DTD [5] for texture classification, and SUN397 [37] for scene recognition. Furthermore, we use the ImageNet and its variants for domain generalization, *i.e.,* the ImageNet is treated as the source domain; ImageNetV2 [31], ImageNet-Sketch [35], ImageNet-A [11] and ImageNet-R [16] are treated as the target domains for evaluation.

Training Details: Our implementation is based on CoOp's [41] [3] and ProGrad's [42] [4] codes with the CLIP model. We conduct the experiments based on the vision backbone with ResNet-50 [14] and Vit-B/16 [8]. Inspired by CoOp, we fix the context length to 4 and initialize the context vectors using the template of “a photo of a []”. The final performance is averaged over three random seeds for a fair comparison. We follow the same training epochs, training schedule, and data augmentation setting in CoOp and ProGrad. The hyperparameter λ is set to 8.0. All experiments are conducted based on RTX 3090.

Baselines: Four type of CoOp-based methods are treated as baselines for comparison:

- CLIP [29] applies the hand-crafted template “a photo of a []” to generate the prompts for knowledge transfer.
- CoOp [41] replaces the hand-crafted prompts with a set of learnable prompts inferred by the downstream datasets, which is our baseline.
- CoCoOp [40] generates the image-conditional prompts by combining the image context of each image and the learnable prompts in CoOp.
- ProGrad [42] uses the same prompts as CoOp while only optimizing the prompt whose gradient is aligned to the “general direction”, which can be treated as CoOp+Grad.
- KgCoOp uses the same prompts as CoOp while optimizing the learnable prompts closed to the fixed prompts in CLIP, which can be treated as CoOp+Kg.

Although the existing VPT [7] and ProDA [24] have been proposed for prompt tuning, they both infer a collection of prompts rather than one learnable prompt used in CoOp-based methods.

4.1. Generalization From Base-to-New Classes

Similar to the previous work CoOp and CoCoop, we split each dataset into two groups: base classes (*Base*) and new classes(*New*). Similar to the zero-shot setting, the new classes disjoint the base classes. To verify the generalization of the CoOp-based methods, all compared methods and the proposed KgCoOp use the base classes for prompt tuning, and conduct evaluation on the new class. The detailed results are shown in Table 2 and Table 3. Table 2 summarizes the average performance among all 11 datasets with different K-shot samples and backbones (ViT-B/16 and ResNet-50). Table 3 gives the detailed performance on all 11 datasets based on the backbone of ViT-B/16 and 16-shot samples.

Total Analysis: As shown in Table 2, the proposed *KgCoOp* obtains a higher average performance in terms of Harmonic mean than existing methods on all settings, demonstrating its superiority for the generalization from base-to-new classes. Among the existing methods, ProGrad obtains the best performance in terms of *Base* classes on all settings while obtaining a worse *New* performance than CoCoOp. The reason is that a higher performance on *Base* classes makes the ProGrad have serious overfitting on the *Base* class, thus producing a biased prompt for the *New*

[3] https://github.com/KaiyangZhou/CoOp

[4] https://github.com/BeierZhu/Prompt-align

Table 3. Comparison with existing methods in the base-to-new generalization setting with ViT-B/16 as the backbone. The context length M is 4 for prompot-based methods with the 16-shots samples from the base classes. H: Harmonic mean.

(a) Average over 11 datasets.

	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.63**	67.99	74.60
CoCoOp	80.47	71.69	75.83
ProGrad	82.48	70.75	76.16
KgCoOp	80.73	73.6	**77.0**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.46	66.31	71.02
CoCoOp	75.98	**70.43**	**73.10**
ProGrad	**77.02**	66.66	71.46
KgCoOp	75.83	69.96	72.78

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	**98.11**	93.52	95.76
CoCoOp	97.96	93.81	95.84
ProGrad	98.02	93.89	95.91
KgCoOp	97.72	**94.39**	**96.03**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	94.24	96.66	95.43
CoCoOp	**95.20**	97.69	**96.43**
ProGrad	95.07	97.63	96.33
KgCoOp	94.65	**97.76**	96.18

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.65
CoOp	76.2	69.14	72.49
CoCoOp	70.49	73.59	72.01
ProGrad	**77.68**	68.63	72.88
KgCoOp	71.76	**75.04**	**73.36**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.63**	69.55	81.23
CoCoOp	94.87	71.75	81.71
ProGrad	95.54	71.87	82.03
KgCoOp	95.00	74.73	**83.65**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	89.44	87.50	88.46
CoCoOp	**90.70**	91.29	90.99
ProGrad	90.37	89.59	89.98
KgCoOp	90.5	**91.7**	**91.09**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	31.09
CoOp	39.24	30.49	34.30
CoCoOp	33.41	23.71	27.74
ProGrad	**40.54**	27.57	32.82
KgCoOp	36.21	33.55	**34.83**

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.85	68.34	74.07
CoCoOp	79.74	**76.86**	78.27
ProGrad	**81.26**	74.17	77.55
KgCoOp	80.29	76.53	**78.36**

(j) DTD.

	Base	New	H
CLIP	53.24	**59.90**	56.37
CoOp	**80.17**	47.54	59.68
CoCoOp	77.01	56.00	**64.85**
ProGrad	77.35	52.35	62.45
KgCoOp	77.55	54.99	64.35

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	**91.54**	54.44	68.27
CoCoOp	87.49	60.04	71.21
ProGrad	90.11	60.89	72.67
KgCoOp	85.64	**64.34**	**73.48**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**85.14**	64.47	73.37
CoCoOp	82.33	73.45	77.64
ProGrad	84.33	74.94	79.35
KgCoOp	82.89	76.67	**79.65**

classes, leading to a worse *New* performance. Compared with CoCoOp, the proposed KgCoOp slightly improves the *Base* classes. For example, based on the backbone of ViT-B/16, KgCoOp achieves the *Base* performance of 78.36% and 80.73% for the 8-shot and 16-shot settings respectively, which are similar to the 78.56% and 80.47% obtained by CoCoOp. However, KgCoOp obtains a significant improvement on the *New* class upon CoCoOp, *e.g.,* obtains the improvement of 1.89% and 1.91% upon CoCoOp for 8-shot and 16-shot setting, respectively. The superior performance on *New* classes demonstrates that the KgCoOp can improve the generability of the wider unseen class without discarding the discriminative ability of the seen classes.

As mentioned above, ProGrad obtains a better performance on the *Base* class and a worse performance on the *New* classes, leading to the generated prompt having serious overfitting on the *Base* classes. Since KgCoOp aims to improve the generability of the *New* class, KgCoOp also obtains a worse performance than ProGrad on the term of *Base* classes. However, KgCoOp obtains a higher performance on the *New* class. By improving the generability of *New* class, KgCoOp obtains a higher performance in terms of H than ProGrad, *e.g.,* improving the harmonic mean (H) from 75.2% and 76.16% to 76.06% and 77.0% for the 8-shot and 16-shot settings, respectively. The superior performance demonstrates that the KgCoOp can effectively adapt the pretrained VLM model on the downstream task with improving the generality of the unseen classes.

Detailed Analysis: We thus give a detailed comparison of each dataset for the prompt-based method with a 16-shot setting with the ViT-B/16 as the backbone. As shown in Table 3, existing CoOp-based methods, *i.e.,* CoOp, CoCoOp, and ProGrad, all significantly improve the *Base* classes compared to CLIP on all 11 datasets. Especially, ProGrad, CoOp, and CoCoOp obtain the best *Base* performance on 4/11 datasets, 5/11 datasets, and 2/11 datasets, respectively. While the CoOp also obtains a better average *Base* performance than ProGrad and CoCoOp. The reason is that CoOp only focuses on inferring a learnable prompt without considering any other constraints, making the generated prompt be discriminative for the *Base* class. Unlike CoOp, CoCoOp considers the instance-conditional

Table 4. Comparison of prompt learning in the domain generalization with 16-shot source samples. where “vp” and “tp” denote the visual prompting and textual prompting, respectively.

	Prompts	Source	Target				
		ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R	Avg.
CLIP [29]	hand-crafted	66.73	60.83	46.15	47.77	73.96	57.17
UPT [39]	vp+tp	**72.63**	64.35	48.66	50.66	76.24	59.98
CoCoOp [40]	vp+tp	71.02	64.07	48.75	50.63	76.18	59.90
CoOp [41]	tp	71.51	64.2	47.99	49.71	75.21	59.28
ProGrad [42]	tp	72.24	**64.73**	47.61	49.39	74.58	59.07
KgCoOp	tp	71.2	64.1	**48.97**	**50.69**	**76.7**	**60.11**

token combined with the learnable context vectors. Using the instance-conditional token can improve the generability on the *New* class, while degrading the discrimination on the *Base* class. Therefore, CoCoOp obtains the best *New* performance on 6/11 datasets, and the best average *New* performance. Specially, CoCoOp obtains an obivous performance improvement of 3.77%, 4.96%, 1.7% and 3.65% on ImageNet, StandfordCars, Food101, and DTD upon ProGrad, respectively, while ProGrad obtains the obvious performance improvement upon CoCoOp for the FGVCAircraft datasets, *e.g.,* 23.71%(CoCoOp) *vs* 27.57%(ProGrad). However, existing methods, *i.e.,* CoOp, CoCoOp, and ProGrad, all obtain a worse performance than the original CLIP in most cases, which indicates that they weaken generability to the *New* classes. Compared with existing methods, the proposed KgCoOp obtains a higher *New* performance on eight datasets among all 11 datasets, *e.g.,* Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, EuroSAT, and UCF101. The superior performance demonstrates that KgCoOp has a better generability to the *New* classes than existing CoOp-based prompt methods. Meanwhile, in most cases, KgCoOp obtains the same performance as CoCoOp on the *Base* classes. Therefore, KgCoOp can improve the generability on *New* classes without degrading the performance of *Base* classes, leading to the best Harmonic mean on all 11 datasets.

4.2. Domain Generalization

Domain Generalization aims to evaluate the generalization by evaluating the trained model on the target dataset, which has the same class but different data distribution from the source domain. Similar to CoCoOp and ProGrad, we conduct the prompt tuning on the few-shot ImageNets, and evaluate the model on the ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. The related results are summarized in Table 4.

From Table 4, we can observe that ProGrad obtains the best performance on the source ImageNet. The superior performance shows that ProGrad can produce a discriminative prompt for the base class, consistent with the conclusion obtained in the base-to-new setting. Similar to the comparison in the base-to-new setting, ProGrad has a weakened generability to the wider unseen classes, *e.g.,* except for the ImageNetV2, ProGrad has achieved weaker performance than CoCoOp on the other three datasets and the mean performance. Among existing methods, CoCoOp is more domain-generalizable than CoOp and ProGrad. Compared with CoCoOp, the proposed KgCoOp obtains a higher performance on the source and target datasets, *e.g.,* improving the average target performance from 59.90% to 60.11%. The above comparison confirms that the learnable prompts in KgCoOp are better domain-generalizable.

Table 5. Accuracy (%) of few-shot(K=4) learning on 11 datasets.

Datasets	CoOp	CoCoOp	ProGrad	KgCoOp
ImageNet	69.38	**70.55**	70.21	70.19
Caltech101	94.44	**94.98**	94.93	94.65
OxfordPets	91.3	**93.01**	93.21	93.2
StanfordCars	**72.73**	69.1	71.75	71.98
Flowers102	**91.14**	82.56	89.98	90.69
Food101	82.58	**86.64**	85.77	86.59
FGVCAircraft	33.18	30.87	**32.93**	32.47
SUN397	70.13	70.5	71.17	**71.79**
DTD	**58.57**	54.79	57.72	58.31
EuroSAT	68.62	63.83	70.84	**71.06**
UCF101	77.41	74.99	77.82	**78.40**
Avg.	73.59	71.98	74.21	**74.48**

4.3. Few-shot Classification

The base-to-new setting assumes that the new classes have different categories from the base classes, which can demonstrate the generability of different classes. To further show the generability of the proposed method, we conduct the few-shot classification, which trains the model based on the few-shot labeled images and evaluates the model on the dataset with the same categories as the training classes. The 4-shot setting results are summarized in Table 5. We can observe that the proposed KgCoOp obtains a higher average performance than existing methods, *i.e.,* CoOp, CoCoOp, and ProGad.

4.4. Analysis

Hyperparameter λ: The critical contribution of the proposed KgCoOp is applying a regularization term to con-

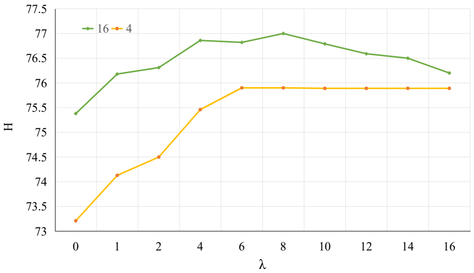

Figure 3. Effect of λ for 4-shot and 16-shot settings on the base-to-new generalization. H: Harmonic mean

Table 6. Effect of $\mathcal{L}_{kg}$ on CoOp, CoCoOp, and ProGrad in the base-to-new generalization setting with 16-shot samples and ViT-B/16 in terms of the average performance among all 11 datasets.

Methods	*Base*	*New*	*H*
CoOp	82.63	67.99	74.6
CoOp+$\mathcal{L}_{kg}$	80.73($\downarrow$ -1.9)	73.6($\uparrow$ 5.61)	77 ($\uparrow$ 2.4)
CoCoOp	80.47	71.69	75.83
CoCoOp +$\mathcal{L}_{kg}$	77.96 ($\downarrow$ -2.50)	74.75($\uparrow$ 3.06)	76.32 ($\uparrow$ 0.49)
ProGrad	82.48	70.75	76.16
ProGrad+$\mathcal{L}_{kg}$	78.64 ($\downarrow$ -3.84)	74.72($\uparrow$ 3.97)	76.63 ($\uparrow$ 0.47)

strain the special knowledge generated by prompt tuning to be closed to the general knowledge, which can improve the generalization on the unseen domain. λ is thus applied to balance the importance of the regularization term during prompt tuning, *e.g.,* the higher λ denotes that the prompt tuning pays more attention to the general knowledge. We thus analyze the effect of λ, and show the results in Figure 3. We can observe that a higher λ can obtain a higher metric of H. For example, setting λ as 8.0 obtains the best performance of 77.0%. By further increasing λ, the performance would be degraded, *e.g.,* setting λ=10.0 obtains a harmonic mean of 76.79%, lower than 77.0% for λ=8.0.

Effect of $\mathcal{L}_{kg}$: The critical of ours is to constrain $\mathcal{L}_{kg}$ to minimize the general textual embedding and specific textual embedding, which can be easily applied to existing CoOp-based methods, *e.g.,* CoOp, CoCoOp, and ProGrad. As shown in Table 6, compared with CoCoOp and ProGrad, considering the additional $\mathcal{L}_{kg}$ constraint improves the performance in terms of *New* and *H*. Especially for the *New* performance, using $\mathcal{L}_{kg}$ achieves more than 3% improvement. The superior performance further proves the effectiveness of using the constraint $\mathcal{L}_{kg}$ for prompt tuning.

Quantitative analysis of $\mathcal{L}_{kg}$: KgCoOp aims to improve the generability of the unseen class by minimizing the distance $\mathcal{L}_{kg}$ between the learnable textual embedding $\mathbf{w}$ and fixed textual embedding$\mathbf{w}_{clip}$. We thus verify the rationality and effectiveness of this motivation and summarize the related results in Table 7. We can observe that a higher λ obtains a lower $\mathcal{L}_{kg}$. Furthermore, the lower $\mathcal{L}_{kg}$, the higher performance H. Therefore, we can conclude that minimizing the distance between the learnable textual embedding $\mathbf{w}$ and fixed textual embedding $\mathbf{w}^{clip}$ can improve the performance.

Table 7. The quantitative analysis of $\mathcal{L}_{kg}$ for different λ on ImageNet.

λ	0.0	1.0	2.0	4.0	6.0	8.0	10.0
$\mathcal{L}_{kg}$	0.18	0.038	0.024	0.015	0.010	0.006	0.005
H	75.38	76.18	76.31	76.86	76.82	77	76.79

Table 8. Traning teim comparison(*ms*/image). The training time is the average time to process one image, *i.e., ms*/image.

	CoOp	CoCoOp	ProGrad	KgCoOp
time	6*ms*	160*ms*	22*ms*	6*ms*
H	74.60	75.83	76.16	77.0

Training efficienty: For the prompt-based method, we calculate the training time on ImageNet datasets with a16-shot setting. Note that the batchsize is 32 for CoOp, ProGrad, and KgCoOp, while CoCoOp uses the batchsize of 1. The training time is the average time to process one image, *i.e., ms*/image. Based on CoOp, the proposed KgCoOp conducts an additional constraint between the $\mathbf{w}$ and $\mathbf{w}_{clip}$ during training. Since $\mathbf{w}_{clip}$ is a pre-computed vector generated by CLIP with the given categories names, the core of KgCoOp is merely to minimize the distance $\mathbf{w}$ and $\mathbf{w}_{clip}$. Compared to the training time, the additional running time of the proposed method can be ignored. As shown in Table 8, KgCoOp has the same training time as the CoOp, which is faster than CoCoOp and ProGrad. Moreover, KgCoOp obtains the best performance. In conclusion, KgCoOp is an efficient model achieving better performance with less training time.

5. Conclusion

To overcome the shortcoming that existing CoOp-based prompt tuning methods weaken the generability of the unseen classes, we introduce a prompt tuning method named Knowledge-guided Context Optimization to boost the generability of the unseen classes by minimizing the discrepancy between the general textual embeddings and the learnable specific textual embeddings. Extensive evaluation of several benchmarks shows that the proposed KgCoOp is an efficient prompt tuning method.

Although using KgCoOp can improve the generability on unseen classes, it may degrade the discriminative ability on the seen class, *e.g.,* KgCoOp obtains a badly *Base* performance on seen classes. We will investigate an effective method for seen and unseen classes in the future.

Acknowledgements: This work was supported by National Key Research and Development Program of China (2021ZD0112200), National Natural Science Foundation of China (62036012,61721004,U21B2044,U20B2070), and Beijing Natural Science Foundation (L201001,4222039).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014. 5

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 2

[4] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 2021. 1

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 5

[7] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees G. M. Snoek, Georgios Tzimiropoulos, and Brais Martínez. Variational prompt tuning improves generalization of vision-language models. *CoRR*, abs/2210.02390, 2022. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5

[9] Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007. 5

[10] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *CoRR*, abs/2210.09263, 2022. 1, 2

[11] Haoran Gao, Hua Zhang, Xingguo Yang, Wenmin Li, Fei Gao, and Qiaoyan Wen. Generating natural adversarial examples with universal perturbations for text classification. *Neurocomputing*, 471:175–182, 2022. 5

[12] 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. *CoRR*, abs/2110.04544, 2021. 3

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. 5

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE, 2021. 5

[17] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 1, 2

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European*

Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII, volume 13693 of *Lecture Notes in Computer Science*, pages 709–727. Springer, 2022. 1

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 2

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013. 5

[22] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021. 1

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. 2

[24] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5196–5205. IEEE, 2022. 3, 5

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 5

[26] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008. 5

[27] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012. 5

[28] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019. 1

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 7

[30] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18061–18070. IEEE, 2022. 1, 3

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019. 5

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 5

[33] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212, 2021. 1

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[35] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518, 2019. 5

[36] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *The Tenth In-*

ternational Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. 2

[37] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society, 2010. 5

[38] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. CPT: colorful prompt tuning for pre-trained vision-language models. *CoRR*, abs/2109.11797, 2021. 1

[39] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *CoRR*, abs/2210.07225, 2022. 2, 7

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 1, 3, 4, 5, 7

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 1, 2, 5, 7

[42] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *CoRR*, abs/2205.14865, 2022. 1, 3, 4, 5, 7