Learning to Learn Better Visual Prompts

Fengxiang Wang[1], Wanrong Huang[1], Shaowu Yang[1*], Fan Qi[2], Long Lan[1]

[1]HPCL, College of Computer Science and Technology, National University of Defense Technology
[2]The Hong Kong University of Science and Technology
{wfx23, huangwanrong12, shaowu.yang}@nudt.edu.cn, fanqics@gmail.com,long.lan@nudt.edu.cn

Abstract

Prompt tuning provides a low-cost way of adapting vision-language models (VLMs) for various downstream vision tasks without requiring updating the huge pre-trained parameters. Dispensing with the conventional manual crafting of prompts, the recent prompt tuning method of Context Optimization (CoOp) introduces adaptable vectors as text prompts. Nevertheless, several previous works point out that the CoOp-based approaches are easy to overfit to the base classes and hard to generalize to novel classes. In this paper, we reckon that the prompt tuning works well only in the base classes because of the limited capacity of the adaptable vectors. In addition, the scale of the pre-trained model is a hundred times the scale of the adaptable vector, thus the learned vector has a very limited ability to absorb the knowledge of novel classes. To minimize this excessive overfitting of textual knowledge on the base class, we view prompt tuning as learning to learn (LoL) and learn the prompt in the way of meta-learning, the training manner of dividing the base classes into many different subclasses could fully exert the limited capacity of prompt tuning and thus transfer its power to recognize the novel classes. To be specific, we initially perform fine-tuning on the base class based on the CoOp method for pre-trained CLIP. Subsequently, predicated on the fine-tuned CLIP model, we carry out further fine-tuning in an N-way K-shot manner from the perspective of meta-learning on the base classes. We finally apply the learned textual vector and VLM for unseen classes. Extensive experiments on benchmark datasets validate the efficacy of our meta-learning-informed prompt tuning, affirming its role as a robust optimization strategy for VLMs.

Introduction

Large-scale vision language pre-training (Zhang et al. 2023; Li et al. 2021; Lu et al. 2019; Bao et al. 2022; Radford et al. 2021) has achieved remarkable progress in zero-shot and few-shot image classification. The large pretrained vision language models (VLMs) encapsulates fundamental universal knowledge, thus endowing models with commendable generalizability to diverse tasks. Despite VLM's efficacy in extracting visual and textual descriptions, their training requires an abundance of high-quality datasets. In real-world vision language tasks, gathering requisite data for task-related model training is a formidable challenge. To address the aforementioned quandary, prompt tuning (Gan et al. 2022; Zhou et al. 2022b,a; Gao et al. 2021; Zhang et al. 2021) has explored the application of pre-trained VLM models to downstream tasks with limited data, and achieved exceptional outcomes in zero-shot visual tasks.

Unlike CLIP utilizes a manually curated handcrafted fixed prompt "a photo of a [Class]" as a text-based class embedding for zero-shot classification, the methods such as CoOp for prompt tuning concatenate learnable text tokens with class labels to acquire specific textual knowledge for prediction. However, this explicit textual understanding becomes excessively tailored to downstream tasks, thus demonstrating poor generalization to new classes. Some previous works (Zhou et al. 2022b,a; Yao, Zhang, and Xu 2023) demonstrate the phenomenon of overfitting with CoOp-based methods: CoOp's accuracy on the base class initially increases then declines, while on the new class, it continually drops and remains unstable. Nevertheless, current efforts to mitigate overfitting within CoOp still possess limitations. For example, distinct from employing conventional anti-overfitting techniques, CoCoOp (Zhou et al. 2022a) seeks to reduce the extent of overfitting by introducing image prompts with Multilayer Perceptron (MLP)-transformed images feature. However, CoCoOp still shows a marked decrease in accuracy in the latter stages of training, manifesting a conspicuous issue of overfitting. Furthermore, KgCoOp (Yao, Zhang, and Xu 2023) introduces specific losses to minimize the disparity between learnable prompts and manual prompts, which alleviates CoOp's overfitting phenomenon to some extent. However, these methods exhibit limitations in alleviating the overfitting within CoOp. We argue that the scale of the pre-trained model is a hundred times more than the scale of the adaptable vector, thus the learned vector has a limited ability to absorb the knowledge of novel classes. We believe that the CoOp-based research lacks consideration for the organization of input data during prompt tuning. They merely use the base class data as a holistic input for training and fitting, thereby resulting in the text knowledge becoming overly tailored to the base class.

Recently, meta-learning (Hospedales et al. 2021), otherwise termed "learning to learn", has emerged as one of the common techniques to address the problem of few-shot learning. Inspired by meta-learning, we explore the advan-

*This is the corresponding author.

tages of both prompt tuning methods and meta-learning by decoupling differences in the prompt tuning of VLM models to predict from base classes to new classes. Diverging from existing CoOp-based techniques, our approach, a prompt tuning method named **learning to learn (LoL)**, delves into the organization of input data, adopting the N-way K-shot framework for prompt learning. This meta-learning task's data input training strategy adeptly minimizes textual knowledge overfitting on the base class while optimally retaining the ubiquitous prior knowledge embedded within the expansive pre-trained CLIP model. The "meta" embodies the network's learning ability for each specific task, furnishing the network with an abstract learning capability through continuous adaptation to each task. Specifically, our approach initially views the base class knowledge acquired by CLIP as the feature-extracting backbone of meta-learning while carrying out classification pre-training on base classes via a CoOp-based approach. The meta-training method is then utilized, and base class data is learned through Episodic Training, followed by prediction on new classes. During Episodic Training the model samples a few classification tasks from the training samples of the base classes and optimizes itself to perform well on these tasks. The task typically assumes an N-way K-shot format, involving N classes, each comprising K support samples and Q query samples. The goal is to classify these $N \times Q$ query samples into N classes, based on $N \times K$ supporting samples. Finally, testing is conducted on the new class data.

In summary, our contributions are as follows:

- We propose a novel visual prompt tuning baseline grounded in the philosophy of meta-learning, neglected in prior works. It yields competitive performance across 7 datasets and is easy to comply with and extend to other prompt tuning methods.
- The method we propose nicely tackles the base-to-new generalization task. We performed comprehensive experimental settings for the generalization capability from base classes to new classes on 7 image classification datasets. Evaluations reveal that our presented meta-learning-based prompt tuning method is effective, obtaining a higher final performance on new classes than existing approaches.

Related Works

Vision-Language Pre-training. Vision-Language pre-training (Zhang et al. 2023) learns generic cross-modal representations from large-scale image-text pairs, then it can be fine-tuned directly on downstream visual-linguistic tasks. The architecture of the vision-language models can be categorized into three ways of encoders: fusion encoder (single-stream (Su et al. 2019; Li et al. 2020, 2019) and dual-stream (Li et al. 2021; Lu et al. 2019; Tan, Bansal, and Assoc Computat 2019)), dual encoder (Lee et al. 2018; Jia et al. 2021; Radford et al. 2021) and the combination of both (Bao et al. 2022; Singh et al. 2022). The representative work of the dual encoder model is CLIP (Radford et al. 2021), which uses a contrastive learning approach to train images and corresponding text descriptions, and then learns the relationship between images and text by comparing the embedding vectors of these images and text descriptions. Since CLIP's inception, a series of subsequent studies and applications have emerged. Enhancements to CLIP-based work are primarily observed in three key areas: data augmentation (Cherti et al. 2023; Gadre et al. 2023), model architecture (Chen et al. 2023; Girdhar et al. 2023; Li et al. 2023; Shen et al. 2022), and the objective function (Yao et al. 2021; Yu et al. 2022). In terms of data augmentation, one strategy (Cherti et al. 2023) involves curating larger datasets and implementing multi-scale training paradigms. Alternatively, maintaining the CLIP training methodology constant while varying the datasets is another approach (Gadre et al. 2023). When it comes to model design, enhancements can be made to the image perspective (Li et al. 2023), the language perspective (Shen et al. 2022), improved interpretability (Chen et al. 2023), and the integration of additional modalities (Girdhar et al. 2023). In terms of the objective function, fine-grained supervision (Yao et al. 2021) and the incorporation of a generative branch (Yu et al. 2022) are key considerations. Moreover, the fusion of CLIP with other learning techniques, such as supervised learning (Wu et al. 2023; Yang et al. 2022; Zhai et al. 2022), image-only contrastive learning (Zhai et al. 2022; Mu et al. 2022; Zhou et al. 2023), and masked image modeling (Fang et al. 2023; Sun et al. 2023; Wei et al. 2022), enlightens our understanding. The basis of our research lies with the vision language model CLIP, striving to deliver an adept solution for adapting pre-trained vision-language models to downstream implementations.

Prompt Tuning. In the process of fine-tuning the pre-trained vision-language models (VLMs) for downstream tasks, prompt tuning (Gan et al. 2022) is seen as a method of extracting useful information for these tasks. Initially, the hand-crafted template in CLIP is used for zero-shot and few-shot prediction. Furthermore, prompt learning aims to automate prompt design with the aid of adequately large labeled datasets. The concept of automatic prompts was introduced by CoOp (Zhou et al. 2022b), which represented the downstream task's prompts as trainable continuous vectors, improving flexibility and adjustability. CoCoOp (Zhou et al. 2022a) creates an image-conditioned context combined with a text-context for prompt tuning. ProGrad (Zhu et al. 2022) introduces prompt-aligned gradient to prevent knowledge forgetting. MaPLe (Khattak et al. 2023) offers a dynamic prompt-building technique for dialogue between text and image prompts during training. PTP (Wang et al. 2023) redefines visual tasks to predict within a block or link to an object's block using fill-in-the-blank queries. KgCoOp (Yao, Zhang, and Xu 2023) boosts learnable prompts' generalizability to new classes by equating embeddings from learned and hand-crafted prompts. CLIP-adapter (Gao et al. 2021) introduced an adapter with a feature blending residual for efficient VLMs transfer learning, while Tip-Adapter (Zhang et al. 2021) proposed a training-free adapter using embeddings from a few labeled images. SVL-Adapter (Pantazis et al. 2022) introduced a self-supervised adapter by performing self-supervised learning on images. Among these methodologies, our work is primarily based on CoOp and its

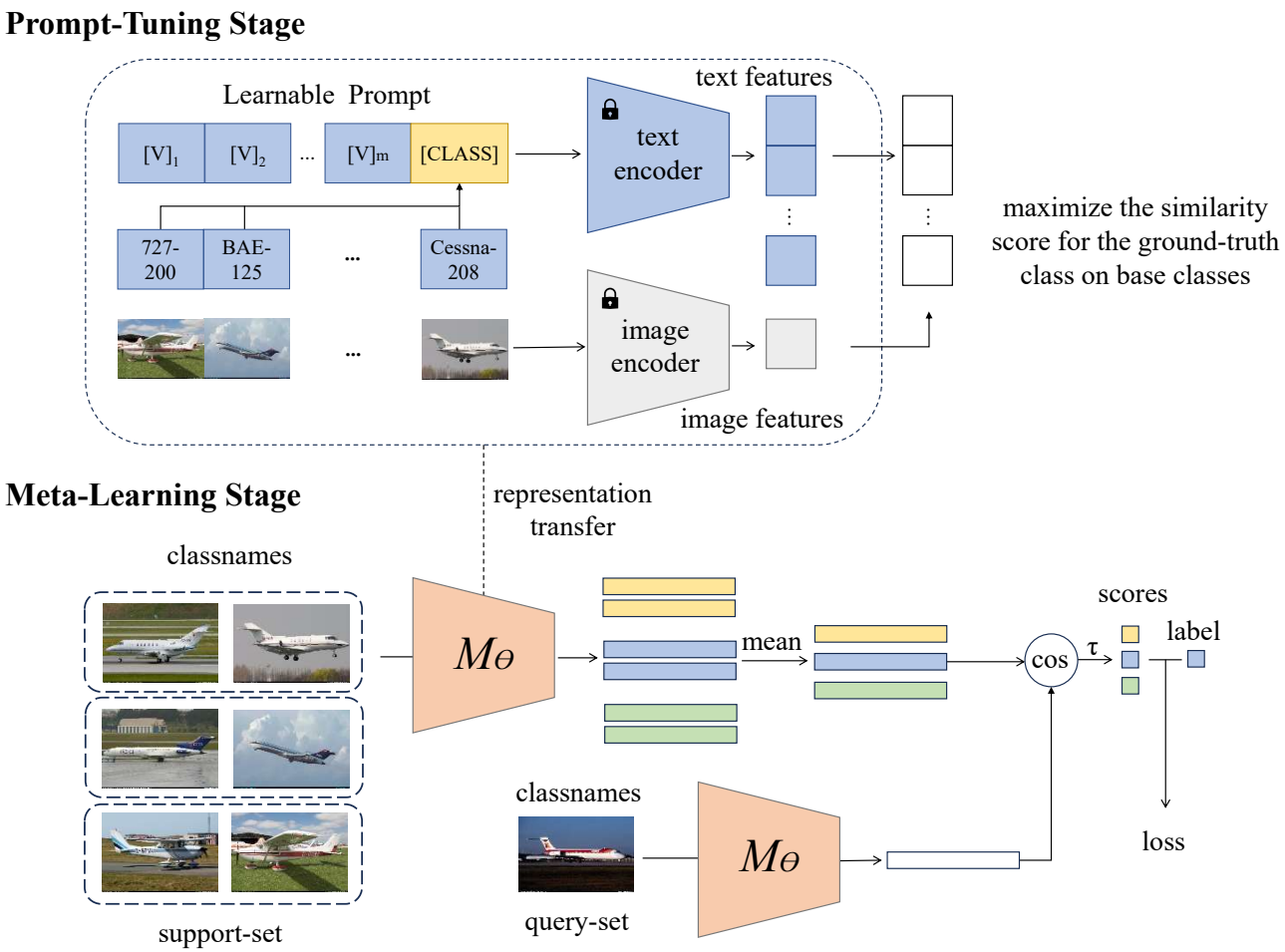

Figure 1: Overview of the proposed method. The main idea is that model uses a learning to learn method for the generation of new classes, with a set of learnable vectors, which can be optimized by minimizing the classification loss.

follow-ups. We look to improve the fine-tuning performance of VLMs for unseen classes by utilizing meta-learning techniques.

Meta Learning. In the few-shot problem (Lake et al. 2017), meta-learning (Hospedales et al. 2021) serves as an instrumental procedure in navigating challenges posed by limited-data scenarios, honing the model's abilities over multiple categories by harnessing an abundance of data before introducing new categories for predictive purposes. Meta-learning techniques can be conveniently divided into three principal sectors: optimization-based techniques, black-box approaches, and metric-based operations. Optimization-based strategies typically derive inspiration from MAML (Finn, Abbeel, and Levine 2017), which pioneers empirical optimization of neural networks through scarce data. Variations of this method take different aspects of optimization into account, encapsulating optimization of model initialization (Rajeswaran et al. 2019; Rusu et al. 2018; Sung et al. 2018; Zintgraf et al. 2018), process optimization (Munkhdalai and Yu 2017; Xu et al. 2020), or both (Baik et al. 2020). Contrarily, black-box methods (Garnelo et al. 2018; Mishra et al. 2017) model the learning process as a neural network without an explicit induction bias. Lastly, the metric-based approach (Liu et al. 2020; Snell, Swersky, and Zemel 2017; Vinyals et al. 2016) cultivates a feature extractor via meta-learning, which yields a well-structured feature space with predefined metrics.

Methods

Prompt-tuning Stage

The contrastive language–image pretraining (CLIP) incorporates two types of encoders: a visual encoder and a text encoder. We adopt the encoders and lock them for prompt training. The image encoder, represented by $W(\cdot)$, transforms an image $x \in R^{3\times H\times W}$ of height H and width W from an d-dimension into a d-dimensional image feature $w_x \in R^{N\times d}$, where N denotes the number of partitioned patches. While CLIP can be seamlessly employed for zero-shot predictions, it resorts solely to fixed handcrafted prompts "a photo of a[]" for generating textual embeddings.

Crafting such prompts for the CLIP model could be labor-intensive, demanding a wealth of time and expertise for word optimization. Following CoOp, we introduces M context vectors $V = \{v_1, v_2, ..., v_M\}$ as learnable prompts. It then amalgamates the class token embeddings c_i pertinent to the i-th class with the learnable context vector V to fashion prompt $p_i = \{v_1, v_2, ...v_M, c_i\}$. Subsequently, the learnable prompt p_i is fed into the text encoder $T(\cdot)$, yielding the textual class embedding $T(p_i)$. The image encoder is then tasked to extract the image feature w_x of image x and obtain the text feature t_y by inputting the prompt description into the text encoder. The prediction task is defined as classifying the image into one of C classes, represented by the set $z\epsilon\{1, ..., C\}$. z is denoted as the predicted class. It refines the learnable context token V by minimizing the negative log-likelihood function between image features and textual encoding. Therefore, we have the predicted probability of the i-th class:

$$P(z = i|x) = \frac{exp(cos(w(x), T(p_i))/\tau)}{\sum_{j=1}^{C} exp(cos(w(x), T(p_i))/\tau)} \quad (1)$$

where $cos(\cdot, \cdot)$ and τ denote the cosine similarity and the temperature parameter of the softmax function, respectively. Throughout this optimization process, the visual encoder and the pretrained text encoder remain static, akin to CLIP. Contrary to CLIP which employs fixed prompts, This stage produces task-specific prompts, augmenting its generalizability and discernment.

Meta-Learning Stage

We use the classifier on the base class trained from the prompt tuning stage as the initialization of our meta-learning stage. The next phase is the meta-learning stage, optimizing the model based on the prompt-tuning stage. In the meta-learning stage, LoL adapts the N-way K-shot manner from the perspective of the N-way K-shot task format on the base classes. N-way signifies N classes in the training data, while K-shot indicates K labeled data under each class. Within an N-way K-shot undertaking, the support set encompasses N classes, each with K samples, and the query set includes samples from those same N classes with Q samples per class. The objective lies in classifying the $N \times Q$ query images into N classes. Specifically, during the meta-learning stage, given the entire feature encoder M_θ trained for classification, N-way K-shot are sampled from the training samples of the base class, constituting $N \times Q$ query samples. Given the support-set S, let S_i denote the samples in class i, it computes the average embedding e_i as the centroid of class i:

$$e_i = \frac{1}{S_i} \sum_{x \epsilon S_i} M_\theta(x) \quad (2)$$

To compute the loss for each task, centroids for the N classes defined in Eq.2 are calculated within the support-set, facilitating the computation of predicted probability distribution for each sample in the query-set. During these training procedures, each training batch may comprise multiple tasks, with the computed loss being the average cross-entropy. A comprehensive explanation of the entire process involved in the meta-learning stage is encapsulated in Algorithm 1. The term $Require1$ represents the distribution of tasks within the base classes in the dataset, which entails the random extraction of data within the base classes, based on the task-unit, for the creation of a task pool. This pool forms the training set for the meta-learning stage. In $Require2$, the step size essentially functions as the learning rate. $Require3$ indicates that the meta-learning stage is predicated upon the parameters learned during the prompt tuning stage.

Algorithm 1: Meta-Learning Stage

Require1: $p(\zeta)$: distribution over tasks
Require2: α, β: step size hyperparameters
Require3: M: Model after prompt-tuning stage
1: initialize θ with M
2: **while** not done **do**
3: Sample batch of tasks $\zeta_i \sim p(\zeta)$
4: **for** ζ_i **do**
5: Evaluate $\bigtriangledown_\theta \varphi_{\zeta_i}(M_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \bigtriangledown_\theta \varphi_{\zeta_i}(M_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \bigtriangledown_\theta \sum_{\zeta_i \sim p(\zeta)} \varphi_{\zeta_i}(M_{\theta_i'})$
9: **end while**

Step 1 indicates the initiation of the model parameters after the learning phase of the prompt tuning stage. The model then commences the cyclical process elaborated in Step 2. Step 3 indicates a random selection of an array of tasks to create a batch. The meta-learning stage heavily relies on the dual-gradient basis, where each iteration includes two parameter updates, often named as gradient by gradient. Steps 4 through 7 outline the process of the first gradient update. Specifically, the gradient parameters of the support set within a task in the batch are calculated. Given an N-way K-shot task, there are supposed to be $N \times K$ in this support set. When adapting to a new task ζ_i, the model's parameters θ' morph into θ. The updated parameter vector θ' in our method is computed utilizing one gradient descent update on task ζ_i, with the step size α fixed as a hyperparameter.

$$\theta_i' = \theta - \alpha \bigtriangledown_\theta \varphi_{\zeta_i}(M_\theta) \quad (3)$$

Step 8 is paralleled with the process of the second gradient update. This secondary update enhances performance by optimizing Model M with parameter θ'. It is noteworthy that the ultimate optimization of the model is achieved on the basis of Model θ parameters.

$$min_\theta \sum_{\zeta_i \sim p(\zeta)} \varphi_{\zeta_i}(M_{\theta_i'}) = \sum_{\zeta_i \sim p(\zeta)} \varphi_{\zeta_i}(M_\theta - \alpha \bigtriangledown_\theta \varphi_{\zeta_i}(M_\theta)) \quad (4)$$

The optimization goal of the second gradient update, however, employs parameters θ'. Therefore, a cross-task meta-optimization of model parameters θ is implemented. For instance, by utilizing an optimizer, the model parameters are updated as follows:

$$\theta \leftarrow \theta - \beta \bigtriangledown_\theta \sum_{\zeta_i \sim p(\zeta)} \varphi_{\zeta_i}(M_{\theta_i'}) \quad (5)$$

Backbones	Methods	K=1			K=2			K=4			K=8			K=16		
		Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
VIT-B/16	CoOp	67.08	72.65	69.75	69.63	73.42	71.47	72.62	71.81	72.21	75.31	72.29	73.77	77.26	70.90	73.94
	CoCoOp	65.72	70.18	67.87	69.54	73.57	71.50	72.93	71.80	72.36	74.09	69.58	71.76	76.40	69.43	72.75
	KgCoOp	68.53	76.63	72.35	70.46	75.95	73.10	70.86	77.00	73.80	72.95	77.10	74.97	74.94	76.52	75.72
	LoL	67.08	**83.25**	**74.30**	69.63	**88.19**	**77.82**	72.62	**90.29**	**80.50**	75.31	**91.87**	**82.77**	77.26	**91.59**	**83.82**
ResNet-50	CoOp	58.22	64.66	61.27	61.40	65.77	63.51	65.66	66.04	65.85	68.70	64.73	66.66	70.86	66.46	68.59
	CoCoOp	56.59	69.13	64.00	61.45	68.91	64.97	65.39	69.01	67.15	67.92	68.46	68.19	70.42	67.68	68.02
	KgCoOp	59.93	70.64	**64.85**	63.10	72.21	67.34	65.40	72.28	68.67	67.51	71.86	69.61	69.12	71.94	70.50
	LoL	58.22	**70.86**	63.92	61.40	**78.78**	**69.01**	65.66	**83.12**	**73.36**	68.70	**80.84**	**74.28**	70.86	**83.72**	**76.76**

Table 1: Comparison in the base-to-new setting with different K-shot samples in terms of the average performance among all 7 datasets and backbones(ViT-B/16 and ResNet-50).

β represents the meta step size. The samples involved in the calculations of Step 8 are the task's query set. The aim is to bolster the model's generalization ability on the task to prevent overfitting of the support set.

As shown in Fig. 1, the model loss is calculated from the cross-entropy loss of the labels of the samples in the query set and p. Additionally, cosine similarity is employed to compute the similarity between the query set and the support set. As the range of cosine similarity is [-1,1], it aids in scaling down the value before the application of the softmax function during loss computation in training. In the training process, we multiply the cosine similarity by a learnable scala τ , thereby transforming the probability prediction into Eq.6 in the training.

$$P(z=i|x) = \frac{exp(cos(M(x), e_i)/\tau)}{\sum_{j=1}^{C} exp(cos(M(x), e_i)/\tau)} \quad (6)$$

As a methodology, even though meta-learning has been expounded in prior work, none of these previous works have excavated the performance potential of meta-learning in the field of VLMs and prompt tuning. Therefore, the meta baseline we proposed in prompt tuning also represents a neglected, yet crucial baseline in this domain.

Experiments

Following CoOp, we evaluate the models' generalization ability from base-to-new classes within various datasets. Every model engaged in our experiments draws its foundation from the publicly available CLIP. Prior to delving into the outcome, we elucidate the specifics of the experimental details.

Experimental Setup

Datasets. We evaluate the methods on 7 image classification datasets, which cover a diverse set of recognition tasks. Specifically, the benchmark includes ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004) for classification on generic objects; Flowers102 (Nilsback and Zisserman 2008) and FGVC Aircraft (Maji et al. 2013) for fine-grained classification; SUN397 (Xiao et al. 2010) for scene recognition; UCF101 (Soomro, Zamir, and Shah 2012) for action recognition; DTD (Cimpoi et al. 2014) for texture classification.

Baselines. Three types of CLIP-based methods are included as baselines for comparison:

- CoOp (Zhou et al. 2022b) replaces the hand-crafted prompts with a set of learnable prompts inferred by the downstream datasets.
- CoCoOp (Zhou et al. 2022a) generates the image-conditional prompts by combining the image context of each image and the learnable prompts in CoOp.
- KgCoOp (Yao, Zhang, and Xu 2023) remedially mitigates the forgetfulness of pivotal knowledge. It achieves this by lessening the inconsistency between the learnable prompt and the manually constructed prompt.

Training Details. Our model's underlying implementation is hinged on the approaches of CoOp and KgCoOp, entwined with the CLIP model. For the task of generalization from base-to-new classes in CoOp, CoCoOp and KgCoOp, we split the dataset categories into a train-test set at a ratio of 3:1. As well, for fair comparison, we divided the dataset into training, validation, and testing sets at a ratio of 2:1:1 in our methodology. The data for test is same. We conduct the experiments based on the vision backbone with ResNet-50 (He et al. 2016) and ViT-B/16 (Dosovitskiy et al. 2020). Gleaning inspiration from CoOp, we determinedly fix the context length at 4 and not initialize the context vectors. And the class token position is end. The data augmentation methods are not adopted in our method. We use the setting of 5-way-K-shot in meta training, K=1,2,4,8,16.

Base-to-Novel Generalization

Analogous to CoOp, each dataset is divided into two groups: the base classes (Base) and the new class (New), with the new class diverging from the categories within the base class. In order to substantiate the generalizability of methods predicated on CoOp-based methods, all comparative methods along with the one we propose have grounded their assessment of the new class on the base class data. The detailed results are depicted in Table 1 and Table 2. Table 1 encapsulates the average performance across all seven datasets with diverse K-shot samples and backbones (ViT-B/16 and ResNet-50). Table 2 offers a detailed account of the performance based on ViT-B/16 backbone and 16-shots setting across all 7 datasets.

(a) Caltech101.

	Base	New	H
CoOp	97.17	98.70	97.93
CoCoOp	96.90	98.30	97.59
KgCoOp	96.73	98.77	97.74
LoL	97.17	**99.43**	**98.29**

(b) FGVCAircraft.

	Base	New	H
CoOp	40.47	39.70	40.08
CoCoOp	38.73	17.47	24.08
KgCoOp	36.60	43.53	39.77
LoL	40.47	**76.03**	**52.82**

(c) DTD.

	Base	New	H
CoOp	72.00	61.10	66.10
CoCoOp	70.57	59.87	64.78
KgCoOp	68.73	66.57	67.63
LoL	72.00	**77.04**	**74.43**

(d) Flowers102.

	Base	New	H
CoOp	97.27	75.30	84.89
CoCoOp	97.00	76.97	85.83
KgCoOp	93.73	83.50	88.32
LoL	97.27	**97.51**	**97.39**

(e) UCF101.

	Base	New	H
CoOp	83.47	63.87	72.37
CoCoOp	82.37	73.20	77.51
KgCoOp	82.60	80.77	81.67
LoL	83.47	**95.29**	**88.99**

(f) SUN397.

	Base	New	H
CoOp	76.67	79.55	78.08
CoCoOp	76.20	81.70	78.85
KgCoOp	75.73	85.57	80.35
LoL	76.67	**98.63**	**86.27**

(g) Imagenet.

	Base	New	H
CoOp	73.79	78.07	75.87
CoCoOp	73.03	78.48	75.66
KgCoOp	70.43	76.90	73.52
LoL	73.79	**97.19**	**83.89**

(h) Average over 7 datasets.

	Base	New	H
CoOp	77.26	70.90	73.94
CoCoOp	76.40	69.43	72.75
KgCoOp	74.94	76.52	75.72
LoL	77.26	**91.59**	**83.82**

Table 2: 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 prompt-based methods with the 16-shot samples from the base classes. H: Harmonic mean.

shots	**Method**	**SUN397**			**Flowers102**			**DTD**		
		Base	**New**	**H**	**Base**	**New**	**H**	**Base**	**New**	**H**
1shot	LoL-CoOp	66.70	93.47	77.85	84.17	89.23	86.63	50.93	61.13	55.57
	LoL-KgCoOp	70.63	**93.57**	80.50	84.47	**89.33**	86.83	53.83	**62.86**	58.00
2shots	LoL-CoOp	69.20	96.22	80.50	86.57	93.08	89.71	54.53	71.26	61.78
	LoL-KgCoOp	72.33	**96.82**	82.80	80.00	**93.14**	86.07	55.60	**75.35**	63.99
4shots	LoL-CoOp	72.00	97.91	82.98	91.33	96.35	93.77	63.50	75.04	68.79
	LoL-KgCoOp	74.03	**98.22**	84.43	85.23	**96.36**	90.45	59.33	**75.42**	66.41
8shots	LoL-CoOp	74.80	98.30	84.95	95.20	97.27	96.22	66.57	75.76	70.87
	LoL-KgCoOp	74.93	**98.55**	85.13	89.90	**97.30**	93.45	64.30	**78.45**	70.67
16shots	LoL-CoOp	76.67	98.63	86.27	97.27	97.51	97.39	72.00	77.04	74.43
	LoL-KgCoOp	75.73	**98.79**	85.74	93.73	**97.69**	95.67	68.73	**77.21**	72.72

Table 3: Generality with other methods. Comparison in the base-to-new setting with different K-shot samples in terms of the performance among 3 datasets and ViT-B/16 backbones. H: Harmonic mean.

Overall Analysis. As depicted in Table 1, the proposed method shows superior performance in terms of Harmonic mean and accuracy in new classes compared to existing methods across all settings, highlighting its efficacy for generalization from base-to-new classes. Our proposed approach chooses to harness the CoOp as the fundamental infrastructure for further making extrapolations within the New classes. Therefore, the performance of LoL in base classes remains commensurate with CoOp.

Compared to CoOp, our method shows obviously improvement for new classes. For instance, using the ViT-B/16 backbone, LoL attains new class performance scores of 91.87% and 91.59% for the 8-shot and 16-shot settings respectively, which significantly surpass the 72.29% and 70.90% achieved by CoOp. In addition, LoL significantly improves on new classes compared to KgCoOp and CoCoOp, for example, achieving an improvement of 22.16% and 15.07% over CoCoOp and KgCoOp for the 16-shot setting, respectively. The superior performance of LoL when applied to new classes corroborates that our method is capable of augmenting the generality of a broader array of New classes.

At the same time, our method obtains a higher performance in terms of the harmonic mean (H) than KgCoOp.

Context initialization	FGVCAircraft			DTD			UCF101		
	Base	New	H	Base	New	H	Base	New	H
Yes, shots=4	32.56	**82.25**	46.65	60.70	**82.51**	69.94	78.52	**95.06**	86.00
No, shots=4	**33.43**	74.10	46.07	**63.50**	76.54	69.41	**78.90**	93.47	85.57
Yes, shots=8	36.53	**83.73**	50.87	65.12	**84.63**	73.60	81.97	**96.11**	88.48
No, shots=8	**37.80**	80.27	51.40	**66.57**	75.76	70.87	**82.87**	95.53	88.75
Yes, shots=16	38.53	**84.27**	52.88	71.14	**86.12**	77.92	83.39	**96.35**	89.40
No, shots=16	**40.47**	76.03	52.82	**72.00**	78.26	75.00	**83.47**	95.29	88.99

Table 4: Context Initialization. Comparison in the base-to-new setting with different K-shot samples in terms of the performance among 3 datasets and ViT-B/16 backbones. H: Harmonic mean.

Illustratively, LoL raises H from 74.97% and 75.72% to an impressive 82.77% and 83.82% within the 8-shot and 16-shot settings respectively. The superior performance of our method is clear evidence of its capacity to efficiently adapt the pre-trained VLM model to downstream tasks, all the while enhancing the generality of unseen classes.

Detailed Analyses. More detailed analyses are conducted on several datasets by implementing the CoOp-based model with ViT-B/16 as the backbone. As illustrated in Table 2, we juxtapose our method (LoL) with the extant CoOp-based strategies, namely CoOp, CoCoOp, and KgCoOp. CoOp outperforms on 6 datasets than both CoCoOp and KgCoOp in terms of accuracy on base classes. This can be attributed to CoOp's exclusive focus on learnable prompts, fostering the creation of distinctive prompts for base classes. However, it is precisely owing to CoOp's overfitting to the base classes that it exhibits an inferior generalization capability on the new class, compared to KgCoOp and CoCoOp.

Compared to existing methods, LoL, which adopts CoOp as the prompt tuning stage, possesses the superior generative capability for new classes over CoOp-based prompt methods, CoCoOp and KgCoOp. Simultaneously, LoL effectively alleviates the overfitting problem of CoOp, and ultimately achieving higher harmonic mean and accuracy in new classes than CoCoOp and KgCoOp across all 7 datasets.

Further Analysis

Generality with other methods. The method we propose boasts superior adeptly adapts to a variety of CoOp-based methods. This is attributable to our strategy of employing various models, post-training on Base classes, as the backbone for meta-learning stage. Compared to CoOp, KgCoOp mitigates the forgetting of fundamental knowledge by minimizing the discrepancy between text embeddings from learnable prompts and handmade prompts. Therefore, KgCoOp obtains a higher performance on New classes than CoOp. Consequently, we embedded our technique within the KgCoOp method, with the experimental results demonstrated in Table 3.

It is evident from Table 1 and Table 2 that under most experimental conditions, KgCoOp delivers superior performance on new classes compared to CoOp. Similarly, under

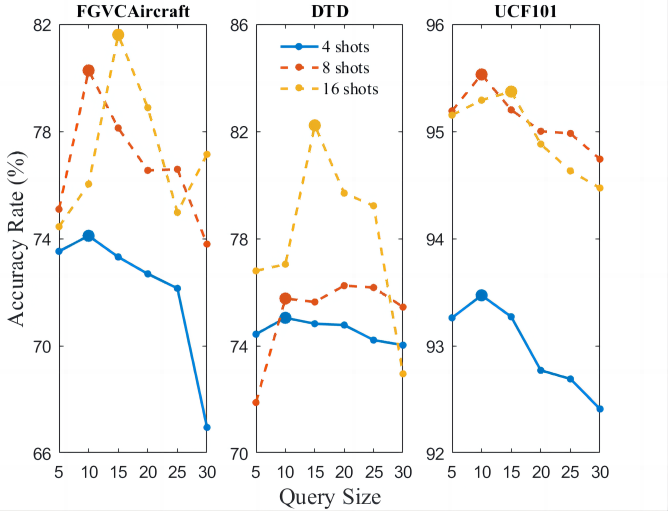

Figure 2: Query Size from 5 to 30. Comparison in the base-to-new setting with different K-shot samples in terms of the performance among 3 datasets and ViT-B/16 backbones.

all experimental conditions, our KgCoOp-based LoL outperforms the CoOp-based LoL when it comes to new classes. For instance, under the 1-shots condition, the KgCoOp-based method surpasses the CoOp-based approach by 0.1%, 0.1%, and 1.73% on the SUN397, Flowers102, and DTD datasets respectively. The experimental results from Table 3 substantiate the effectiveness of our method for new-class generalization and its commendable generality with other methods.

Initialization. To comprehend the ramifications of initialization, we conducted an ablation study, comparing word-embedding-based initialization with random initialization, whilst maintaining consistency with other settings. Implementing CoOp (Zhou et al. 2022b) for word-embedding-based initialization, we designated "a photo of a" for the initialization of FGVCAircraft and UCF101 datasets, and "a texture of a" for the DTD dataset. In the case of random initialization, we adopted CoOp, drawing samples from a Gaussian distribution with zero mean and a standard deviation of 0.02. To establish an equitable comparison, we also

configured the context length to 4 during random initialization.

Table 4 signifies that the meticulously architected word-embedding-based initialization significantly amplify the efficacy of our method. We posit that crafted prompt words, when generalized to new classes, can assimilate the universal textual knowledge embedded within the CLIP pre-trained model, and partially mitigate the overfitting phenomenon of CoOp. Although the refinement of initial words may engender certain advantages, in essence, it is an intricate and laborious task to design appropriate prompt words for different models. Therefore, in practice, a delicate balance needs to be achieved between performance and model training complexity with respect to context length.

Query size. Our methodology harnesses the meta-learning, commonly employing an approach referred to as Episodic Training for instruction. Each episode typically refers to the training samples in the training set as the "support set" and those in the testing set as the "query set". In Fig.2, while keeping all other parameters consistent, we adjusted the size of the query in each episode. All experiments utilized random initialization with a context length set to 4. The experiment manifests that either excessively large or small query sizes impair the model's generalization performance on novel classes. To be more precise, under the condition of 16 shots, a query value of 15 yields the most optimal results. Whereas, for 4 shots and 8 shots, a query value of 10 is most efficacious. We surmise that, due to the adoption of the N-way K-shot sampling method, there might be a direct correlation between the increase in shots and the magnitude of the query. Nonetheless, considering the specific conditions for each dataset, there remains room for optimization of the query size.

Conclusion

In this paper, we proposed a new "Learning to Learn" (LoL) approach for learning better visual prompts. Instead of the conventional way of manually creating prompts like in CLIP, we adopt adaptable vectors as text prompts and learn the prompt by an N-way K-shot task inspired by meta-learning. We initially perform fine-tuning on the base class and then carry out further fine-tuning in an N-way K-shot training on the base classes. Our proposed method not only excels in New class representation but also seamlessly integrates into extant prompt tuning frameworks. Comprehensive evaluations across multiple benchmark datasets attest to the efficacy of our proposed LoL method as a potent prompt refinement strategy.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 623762829).

References

Baik, S.; Choi, M.; Choi, J.; Kim, H.; and Lee, K. M. 2020. Meta-learning with adaptive hyperparameters. *Advances in neural information processing systems*, 33: 20755–20765.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35: 32897–32912.

Chen, C.; Zhang, B.; Cao, L.; Shen, J.; Gunter, T.; Jose, A. M.; Toshev, A.; Shlens, J.; Pang, R.; and Yang, Y. 2023. STAIR: Learning Sparse Text and Image Representation in Grounded Tokens. *arXiv preprint arXiv:2301.13081*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 19358–19369.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 178–178. IEEE.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1126–1135. PMLR.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2023. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; Gao, J.; et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4): 163–352.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Garnelo, M.; Rosenbaum, D.; Maddison, C.; Ramalho, T.; Saxton, D.; Shanahan, M.; Teh, Y. W.; Rezende, D.; and Eslami, S. A. 2018. Conditional neural processes. In *ICML*, 1704–1713. PMLR.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, 15180–15190.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hospedales, T.; Antoniou, A.; Micaelli, P.; and Storkey, A. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 4904–4916. PMLR.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40: e253.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*, 201–216.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 121–137. Springer.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023. Scaling language-image pre-training via masking. In *CVPR*, 23390–23400.

Liu, B.; Cao, Y.; Lin, Y.; Li, Q.; Zhang, Z.; Long, M.; and Hu, H. 2020. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, 438–455. Springer.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mishra, N.; Rohaninejad, M.; Chen, X.; and Abbeel, P. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 529–544. Springer.

Munkhdalai, T.; and Yu, H. 2017. Meta networks. In *ICML*, 2554–2563. PMLR.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*, 722–729. IEEE.

Pantazis, O.; Brostow, G.; Jones, K.; and Mac Aodha, O. 2022. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rajeswaran, A.; Finn, C.; Kakade, S. M.; and Levine, S. 2019. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2018. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35: 15558–15573.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *CVPR*, 15638–15650.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.

Tan, H.; Bansal, M.; and Assoc Computat, L. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*, 5100–5111. ISBN 978-1-950737-90-1.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, J.; Zhou, P.; Shou, M. Z.; and Yan, S. 2023. Position-guided Text Prompt for Vision-Language Pre-training. In *CVPR*, 23242–23251.

Wei, L.; Xie, L.; Zhou, W.; Li, H.; and Tian, Q. 2022. Mvp: Multimodality-guided visual pre-training. In *ECCV*, 337–353. Springer.

Wu, W.; Timofeev, A.; Chen, C.; Zhang, B.; Duan, K.; Liu, S.; Zheng, Y.; Shlens, J.; Du, X.; Gan, Z.; et al. 2023. MOFI: Learning Image Representations from Noisy Entity Annotated Images. *arXiv preprint arXiv:2306.07952*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492. IEEE.

Xu, J.; Ton, J.-F.; Kim, H.; Kosiorek, A.; and Teh, Y. W. 2020. Metafun: Meta-learning with iterative functional updates. In *ICML*, 10617–10627. PMLR.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022. Unified contrastive learning in image-text-label space. In *CVPR*, 19163–19173.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 6757–6767.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 18123–18133.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2023. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Zhou, J.; Dong, L.; Gan, Z.; Wang, L.; and Wei, F. 2023. Non-contrastive learning meets language-image pre-training. In *CVPR*, 11028–11038.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2022. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*.

Zintgraf, L. M.; Shiarlis, K.; Kurin, V.; Hofmann, K.; and Whiteson, S. 2018. CAML: Fast Context Adaptation via Meta-Learning. *arXiv preprint*, arXiv:1810.03642.