Diversity-Aware Meta Visual Prompting

Qidong Huang[1]   Xiaoyi Dong[1]   Dongdong Chen[2]   Weiming Zhang[1,*]

Feifei Wang[1]   Gang Hua[3]   Nenghai Yu[1]

[1]University of Science and Technology of China   [2]Microsoft Cloud AI   [3]Wormpex AI Research

{hqd0037@mail., dlight@mail., zhangwm@, wangfeifei@mail., ynh@}ustc.edu.cn

{cddlyf@, ganghua@}gmail.com

Abstract

*We present Diversity-Aware Meta Visual Prompting (DAM-VP), an efficient and effective prompting method for transferring pre-trained models to downstream tasks with frozen backbone. A challenging issue in visual prompting is that image datasets sometimes have a large data diversity whereas a per-dataset generic prompt can hardly handle the complex distribution shift toward the original pretraining data distribution properly. To address this issue, we propose a dataset **D**iversity-**A**ware prompting strategy whose initialization is realized by a **M**eta-prompt. Specifically, we cluster the downstream dataset into small homogeneity subsets in a diversity-adaptive way, with each subset has its own prompt optimized separately. Such a divide-and-conquer design reduces the optimization difficulty greatly and significantly boosts the prompting performance. Furthermore, all the prompts are initialized with a meta-prompt, which is learned across several datasets. It is a bootstrapped paradigm, with the key observation that the prompting knowledge learned from previous datasets could help the prompt to converge faster and perform better on a new dataset. During inference, we dynamically select a proper prompt for each input, based on the feature distance between the input and each subset. Through extensive experiments, our DAM-VP demonstrates superior efficiency and effectiveness, clearly surpassing previous prompting methods in a series of downstream datasets for different pretraining models. Our code is available at: https://github.com/shikiw/DAM-VP.*

1. Introduction

With the increasing scale of training data and model size, the pretraining-finetuning paradigm has shown remarkable achievement in many areas, including natural language processing (NLP) [4,13] and computer vision (CV) [2,7,8,19].
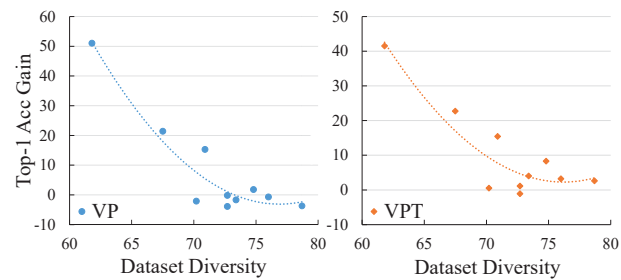
*Corresponding author.

Figure 1. Relation between dataset diversity and the performance gain got by using prompting. The gain is the performance improvement when compared with the linear-probing accuracy, under the head-tuning setting. Both previous methods get a large performance gain on low-diversity datasets, while failing to boost the transfer performance on high-diversity datasets.

However, fully finetuning a large pre-trained model for each small downstream task still has some problems in real-world usage. The most practical one is the storage and distribution problem that we have to maintain an independent copy of the model for each task, which is quite expensive and inflexible, especially for increasing numbers of downstream tasks [9].

To break the dilemma, many efforts [6,17,18,25,51] have been paid to efficiently transfer the given pre-trained models into a particular dataset. Prompting is an extensively studied method in the NLP area, which appends a few tokens before the input sequence to provide some task-specific knowledge to the pre-trained model, so that the model could adapt well on the downstream tasks without the fully-finetuning. Inspired by the success of prompting in NLP, some recent works [1, 26] propose visual prompting for vision models. By adding some learnable noise onto the input image or appending some learnable tokens to the model input sequence, the pre-trained models show promising results on different kinds of downstream tasks.

However, we argue that these methods ignore the diverse distribution property of the image dataset and using a sin-

gle prompt for all the images in each dataset is not optimal. In Figure 1, we show the relationship between the gain from prompting and the diversity of the dataset. Here the gain represents the accuracy improvement compared with the linear probing setting. We find that both VP [1] and VPT [26] improve the model accuracy by a large margin on the low-diversity dataset, but relatively small gains on the high-diversity datasets, which is intuitively sensible. For low-diversity datasets, such as the street view house number dataset (SVHN) [37], all the images have similar content so a unified prompt is sufficient. On the contrary, when it comes to high-diversity datasets, such as the ImageNet [12] dataset, it covers very diverse classes from the wordnet and there is not any pre-defined relationship between the classes, so it is hard to use a single prompt to provide the prior for all the images, such as for "car" and "dog".

Motivated by this observation, we propose our Diversity-Aware Meta Visual Prompting (DAM-VP). It has two core designs. Firstly, to provide a proper prompt for each image from high-diversity datasets, we propose a clustering-based prompt selection method. In detail, given a pre-trained visual model and a downstream dataset, we use the off-the-shelf clustering method to cluster the feature of the downstream data into several coarse-grained subsets, and guide each cluster to learn its own prompt separately. Based on the strong homogeneity of the same clustered data, the optimization of cluster-specific visual prompts can be greatly facilitated and the data commonalities can be also easily covered. Secondly, we argue the prompt across different clusters or datasets may have some shared pattern, from which the model can be adapted to a new dataset faster and get better performance. This motivates us to introduce a meta-learning-based method that learns a meta prompt and initializes the prompt of each cluster with it.

We conduct our experiments on datasets with different data diversity and evaluate the transfer performance with different pre-trained models. We report the performance on both the widely used head-tuning setting and a more challenging head-freezing/missing setting. Our DAM-VP outperforms previous methods by a large margin, especially on high-diversity datasets. For example, with the ImageNet-22k pre-trained ViT-B model, DAM-VP gets 73.1% top-1 accuracy under the head-tuning setting on the diverse DTD [10] dataset, surpassing previous methods VP [1] and VPT [26] with +13.6% and +7.3% respectively. Meanwhile, we find DAM-VP is quite efficient that with only 10 epoch tuning, it gets 85.7% average top-1 accuracy over the 10 datasets, comparable with previous methods that tunes 100 epochs (83.4% for VP [1] and 85.5% for VPT [26]). Our contributions can be summarized as follows:

- We analyze the limitation of previous visual prompting methods, and point out that vision-suitable prompting should consider the dataset diversity.

- Accordingly, we propose a novel Diversity-Aware Meta Visual Prompting (DAM-VP) method. It uses the divide-and-conquer idea by clustering high-diversity datasets into subsets and learning separate prompts for each subset, in cooperation with a meta-prompt learning design.

- Through extensive experiments, our DAM-VP demonstrates superior performance, achieving SOTA performance in a series of downstream datasets for different pretraining models.

2. Related Work

Prompt learning. Served as a new paradigm, prompting [33] originally emerges in NLP for adapting pre-trained language models (PLM) [4, 13] to downstream tasks. Its principle idea is to reformulate downstream data into the model knowledge learned during the pretraining phase, enabling the frozen pre-trained model to understand the task rather than tuning the model parameters for adaption. This goal has been initially reached through constructing pure text prompts that contains task-specific templates and label words to perform cloze test, *e.g.*, hand-craft prompts [16] and generative text prompts [27, 45], but unfortunately, still requiring specialized linguistic expertise for preparation. To alleviate this, recent efforts have been paid on prompt tuning (PT) [30, 31] that learns a task-specific continuous vector as tunable prefix tokens. These tokens can be optimized via gradients to act as prompts in task adaption while maintaining the pre-trained model untouched. Driven by the success of language prompts, a lot of works [35, 38, 43, 47, 50, 55], like CoOP [57] and CoCoOP [56], have been mushroomed to explore vision-related prompting especially in multi-modal scenarios, while still concentrating on text prompting in practice. Due to the gap of information density [19] between languages and images, prompting for vision models is more challenging and complex. Inspired by prefix tuning, VPT [26] takes the first step to adapt vision transformers to downstream tasks by prepending a set of learnable tokens at the model input. Concurrently, VP [1] follows the pixel-level perspective to optimize task-specific patches that are incorporated with input images. Despite the pioneering successes of VP and VPT, we find that they need to pre-assign the number of prompts, which is not flexible to handle the datasets with different diversities. In contrast, our method uses a diversity-adaptive solution to well address this issue.

Transfer learning. Typically, transfer learning focuses on how to efficiently fine-tune the supervised or un-/self-supervised pre-trained model when tackling with a new task. A conventional art of transfer learning is to fully fine-tune all of the model parameters on training data of the new task, using the pretraining knowledge as model initializa-
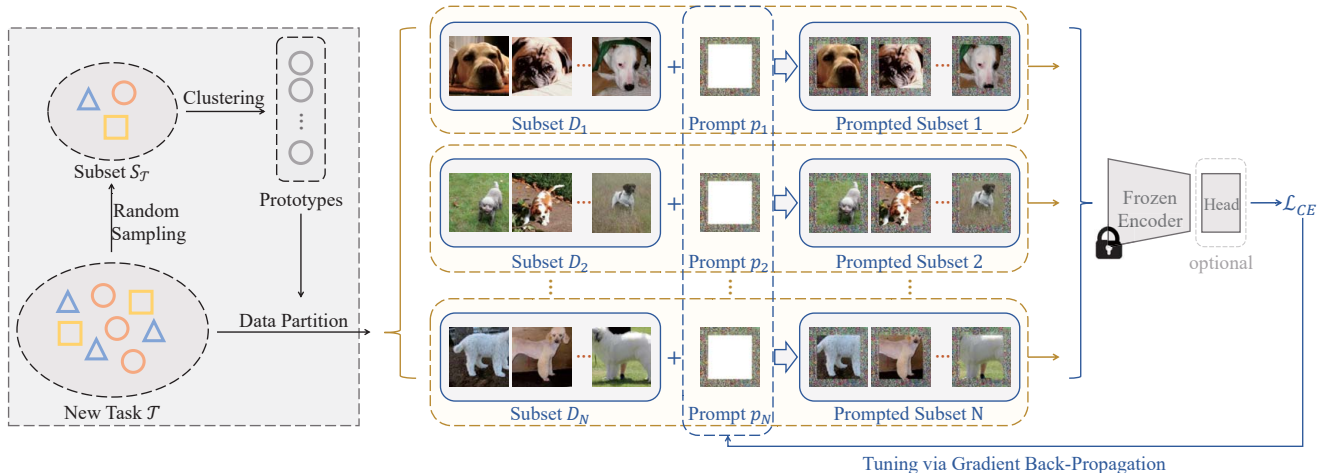
Figure 2. The pipeline of our diversity-aware adaption for frozen pre-trained encoder $\mathcal{M}$ on new task $\mathcal{T}$. We randomly select a little subset $\mathcal{S}_\mathcal{T}$ from the training data of task $\mathcal{T}$ to implement agglomerative clustering and simulate prototypes of clusters. These prototypes are utilized to partition the whole training set into different subsets, so that we can optimize the prompt for each subset separately.

tion. However, the growing model capacity has exposed the inefficiency of fully fine-tuning, simulating the demand of parameter-efficient learning on downstream tasks, *i.e.*, selecting or appending a few parameters for tuning and freezing the remaining of the model meanwhile. This topic has been widely explored in NLP by prepending extra learnable tokens or feature vectors [30, 31] at transformer input, whereas limited in vision which still focuses on ConvNets [5, 44, 53] and rarely on the emerging vision transformers. To mitigate this gap, recent efforts have explored to efficiently transfer vision transformers by introducing a parallel trainable down-to-up branch into the MLP module of self-attention blocks [6] or scaling and shifting the learned features [32]. VPT [26] is the pioneer work to leverage learnable prefix tokens/features for visual prompting, but still not efficient in terms of convergence time. Driven by this, our DAM-VP strives to learn faster during tuning with comparable amount of learnable parameters introduced.

3. Method

We introduce our Diversity-Aware Meta Visual Prompting (DAM-VP), a novel prompting method that is adaptive to diverse downstream datasets effectively and efficiently. As shown in Figure 2, with a given dataset, DAM-VP first extracts its specific prototypes in an unsupervised and adaptive manner as the pre-processing. Then we split the dataset into different subsets, according to the prototypes. For each subset, we assign a specific prompt to it and optimize it with the tuning loss. Rather than random initialization, all of the prompts are initialized by a meta prompt learned across different datasets, as shown in Figure 4.

Diversity-adaptive dataset partition. As we briefly introduced in Sec. 1, different image datasets have different dis-

Figure 3. Examples of GTSRB (**top**) and SUN397 (**bottom**).

tribution diversity. Take Figure 3 as an example, when comparing with the traffic sign dataset [46], the data in the scene dataset [49] is more diverse in terms of angle, illumination, the complexity of content, *etc*. The prompting is designed to reduce the distribution gap between the target downstream dataset and the model pretraining data, and it is intuitive that a dataset with similar content is easy to transfer. So it is a straightforward idea to split a diverse dataset into small subsets and apply different prompts to each subset.

To this end, we consider the diversity property in our visual prompting design and propose an adaptive dataset partition strategy to suit the diversity of task data automatically. Specifically, when adapting the frozen pre-trained backbone $\mathcal{M}$ to the new task $\mathcal{T}$, we first randomly sample a small subset $\mathcal{S}_\mathcal{T}$ from the training set of $\mathcal{T}$. Then we extract the feature of subset $\mathcal{S}_\mathcal{T}$ with the frozen $\mathcal{M}$ without any prompting. We denote the features as $\{\mathcal{M}(s)|s \in \mathcal{S}_\mathcal{T}\}$ and use an off-the-shelf clustering method [36] to aggregate them into several clusters. The clustering procedure is time-efficient (less than 1% of total tuning time) and the number of clusters N is auto-adaptive to the dataset diversity by a pre-defined threshold. Once the clusters are constructed, we compute the average values for features of each cluster
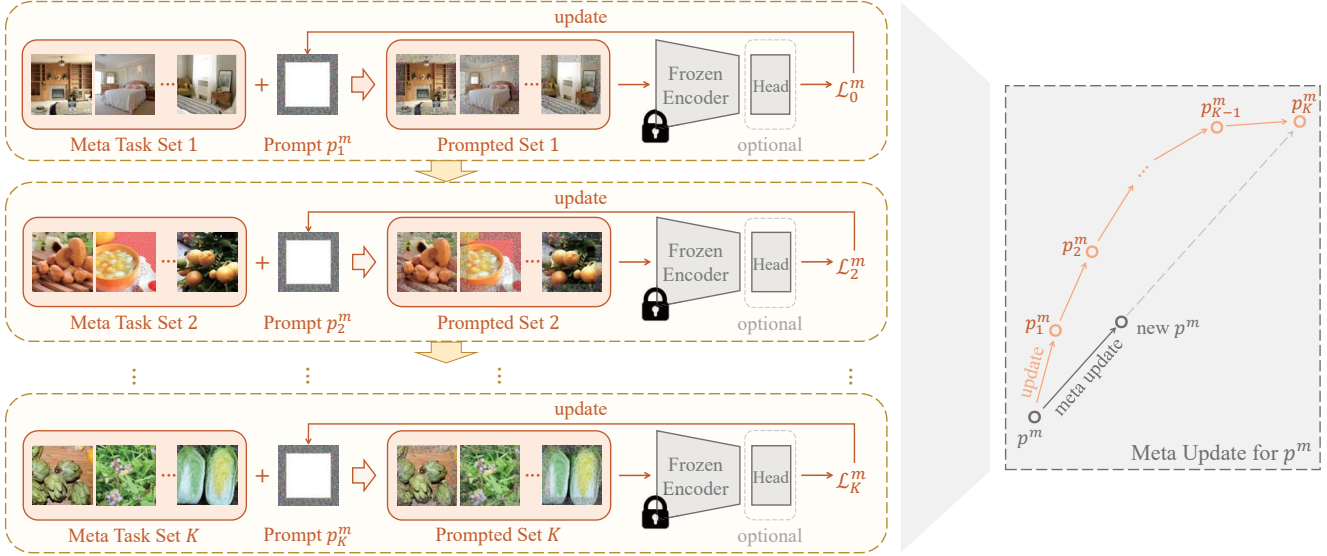
Figure 4. The pipeline of meta-learning-based prompt initialization. We partition each task dataset into subsets and regard each subset as a single meta task. During meta training on each meta batch, we continuously update the temporary prompt on each meta task set and finally apply moving average to get the updated meta prompt.

as the cluster-specific prototypes $\{c_i\}_{i=1}^{N}$, *i.e.*,

$$c_i = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} \mathcal{M}(s), \quad i = 1, \cdots, N, \qquad (1)$$

where $\mathcal{S}_i$ is the data samples corresponding to the ith cluster that satisfies $|\mathcal{S}_1| + |\mathcal{S}_2| + \cdots + |\mathcal{S}_N| = |\mathcal{S}_{\mathcal{T}}|$. The unsupervised mechanism naturally guarantees that the dataset with a higher diversity can be divided into more clusters, and the dataset with a lower diversity can be divided into fewer clusters or even just a single cluster. In practice, we configure the threshold of clustering properly to keep N in a reasonable range, usually less than the data category number.

Diversity-aware prompt selection. With the simulated prototypes $\{c_i\}_{i=1}^{N}$, it is easy to partition the whole training dataset $\mathcal{D}_{\mathcal{T}}$ of task $\mathcal{T}$ into small subsets $\{\mathcal{D}_i\}_{i=1}^{N}$. Correspondingly, we assign one visual prompt for each subset and get N visual prompts $\{p_i\}_{i=1}^{N}$ in total. Similar to the design in [1], we use a photo-frame-like pixel-level prompting. It has the same size as the model input and we add it to the input image directly. Such a design has two advantages: Firstly, it does not introduce additional computation cost during inference, while the prefix-token prompting used in VPT [26] increases the input length and leads to a larger computation cost. Secondly, such pixel-level prompting is irrelevant to the model type, so it could be used in both recent popular Vision Transformer models and traditional convolution networks. On the contrary, the prefix-token prompting is specifically designed for token-list-type input so that could only be used for Vision Transformer.

Given an image-label pair (x, y) from the training set,

we forward x on the frozen $\mathcal{M}$ without prompting to get its original feature and compute the euclidean distances between the feature and each aforementioned prototype. The prompt that corresponds to the minimal distance is considered as the prompt added on x. Formally, prompted image x^p is defined as

$$x^p \triangleq x + p_t, \quad \text{s.t.} \quad t = \arg\min_i \|\mathcal{M}(x) - c_i\|_2^2. \quad (2)$$

which indicates the input image x is assigned into the tth image subset $\mathcal{D}_t$.

Prompt learning. In this paper, we seek a more general prompting format that could be utilized in different settings. 1) The head-tuning setting used in VPT [26] that a learnable classification head is optimized with the prompt jointly. 2) The head-freezing/missing setting that only the prompt is learnable. The second one is a more challenging task, but it is also a more flexible real-world usage format: we only need to add different prompts on the input for different tasks, maintaining an end-to-end frozen pre-trained model.

For the head-tuning setting, we follow the design in VPT [26] that optimize a new k-class classification head for the target task with k categories. Similarly, for the head-freezing case, we assign the first k classes in the frozen head to the new task. When it comes to the head-missing case that the pre-trained model does not have a classification head (such as self-supervised pre-trained models), a simple hard-coded mapping solution [1] is convert the output feature (*e.g.*, 768 channels output feature of ViT/16-Base [14] encoder) to classification logits. However, we argue such hard-coded mapping is inefficient, because the optimization

might be limited if the neurons at these fixed positions are not active enough. So we propose an active-based mapping method, which selects the most active top-k neurons of the output layer of $\mathcal{M}$ by measuring the variances of each position output when confronted with random noise inputs.

Based on the above designs, we can minimize the cross-entropy loss between the logits of prompted image x^p and ground-truth label y to tune our visual prompts on $\mathcal{D}_{\mathcal{T}}$, *i.e.*,

$$p_1^*, \ldots, p_N^* = \arg\min_{p_1, \ldots, p_N} \frac{1}{|\mathcal{D}_{\mathcal{T}}|} \sum_{i=1}^{N} \sum_{x \in \mathcal{D}_i} \mathcal{L}_{CE}(\mathcal{M}(x + p_i), y).$$
(3)

Once the prompts are well-optimized, we can utilize prototypes to categorize the input image and assign it to its belonging subset during inference. The corresponding prompt is subsequently incorporated with the input image as prompted input to get the classification logits.

Prompt boosting via Meta-learning. In real-world usage, a frozen model $\mathcal{M}$ needs to be transferred to a bunch of downstream tasks, this leads to two desired favorable properties of the prompting method. First, it should be efficient that only a few epochs tuning could get a good result. Second, the prompts learned from previous tasks could help the new tasks learn better prompts so that the method is bootstrapped.

In light of meta-learning [15, 39], we integrate a quick algorithm Reptile [39] into our diversity-aware learning to boost the prompting learning. The principle idea is to learn a meta prompt p^m on several task datasets $\{\mathcal{D}_i^m\}_{i=1}^{M}$ that are prepared as the meta training data, and adopt the well-trained p^m as the initial prompt for diversity-aware adaption. Specifically, we first partition each task dataset into subsets, *e.g.*, dataset $\mathcal{D}_i^m$ is divided into $\mathcal{D}_{i,1}^m, \mathcal{D}_{i,2}^m, \ldots, \mathcal{D}_{i,K_i}^m$. Then, we regard each subset as a single meta task and sample images from each subset to form the meta training batch $\mathcal{B}$, *i.e.*, dataset $\mathcal{D}_i^m$ contributes total K_i meta task sets for $\mathcal{B}$. Formally, a meta training batch is constructed by

$$\mathcal{B} = \bigcup_{j=1}^{K} \mathcal{B}_j \quad \text{s.t.} \quad \mathcal{B}_j \in \mathcal{G}_j.$$
(4)

where K is the total number of subsets that satisfies $K = K_1 + K_2 + \cdots + K_M$ and we rename each subset as group $\mathcal{G}_j$ for convenience, satisfying $\bigcup_{i=1}^{M} \bigcup_{k=1}^{K_i} \mathcal{D}_{i,k}^m = \bigcup_{j=1}^{K} \mathcal{G}_j$.

With the sampled meta training batch $\mathcal{B}$, we can update the temporary prompt p_j^m on each meta task set $\mathcal{B}_j$ by minimizing cross-entropy loss $\mathcal{L}_j^m$, *i.e.*,

$$p_j^m = p_{j-1}^m - \eta \nabla_{p_{j-1}^m} \frac{1}{|\mathcal{B}_j|} \sum_{x \in \mathcal{B}_j} \mathcal{L}_j^m,$$
$$\text{s.t.} \quad \mathcal{L}_j^m = \mathcal{L}_{CE}(\mathcal{M}(x + p_{j-1}^m), y),$$
(5)

where y is the ground-truth label of x. Finally, we implement meta update to get the new meta prompt after training on meta batch $\mathcal{B}$ by moving average,

$$p^m \leftarrow p^m + \gamma \frac{1}{K} \sum_{j=1}^{K} (p_j^m - p^m),$$
(6)

where γ is the meta update step that varies within $(0, 1)$.

4. Experiments

4.1. Setup

Datasets. Here we select 16 popular image datasets for the experiments including CIFAR10 [29], CIFAR100 [29], DTD [10], CUB200 [48], NABirds [22], Stanford-Dogs [28], Oxford-Flowers [40], Food101 [3], GTSRB [46], SVHN [37], SUN397 [49], STL10 [11], Fru92 [23], Veg200 [23], Oxford-IIIT Pet [41] and EuroSAT [21], where the first 10 datasets are used for prompt evaluation and the remaining 6 are prepared for our meta prompt initialization. For data preprocessing, we randomly resize the input image into the size of 256×256 and subsequently crop it into 224×224. More details including the performance on VTAB-1k benchmark [52] are provided in supplementary.

Models. Our experiments involve six pre-trained vision models including ImageNet-1k [12] supervised ViT-B/16 [14], supervised ResNet-50 [20], MoCo v3 [8] learned ViT-B/16; ImageNet-22k supervised ViT-B/16 and Swin-Base [34]; 400m web data contrastive learning ViT-B/16 model CLIP [43]. As we illustrated above, our DAM-VP could be used for models without classification heads and traditional convolution networks, we discuss these settings in the supplementary materials.

Baselines. We compare our method with both parameter-tuning methods and prompt-tuning methods. For parameter-tuning, we report the fully-tuning, linear probing results as baseline, and the efficient-tuning method Adapter [24, 42]. For prompt-tuning, we compare with the VP [1] and VPT [26].

Diversity metrics. To quantitatively measure the data diversity of a given dataset, we follow [1] to randomly sample 10,000 image pairs in each dataset and compute the LPIPS distance [54] of each pair. The average LPIPS is regarded to measure the perceptual diversity of the given dataset.

Implementation details. For meta learning based prompt initialization, we unify update learning rate η (in Eq. (5)) as 0.5 and meta step size γ (in Eq. (6)) as 0.5. Our meta prompts of different pre-trained backbones are equally trained for 200 epochs by Adam optimizer with a cosine annealing schedule. The step number for updating the temporary prompt on each meta task is set as 4. For diversity-aware adaption, we set the subset size $|\mathcal{S}_{\mathcal{T}}|$ as 1000 by default. More implementation details, such as training or clustering configurations, are provided in our supplementary.

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	70.6	84.7	72.3	84.6	98.3	83.0	87.5	97.4	96.8	96.9	87.2
Linear	✓	68.7	83.5	69.3	84.4	97.7	78.5	77.6	92.9	65.6	61.1	77.9
Adapter [24, 42]	✗	47.7	17.5	3.8	32.0	40.1	46.1	43.0	72.8	82.2	19.6	40.5
VP [1]	✗	47.8	40.6	13.8	61.9	56.5	55.7	54.4	92.9	86.0	87.8	59.7
VPT [26]	✗	27.8	10.9	1.3	36.7	9.3	63.0	31.8	46.1	84.3	28.2	33.9
DAM-VP (10 epochs)	✗	51.3	43.6	22.3	70.8	65.9	61.5	61.5	90.5	79.7	85.7	63.3
DAM-VP (50 epochs)	✗	53.9	64.6	38.6	75.5	84.1	66.6	67.2	92.4	86.2	88.4	71.8

Table 1. Head-freezing/missing adaption performance of different methods on ViT-B-1K, where we report image classification accuracy and all of baseline methods are trained for 50 epochs.

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	78.6	81.9	72.6	80.5	97.3	91.8	80.9	96.3	95.3	95.7	87.1
Linear	✓	77.0	81.5	72.4	79.5	95.8	92.2	79.8	95.0	85.6	69.2	82.8
TP [43]	✗	41.8	55.5	44.7	62.5	65.5	87.7	64.6	87.5	40.2	13.7	56.4
VP [1]	✗	54.3	56.8	46.3	63.5	71.7	86.5	70.7	93.2	90.5	90.4	73.4
DAM-VP (10 epochs)	✗	58.4	61.1	49.6	68.6	84.5	85.0	68.3	92.7	87.6	87.9	74.4
DAM-VP (50 epochs)	✗	63.7	65.9	54.8	71.5	87.0	87.7	72.2	93.7	92.3	90.4	77.9

Table 2. Head-freezing/missing adaption performance of different methods on CLIP-ViT-B, where we report image classification accuracy and all of baseline methods are trained for 50 epochs. "TP" denotes text prompting that directly adopts zero-shot classification head of CLIP. Here "VP" and our method also use this fixed head rather than particular mapping to get output logits.

4.2. Comparison with Baseline Methods

Quantitative results. We comprehensively compare our method with the baselines mentioned in Section 4.1. Two scenarios are taken into our consideration: 1) **head-freezing/missing adaption**, *i.e.*, only tuning the introduced modules like prompts without any extra task-specific head. 2) **head-tuning adaption**, *i.e.*, tuning the introduced modules like prompts along with learning a task-specific head. Table 1 and 2 show the quantitative comparison results of the head-freezing/missing setting between our DAM-VP and other adaption methods. We can find that our method significantly outperforms baseline methods even when our prompts are just trained for 10 epochs. Table 4 and 5 show the quantitative comparison results of the head-tuning setting between our DAM-VP and other adaption methods. It is easy to discover that DAM-VP presents its strong ability to help pre-trained models generalize on various image datasets, surpassing other methods with fewer training epochs and higher recognition accuracy. Furthermore, our DAM-VP even outperforms the fully-tuning setting on both the ViT-B-22K model and the Swin-B-22K model with only 50 epochs tuning.

Qualitative results. To better observe the adaption performance of each method, we depict the Top-1 accuracy curve of the first 50 training epochs to investigate the differences. Figure 7 shows the curve results of DAM-VP and the other three baselines in both head-freezing/missing and

Setting	Meta	Diversity	CUB200 [48]	Flowers [40]	CIFAR100 [29]	SVHN [37]
A	✗	✗	41.2	59.9	55.1	87.9
B	✓	✗	43.9	65.7	59.4	88.1
C	✗	✓	63.3	81.5	66.9	88.2
D	✓	✓	64.6	84.1	67.2	88.4

Table 3. Ablation results of diversity-aware strategy and meta-prompt initialization. We report Top-1 accuracy on ViT-B-1K.

head-tuning scenarios, where four datasets with different diversities are selected. It is obvious that, the performance of DAM-VP is far ahead in the early stage, especially in the first 10 epochs. This phenomenon indicates our diversity-aware strategy can boost the efficiency of prompt optimization, since each prompt just need to learn from a group of images that already have considerable homogeneity. The design of meta-prompt initialization also benefits this quick converging of our method with a good start point.

4.3. Ablation Study

Component ablation. We first verify the significance of the proposed diversity-aware strategy and the meta-prompt initialization on four aforementioned datasets that have different diversities. As listed in Table 3, both two components contribute a lot to boosting prompting performance, especially when dealing with task data that has high diversity.

Prompt learning stability. Previous methods mainly ini-

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	64.3	87.3	82.7	89.4	98.8	84.9	68.9	97.4	97.1	87.4	85.8
Linear	✓	63.2	85.3	75.9	86.2	97.9	84.4	63.4	96.3	68.0	36.6	75.7
Adapter [24,42]	✓	62.7	87.1	84.3	89.8	98.5	86.0	74.2	97.7	91.1	36.3	80.8
VP [1]	✓	59.5	84.6	77.7	84.5	97.7	80.5	78.7	94.2	89.4	87.6	83.4
VPT [26]	✓	65.8	88.5	84.2	90.2	99.0	83.3	78.8	96.8	90.7	78.1	85.5
DAM-VP (10 epochs)	✓	72.4	86.3	81.5	92.2	98.6	86.2	80.1	90.5	87.8	81.1	85.7
DAM-VP (50 epochs)	✓	73.1	87.5	82.1	92.3	99.2	86.9	88.1	97.3	90.6	87.9	**88.5**

Table 4. Head-tuning adaption performance of different methods on ViT-B-22K, where we report image classification accuracy and all of baseline methods are trained for **100 epochs**.

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	72.4	89.7	86.8	86.2	98.3	91.7	73.3	98.3	97.1	91.2	88.5
Linear	✓	73.6	88.6	85.2	85.9	99.4	88.2	61.6	96.3	83.8	43.5	80.6
Adapter [24,42]	✓	73.9	88.5	84.6	86.8	98.9	88.7	85.7	96.5	83.6	71.3	85.9
VP [1]	✓	75.1	86.5	82.9	81.3	98.6	83.4	80.6	94.8	82.4	80.3	84.6
VPT [26]	✓	78.5	90.0	85.4	84.8	99.3	90.1	80.5	96.9	86.2	87.8	87.9
DAM-VP (10 epochs)	✓	77.0	89.4	86.8	88.3	99.6	90.2	85.5	96.4	84.7	79.0	87.7
DAM-VP (50 epochs)	✓	80.0	90.4	86.9	88.5	99.6	90.5	88.1	97.3	86.8	81.7	**89.0**

Table 5. Head-tuning adaption performance of different methods on Swin-B-22K, where we report image classification accuracy and all of baseline methods are trained for **100 epochs**.
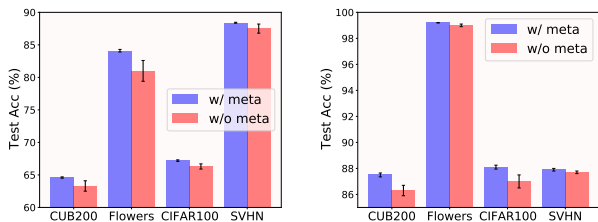

Figure 5. Meta-prompt initialization makes prompt tuning more robust to the random factor. We test in both head-freezing/missing (ViT-B-1K, **left**) and head-tuning (ViT-B-22K, **right**) cases.
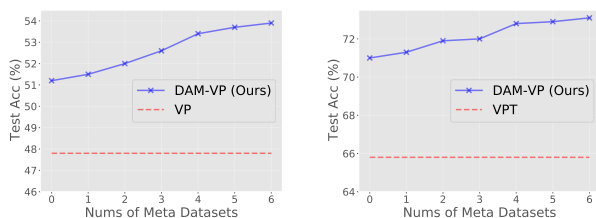

Figure 6. Ablating meta-prompt dataset number on DTD in the head-freezing/missing setting (ViT-B-1K, **left**) and the head-tuning (ViT-B-22K, **right**) setting.

tialize the prompt randomly, this leads to the instability of the performance caused by different random seeds. On the contrary, our meta-prompt design provides a good initialization point for optimization, which not only boosts the prompting performance but also improves the training stability. To eliminate the impact of random seeds, we test on 5 random seeds in experiments and report the results in Figure 5. It can be found that our meta-prompt initialization provides superb robustness across different training random seeds, while the randomly initialized prompt is unstable.

Meta-prompt dataset number. As we elaborated in Sec. 4.1, there are 6 datasets prepared for our meta-prompt learning, which is also the default setting we used. Here we ablate this setting by reducing the number of meta datasets. As shown in Figure 6, when the meta dataset number is 0, *i.e.*, no meta-prompt is used, our DVM-VP gets a rea-

sonable result that performs better than previous baseline methods. With the number of meta datasets increases, the prompting performance in both head-freezing/missing and head-tuning scenarios generally gets boosted. It proves that more prompting knowledge obtained from previous data is quite helpful for visual prompts to reduce the data distribution gap between downstream tasks and pretraining tasks.

Meta-prompt update step size η. The step size η is a crucial hyper-parameter applied in the meta-prompt update. Here we ablate different step sizes that vary from 0.1 to 0.7 with the results given in Figure 8. Compared with baselines, the performance is relatively robust to different step sizes. We choose $\eta = 0.5$ as the default configuration.

Different size of subset $\mathcal{S}_{\mathcal{T}}$. The size of subset $\mathcal{S}_{\mathcal{T}}$ can directly influence the clustering result and further affects the partition result during our diversity-adaptive dataset parti-
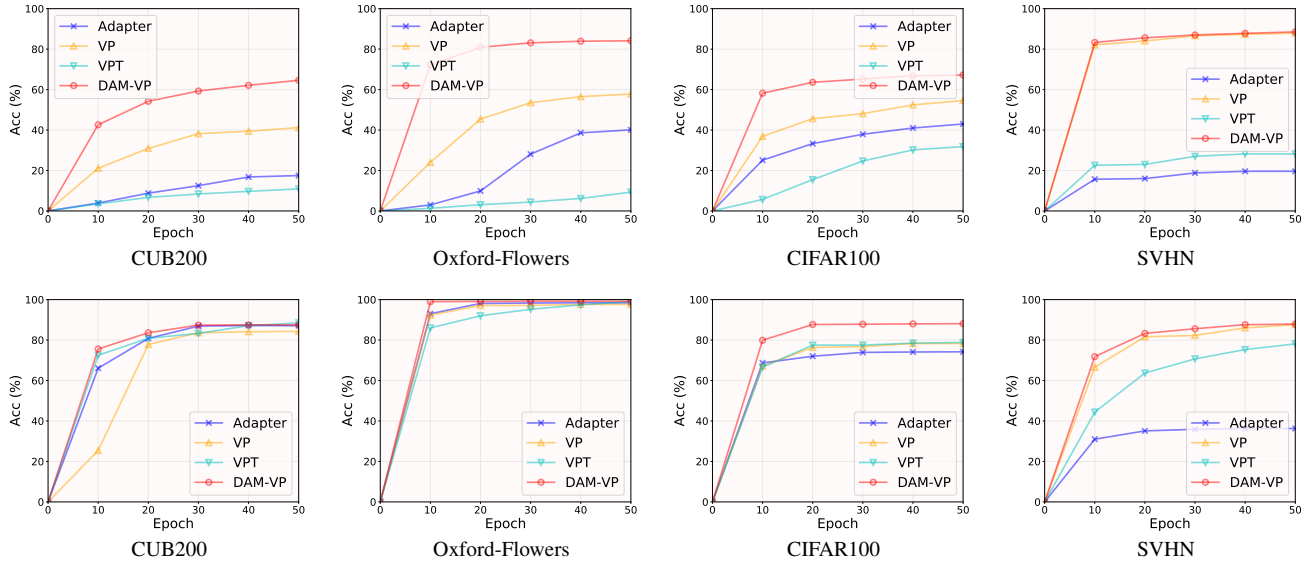
Figure 7. Test accuracy curves of different adaption methods when adapting pre-trained ViT-B-1K to different datasets in head-freezing/missing scenario (**top row**) and adapting pre-trained ViT-B-22K to different datasets head-tuning scenario (**bottom row**).
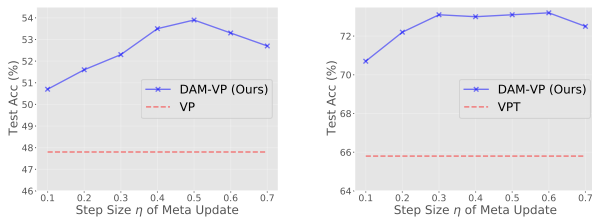

Figure 8. Ablating step size η on DTD in head-freezing/missing (ViT-B-1K, **left**) and head-tuning (ViT-B-22K, **right**) cases.
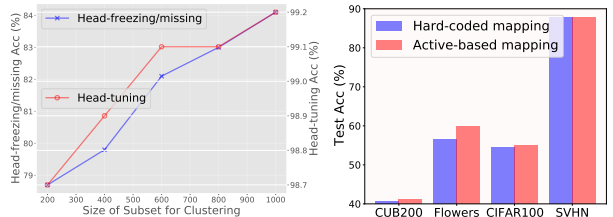

Figure 9. (**Left**) Ablating the size of subset $\mathcal{S}_{\mathcal{T}}$ in both head-freezing/missing (ViT-B-1K) and head-tuning (ViT-B-22K) cases. (**Right**) Comparison between hard-coded mapping and our active-based mapping used for head-freezing/missing case.

tion. Therefore, we apply different subset size that varies from 200 to 1000 to explore how it affects the prompting ability. From Figure 9 (left), we find that as the size of $\mathcal{S}_{\mathcal{T}}$ increases, we can get higher performance in both scenarios. Meanwhile, we find our divide-and-conquer design is quite efficient that even with only 200 images, our DAM-VP still outperforms previous methods by a large margin. Considering the training images of some datasets are limited, we adopt 1,000 as the default size of $\mathcal{S}_{\mathcal{T}}$ in this work.

Hard-coded mapping *vs*. Active-based mapping. As we discussed in Sec. 3, we argue the hard-coded mapping used by [1] for head-freezing/missing scenario is inefficient, since it might optimize some not active enough channels of the output feature (*i.e*., relatively robust to diverse model input), thus we propose active-based mapping to alleviate this issue. Figure 9 (right) show the comparison between these two mapping methods on VP, tested on ViT-B-1K.

5. Conclusion

This paper considers the data diversity property in downstream task adaption for prompting pre-trained vision mod-

els. We argue that the per-dataset generic prompt adopt in previous methods can hardly handle the dataset of large data diversity. To address this, we propose DAM-VP based on diversity-adaptive dataset partition and prompt selection, where our prompts is initialized by a meta-prompt that learns through a quick meta learning algorithm. Extensive experiments prove the superior performance of DAM-VP in both head-freezing/missing and head-tuning cases.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 1, 2, 4, 5, 6, 7, 8

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 1

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, 2014. 5, 6, 7

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 2

[5] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *NeurIPS*, 2020. 3

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, 2022. 1, 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1, 5

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *CVPR*, 2021. 1

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2, 5, 6, 7

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 5

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[14] 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 *ICLR*, 2021. 4, 5

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 5

[16] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, 2021. 2

[17] Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In *ACL*, 2021. 1

[18] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *ICLR*, 2021. 1

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[21] 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-STARS*, 2019. 5

[22] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015. 5, 6, 7

[23] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *ICCV*, 2017. 5

[24] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019. 5, 6, 7

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 1

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 7

[27] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 2020. 2

[28] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshops*, 2011. 5, 6, 7

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6, 7

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 2, 3

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021. 2, 3

[32] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*, 2022. 3

[33] 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. *arXiv preprint arXiv:2107.13586*, 2021. 2

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[35] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022. 2

[36] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011. 3

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2, 5, 6, 7

[38] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2

[39] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018. 5

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5, 6, 7

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[42] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *EMNLP*, 2020. 5, 6, 7

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5, 6

[44] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017. 3

[45] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. 2

[46] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 2012. 3, 5, 6, 7

[47] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 2022. 2

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 6, 7

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3, 5

[50] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 2

[51] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*, 2022. 1

[52] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 5

[53] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *ECCV*, 2020. 3

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[55] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022. 2

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2