DePT: Decoupled Prompt Tuning

Ji Zhang[1*] Shihan Wu[1*] Lianli Gao[2] Heng Tao Shen[3] Jingkuan Song[3,1†]
[1]University of Electronic Science and Technology of China (UESTC)
[2]Shenzhen Institute for Advanced Study, UESTC
[3]Tongji University
{jizhang.jim,jingkuan.song}@gmail.com

Abstract

*This work breaks through the Base-New Trade dilemma in prompt tuning, i.e., the better the tu generalizes to the base (or target) task, the wors alizes to new tasks, and vice versa. Specifically, t in-depth analysis of the learned features of the bas tasks, we observe that the BNT stems from a ch issue – the vast majority of feature channels are by base-specific knowledge, leading to the collap shared knowledge important to new tasks. To ad we propose the **De**coupled **P**rompt **T**uning (**DeP** work, which decouples base-specific knowledge ture channels into an isolated feature space duri tuning, so as to maximally preserve task-shar edge in the original feature space for achieving b shot generalization on new tasks. Importantly, is orthogonal to existing prompt tuning approa can enhance them with negligible additional compu cost. Extensive experiments on several datasets show the flexibility and effectiveness of DePT. Code is available at* https://github.com/Koorye/DePT.

1. Introduction

Recent years have witnessed remarkable progress in large vision-language pre-trained models (VLPMs). One of the striking successes has been achieved by the contrastive language-image pretraining (CLIP) [39] model, which formulates the learning objective as a contrastive loss to establish alignment between images and their textual descriptions in a common feature space. Despite the ability to capture open-set visual concepts, the zero-shot generalization performance of VLPMs is greatly reduced when there is a severe *category shift*, *distribution shift*, or *domain shift* between upstream training data and downstream tasks.

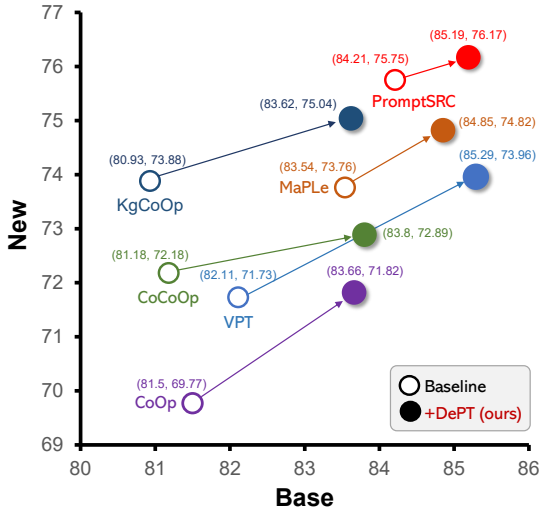

Figure 1. Classification ACCs of six prompt tuning methods w/ or w/o our DePT framework on **Base** (or seen) and **New** (or unseen) tasks. The results are the average on 11 datasets in Table 3.

Inspired by the success of prompt engineering in NLP, *Prompt Tuning* (or, *Context Optimization* [59]) has emerged as a parameter-efficient learning paradigm to adapt powerful VLPMs to downstream tasks, by optimizing a task-specific prompt (i.e., a set of trainable vectors) with a handful of training data from the base (target) task while keeping the weights of VLPMs frozen. Although the advantages are remarkable, existing prompt tuning methods usually fail to escape the Base-New Tradeoff (BNT) dilemma, i.e., the better the tuned/adapted model generalizes to the base task, the worse it generalizes to new tasks (with unseen classes), and vice versa. Numerous efforts [50, 58, 60] have been devoted in recent years to alleviate the performance degradation of tuned models on new tasks by developing anti-overfitting strategies in the process of prompt tuning. Nevertheless, the BNT problem is still far from being resolved and its underlying mechanisms are poorly understood.

In this work, we bridge the gap by proposing **De**coupled **P**rompt **T**uning (**DePT**), a first framework tackling the BNT

*Equal contribution.
†Corresponding author.

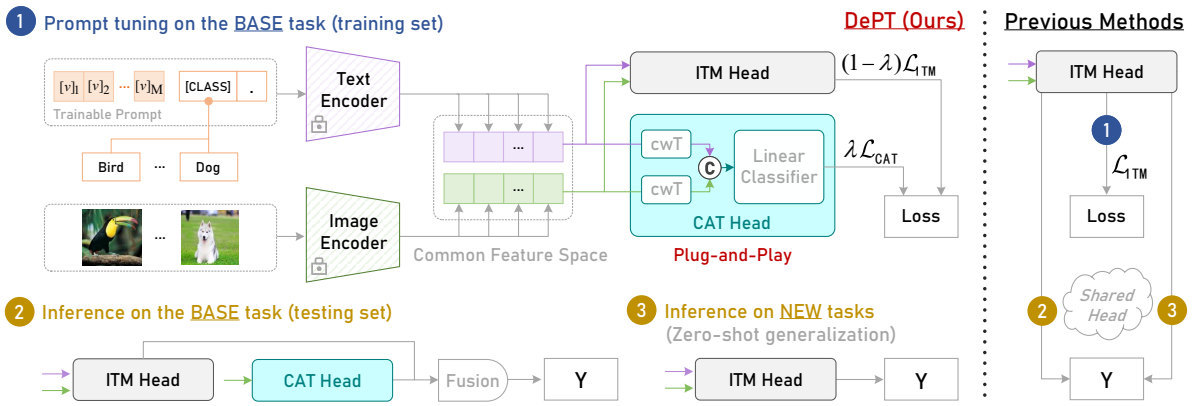

Figure 2. Illustration of our DePT framework (in a CoOp [59] style). Unlike previous methods (*right*) that use the same Image Text Matching (**ITM**) head for training/inference on the base task and zero-shot generalization on new tasks, our DePT (*left*) employs a Channel Adjusted Transfer (**CAT**) head to capture *base-specific* knowledge in an isolated feature space, so as to maximally preserve *task-shared* knowledge in the original feature space for improving zero-shot generalization on new tasks. At inference, we further boost the performance on the base task by simply fusing base-specific and task-shard knowledge obtained by the two heads. © denotes the concatenation operation.

problem in prompt tuning from a feature decoupling perspective. Specifically, through an in-depth analysis of the feature channels of base and new tasks learned by the standard Image Text Matching (ITM) head, we discern that the BNT stems from a *channel bias* issue: the vast majority of feature channels are occupied by *base-specific* knowledge (i.e., task-specific knowledge of the base task), resulting in the collapse of *task-shared* knowledge important to new tasks (Section 2.2). Inspired by this, the direct strategy to tackle the BNT problem is to decouple base-specific knowledge and task-shared knowledge in feature channels during prompt tuning. To accomplish this, we introduce a Channel Adjusted Transfer (CAT) head to encourage the mining of base-specific knowledge from feature channels in an isolated feature space, thereby facilitating the preservation of task-shared knowledge in the original feature space and enhancing zero-shot generalization performance on new tasks (Section 2.3). Furthermore, by simply fusing base-specific knowledge and task-shard knowledge in the two feature spaces at inference, we boost the performance on the base task remarkably (Section 3.2).

Flexibility and Effectiveness. Our DePT framework is orthogonal to existing prompt tuning methods, hence it can be flexibly used to overcome the BNT problem for them. We evaluate our DePT using a broad spectrum of baseline methods, including the *visual* prompt tuning method VPT [21], *textual* prompt tuning methods CoOp [59], CoCoOp [58] and KgCoOp [50], and *multi-model* prompt tuning methods MaPLe [25], PromptSRC [26]. Experimental results on 11 diverse datasets show that DePT consistently improves the performance of those methods, regardless of whether there is a *category shift distribution shift* or *domain shift* between base and new tasks, demonstrating the strong flexibility and effectiveness of DePT (Section 3.3). Notably, DePT enhances the six baselines without performance tradeoffs on base and new tasks – DePT achieves absolute gains of **1.31%∼3.17%** (resp. **0.71%∼2.23%**) on base (resp. new) tasks, averaged on the 11 datasets (Figure 1).

Contributions. Our main contributions are threefold. **1)** We provide an insightful analysis of the BNT problem in prompt tuning, revealing for the first time that the BNT stems from the channel bias issue. **2)** We propose the DePT framework to tackle the BNT problem from a feature decoupling perspective, and DePT is orthogonal to existing prompt tuning methods. **3)** We perform experiments on 11 diverse datasets and show that DePT consistently enhances the performance of a broad spectrum of baseline methods.[1]

2. Methodology

In this section, we initially offer an insightful examination of the BNT problem in prompt tuning, followed by a detailed exposition of our proposed DePT framework.

2.1. Preliminaries

Contrastive Language-Image Pre-training (CLIP) [46]. CLIP targets learning an alignment between image and text features produced by an image encoder and a text encoder, respectively. After seeing 400 million image-text association pairs and performing a contrastive learning paradigm in a common feature space, CLIP captures diverse open-set visual concepts that can readily be generalized to downstream applications. For example, we can achieve zero-shot classification by formulating the classification task as an image-text matching problem. Specifically, we first craft a prompt (e.g., "`a photo of a`") to obtain the text features of all inner-task classes by feeding the class-extended prompt (e.g., "`a photo of a [CLASS]`") to the text

[1] Our proposed DePT can also be used as a plugin to improve existing *adapter* tuning methods, as proven in **Sup.Mat. (E)**.

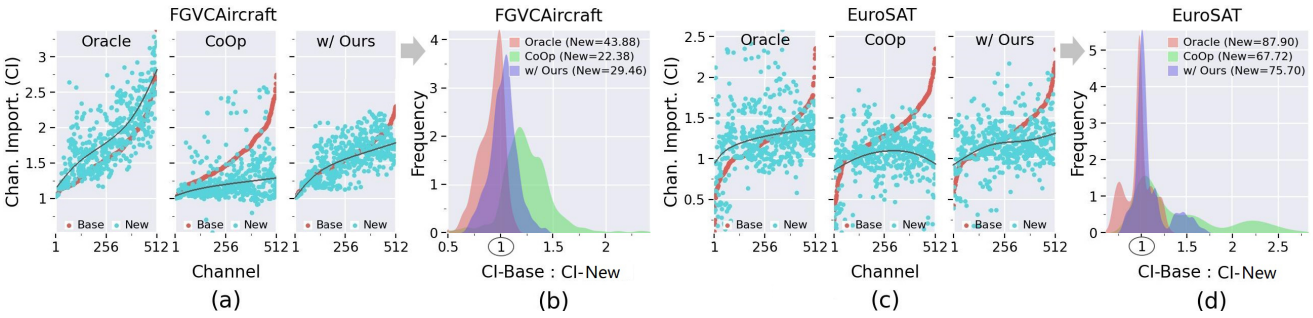

Figure 3. Channel Importance (**CI**) distributions of base and new tasks learned by the Oracle model and CoOp [59] w/ or w/o our DePT on the datasets FGVCAircraft [33] and EuroSAT [14]. In (**a**)(**c**), the indexes of channels in the x-axis are reordered based on the CI of the base task, a blue/red point indicates a channel. In (**b**)(**d**), the frequency distributions of CI-Base : CI-New are presented, where CI-Base and CI-New are the CI of the base and new tasks, respectively; "H" denotes the harmonic mean [58] of base-task and new-task accuracies.

encoder. Then, we use the image encoder to obtain the image feature of an input example, and predict the class of the example by comparing the cosine distances between the image feature and the text features of classes.

Prompt Tuning with the Image-Text Matching Head. Instead of using a hand-crafted prompt (e.g., "`a photo of a`"), prompt tuning aims to learn a *task-specific* prompt using a handful of training data from the base (or target) task. Let $[\boldsymbol{v}]_1[\boldsymbol{v}]_2...[\boldsymbol{v}]_l$ denote l trainable vectors, we forward the class-extended prompt $\boldsymbol{c}_i = [\boldsymbol{v}]_1[\boldsymbol{v}]_2...[\boldsymbol{v}]_l[\texttt{CLASS}]$ to the text encoder $g(\cdot)$ to obtain the text feature of the i-th class: $g(\boldsymbol{c}_i)$. Let $\boldsymbol{f}$ denote the image feature of an example $\boldsymbol{x}$ obtained by the image encoder, the task-specific prompt can be optimized using a parameter-free Image-Text Matching (ITM) head, which formulates the learning objective as:

$$\mathcal{L}_{\texttt{ITM}} = -\sum_i \boldsymbol{y}_i \log \mathcal{P}_{\texttt{ITM}}(\boldsymbol{c}_i|\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{y}$ is the one-hot label,

$$\mathcal{P}_{\texttt{ITM}}(\boldsymbol{c}_i|\boldsymbol{x}) = \frac{\exp(< g(\boldsymbol{c}_i), \boldsymbol{f} > /\tau)}{\sum_{i=1}^{M} \exp(< g(\boldsymbol{c}_i), \boldsymbol{f} > /\tau)}, \tag{2}$$

$< \cdot >$ denotes cosine similarity, M is the number of classes, and τ is the temperature learned by CLIP. During training, the gradients calculated in the ITM head can be back-propagated all the way through the text encoder $g(\cdot)$ to optimize the trainable vectors in the prompt.

2.2. A Closer Look at the BNT Problem

Due to the BNT problem, adapting a pretrained model to the base task $\mathcal{T}_{\text{base}}$ will decrease the generalization of the model on the new task $\mathcal{T}_{\text{new}}$, and vice versa. In this part, we provide an insightful view to analyze the BNT problem.

Deriving an Oracle Model on $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$. We start the investigation of the BNT problem by deriving an *oracle* model on $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$. Specifically, we adapt the pretrained model to both $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$ by jointly training the model on the data of the two tasks during prompt tuning. The derived oracle model thus can be seen as an approximation of a *BNT-free* model, since it avoids overfitting to either $\mathcal{T}_{\text{base}}$ or $\mathcal{T}_{\text{new}}$. Here, we use the word "oracle", because the model is derived by leveraging data from the new task, which is not accessible in prompt tuning.

Calculating Channel Importance for $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$. Denote $\boldsymbol{f}_j$ and $\boldsymbol{e}_* \in \{\boldsymbol{e}_i = g(\boldsymbol{c}_i)\}_{i=1}^{M}$ the d-dimensional image and text features of the example $\boldsymbol{x}_j$ in the learned feature space, respectively. We calculate the Channel Importance (**CI**) of the r-th ($r = 1, ..., d$) feature channel for each task of $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$ as follows:

$$\textbf{CI}^{(r)} = \frac{1}{N}\sum_{j=1}^{N} \frac{\text{ReLU}(\bar{\boldsymbol{e}}_*^{(r)} \bar{\boldsymbol{f}}_j^{(r)})}{1/M \sum_{i=1}^{M} \text{ReLU}(\bar{\boldsymbol{e}}_i^{(r)} \bar{\boldsymbol{f}}_j^{(r)})}, \tag{3}$$

where $\bar{\cdot} = \cdot / ||\cdot||_2$, N is the number of examples in the task. ReLU [1] is used to avoid the denominator being equal to 0. The derived Eq. (3) has an intuitive explanation: a feature channel is of greater importance if it can better distinguish the classes in the task, i.e., the image features are close to the ground-truth text features and far away from the text features of other classes at this channel.

Analysis. What are the differences between the derived oracle model and the model learned through the standard prompt tuning paradigm w.r.t. the calculated CI distributions of $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$? To answer this question, we take CoOp [59] as the baseline method, and plot the CI distributions of the testing data of $\mathcal{T}_{\text{base}}$ and $\mathcal{T}_{\text{new}}$ for CoOp and the oracle model on the datasets FGVCAircraft [33] and EuroSAT [14] in Figure 3 (see **Sup. Mat.(A)** for details). As observed in the figure, the CI distributions of base and new tasks obtained by the oracle model show greater consistency compared to that obtained by CoOp. Concretely, from the results of CoOp in **(a)(c)**, the achieved CI values of new tasks are significantly lower than that of base tasks at the vast majority of feature channels, which is further confirmed in **(b)(d)**, where the computed values of "CI-Base :

CI-New" are larger than (resp. close to) **1.0** in most cases for CoOp (resp. the oracle model). In **(b)(d)**, we present the classification accuracies of CoOp and the oracle model on new tasks, where the oracle model outperforms CoOp by large margins, suggesting that most feature channels produced by the oracle model contain *task-shared* knowledge that is valuable for the generalization of new tasks. In a nutshell, after prompt tuning, the vast majority of learned feature channels are occupied by *base-specific* knowledge, resulting in the collapse (or catastrophic forgetting) of task-shared knowledge important to new tasks – we refer to this as a *channel bias* issue in this work. Inspired by the above observations, we raise the following question:

> *Can we simultaneously preserve base-specific and task-shared knowledge in feature channels to overcome the BNT problem in prompt tuning?*

2.3. Decoupled Prompt Tuning

In this work, we answer the above question by proposing Decoupled Prompt Tuning (DePT), a first framework overcoming the BNT problem in prompt tuning from a feature decoupling perspective. An overview of the DePT framework is presented in Figure 2.

A Plug-and-play Channel Adjusted Transfer Head. Due to the channel bias issue, striving for base-specific knowledge during prompt tuning will inevitably trigger the catastrophic forgetting of task-shared knowledge in the learned feature channels. To address this, DePT employs a Channel Adjusted Transfer (CAT) head to decouple base-specific knowledge from feature channels into an isolated feature space, so as to maximally preserve task-shared knowledge in the original feature space. Denote $\mathcal{S}_{\texttt{img}} = \{\boldsymbol{f}_j\}_{j=1}^{J}$ and $\mathcal{S}_{\texttt{text}} = \{\boldsymbol{e}_j\}_{j=1}^{J}$[2] the sets of image and text features for a batch of training examples respectively, and $\boldsymbol{f}_j, \boldsymbol{e}_j \in \mathbb{R}^d$. First, the CAT head leverages a channel-wise Transformation (cwT) layer to transform both $\mathcal{S}_{\texttt{img}}$ and $\mathcal{S}_{\texttt{text}}$ to a new feature space. Formally, $\mathcal{S}'_{\texttt{img}} = \{\boldsymbol{f}'_j\}_{j=1}^{J}$, and

$$\boldsymbol{f}'_j = \boldsymbol{\gamma} \odot \boldsymbol{f}_j + \boldsymbol{\beta}, \quad j = 1, ..., J, \tag{4}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^d$ are trainable scaling and shift vectors. Denote $\mathcal{S}'_{\texttt{text}} = \{\boldsymbol{e}'_j\}_{j=1}^{J}$ similar to $\mathcal{S}'_{\texttt{img}} = \{\boldsymbol{f}'_j\}_{j=1}^{J}$. Next, a linear classifier takes $\mathcal{S}_{\cup}$ and $\mathcal{Y}_{\cup}$ as input to encourage the mining of base-specific knowledge in the isolated feature space, where $\mathcal{S}_{\cup} = \mathcal{S}'_{\texttt{img}} \cup \mathcal{S}'_{\texttt{text}} = \{\boldsymbol{s}_j\}_{j=1}^{2J}$ and $\mathcal{Y}_{\cup} = \{\boldsymbol{y}_j\}_{j=1}^{2J}$, $\boldsymbol{y}_j \in \mathbb{R}^M$ is the one-hot label for $\boldsymbol{s}_j$, and M is the number of classes of the task. For each pair of $(\boldsymbol{s}, \boldsymbol{y})$, the CAT head minimizes the following cross-entropy loss:

$$\mathcal{L}_{\texttt{CAT}} = -\sum_i \boldsymbol{y}_i \log \mathcal{P}_{\texttt{CAT}}(\boldsymbol{c}_i|\boldsymbol{x}), \tag{5}$$

[2]Here, $\boldsymbol{e}_i$ may equal to $\boldsymbol{e}_j$ for $i \neq j$, we ignore it for simplification.

where

$$\mathcal{P}_{\texttt{CAT}}(\boldsymbol{c}_i|\boldsymbol{x}) = \sigma(\boldsymbol{W} \cdot \boldsymbol{s})[i], \tag{6}$$

$\boldsymbol{W} \in \mathbb{R}^{M \times d}$ denotes the projection matrix for classification, σ denotes the softmax operation. During training, the gradients calculated by $\mathcal{L}_{\texttt{CAT}}$ are back-propagated to update the weights in the CAT head (i.e., $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{W}$) as well as the trainable prompt (i.e., $[\boldsymbol{v}]_1[\boldsymbol{v}]_2...[\boldsymbol{v}]_l$). Ablation studies in Section 3.2 show that employing two independent cwT layers (one for each modality) is more effective than using a shared cwT layer in the CAT head.

Prompt Tuning with Dual Heads. Rather than solely using the designed CAT head to facilitate the preservation of task-shared knowledge during prompt tuning, our DePT also retains the standard ITM head to learn an alignment of image-text pairs in the original feature space, which is of great importance for establishing better zero-shot generalization on new tasks (as proven in Section 3.2). Thus, the overall learning objective of DePT is expressed as:

$$\mathcal{L} = \lambda \mathcal{L}_{\texttt{CAT}} + (1 - \lambda)\mathcal{L}_{\texttt{ITM}}, \tag{7}$$

where λ is a balance weight controlling the relative importance of the two losses.

Test-time Knowledge Fusion for the Base Task. At inference, the standard ITM head is used to achieve zero-shot generalization/prediction on new tasks in the original feature space. For the base task, our CAT head directly takes the image feature of a testing example as input to predict the in-distribution class label with the linear classifier. Notably, we can further boost the performance on the base task by simply fusing base-specific knowledge in the CAT head as well as task-shard knowledge in the ITM head at inference. By connecting Eq. (2) and Eq. (6), the prediction probability of the in-distribution testing example $\boldsymbol{x}$ belonging to the i-th class can be computed as:

$$\mathrm{p}(\boldsymbol{c}_i|\boldsymbol{x}) = \lambda \mathcal{P}_{\texttt{CAT}}(\boldsymbol{c}_i|\boldsymbol{x}) + (1 - \lambda)\mathcal{P}_{\texttt{ITM}}(\boldsymbol{c}_i|\boldsymbol{x}), \tag{8}$$

where the balance weight λ in Eq. (7) is directly used to control the contributions of the two heads for simplification. Pytorch-like pseudocode for the implementation of DePT is presented in **Sup. Mat.(B)**.

3. Experiments

In this section, we first present ablation studies to analyze the impacts of different factors on DePT. Next, we validate the flexibility and effectiveness of DePT by applying it to several baseline schemes. We start with an introduction of the experimental setup below.

3.1. Experimental Setup

Baselines. We apply our DePT to a broad spectrum of baseline approaches, including the *visual* prompt tuning method

Setting	ITM Head	CAT Head		Test-time fusion for the *Base* task	Average accuracy over 11 datasets (%)		
		cwT+LC	cwT+ITM		Base	New	H
① ITM only (**Baseline**)	✓	✗	✗	✗	81.50	69.77	75.18
② ITM+CAT (CAT=cwT+LC)	✓	✓	✗	✗	82.14 (+0.64)	**71.82** (+2.05)	76.63 (+1.45)
v1. Use a shared cwT in CAT	✓	✓	✗	✗	82.24 (+0.74)	70.85 (+1.08)	76.12 (+0.94)
v2. Use an ITM classifier in CAT	✓	✗	✓	✗	82.16 (+0.66)	71.31 (+1.54)	76.35 (+1.17)
v3. Only use image features in CAT	✓	✓	✗	✗	81.11 (- 0.39)	70.93 (+1.16)	75.68 (+0.50)
③ ITM+CAT+Fusion (**Our DePT**)	✓	✓	✗	✓	**83.66** (+2.16)	**71.82** (+2.05)	**77.29** (+2.11)

Table 1. Ablation study for the designed components of DePT. The baseline method is CoOp [59], and the average accuracy on 11 datasets are reported. The metric "H" indicates the harmonic mean [58] of base-task and new-task accuracies. "LC": Linear Classifier.

VPT [21], *textual* prompt tuning methods CoOp [59], CoCoOp [58] and KgCoOp [50], and *multi-model* prompt tuning methods MaPLe [25], PromptSRC [26].

Datasets. We conduct experiments on several datasets from diverse sources. Concretely, for the settings of *base-to-new generalization* and *cross-dataset generalization*, we use **11** datasets: ImgNet [6], Caltech [8], OxfordPets [37], StanfordCars [28], Flowers [36], Food101 [2], FGVCAircraft [33], EuroSAT [14], UCF101 [45], DTD [5], and SUN397 [49]; for the *domain generalization* setting, we use ImgNet as the source domain (i.e. the base task), and its four variants ImgNet-V2 [41], ImgNet-Sketch [48], ImgNet-A [11] and ImgNet-R [15] as target domains (i.e. new tasks).

Implementation Details. Our implementations are based on the open-source repository of MaPLe [25][3]. For each baseline method, we use the same experimental setup (e.g., feature backbone, prompt length and learning rate) as used in the original implementation. For DePT, the learning rate for updating the parameters in the devised CAT head is set to $6.5 \times \delta$, where δ is the adopted learning rate of each baseline for prompt tuning. We adjust the value of λ and the training epoch for our DePT in ablation studies. The above hyperparameters are fixed across all datasets. Unless stated otherwise, the base task is constructed as a many-way 16-shot task. We report base-task and new-task accuracies as well as their harmonic-mean (H) [58] averaged over 3 runs to compare the performance of different methods. Code: https://github.com/Koorye/DePT.

3.2. Ablation Studies

Here, we first conduct an ablative analysis of the designed components of DePT in Table 1. Then, we investigate the impact of the balance weight λ on DePT in Figure 4 (**Left**). Next, we scrutinize the performance of DePT on different training epochs in Figure 4 (**Right**). Finally, we validate the robustness of DePT under different shots in Figure 5. We perform experiments using the baseline method CoOp [59] in the base-to-new generalization setting, results averaged on the **11** datasets are reported.

Effectiveness of the Devised Components in DePT. Our DePT contains two key components, including a plug-and-play CAT head for capturing base-specific knowledge in an isolated feature space, as well as a test-time knowledge fusion strategy for exploring both base-specific and task-shard knowledge to improve the performance on the base task. We conduct component-wise analysis on the two components by progressively adding one of them to the baseline method CoOp [59] in Table 1, where the results are averaged over 11 datasets. From ① and ② in the table, we observe that integrating our CAT head with the standard ITM head for prompt tuning improves both base-task and new-task accuracies of the baseline method, achieving a clear enhancement of the harmonic-mean (by **1.45**%). Notably, ② outperforms ① by up to **2.05**% in terms of new-task accuracy, which demonstrates the effectiveness of our CAT head in facilitating the preservation of task-shared knowledge during prompt tuning. Besides, we also compare the CAT head with its three variants. Concretely, we replace the two independent cwT layers (one for each modality) with a shared cwT layer in *v1*, we replace the linear classifier with an ITM classifier in *v2*, and we only feed image features to the liner classifier in *v3* (more details are in **Sup. Mat.(C)**). As shown, all the three variants underperform our designed CAT head. What is noteworthy is that directly appending a standard ITM classifier in the cwT-transformed feature space also considerably improves the performance of the baseline on new tasks (see *v2*), showing the effectiveness of the CAT head for decoupling base-specific knowledge and task-shared knowledge during prompt tuning. Besides, we see that using only image features in the CAT head damages the performance on the base task (see *v3*). This is possibly due to that relying on a limited number of examples for model optimization, the parameters in the CAT head may overfit to the training data of the base task when the gradients of $\mathcal{L}_{\texttt{CAT}}$ can not be back-propagated to the text encoder to optimize the parameters of the prompt. What's more, by simply fusing base-specific knowledge and task-shard knowledge in the two heads at inference, the performance on the base task can be enhanced considerably, achieving an absolute gain of **2.16**% in accuracy, as shown in ③.

Impact of the Balance Weight λ on DePT. In the proposed DePT, we employ the balance weight λ to control the relative importance of the standard ITM head and our devised CAT head in Eq. (7)/(8). It is necessary to investigate the

[3]https://github.com/muzairkhattak/multimodal-prompt-learning

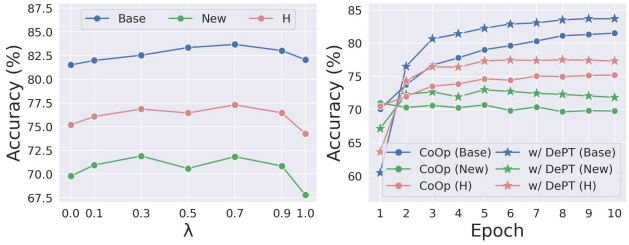

Figure 4. **Left**: Impact of the balance weight λ in Eq. (7)/(8) on DePT. **Right**: Performance of DePT at different training epochs.

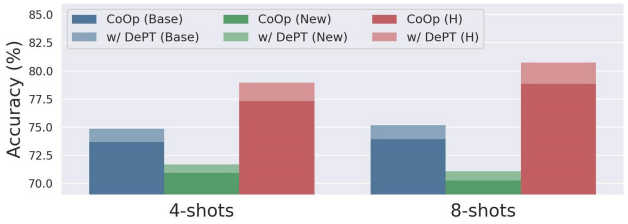

Figure 5. Robustness of DePT under different shots.

impact of λ on the performance of DePT. To this end, we set λ to the values of $\{0.0, 0.1, 0.2, ..., 1.0\}$, and report the average testing results on the 11 datasets in Figure 4 (**Left**). Overall, the performance of DePT gradually increases as the λ value grows from 0.0 to 0.7, after which the performance of DePT gradually decreases and reaches the lowest value when λ=1.0. In particular, when λ=0.7 DePT establishes the best performance on both base and new tasks. What is noteworthy is that when λ=1.0, i.e., only our CAT head is used for training, the performance of DePT on new tasks sharply decreases, which suggests that retaining the ITM head to learn an alignment of positive image-text features in the original feature space is of great importance for achieving better zero-shot prediction performance on new tasks.

Performance of DePT at Different Training Epochs. In Figure 4 (**Right**), we report the obtained results of the baseline method w/ or w/o our DePT at different training epochs. As can be observed, our DePT consistently improves the baseline method after epoch 2, in terms of base-task, new-task, and harmonic mean (H) accuracies. One possible reason for the failure case at epoch 1 is that the weights in the CAL head (i.e., $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{W}$) are initialized randomly, thus it is difficult for the CAL head to fully capture base-specific knowledge with only one training epoch. We also see the baseline fails to address the BNT problem during prompt tuning – overall, the accuracy of the baseline on new tasks decreases as its performance on base tasks increases from epoch 1 to epoch 10. It is obvious that DePT mitigates the BNT problem effectively. The performance of the baseline method and DePT is saturated at epoch 10.

Robustness of DePT under Different Shots. All the aforementioned results are obtained on many-way 16-shot base tasks – for every base task, 16 training examples are sampled from each class for prompt tuning. It is interesting to further scrutinize the robustness of our DePT under different shots. To achieve this, we set the shots to $\{4, 8, 16\}$, and report the average testing results of the baseline method w/ or w/o our DePT on the 11 datasets in Figure 5. As can be observed, our DePT consistently improves the baseline method across all 4-shot, 8-shot, and 16-shot settings, in terms of base-task, new-task, and harmonic mean (H) accuracies, showing the robustness of DePT under few shots. In the following section, we follow [25, 50, 58, 59] to evaluate methods in the 16-shot setting.

Method	Learnable para.	Train.time	Infer.time	Memory	H (*avg*)
CoOp	8K	105min	2.78ms	11264MB	75.18
+DePT	+(2+N/2)K	+0min	+0ms	+2MB	**77.29** (+2.11)

Table 2. Computational cost of DePT. "N": the num of classes in the base task, "ms": millisecond per image. Experiments are performed on a V100 GPU.

Computational Cost. The computational cost of our DePT is shown in Table 2. As observed, the additional computational cost is low (even *negligible*) compared to the performance improvement established by DePT.

3.3. Experimental Results

In this part, we demonstrate the flexibility and effectiveness of DePT in the base-to-new generalization (in Table 3) and cross-dataset generalization (in Table 4) settings.

Base-to-New Generalization. The base-to-new generalization setting evaluates whether the models learned on base tasks can generalize to new tasks with unseen classes, i.e., there is a *category shift* between base and new tasks. Following the baseline methods, for each dataset, we first construct a base task and a new task by equally dividing the dataset into two sets of classes, then we perform prompt tuning on the base task and test the learned model on both the base and new tasks. Table 3 presents the base-to-new generalization performance of the six baselines w/ or w/o our DePT framework over 11 datasets. From the average results in the table, we observe a tradeoff between base-task and new-task accuracies for most of the baseline methods, e.g., CoCoOp outperforms CoOp on new tasks but underperforms CoOp on base tasks. Notably, DePT consistently improves the performance of all baselines without performance tradeoffs on base and new tasks. Specifically, DePT improves each baseline in terms of all base-task, new-task and harmonic-mean accuracies. From the results on 11 datasets, we also observe some failure cases, e.g., on the OxfordPets dataset, DePT fails to bring clear performance gains on most baseline methods. Possible reasons are as following. **1)** The optimal hyperparameters of DePT for different datasets and baselines are quite different, while we fix them across all datasets and baselines. **2)** When the *category shift* between downstream tasks and the upstream data for model (i.e. CLIP) pretraining is minimal, the advantages of our DePT for task adaptation become less significant.

Method	Avg over 11 datasets			ImageNet			Caltech101			OxfordPets		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp [59]	81.50	69.77	75.18	76.57	69.97	73.12	98.17	**94.83**	**96.47**	**95.57**	97.53	**96.54**
+DePT	**83.66**	**71.82**	**77.29**	**77.13**	**70.10**	**73.45**	**98.33**	94.33	96.29	94.70	**97.63**	96.14
CoCoOp [58]	81.18	72.18	76.42	75.90	**70.73**	73.23	97.70	93.20	95.40	**94.93**	**97.90**	**96.39**
+DePT	**83.80**	**72.89**	**77.97**	**76.87**	70.47	**73.53**	**98.37**	**93.87**	**96.06**	94.03	97.20	95.59
KgCoOp [50]	80.93	73.88	77.25	76.17	**70.53**	73.24	97.87	94.03	95.91	**95.47**	**97.80**	**96.62**
+DePT	**83.62**	**75.04**	**79.10**	**77.03**	70.13	**73.42**	**98.30**	**94.60**	**96.41**	94.33	97.23	95.76
MaPLe [25]	83.54	73.76	78.35	77.23	69.63	73.24	98.30	93.70	95.94	**95.17**	97.77	**96.45**
+DePT	**84.85**	**74.82**	**79.52**	**77.87**	**70.23**	**73.85**	**98.53**	**95.03**	**96.75**	95.03	**97.83**	96.41
VPT [21]	82.11	71.73	76.57	75.90	68.10	71.79	98.03	94.30	96.13	95.13	**96.47**	95.80
+DePT	**85.28**	**73.96**	**79.22**	**78.40**	**68.90**	**73.34**	**98.67**	**94.33**	**96.45**	**95.50**	96.33	**95.91**
PromptSRC [26]	84.21	75.75	79.76	77.63	70.23	73.75	98.10	93.87	95.94	95.27	97.23	96.24
+DePT	**85.19**	**76.17**	**80.43**	**78.20**	**70.27**	**74.02**	**98.57**	**94.10**	**96.28**	**95.43**	**97.33**	**96.37**

Method	StanfodCars			Flowers102			Food101			FGVCAircraft		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp [59]	74.30	72.10	73.18	97.07	**74.33**	**84.19**	90.43	90.97	90.70	31.70	17.30	22.38
+DePT	**79.67**	**72.40**	**75.86**	**98.20**	72.00	83.08	**90.43**	**91.33**	**90.88**	**42.53**	**22.53**	**29.46**
CoCoOp [58]	70.77	72.50	71.62	95.03	69.07	80.00	**90.57**	91.20	**90.88**	35.63	**32.70**	34.10
+DePT	**79.87**	**73.33**	**76.46**	**98.33**	**72.57**	**83.51**	90.30	**91.30**	90.80	**43.07**	31.30	**36.25**
KgCoOp [50]	71.13	74.67	72.86	95.90	74.83	84.07	90.47	**91.60**	91.03	35.10	**35.20**	35.15
+DePT	**79.13**	**75.47**	**77.26**	**98.00**	**76.37**	**85.84**	**90.50**	**91.60**	**91.05**	**43.20**	34.83	**38.57**
MaPLe [25]	76.30	**72.53**	74.37	97.23	72.07	82.78	90.30	**91.53**	90.91	40.57	**36.47**	**38.31**
+DePT	**80.93**	71.73	**76.06**	**98.03**	**73.17**	**83.79**	**90.33**	**91.53**	**90.93**	**44.53**	32.80	37.78
VPT [21]	71.63	**72.20**	71.92	95.93	70.37	81.18	89.80	90.37	90.08	35.90	30.37	32.90
+DePT	**82.13**	72.17	**76.83**	**98.17**	**73.20**	**83.86**	**90.27**	**91.03**	**90.65**	**45.30**	**31.87**	**37.41**
PromptSRC [26]	78.37	74.97	76.63	97.90	76.97	86.18	90.63	91.53	91.08	42.53	**36.87**	39.50
+DePT	**80.80**	**75.00**	**77.79**	**98.40**	**77.10**	**86.46**	**90.87**	**91.57**	**91.22**	**45.70**	36.73	**40.73**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp [59]	81.13	**76.07**	78.52	79.33	49.70	61.11	**89.35**	57.30	69.82	83.87	69.80	76.19
+DePT	**82.37**	75.07	**78.55**	**83.20**	**56.13**	**67.04**	88.27	**66.27**	**75.70**	**85.43**	**72.17**	**78.24**
CoCoOp [58]	79.50	76.27	77.85	77.37	52.97	62.88	87.97	63.63	73.85	82.33	72.40	77.05
+DePT	**82.20**	**76.73**	**79.37**	**82.77**	**55.40**	**66.37**	**90.27**	**66.87**	**76.82**	**85.70**	**72.80**	**78.73**
KgCoOp [50]	80.40	77.30	78.82	78.27	57.93	66.58	85.77	63.40	72.91	83.73	75.40	79.35
+DePT	**82.33**	**77.80**	**80.00**	**82.20**	**59.13**	**68.78**	**89.03**	**71.07**	**79.04**	**85.80**	**77.23**	**81.29**
MaPLe [25]	81.93	**76.50**	79.12	81.93	58.20	68.06	**94.67**	66.73	78.28	85.30	76.23	80.51
+DePT	**82.90**	76.40	**79.52**	**83.87**	**59.93**	**69.91**	94.43	**76.23**	**84.36**	**86.87**	**78.10**	**82.25**
VPT [21]	79.50	76.17	77.80	80.90	52.73	63.85	**95.83**	65.03	77.48	84.63	72.90	78.33
+DePT	**83.03**	**77.77**	**80.31**	**85.07**	**56.60**	**67.97**	93.77	**76.30**	**84.14**	**87.73**	**75.10**	**80.93**
PromptSRC [26]	82.63	**78.97**	80.76	83.43	**62.53**	**71.49**	92.80	72.07	81.13	87.03	**78.07**	82.31
+DePT	**83.27**	**78.97**	**81.06**	**84.80**	61.20	71.09	**93.23**	**77.90**	**84.88**	**87.73**	77.70	**82.46**

Table 3. Base-to-new generalization performance of six prompt tuning approaches w/ or w/o our DePT on 11 datasets.

Method	Source (ImgNet)	Target										
		Avg	Caltech101	OxfordPets	StanfCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOp [59]	71.80	64.40	**93.97**	89.60	64.60	69.13	85.47	20.70	65.70	43.07	44.50	67.23
+DePT	**72.63**	**65.02**	93.30	**90.00**	**65.53**	**70.50**	**85.97**	**21.90**	**66.07**	**43.17**	**44.97**	**68.80**
CoCoOp [58]	71.17	65.73	**94.30**	**90.80**	65.53	71.80	86.13	22.83	**67.73**	**45.57**	43.47	69.10
+DePT	**72.77**	**65.88**	94.10	90.63	**66.23**	**72.17**	**86.27**	**22.90**	67.30	45.50	**44.17**	**69.53**
KgCoOp [50]	71.17	65.06	94.17	89.70	64.77	70.30	**86.47**	22.43	**66.83**	44.43	**43.53**	68.00
+DePT	**72.77**	**65.55**	**94.23**	**90.03**	**65.57**	**70.57**	86.37	**23.27**	66.67	**45.97**	**43.53**	**69.30**
MaPLe [25]	72.47	64.17	**92.97**	**90.20**	63.97	70.03	84.83	23.23	66.00	43.23	40.03	67.23
+DePT	**73.27**	**64.56**	92.53	90.10	**64.60**	**70.10**	**85.57**	**23.63**	**66.40**	**45.03**	**40.13**	**67.53**
VPT [21]	70.80	62.61	**91.67**	**90.03**	62.47	66.03	81.70	**24.07**	65.27	44.27	35.77	64.83
+DePT	**71.97**	**63.01**	91.30	**90.03**	**62.63**	**66.77**	**83.03**	23.73	**65.57**	**44.57**	**37.03**	**65.40**
PromptSRC [26]	71.33	65.71	93.77	**90.40**	65.77	70.80	**86.30**	23.67	66.93	46.07	44.23	**69.20**
+DePT	**71.60**	**66.02**	**93.80**	90.13	**66.00**	**70.93**	86.27	**24.30**	**67.23**	**46.60**	**45.83**	69.10

Table 4. Cross-dataset generalization performance of six prompt tuning approaches w/ or w/o our DePT on 11 datasets.

Cross-Dataset Generalization. The cross-dataset generalization setting evaluates whether the model learned on the source dataset can generalize to unseen target datasets, i.e., there is a *distribution shift* between base and new tasks. In this experiment, we follow the baselines to regard ImgNet as the source dataset and the other 10 datasets as target datasets. Table 4 presents the performance of the six baselines w/ or w/o DePT on the 11 datasets. As can be seen,

our DePT consistently improves the accuracy on the source dataset for all baselines, without compromising the performance on 10 target datasets in most cases. Notably, on average our DePT consistently enhances the performance of all baselines on both the source and target datasets, suggesting DePT is robust to the distribution shift. Moreover, we see that the previous state-of-the-art method MaPLe establishes the best base-to-new generalization performance among the six baseline methods in Table 3, but in Table 4 it achieves inferior cross-dataset generalization performance to CoCoOp and KgCoOp. This is probably due to that without decoupling base-specific knowledge and task-shared knowledge during prompt tuning, the learned prompts in both the image encoder and the text encoder for MaPLe are not generalizable enough under distribution shift.

Domain Generalization. We also evaluate DePT in the domain generalization setting using the six baseline methods in **Sup.Mat (D)**, where DePT still maintains the advantages as in previous settings.

Effectiveness of DePT on Adapter Tuning Methods. We also apply DePT to the adapter tuning method CLIP-adapter [12] in **Sup.Mat (E)**, where DePT still maintains the advantages as on the six prompt tuning methods.

4. Related Work

Vision-Language Pre-training. Deep learning algorithms have been reported to exhibit or even surpass human-level performance on computer vision and natural language processing tasks [22, 23, 42–44, 52, 54]. By modeling a connection of image-text pairs, large vision-language pre-trained models (VLPMs) have shown strong zero-shot generalization on various downstream tasks. Generally, VLPMs leverage three types of pre-text tasks for modeling the semantic correspondence between the vision and language modalities, including **1)** image-text matching [20, 27], **2)** contrastive learning [19, 29, 34, 35], and **3)** masked vision/language prediction [27, 30]. Beyond image domain, 3D-VLP [24] initially achieves VLPM on sparse and irregular point clouds using a novel mutual mask modeling. In this work, we mainly focus on VLPMs establishing image-text alignment with contrastive learning, motivated by their excellent generalization ability to downstream tasks. For example, after seeing 400 million text-image pairs, CLIP [39] learns an alignment between visual and textual features output by an image encoder and a text encoder respectively. Beyond recognition [7, 25, 56, 58], CLIP also demonstrates great potential for other downstream applications, such as image manipulation [38, 46], video-text retrieval [4, 32, 61], and dense prediction [40, 57].

Task Adaptation on VLPMs. The remarkable success of VLPMs have brought new light but also pose a new question: how to efficiently adapt the knowledge from VLPMs to different downstream tasks? The most direct solution is *full-finetuning*, which fixes the architecture of VLPMs and tunes all the parameters on the target task. While the results are impressive, this line of work becomes prohibitively expensive with the ever-increasing size of parameters of VLPMs. To remedy this, *partial-finetuning* has been proposed to update only a small number of extra parameters (a.k.a. *adapters*) while keeping most pre-trained parameters frozen. Representative schemes are Adapters [16], CLIP-adapter [12], LoRA [17], BitFit [51] and Diff-pruning [13].

Prompt Tuning. Inspired from the field of NLP, a rich line of recent works adapt VLPMs to downstream tasks by learning task-specific prompts in an end-to-end manner [3, 47, 60]. Since only a handful of labeled examples are available during training, prompt tuning can be regarded as few-shot learning task [9, 10, 31, 53, 55]. In particular, CoOp [59] performs task adaptation by optimizing a set of prompt vectors at the language branch of CLIP. While simple and effective, CoOp tends to achieve poor generalization on new tasks after overfitting to the base (or target) task. To overcome this issue, CoCoOp [58] learns a lightweight meta-net to generate an input-conditional token for each input image. By reducing the discrepancy between the hand-crafted prompt and the trainable prompt tokens, KgCoOp [58] significantly improves the generalization of the adapted models on new tasks. ProGrad [60] mitigates the overfitting issue by regularizing each tuning step that is not to conflict with the general knowledge of the hand-crafted prompt. Unlike the aforementioned methods that mainly focus on developing efficient textual prompts, a rich line of works also explores visual prompts for task adaptation [18, 21]. By adding trainable prompts at both the language and text branches of CLIP, multi-model prompt tuning methods MaPLe[25] and PromptSRC[26] yield remarkable performance on both the base task and new tasks.

5. Conclusions

In this work, we propose the DePT framework to tackle the Base-New Tradeoff (BNT) problem in prompt tuning. First, we offer an insightful view to analyze the BNT problem, and reveal that the BNT stems from the channel bias issue. Second, we present the DePT framework to tackle the BNT problem, and DePT is orthogonal to existing prompt tuning methods. Third, we apply DePT to a broad spectrum of baselines, and the results on 11 datasets demonstrate the strong flexibility and effectiveness of DePT. We hope this work can bring some inspiration to related fields.

Acknowledgements. This work is supported by grants from the National Natural Science Foundation of China (Grant No. 62122018, No. 62020106008, No. U22A2097, No. U23A20315), Kuaishou Tech. and SongShan Laboratory YYJC012022019.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 3

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 5

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *ICLR*, 2022. 8

[4] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 8

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[7] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. *CVPR*, 2024. 8

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178. IEEE, 2004. 5

[9] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *ACM MM*, pages 5326–5334, 2021. 8

[10] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *CVPR*, pages 24575–24584, 2023. 8

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

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 8

[13] Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *ACL*, 2020. 8

[14] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 3, 5

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 5

[16] 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*, pages 2790–2799, 2019. 8

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[18] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, pages 10878–10887, 2023. 8

[19] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021. 8

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 8

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 2, 5, 7, 8

[22] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *CVPR*, pages 2456–2465. IEEE, 2022. 8

[23] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Sdn: Semantic decoupling network for temporal language grounding. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. 8

[24] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3D-language pre-training. In *CVPR*, pages 10984–10994, 2023. 8

[25] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2, 5, 6, 7, 8

[26] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 2, 5, 7, 8

[27] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 8

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 5

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. 8

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 8

[31] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *ICML*, pages 23103–23123. PMLR, 2023. 8

[32] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, pages 638–647, 2022. 8

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3, 5

[34] Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Disentangled multiplex graph representation learning. In *ICML*, pages 24983–25005, 2023. 8

[35] Yujie Mo, Feiping Nie, Zheng Zhang, Ping Hu, Heng Tao Shen, Xinchao Wang, and Xiaofeng Zhu. Self-supervised heterogeneous graph learning: a homogeneity and heterogeneity perspective. In *ICLR*, 2024. 8

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008. 5

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 5

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, pages 2085–2094, 2021. 8

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 8

[40] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 8

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 5

[42] Shuai Shao, Yan Wang, Bin Liu, Weifeng Liu, Yanjiang Wang, and Baodi Liu. Fads: Fourier-augmentation based data-shunting for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 8

[43] Shuai Shao, Lei Xing, Yanjiang Wang, Baodi Liu, Weifeng Liu, and Yicong Zhou. Attention-based multi-view feature collaboration for decoupled few-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[44] Shao Shuai, Bai Yu, Wang Yan, Liu Baodi, and Liu Bin. Collaborative consortium of foundation models for open-world few-shot learning. In *AAAI*, 2024. 8

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[46] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, pages 3835–3844, 2022. 2, 8

[47] Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alexander G Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. *ICLR*, 2022. 8

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 5

[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*, pages 3485–3492. IEEE, 2010. 5

[50] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 1, 2, 5, 6, 7

[51] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 8

[52] Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. Progressive meta-learning with curriculum. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5916–5930, 2022. 8

[53] Ji Zhang, Jingkuan Song, Lianli Gao, and Hengtao Shen. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In *ACM MM*, pages 2586–2594, 2022. 8

[54] Ji Zhang, Lianli Gao, Bingguang Hao, Hao Huang, Jingkuan Song, and Hengtao Shen. From global to local: Multi-scale out-of-distribution detection. *IEEE Transactions on Image Processing*, 2023. 8

[55] Ji Zhang, Lianli Gao, Xu Luo, Hengtao Shen, and Jingkuan Song. Deta: Denoised task adaptation for few-shot learning. In *ICCV*, pages 11541–11551, 2023. 8

[56] Renrui Zhang, Wei Zhang, 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*, pages 493–510. Springer, 2022. 8

[57] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. 8

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2, 3, 5, 6, 7, 8

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 5, 6, 7, 8

[60] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *ICCV*, 2023. 1, 8

[61] Jinkuan Zhu, Pengpeng Zeng, Lianli Gao, Gongfu Li, Dongliang Liao, and Jingkuan Song. Complementarity-aware space learning for video-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 8