Prompt Learning via Meta-Regularization

Jinyoung Park, Juyeon Ko, Hyunwoo J. Kim*
Department of Computer Science and Engineering, Korea University
{lpmn678, juyon98, hyunwoojkim}@korea.ac.kr

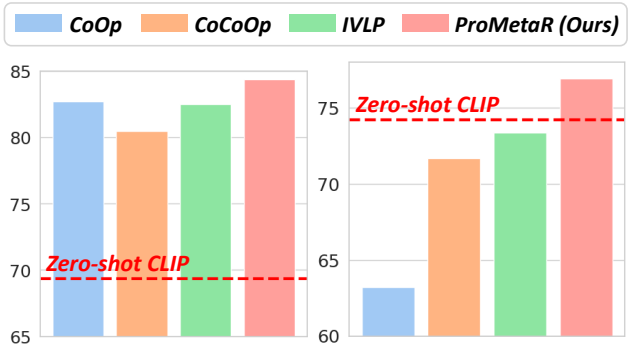
Abstract

*Pre-trained vision-language models have shown impressive success on various computer vision tasks with their zero-shot generalizability. Recently, prompt learning approaches have been explored to efficiently and effectively adapt the vision-language models to a variety of downstream tasks. However, most existing prompt learning methods suffer from task overfitting since the general knowledge of the pre-trained vision language models is forgotten while the prompts are finetuned on a small data set from a specific target task. To address this issue, we propose a **Pro**mpt **Meta**-**R**egularization (ProMetaR) to improve the generalizability of prompt learning for vision-language models. Specifically, ProMetaR meta-learns both the regularizer and the soft prompts to harness the task-specific knowledge from the downstream tasks and task-agnostic general knowledge from the vision-language models. Further, ProMetaR augments the task to generate multiple virtual tasks to alleviate the meta-overfitting. In addition, we provide the analysis to comprehend how ProMetaR improves the generalizability of prompt tuning in the perspective of the gradient alignment. Our extensive experiments demonstrate that our ProMetaR improves the generalizability of conventional prompt learning methods under base-to-base/base-to-new and domain generalization settings. The code of ProMetaR is available at https://github.com/mlvlab/ProMetaR.*

1. Introduction

Foundational vision-language models (VLMs) have established their precedence in various computer vision applications such as object detection [10, 12, 15, 76], image classification [52, 57, 72], segmentation [42], and captioning [37, 45, 75]. Represented by CLIP [52] and ALIGN [24], these models are pre-trained on millions of image-text pairs with contrastive loss, creating a shared, well-aligned joint embedding space for vision and language. They have demonstrated their generalization abilities in zero-shot image recognition and object detection.

*is the corresponding author.



(a) Base class performance. (b) New class performance.

Figure 1. Performance comparison of ProMetaR with prompt learning methods (Zero-shot CLIP, CoOp, CoCoOp, IVLP (base method), and ProMetaR (Ours)) under the base-to-base/base-to-new setting. We measure average accuracy on the base classes (a) and new classes (b) over 11 datasets. The red dotted line indicates the performance of the zero-shot CLIP.

Despite the effectiveness of VLMs on zero-shot image recognition, they suffer from time-consuming manual text prompting for each task, which is inefficient and requires human efforts and prior knowledge. Prompt tuning methods such as Context Optimization (CoOp) [78] have arisen as a new paradigm that uses a small number of learnable vectors (soft prompts) instead of manual prompting. They efficiently and effectively adapt models to downstream tasks by optimizing only a small number of learnable vectors (soft prompts) while keeping VLMs frozen. In recent, some works [26, 32] further enhance the performance by applying prompt tuning to both image and text modalities. Prompt tuning methods enhance *traditional* generalization capabilities showing good performance on trained tasks with only a few samples. However, as the soft prompts tend to prioritize task-specific knowledge, they easily overfit the target task and show poor *task* generalization abilities. In other words, they have difficulty in generalizing on new tasks, resulting in worse performance than CLIP in data-deficient settings. From Figure 1, standard prompt learning methods (CoOp, CoCoOp, and IVLP) show worse performance

than zero-shot CLIP on the unseen (new) classes during the training, while they perform well on the seen (base) classes.

One remedy to alleviate the *task* overfitting is learning the learnable prompts with the regularizer. However, the regularizers are not always beneficial for all the tasks, and it is nontrivial to manually balance the strength of the downstream loss (*i.e.*, contrastive loss) and regularizer for each task. So, we propose a framework named ProMetaR (**Pro**mpt learning via **Meta R**egularization) that jointly meta-learns the regularizer and soft prompts to improve the generalizability of the prompt tuning. Specifically, ProMetaR learns to modulate the gradients of the regularizer to automatically learn effective regularization with a learnable gradient modulation function. This can be viewed as a bi-level optimization, which can be solved with the meta-learning algorithm. The representations learned through the meta-learning algorithms are at a high risk of suffering from *meta-overfitting*, meaning that the meta-parameters are overfitted to a small set of validation data (also referred to as meta-data). To address this issue, we present task augmentation to generate diverse virtual tasks by augmenting the validation set. We also show how ProMetaR improves the generalizability of existing prompting methods from the perspective of gradient alignments.

Our extensive experiments validate the effectiveness of ProMetaR under the base-to-base/base-to-new generalization and domain generalization settings over 11 image recognition datasets and four variants of Imagenet datasets. In the base-to-base/base-to-new generalization settings (Figure 1), our ProMetaR outperforms existing prompt learning methods on 11 image recognition datasets on the both base classes and new classes. It also outperforms CLIP on the new classes while improving the performance on the base classes. These indicate that ProMetaR is effective in both *traditional* generalization and *task* generalization. Further, ProMetaR demonstrates its competitive performance under the domain generalization setting. We also show that our ProMetaR is applicable to various prompting methods as a general training scheme.

The **contribution** of our work can be summarized as:

- We propose ProMetaR, a prompt learning framework for improving the generalizability of the prompt optimization methods. ProMetaR meta-learns both the regularizer and learnable prompts, incorporating task augmentation for more effective meta-learning.
- We provide the theoretical analysis of how our ProMetaR improves the generalizability of prompt learning approaches.
- Our experiments demonstrate the effectiveness and robustness of ProMetaR under the base-to-base/base-to-new settings and domain generalization. Our ProMetaR significantly improves the base prompting methods on the seen (base) and unseen (new) tasks.

2. Related works

Meta-Learning. The goal of meta-learning, as known as *learning to learn*, is to efficiently and effectively adapt to new tasks by leveraging past learning experiences [20]. Applications of learning to learn include learning loss functions [2, 3, 56], learning initialization for task adaptation [13], and few-shot learning [30, 58, 61]. Meta-learning algorithms are typically categorized into three types: metric-based methods [33, 58, 61, 64], memory-based methods [19, 44, 46, 47, 55], and gradient-based methods [14, 35, 48, 53]. After Model-agnostic meta-learning (MAML) [13] has been proposed, gradient-based approaches have been actively explored. But, the gradient-based approaches are often prone to meta-overfitting due to insufficient meta-training tasks [1, 21, 22, 29, 69, 80]. Inspired by these works, ProMetaR automatically learns the effective regularization in a meta-learning manner for the generalizability of the prompting methods and address the meta-overfitting via task augmentation.

Regularization. Regularization is a conventional technique to prevent neural networks from overfitting and enhance the generalization. Conventional regularization methods include constraint-based approaches like weight decay [40, 73], and input-dependent or parameter-dependent approaches such as ensembling [23, 66], dropout [60], and data augmentation [6, 28, 50, 62, 63, 70, 74]. In this work, we present learning to regularize the soft prompts and task augmentation to improve the *traditional* generalization and *task* generalization abilities.

Prompt Learning in Vision-Language Models. Prompt learning has proven to be an effective technique in various natural language processing tasks [34, 38, 39]. Inspired by this success, prompt learning in vision-language models has also been explored [41, 77]. Specifically, CoOp [78] introduces learnable prompts, or soft prompting, which enables efficient finetuning and adaptation of CLIP [52] text encoder. VPT [25] proposes to optimize the prompts in the Vision Transformer (ViT) [9]. Recently, a line of works [5, 26, 71] presents multimodal prompt tuning methods by combining the vision and language prompts. However, many prompt learning approaches in VLMs suffer from the over-fitting issue. Some works have been proposed to address it. For example, ProGrad [79] regularizes the learning process by aligning the update of the soft prompts to the the task-agnostic general knowledge of the VLMs with the gradient alignment. UNIGRAM [36] meta-learns the prompt initialization with a large scale of external data to alleviate the generalizability degradation. PromptSRC [27] regulates the prompt with mutual agreement maximization and self-ensemble. Our ProMetaR meta-learns both learnable prompts and regularization to improve the generalizability without using any external data.

3. Method

We present our ProMetaR (**Pro**mpt learning via **Meta** **R**egularization) to address the limitations of prompt learning in small data regimes. Our novel framework automatically learns effective regularization via meta-learning. We will refer to it as Meta Regularization. Remarkably, the proposed method effectively improves the performance in not only base tasks (*traditional* generalization) but also new tasks (*task* generalization) to address the *task* overfitting problem. We first introduce the background of prompt tuning methods for the vision-language models and the meta-learning. Second, we propose a prompt learning mechanism via meta-regularization to address the over-fitting problems of the prompting approaches. Finally, we provide the theoretical analysis of our ProMetaR to demonstrate how it enhances the prompt tuning methods.

3.1. Preliminaries

Prompt tuning for VLMs. CLIP [52] provides a well-aligned image-text joint embedding space. The pre-trained CLIP image encoder f and text encoder g can be used for zero-shot visual recognition by constructing the *hard prompt*. Specifically, CLIP employs text prompts $\mathbf{p}_y$ generated by hand-crafted templates (*e.g.*, "`A photo of a [CLASS]`"). Then, the prediction probability can be calculated using the visual embeddings $\mathbf{z} = f(\mathbf{x})$ and textual embeddings $\mathbf{w}_y = g(\mathbf{p}_y)$. Given N_c classes, the predicted probability of image $\mathbf{x}$ to be class y is given as:

$$p\left(y|\mathbf{x}\right) = \frac{\exp\left(\text{sim}\left(\mathbf{z}, \mathbf{w}_y\right)/\tau\right)}{\sum_{j=1}^{N_c} \exp\left(\text{sim}\left(\mathbf{z}, \mathbf{w}_j\right)/\tau\right)}, \tag{1}$$

where $\text{sim}\left(\cdot,\cdot\right)$ denotes the cosine similarity, $\mathbf{w}_y$ is the textual embedding of the class y, and τ is the temperature. Even though hard prompts considerably improve CLIP's performance, this technique requires manual efforts to find effective hand-crafted templates for each task, namely, 'prompt engineering'. Instead of manually optimizing hard prompts, 'prompt tuning', also known as 'prompt learning', approaches have been proposed to learn context vectors for the textual and/or visual prompts, namely *soft prompts* [25, 78]. Concretely, by inserting N_t learnable textual prompts $\boldsymbol{\theta}^{\text{txt}} = \left\{\theta_1^{\text{txt}}, \cdots, \theta_{N_t}^{\text{txt}}\right\}$ and N_v visual prompts $\boldsymbol{\theta}^{\text{vis}} = \left\{\theta_1^{\text{vis}}, \cdots, \theta_{N_v}^{\text{vis}}\right\}$, the textual embedding $\tilde{\mathbf{w}}_y$ for class y and visual embedding $\tilde{\mathbf{z}}$ are obtained as follows:

$$\tilde{\mathbf{w}}_y = f\left(\left[\theta_1^{\text{txt}}, \ldots, \theta_{N_t}^{\text{txt}}, \boldsymbol{c}_y\right]\right), \tag{2}$$

$$\tilde{\mathbf{z}} = g\left(\left[\texttt{CLS}, \theta_1^{\text{vis}}, \ldots, \theta_{N_v}^{\text{vis}}, \boldsymbol{E}\right]\right), \tag{3}$$

where $\boldsymbol{c}_y$ is the word embedding of class y, `CLS` denotes the class token, and $\boldsymbol{E}$ is the image patch embeddings. With the weights of the visual encoder f and text encoder g frozen, the prompts are optimized with the contrastive loss:

$$\mathcal{L} = -\sum_i \mathbf{y}_i \log \frac{\exp\left(\text{sim}\left(\tilde{\mathbf{z}}_i, \tilde{\mathbf{w}}_i\right)/\tau\right)}{\sum_{j=1}^{N_c} \exp\left(\text{sim}\left(\tilde{\mathbf{z}}_i, \tilde{\mathbf{w}}_j\right)/\tau\right)}, \tag{4}$$

where $\mathbf{y}_i$ denotes the one-hot vector for the class of the input $\mathbf{x}_i$. With the soft prompts, prompt tuning minimizes manual efforts, and it improves CLIP's performance in the downstream tasks. However, since existing prompt tuning methods tend to focus on *task-specific knowledge*, they often suffer from the overfitting problem, necessitating proper regularization, especially in small data regimes.

Meta-learning. The goal of meta-learning, commonly referred to as 'learning-to-learn,' is to design models that can quickly adapt to new tasks with small data by leveraging past learning experiences across multiple tasks [20]. Let $\mathcal{D}$ denote a meta-training set that consists of training and validation sets across tasks $\mathcal{T}$, *i.e.*, $\mathcal{D} = \{\{D_i^{\text{tr}}, D_i^{\text{val}}\}\}_{i\in\mathcal{T}}$, where D_i^{tr}, and D_i^{valid} are the *traditional* training and validation sets of i-th task. Then, meta-learning can be formulated as a bi-level optimization problem given as:

$$\min_{\phi} \sum_{i\in\mathcal{T}} \mathcal{L}_{\text{valid}}(\theta_i^*(\phi); D_i^{\text{val}}) \tag{5}$$

$$\text{s.t.}\ \ \theta_i^*(\phi) = \arg\min_{\theta_i} \mathcal{L}_{\text{train}}(\theta_i; \phi, D_i^{\text{tr}}), \forall i \in \mathcal{T}, \tag{6}$$

where $\mathcal{L}_{\text{valid}}$ and $\mathcal{L}_{\text{train}}$ denote the losses for the upper- and lower-level optimization problems, and ϕ_i, θ_i are task-specific parameters for i-th task and meta-parameters, respectively. The lower-level optimization in Eq. (6) does the task-specific adaption/training leveraging learning experiences encoded in the meta-parameters ϕ and training set D_i^{tr}. The upper-level optimization in Eq. (5) searches for meta-parameters ϕ that improve the overall validation losses of trained task-specific parameters $\theta_i^*(\phi)$.

A seminal work, MAML [13], can be derived from the formulation above. MAML aims at learning good initialization that is efficiently adaptable to new tasks. Let ϕ denote the initialization of model parameters. With task-specific loss function $\mathcal{L}_i$ and the approximation of lower optimization by one-step update (Eq. (8)), the meta-learning formulation in (Eq. (5)) is converted into MAML's formulation given as:

$$\min_{\phi} \sum_i \mathcal{L}_i(\hat{\theta}_i(\phi); D_i^{\text{val}}) \tag{7}$$

$$\text{s.t.}\ \ \hat{\theta}_i(\phi) = \phi - \alpha\nabla_{\phi}\mathcal{L}_i(\phi; D_i^{\text{tr}}), \forall i \in \mathcal{T}, \tag{8}$$

where α denotes the step size to adapt the initialization ϕ to i-th task. The approximation by one-step update in Eq. (8) enables efficient optimization without the necessity of iterative optimization for lower optimization.

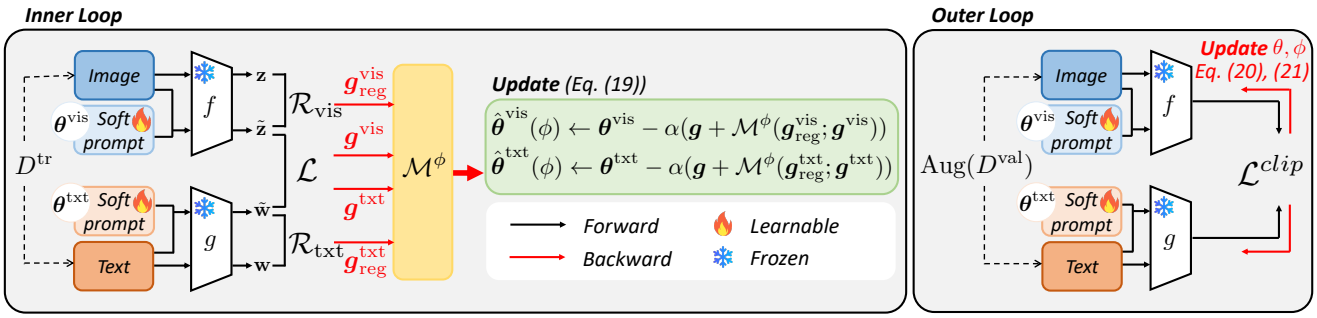


Figure 2. ProMetaR learns the soft prompts $\Theta = \{\boldsymbol{\theta}^{\text{vis}}, \boldsymbol{\theta}^{\text{txt}}\}$ with meta-regularization to generalize well on the new tasks without losing the generalizability of the pretrained VLMs (*e.g.*, CLIP). In the inner-loop (Eq. (19)), we adapt the soft prompts Θ with the gradients $\boldsymbol{g}$ of the loss $\mathcal{L}$ and modulated gradients $\boldsymbol{g}_{\text{reg}} = \mathcal{M}^{\phi}\left(\boldsymbol{g}_{\text{reg}}; \boldsymbol{g}\right)$. In the outer-loop (Eq. (20), (21)), the soft prompts $\boldsymbol{\Theta}$ and the gradient modulation function ϕ are updated on the augmented validation set D^{val}. The image encoder f and text encoder g of the pretrained vision-language models are frozen during the training phase.

3.2. Prompt learning via meta-regularization

We propose a novel framework named Prompt Learning via Meta-Regularization (ProMetaR) to improve the generalizability of existing prompt tuning methods. Our framework adopts meta-learning approaches to learn both soft prompts and regularizers. In addition, we incorporate *task augmentation* into our framework to generate diverse tasks to alleviate the *meta-overfitting*. Figure 2 delineates the overall meta-learning pipeline of the proposed method.

Prompt tuning optimizes prompts to adapt pre-trained models, *e.g.*, a Vision-Language Models (VLMs), to the specific tasks by minimizing a loss:

$$\min_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right), \tag{9}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{\text{txt}}, \boldsymbol{\theta}^{\text{vis}}\}$ denotes the learnable prompts and D^{tr} is the training set of the target downstream task. Since the goal of prompt tuning is the sample-efficient adaptation of the pre-trained models, the training set for prompt tuning is usually small. Thus, prompt tuning methods often suffer from overfitting, showing inferior performance compared to even zero-shot VLMs. To address this problem, we introduce a regularizer $\mathcal{R}$ that penalizes large changes in representations as

$$\mathcal{R}_{\text{vis}} = \sum_{i} \left|\tilde{\mathbf{z}}_i - \mathbf{z}_i\right|, \quad \mathcal{R}_{\text{txt}} = \sum_{j} \left|\tilde{\mathbf{w}}_j - \mathbf{w}_j\right|, \tag{10}$$

where $\mathbf{z}, \mathbf{w}$ denote original visual and textual embeddings, while $\tilde{\mathbf{z}}, \tilde{\mathbf{w}}$ represent corresponding embeddings obtained with prompts Θ. Then, we have:

$$\min_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right) + \lambda \mathcal{R}(\boldsymbol{\Theta}; D^{\text{tr}}), \tag{11}$$

where $\lambda \in \mathbb{R}_{+}$ is the regularization strength and $\mathcal{R}$ unifies $\mathcal{R}_{\text{vis}}$, and $\mathcal{R}_{\text{txt}}$.

However, the regularizer may not always be helpful and manually adjusting the strength of the regularizer, is non-trivial. So, we learn the regularizer to automatically balance it with the main loss, which can be formulated as a bi-level optimization given as:

$$\min_{\boldsymbol{\Theta}, \phi} \mathcal{L}\left(\boldsymbol{\Theta}^{*}\left(\phi\right); D^{\text{val}}\right) \tag{12}$$
$$\text{s.t. } \boldsymbol{\Theta}^{*}\left(\phi\right) = \arg\min_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right) + \mathcal{R}^{\phi}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right),$$

where $\boldsymbol{\Theta}$ is a meta-parameter for the better adaptation, and ϕ is a also meta-parameter to learn the strength of regularizer $\mathcal{R}$. Similar to Eq. (8) in MAML, using the one-step update approximation, Eq. (12) can be rewritten as

$$\min_{\boldsymbol{\Theta}, \phi} \mathcal{L}\left(\hat{\boldsymbol{\Theta}}\left(\phi\right); D^{\text{val}}\right) \tag{13}$$
$$\text{s.t. } \hat{\boldsymbol{\Theta}}\left(\phi\right) = \boldsymbol{\Theta} - \alpha\left(\boldsymbol{g} + \mathcal{M}^{\phi}\left(\boldsymbol{g}_{\text{reg}}; \boldsymbol{g}\right)\right), \tag{14}$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right)$ and $\boldsymbol{g}_{\text{reg}} = \nabla_{\boldsymbol{\Theta}} \mathcal{R}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right)$, and $\mathcal{M}^{\phi}$ is the gradient modulation function with the parameter ϕ that adaptively adjusts $\boldsymbol{g}_{\text{reg}}$ considering $\boldsymbol{g}$ as:

$$\mathcal{M}^{\phi}\left(\boldsymbol{g}_{\text{reg}}; \boldsymbol{g}\right) = \sigma\left(\boldsymbol{m}^{\phi}\right) \odot \boldsymbol{g}_{\text{reg}}, \tag{15}$$

where σ is the sigmoid function and $\odot$ is Hadamard product. The modulation vectors $\boldsymbol{m}^{\phi}$ is computed by $\text{MLP}_{\phi}\left(\left[\boldsymbol{g} || \boldsymbol{g}_{\text{reg}}\right]\right)$ considering the gradients of both the loss and the regularizer.

By learning the regularizer, we have addressed the overfitting problem of the prompt learning methods. We further extend our framework to boost generalization performance in new tasks (*task* generalization) by generating diverse tasks. To this extent, we incorporate *task augmentation* into our framework as:

$$\min_{\boldsymbol{\Theta}, \phi} \mathbb{E}\, \mathcal{L}\left(\hat{\boldsymbol{\Theta}}\left(\phi\right); \text{Aug}\left(D^{\text{val}}\right)\right)$$
$$\text{s.t. } \hat{\boldsymbol{\Theta}}\left(\phi\right) = \boldsymbol{\Theta} - \alpha\left(\boldsymbol{g} + \mathcal{M}^{\phi}\left(\boldsymbol{g}_{\text{reg}}; \boldsymbol{g}\right)\right), \tag{16}$$

where Aug $(\cdot)$ is the task augmentation operation. The task augmentation generates new labels to augment many tasks, which encourages the parameters to be optimized for diverse tasks. The augmented task can be viewed as a virtually large meta-validation set with many tasks. This helps the model generalize on new tasks.

Mixup-based augmentation is one of the augmentation operations that generate new interpolated labels. In our experiments, task augmentation randomly draws samples from train and validation sets and employs manifold mix [63] for augmentation. Specifically, given a pair of random samples $\mathbf{x}_i \in D^{\text{val}}$ and $\mathbf{x}_j \in D^{\text{tr}}$ from validation and training sets, we interpolate the last layer features $(\boldsymbol{h}_{\text{val}}^{(i)}, \boldsymbol{h}_{\text{tr}}^{(j)})$ and their labels $(y_{\text{val}}^{(i)}, y_{\text{tr}}^{(j)})$ as:

$$\hat{\boldsymbol{h}}_{\text{val}}^{(i)} = \rho \boldsymbol{h}_{\text{val}}^{(i)} + (1-\rho)\boldsymbol{h}_{\text{tr}}^{(j)}, \tag{17}$$

$$\hat{y}_{\text{val}}^{(i)} = \rho y_{\text{val}}^{(i)} + (1-\rho) y_{\text{tr}}^{(j)}, \tag{18}$$

where $\rho \in [0,1]$ is a mixture ratio, which is sampled from the Beta distribution Beta (μ, ν).

Remarks. Note that similar to overfitting, meta-learning algorithms often suffer from *meta-overfitting*, especially when the size of the meta-validation set $\{D_i^{\text{val}}\}_{i\in\mathcal{T}}$ in Eq. (5) is small [1, 69, 80]. The size is related to the quantity of both samples and tasks, and their diversity. Unfortunately, in prompt tuning benchmarks such as base-to-base/base-to-new generalization and domain generalization settings, only one task is available for training with a small number of samples. This setting is challenging and can be seen as 'single-task meta-learning'. In our framework, task augmentation effectively addresses the scarcity of tasks/samples and boosts generalization performance in both a base (seen) task and a new task.

Overall procedure of ProMetaR. Motivated by the episodic training scheme [64], we divide the batch into the training and validation set based on the class of the sample. To maintain the in-domain performance, we first update the parameters with the conventional gradient descent. Then, we update the parameters with the meta-learning. The learnable prompts $\boldsymbol{\Theta}$ are adapted with the gradients of the loss and modulated gradients (inner-loop):

$$\hat{\boldsymbol{\Theta}}(\boldsymbol{\phi}) \leftarrow \boldsymbol{\Theta} - \alpha \left(\boldsymbol{g} + \mathcal{M}^{\boldsymbol{\phi}}\left(\boldsymbol{g}_{\text{reg}}; \boldsymbol{g}\right)\right) \tag{19}$$

After the update, the learnable prompts $\boldsymbol{\Theta}$ and gradient modulation function $\boldsymbol{\phi}$ are optimized for performing well on the augmented set (outer-loop):

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \beta \nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\hat{\boldsymbol{\Theta}}(\boldsymbol{\phi}); \text{Aug}\left(D^{\text{val}}\right)\right), \tag{20}$$

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta \nabla_{\boldsymbol{\phi}} \mathcal{L}\left(\hat{\boldsymbol{\Theta}}(\boldsymbol{\phi}); \text{Aug}\left(D^{\text{val}}\right)\right), \tag{21}$$

where α, β are hyperparameters, respectively.

3.3. Analysis of ProMetaR

We provide the analysis to elucidate how our proposed ProMetaR enhances the generalizability of prompt learning from the standpoint of gradient alignment [64]. The objective of ProMetaR is to find the optimal soft prompts as follows:

$$\min_{\boldsymbol{\Theta},\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\Theta} - \alpha\left(\boldsymbol{g} + \mathcal{M}^{\boldsymbol{\phi}}\left(\boldsymbol{g}_{\text{reg}}\right)\right); D^{\text{val}}\right), \tag{22}$$

where $\boldsymbol{g} = \nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right), \boldsymbol{g}_{\text{reg}} = \nabla_{\boldsymbol{\Theta}} \mathcal{R}\left(\boldsymbol{\Theta}; D^{\text{tr}}\right)$ are the gradients of loss $\mathcal{L}$ and regularizer $\mathcal{R}$, respectively.

We can approximate $\mathcal{L}(\boldsymbol{x})$ with first-order Taylor expansion. Given loss $\mathcal{L}(\boldsymbol{x})$, its first-order approximation via Taylor expansion is as follows:

$$\mathcal{L}(\boldsymbol{x}) \approx \mathcal{L}(\boldsymbol{x}_0) + \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}_0)^{\top}(\boldsymbol{x} - \boldsymbol{x}_0), \tag{23}$$

where $\boldsymbol{x}_0$ is an arbitrary point and $\boldsymbol{x}$ is a point close to $\boldsymbol{x}_0$. Assume that we have $\boldsymbol{x} = \boldsymbol{\Theta} - \alpha\left(\boldsymbol{g} + \mathcal{M}^{\boldsymbol{\phi}}\left(\boldsymbol{g}_{\text{reg}}\right)\right)$ and $\boldsymbol{x}_0 = \boldsymbol{\Theta}$. Then, our objective (Eq. (22)) can be written as:

$$\begin{aligned}\min_{\boldsymbol{\Theta},\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{val}}\right) + \\ \nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}\right)^{\top}\left(-\alpha\left(\boldsymbol{g} + \mathcal{M}^{\boldsymbol{\phi}}\left(\boldsymbol{g}_{\text{reg}}\right)\right)\right).\end{aligned} \tag{24}$$

Since $\mathcal{M}^{\boldsymbol{\phi}}\left(\boldsymbol{g}_{\text{reg}}\right) = \sigma\left(\boldsymbol{m}^{\boldsymbol{\phi}}\right) \odot \boldsymbol{g}_{\text{reg}}$, we can rewrite Eq. (24) as below:

$$\begin{aligned}\min_{\boldsymbol{\Theta},\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\Theta}; D^{\text{val}}\right) - \alpha\left(\nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}\right)^{\top} \boldsymbol{g}\right) \\ - \alpha\left(\nabla_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta}\right)^{\top}\left(\sigma\left(\boldsymbol{m}^{\boldsymbol{\phi}}\right) \odot \boldsymbol{g}_{\text{reg}}\right)\right).\end{aligned} \tag{25}$$

This equation has three terms. The optimization above implies minimizing (i) the loss on the validation set, (ii) maximizing the inner product between the gradients of the losses on the validation set and the training set, and (iii) maximizing the inner product between the gradient of the validation loss and the regularizer on the training set. So, these indicate that this optimization prefers a solution/direction where the training and validation gradients agree, which leads to better generalization on new tasks. In addition, the third term in Eq. (25) plays a role in avoiding the conflict of the update between the task-specific knowledge by tuned prompts and task-agnostic general knowledge provided by original prompts. From the perspective of the gradient alignment [79], the third term leads to a reduction in the generalization error by aligning the gradients induced by tuned prompts and general knowledge from the original prompts. So, our proposed ProMetaR enhances the *task* generalization ability as well as *traditional* generalization capability.

4. Experiments

In this section, we demonstrate the effectiveness of our proposed ProMetaR. We first introduce datasets, baselines, and implementation details. Next, we provide the ablation studies to explore the contribution of each component in ProMetaR. Then, we compare the proposed method with other prompting-based methods to evaluate the ability of traditional generalization on seen categories (base-to-base), and task generalization to unseen categories (base-to-new) and new datasets (domain generalization). We also design a task overfitting score and provide analysis to show the efficacy of the proposed method.

4.1. Experimental settings

We evaluate ProMetaR on base-to-base/base-to-new generalization and domain generalization following other prompting works [26].

Base-to-base/Base-to-new generalization. We train the prompts only on the base classes in a 16-shot (16 images per class) setting and measure the performance of the prompting methods on base and new classes. In this setting, the model cannot see new classes in the training phase.

Domain generalization. We also validate the effectiveness of our model in a 16-shot on out-of-distribution datasets. We train the model only using ImageNet dataset (source) and perform inference on four other variants (target) of ImageNet dataset. In other words, the model cannot see target domains in the training phase.

Datasets. For base-to-base/base-to-new class generalization, we evaluate our method on 11 image recognition datasets: ImageNet [8], Caltech101 [11], OxfordPets [51], StanfordCars [31], Flowers102 [49], Food101 [4], FGVCAircraft [43], SUN397 [68], UCF101 [59], DTD [7], and EuroSAT [16], following other prompting methods [26, 77]. We also evaluate our method on domain generalization settings by setting ImageNet [8] as the source dataset. The target datasets contain four ImageNet variants: ImageNetV2 [54], ImageNet-Sketch [65], ImageNet-A [18], and ImageNet-R [17].

Baselines. To validate the effectiveness of our ProMetaR, we use the following baselines: (1) zero-shot CLIP [52], (2) textual prompt learning approaches: CoOp [78] and CoCoOp [77], (3) multimodal prompt learning approaches: MAPLE [26] and RPO [32], (4) prompt learning with regularization and ensemble methods: PromptSRC [27], (5) prompt learning with the meta-learning: UNIGRAM [36], and (6) our base prompting method: IVLP.

Experimental details. Following other prompt learning works [26, 27, 77], we use CLIP-ViT-B/16 as the pretrained backbone model and four soft prompting tokens for each modality. For the base prompt learning method, we use Independent Vision-Language Prompting as a base prompt learning method that optimizes hierarchical prompts on both image and text modalities [26]. In all experiments, we evaluate the performance of the methods in three independent runs (seed 1, 2, and 3) and report average performance following other prompt learning works [26, 27, 77].

	MetaLearn	TaskAug	MetaReg	Base	New	H
(a)				82.51	73.36	77.66
(b)	✓			83.51	73.15	77.99
(c)	✓	✓		84.04	75.37	79.47
(d)	✓		✓	84.27	75.06	79.40
(e)	✓	✓	✓	**84.39**	**76.93**	**80.49**

Table 1. Contribution of each component of our ProMetaR. Results are averaged over 11 datasets. H refers to harmonic mean. MetaLearn: meta-learning, TaskAug: Task augmentation to alleviate the meta-overfitting, MetaReg: meta-regularization to learn the regularizer.

4.2. Effectiveness of ProMetaR

We validate the effectiveness of each component of the proposed ProMetaR under the base-to-base/base-to-new setting. Table 1 provides the ablation study on our components, and the results are averaged over 11 datasets. MetaLearn denotes meta-learning, TaskAug indicates task augmentation to alleviate the meta-overfitting, and MetaReg refers to meta-regularization. Eliminating all of our components, or (a), corresponds to using only IVLP, which is the base prompt learning method of ProMetaR. By adopting meta-learning to IVLP ((a) $\rightarrow$ (b)), the base class performance improves (+1.0%) but it impairs generalization to new classes (-0.21%). However, our task augmentation ((b) $\rightarrow$ (c)) significantly enhances the average accuracy on new classes and harmonic mean with gains of +2.22% and +1.48%, respectively, compared to IVLP+meta-learning. Additionally, our meta-regularization ((b) $\rightarrow$ (d)) improves accuracy for both base and new classes by +0.76% and +1.91%, respectively. This indicates that both task augmentation and meta-regularization clearly ameliorate the meta-overfitting caused by meta-learning and contribute to strong generalization. Furthermore, by adding meta-regularization to (c), *i.e.*, (c) $\rightarrow$ (e), all three accuracies increase to +0.35% (base class), +1.56% (new class), and +1.02% (harmonic mean). Employing task augmentation to (d), *i.e.*, (d) $\rightarrow$ (e), leads to an additional +1.87% growth in new class accuracy. Our ProMetaR significantly improves over IVLP for both base and new classes ((a) $\rightarrow$ (e)), achieving performance gains of +1.88%, +3.57%, and +2.83% on the base class, new class accuracy, and harmonic mean, respectively.

Dataset		CLIP [52]	CoOp [78]	CoCoOp [77]	MaPLe [26]	RPO [32]	PromptSRC [27]	UNIGRAM [36]	IVLP (Base)	**ProMetaR (Ours)**	Gain Δ
Avg. Rank		8.18	8.55	6.73	3.64	4.55	2.73	3.82	5.27	**1.36**	-
Average on 11 datasets	Base	69.34	82.69	80.47	82.28	81.13	84.26	80.34	82.51	**84.39**	+1.88
	New	74.22	63.22	71.69	75.14	75.00	76.10	75.92	73.35	**76.93**	+3.58
	H	71.70	71.66	75.83	78.55	77.78	79.97	78.07	77.66	**80.49**	+2.83
ImageNet	Base	72.43	76.47	75.98	76.66	76.60	77.60	76.60	77.39	**77.76**	+0.37
	New	68.14	67.88	70.43	70.54	**71.57**	70.73	70.69	70.04	70.75	+0.71
	H	70.22	71.92	73.10	73.47	74.00	74.01	73.53	73.53	**74.09**	+0.56
Caltech 101	Base	96.84	98.00	97.96	97.74	97.97	98.10	98.07	**98.28**	98.11	-0.17
	New	94.00	89.81	93.81	94.36	94.37	94.03	**95.11**	93.65	94.29	+0.64
	H	95.40	93.73	95.84	96.02	96.03	96.02	**96.57**	95.91	96.16	+0.25
Oxford Pets	Base	91.17	93.67	95.20	95.43	94.63	95.33	94.94	95.41	**95.57**	+0.16
	New	97.26	95.29	97.69	97.76	97.50	97.30	**97.94**	96.31	97.43	+1.12
	H	94.12	94.47	96.43	**96.58**	96.05	96.30	96.42	95.86	96.49	+0.63
Stanford Cars	Base	63.37	78.12	70.49	72.94	73.87	78.27	73.50	72.39	**78.32**	+5.93
	New	74.89	60.40	73.59	74.00	**75.53**	74.97	75.38	73.31	75.18	+1.87
	H	68.65	68.13	72.01	73.47	74.69	76.58	74.43	72.85	**76.72**	+3.87
Flowers 102	Base	72.08	97.60	94.87	95.92	94.13	98.07	95.20	96.17	**98.13**	+1.96
	New	**77.80**	59.67	71.75	72.46	76.67	76.50	76.21	73.64	77.66	+4.02
	H	74.83	74.06	81.71	82.56	84.50	85.95	84.65	83.41	**86.70**	+3.29
Food101	Base	90.10	88.33	90.70	90.71	90.33	90.67	**90.84**	90.53	90.80	+0.27
	New	91.22	82.26	91.29	92.05	90.83	91.53	**92.12**	91.66	91.89	+0.23
	H	90.66	85.19	90.99	91.38	90.58	91.10	**91.48**	91.09	91.34	+0.25
FGVC Aircraft	Base	27.19	40.44	33.41	37.44	37.33	**42.73**	32.25	37.24	42.02	+4.78
	New	36.29	22.30	23.71	35.61	34.20	37.87	38.00	34.47	**38.63**	+4.16
	H	31.09	28.75	27.74	36.50	35.70	40.15	34.89	35.80	**40.25**	+4.45
SUN397	Base	69.36	80.60	79.74	80.82	80.60	82.67	80.43	82.63	**82.70**	+0.07
	New	75.35	65.89	76.86	78.70	77.80	78.57	77.91	78.40	**79.02**	+0.62
	H	72.23	72.51	78.27	79.75	79.18	80.52	79.15	80.46	**80.82**	+0.36
DTD	Base	53.24	79.44	77.01	80.36	76.70	**83.37**	73.62	80.67	83.02	+2.35
	New	59.90	41.18	56.00	59.18	62.13	62.97	62.38	55.31	**64.05**	+8.74
	H	56.37	54.24	64.85	68.16	68.61	71.75	67.56	65.63	**72.31**	+6.68
EuroSAT	Base	56.48	92.19	87.49	94.07	86.63	92.90	86.26	92.64	**94.94**	+2.30
	New	64.05	54.74	60.04	73.23	68.97	73.90	71.38	63.33	**77.44**	+14.11
	H	60.03	68.69	71.21	82.35	76.79	82.32	78.12	75.23	**85.30**	+10.07
UCF101	Base	70.53	84.69	82.33	83.00	83.67	**87.10**	82.00	84.23	86.97	+2.74
	New	77.50	56.05	73.45	78.66	75.43	78.80	78.06	76.78	**79.84**	+3.06
	H	73.85	67.46	77.64	80.77	79.34	82.74	79.98	80.33	**83.25**	+2.92

Table 2. Performance comparison on the base-to-new generalization setting. We train our model with a subset of the classes (base classes) in a 16-shot setting and evaluate on the test set including base classes and new classes. H denotes the harmonic mean of base and novel performance to show the generalization trade-off [67]. **Avg. Rank** is the average rank of the harmonic mean on each dataset among the baselines. Δ denotes the performance gain of ProMetaR from IVLP (our base prompting method).

4.3. Base-to-base/Base-to-new generalization

We compare the performance of ProMetaR with other recent prompting approaches in the base-to-base/base-to-new generalization setting to demonstrate the effectiveness of the proposed learning framework. Following [26, 77], we report the average accuracy of three different data splits used in CoCoOp [77] for a fair comparison. The results are reported in Table 2.

Our ProMetaR shows the best performance on the average accuracy over 11 datasets among baselines. In particular, ProMetaR achieves a significant improvement on new classes from 76.10 to 76.93 compared to the best baseline method PromptSRC. Also, ProMetaR substantially improves the average accuracy of the base model IVLP by 3.58 on new classes. This result indicates that our ProMetaR enhances the generalizability of existing prompting methods by meta-learning the regularization. In comparison with UNIGRAM, which applies meta-learning with a large scale of external data, ProMetaR shows impressive performance improvement on both base and new categories without any external data for the meta-learning.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
RPO	**71.67**	**65.13**	49.27	50.13	76.57	60.28
PromptSRC	71.27	64.35	**49.55**	50.90	77.80	60.65
UNIGRAM	71.65	64.81	49.54	51.51	77.34	60.80
ProMetaR	71.29	64.39	**49.55**	**51.25**	**77.89**	**60.77**

Table 3. Performance comparison on the domain generalization.

	Top-3			Bottom-3	
Dataset	$\boldsymbol{tos}^{\text{IVLP}}$	Gain Δ	Dataset	$\boldsymbol{tos}^{\text{IVLP}}$	Gain Δ
EuroSAT	36.88	10.07	Food101	-0.01	0.25
DTD	32.02	6.68	Caltech101	1.79	0.25
Flowers	28.25	3.29	Imagenet	3.06	0.56

Table 4. Task overfitting score $\boldsymbol{tos}^{\text{IVLP}} = \delta_{\text{base}}^{\text{IVLP}} - \delta_{\text{new}}^{\text{IVLP}}$ and the gain Δ. Δ denotes the performance gain (H) by ProMetaR on IVLP (Table 2).

4.4. Domain generalization

In the domain generalization setting, the performance comparison of ImageNet-trained models, evaluated with four out-of-distribution variants, is reported in Table 3. For a fair comparison, we exclude UNIGRAM since it employs a large scale of extra datasets to pre-train the learnable prompts. ProMetaR successfully generalizes to out-of-domain datasets showing the best average accuracy. This demonstrates that our meta-regularizer and task augmentation clearly enhance the robustness to domain shifts.

4.5. Analysis

Task overfitting score. We analyze when our ProMetaR provides a relatively large (or small) performance improvement compared to the base model (IVLP). To quantify the room for improvement, we define Task Overfitting Score ($\boldsymbol{tos}$) of the prompting method <pr> as

$$\boldsymbol{tos}^{<pr>} = \delta_{\text{base}}^{<\text{pr}>} - \delta_{\text{new}}^{<\text{pr}>}, \tag{26}$$

where $\delta_{\text{base}}^{<\text{pr}>} = \max(0, s_{\text{base}}^{<\text{pr}>} - s_{\text{base}}^{\text{CLIP}}), \delta_{\text{new}}^{<\text{pr}>} = s_{\text{new}}^{<\text{pr}>} - s_{\text{new}}^{\text{CLIP}}$ be the performance difference between prompting method <pr> and zero-shot CLIP on the base and new classes, respectively. $s_{\text{base}}^{<\text{pr}>}, s_{\text{new}}^{<\text{pr}>}$ indicate the accuracy of the prompting method <pr> on base and new classes, respectively. As the task overfitting score is lower, the method <pr> tends to generalize well on new tasks. Table 4 reports the task overfitting score and performance gain Δ of ProMetaR from IVLP (Table 2) on the datasets with top-3 (left) and bottom-3 (right) task overfitting scores. The table shows that gains of ProMetaR are relatively high when the task overfitting score is high. It demonstrates that ProMetaR is more effective when prompting method IVLP suffers from overfitting.

Methods	Base	New	H
CoOp	82.69	63.22	71.66
+ ProMetaR	83.35	71.20	76.80
VPT	82.75	71.00	76.43
+ ProMetaR	83.18	73.19	77.87

Table 5. Performance comparison of ProMetaR with different prompting approaches (CoOp [78] and VPT [25]) under the base-to-base/base-to-new generalization setting.

Method	Base	New	H
Loss+Reg.	83.96	75.70	79.62
ProMetaR (**Ours**)	**84.39**	**76.93**	**80.49**
Performance Gain (Δ)	+0.43	+1.23	+0.87

Table 6. Performance comparison of ProMetaR with IVLP trained with the loss and regularizer under the base-to-base/base-to-new generalization setting.

ProMetaR with diverse methods. ProMetaR can be applied to any existing prompting methods in a plug-and-play manner. We elucidate the effectiveness of ProMetaR by comparing the performance of various methods, such as CoOp and VPT, with our method plugged in (Table 5). ProMetaR consistently improves all the other prompt learning methods with harmonic mean gains of +5.14% and +1.44% over CoOp and VPT, respectively. Moreover, the performance is enhanced, especially in new classes, indicating that our ProMetaR effectively prevents the prompts from overfitting to downstream tasks.

Meta-Regularization. In Table 6, we also compare ProMetaR with IVLP trained with the loss and the regularizer (Loss+Reg) in (11) with manually tuned hyperparameters (*e.g.*, a regularization strength). The experimental results show that our ProMetaR outperforms standard IVLP training with regularization (Loss+Reg). This result indicates that our ProMetaR automatically learns more effective regularization via meta-learning.

5. Conclusion

We propose ProMetaR to encourage both traditional generalization and task generalization, yielding a significant performance improvement in base-to-base/base-to-new and domain generalization settings. Specifically, we adopt meta-learning to learn both soft prompts and regularizers. We further incorporate task augmentation to generate diverse tasks and address the meta-overfitting. Extensive experiments and analyses demonstrate that our ProMetaR enhances the generalizability of prompt learning.

Acknowledgements. This work was partly supported by ICT Creative Consilience Program through the IITP, NRF of Korea grants funded by the Korea government (MSIT) (IITP-2024-2020-0-01819, NRF-2023R1A2C2005373), and the NVIDIA academic grant. We thank Jongha Kim for the suggestions on the analysis.

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *ICLR*, 2019. 2, 5

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018. 2

[3] Sarah Bechtle, Artem Molchanov, Yevgen Chebotar, Edward Grefenstette, Ludovic Righetti, Gaurav S. Sukhatme, and Franziska Meier. Meta learning via learned loss. In *ICPR*, 2020. 2

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

[5] Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. Distribution-aware prompt tuning for vision-language models. In *ICCV*, 2023. 2

[6] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J Kim. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. In *NeurIPS*, 2023. 2

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

[8] 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. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 1

[11] 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*, 2004. 6

[12] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 1

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2, 3

[14] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *ICLR*, 2018. 2

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTARS*, 12(7):2217–2226, 2019. 6

[17] 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*, 2021. 6

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 6

[19] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *ICANN*, 2001. 2

[20] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *TPAMI*, 44(9):5149–5169, 2021. 2, 3

[21] Dasol Hwang, Jinyoung Park, Sunyoung Kwon, KyungMin Kim, Jung-Woo Ha, and Hyunwoo J Kim. Self-supervised auxiliary learning with meta-paths for heterogeneous graphs. In *NeurIPS*, 2020. 2

[22] Dasol Hwang, Jinyoung Park, Sunyoung Kwon, Kyung-Min Kim, Jung-Woo Ha, and Hyunwoo J Kim. Self-supervised auxiliary learning for graph neural networks via meta-learning. *arXiv:2103.00771*, 2021. 2

[23] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *NeurIPS*, 2022. 2

[24] 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*, 2021. 1

[25] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2, 3, 8

[26] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 1, 2, 6, 7

[27] 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*, 2023. 2, 6, 7

[28] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *ICLR*, 2021. 2

[29] Dohwan Ko, Joonmyung Choi, Hyeong Kyu Choi, Kyoung-Woon On, Byungseok Roh, and Hyunwoo J Kim. Meltr: Meta loss transformer for learning to fine-tune video foundation models. In *CVPR*, 2023. 2

[30] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICMLW*, 2015. 2

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 6

[32] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, 2023. 1, 6, 7

[33] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019. 2

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

[35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2

[36] Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *ICCV*, 2023. 2, 6, 7

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

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

[39] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

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

[42] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022. 1

[43] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 6

[44] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018. 2

[45] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv:2111.09734*, 2021. 1

[46] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017. 2

[47] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018. 2

[48] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv:1803.02999*, 2018. 2

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

[50] Hyeonjin Park, Seunghun Lee, Sihyeon Kim, Jinyoung Park, Jisu Jeong, Kyung-Min Kim, Jung-Woo Ha, and Hyunwoo J Kim. Metropolis-hastings data augmentation for graph neural networks. In *NeurIPS*, 2022. 2

[51] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 6

[52] 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. 1, 2, 3, 6, 7

[53] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016. 2

[54] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 6

[55] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016. 2

[56] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 2

[57] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 1

[58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2

[59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *ICCVW*, 2013. 6

[60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15 (1):1929–1958, 2014. 2

[61] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 2

[62] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *ICLR*, 2021. 2

[63] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Aaron Courville, Ioannis Mitliagkas, and Yoshua Bengio. Manifold mixup: learning better representations by interpolating hidden states. In *ICML*, 2019. 2, 5

[64] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 2, 5

[65] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 6

[66] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 2

[67] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 7

[68] 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. 6

[69] Huaxiu Yao, Linjun Zhang, and Chelsea Finn. Meta-learning with fewer tasks through task interpolation. In *ICLR*, 2022. 2, 5

[70] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regu-

larization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2

[71] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv:2210.07225*, 2022. 2

[72] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 1

[73] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *ICLR*, 2019. 2

[74] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[75] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024. 1

[76] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1

[77] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2, 6, 7

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

[79] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, 2023. 2, 5

[80] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *ICML*, 2019. 2, 5