# Learning Domain Invariant Prompt for Vision-Language Models

Cairong Zhao, *Member, IEEE*, Yubin Wang, Xinyang Jiang, Yifei Shen, *Graduate Student Member, IEEE*, Kaitao Song, Dongsheng Li, and Duoqian Miao

***Abstract*— Prompt learning stands out as one of the most efficient approaches for adapting powerful vision-language foundational models like CLIP to downstream datasets by tuning learnable prompt vectors with very few samples. However, despite its success in achieving remarkable performance on in-domain data, prompt learning still faces the significant challenge of effectively generalizing to novel classes and domains. Some existing methods address this concern by dynamically generating distinct prompts for different domains. Yet, they overlook the inherent potential of prompts to generalize across unseen domains. To address these limitations, our study introduces an innovative prompt learning paradigm, called MetaPrompt, aiming to directly learn *domain invariant* prompt in few-shot scenarios. To facilitate learning prompts for image and text inputs independently, we present a dual-modality prompt tuning network comprising two pairs of coupled encoders. Our study centers on an alternate episodic training algorithm to enrich the generalization capacity of the learned prompts. In contrast to traditional episodic training algorithms, our approach incorporates both in-domain updates and domain-split updates in a batch-wise manner. For in-domain updates, we introduce a novel asymmetric contrastive learning paradigm, where representations from the pre-trained encoder assume supervision to regularize prompts from the prompted encoder. To enhance performance on out-of-domain distribution, we propose a domain-split optimization on visual prompts for cross-domain tasks or textual prompts for cross-class tasks during domain-split updates. Extensive experiments across 11 datasets for base-to-new generalization and 4 datasets for domain generalization exhibit favorable performance. Compared with the state-of-the-art method, MetaPrompt achieves an absolute gain of 1.02% on the overall harmonic mean in base-to-new generalization and consistently demonstrates superiority over all benchmarks in domain generalization.**

Manuscript received 7 September 2023; revised 21 December 2023; accepted 29 January 2024. Date of publication 9 February 2024; date of current version 14 February 2024. This work was supported in part by the National Natural Science Fund of China under Grant 62076184, Grant 61976158, Grant 61976160, Grant 62076182, and Grant 62276190; in part by the Fundamental Research Funds for the Central Universities and the State Key Laboratory of Integrated Services Networks (Xidian University); in part by the Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; and in part by the Shanghai Natural Science Foundation under Grant 22ZR1466700. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikos Deligiannis. *(Cairong Zhao and Yubin Wang contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

Cairong Zhao, Yubin Wang, and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhaocairong@tongji.edu.cn; wangyubin2018@tongji.edu.cn; dqmiao@tongji.edu.cn).

Xinyang Jiang, Yifei Shen, Kaitao Song, and Dongsheng Li are with Microsoft Research Asia, Shanghai 200232, China (e-mail: xinyangjiang@microsoft.com; yifeishen@microsoft.com; kaitaosong@microsoft.com; dongsheng.li@microsoft.com).

Digital Object Identifier 10.1109/TIP.2024.3362062

***Index Terms*— Prompt learning, meta-learning, few-shot learning, domain generalization.**

## I. Introduction

RECENT research in pre-training large Vision-Language Models (VLM) using web-scale data has shown remarkable progress in learning transferable representations [23], [47]. In contrast to conventional supervised learning approaches that acquire closed-set visual concepts through discrete labels, these models align images within a shared embedding space using contrastive learning, presenting a promising prospect for harnessing human language to guide visual recognition tasks. Benefiting from this paradigm, pre-trained vision-language models can conduct zero-shot or few-shot transfer to downstream tasks with open-set visual concepts learned from natural language supervision. Consequently, how to effectively leverage these powerful foundation models emerges as a pivotal direction of research. Recent studies [62], [73] have employed a simple yet effective way to adapt pre-trained vision-language models to downstream tasks, called prompting. Manually designing an appropriate prompt constitutes a nontrivial endeavor due to its inherent ambiguity, thereby rendering automatic prompt tuning the current mainstream approach. Drawing inspiration from recent progress in prompt learning [30], [34], [37] within the domain of natural language processing, methods like CoOp [73], CoCoOp [62] and MaPLe [26] learn a set of continuous vectors as the context (i.e., prompt vector) with the pre-trained parameters fixed. This approach leads to noteworthy enhancements even when utilizing a limited number of training samples.

Despite demonstrating promising performance in i.i.d. samples, as discussed in prior research [62], prompt learning still encounters a significant challenge in terms of domain generalization. Similar to other machine learning methods, conventional prompt tuning approaches [73] often tend to overfit the distribution of the training set. When transferred to unseen domains, the strong generalization capacity of learned prompt vectors becomes compromised, leading to a substantial reduction in performance. Even with massive tuning, ensuring an optimal prompt for downstream tasks remains elusive. Recently, several methods [62], [71] have addressed this challenge through the adaptive generation of prompts for different tokens or domains, known as conditional prompt learning. Nevertheless, they fall short in enhancing the

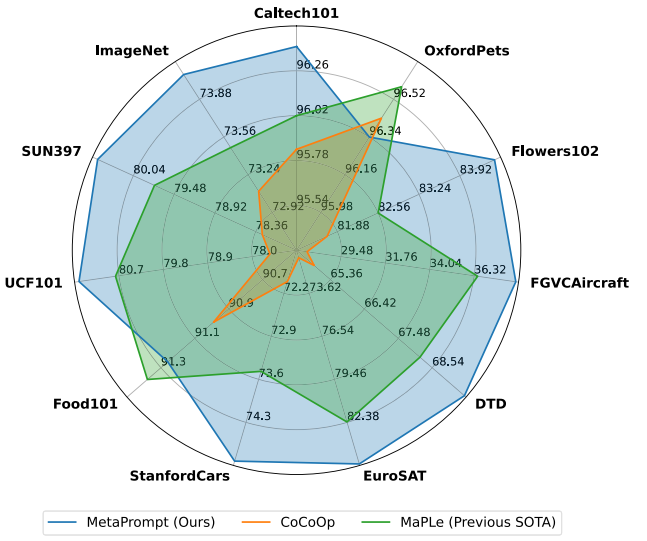



Fig. 1. Comprehensive comparison of the harmonic mean of previous methods CoCoOp, MaPLe, and our method MetaPrompt on 11 diverse image recognition datasets for base-to-new generalization. MetaPrompt surpasses state-of-the-art methods on 9 of 11 datasets.

generalization ability of learned prompts and cannot enforce the prompts to generalize to unseen domains.

In this paper, our goal is to explicitly learn the domain invariant prompt for vision-language models, which is independent of the input and exhibits a low bias toward visual representations of various downstream tasks. Due to the significant distribution shift, our emphasis is directed towards cross-domain tasks, wherein the test samples are out-of-domain. As discussed in previous literature [21], [27], [65], input samples are composed of attributes (i.e., factors of variation), such as color, shape, texture, etc., and different domains are defined by different distributions of each attribute. As a result, there exists a unified meta-domain containing all possible attributes, where data domains are attribute distributions sampled from this meta-domain. Under this assumption, our theoretical analysis, in alignment with [8], demonstrates that tuning prompts via an episodic training strategy provides a robust generalization guarantee. Specifically, this approach has the generalization bound of $O(1/\sqrt{N})$, where $N$ represents the number of tasks, independent of the sample size within each domain. This observation drives our proposal of an episodic prompt tuning method in few-shot scenarios.

Consequently, to better leverage the potency of episodic training and maintain good performance on unseen domains, we introduce MetaPrompt, a simple but effective few-shot approach that generates the domain invariant prompt for vision-language models. Aiming at addressing the overfitting issue of prompts learned on in-domain data, we introduce an alternated episodic training algorithm designed to improve generalization when applied to out-of-domain data. To facilitate this algorithm, we propose a dual-modality prompt tuning network as our framework, which learns prompt vectors from both vision and text modalities, respectively, using two distinct pairs of coupled encoders.

In contrast to conventional meta-learning-based episodic training strategies, our alternate algorithm, as a batch-wise algorithm, performs two distinct updates on a single batch, i.e. an in-domain update following a domain-split update. During in-domain updates, a novel asymmetric contrastive learning paradigm is elaborated, aiming to exploit the robust generalization capacity of the pre-trained vision-language model. For instance, representations from the pre-trained text encoder assume guidance for tuning the prompted image encoder with contrastive learning, and vice versa. To explicitly enhance the generalization ability of prompts on unseen domains, we additionally present a domain-split optimization for prompt tuning. With a modality-specific optimization strategy, we impose a constraint on visual prompts for cross-domain tasks and textual prompts for cross-class tasks during domain-split updates. During training on a specific distribution, this constraint optimizes prompts for achieving good performance on out-of-distribution samples.

In this paper, the ability of generalization is evaluated from two perspectives, new image domains and new class domains. Our MetaPrompt is applicable for both out-of-domain classes (i.e., base-to-new generalization) and images (i.e., conventional domain generalization). As shown in Fig. 1, for base-to-new generalization, MetaPrompt obtains an overall improvement of harmonic mean accuracy by an average gain of 1.02% over the previous state-of-the-art method MaPLe on 11 image recognition benchmark datasets. For domain generalization, our few-shot method achieves comparable performance over other methods training on full samples and outperforms other zero-shot or few-shot methods on all domain generalization benchmark datasets. These experimental results demonstrate the effectiveness of MetaPrompt and show its superiority in generalization capacity to other prompt tuning approaches.

The contributions of our work are summarized as follows. 1) We introduce an innovative prompt learning paradigm, called MetaPrompt, which directly learns domain invariant prompt in few-shot scenarios. This paradigm aims to tackle the major challenge of generalizing to unseen classes or domains in prompt learning with vision-language models. 2) We present a dual-modality prompt tuning network comprising two pairs of coupled encoders to facilitate learning prompts for image and text inputs independently. 3) We center on an alternate episodic training algorithm to enrich the generalization capacity of the learned prompts, which alternates between in-domain updates and domain-split updates for prompt tuning.

## II. Related Work

### A. Prompt Learning

Prompt learning emerges from recent advances in natural language processing. The core idea of prompt learning is to formalize various tasks [11], [47], [48] to masked language modeling problems with different prompt templates. A prompt can be seen as a function of the input tokens, providing instruction for adapting pre-trained language models such

as BERT [11] or GPT [48] to downstream tasks. Earlier work [36] has enabled the model to understand the task and make better predictions by manually designing discrete natural language prompts. Nonetheless, some hand-crafted prompt templates prove inappropriate in many cases due to their inherent ambiguity, while the performance of recognition remains sensitive to the form of the provided content. Based on LLMs, some works in the field of multi-modal comprehension solve this problem by designing or generating discrete text prompts using answers [29], reasoning questions [53], and structure-driven contexts [70] instead of vanilla task-specific templates. However, a paradigm for automated prompt learning is urgently needed. Recent methods [30], [34], [37] learn continuous contexts to automate prompt engineering and explore optimal prompts, called prompt tuning. This paradigm can also be applied to vision-language models [23], [47]. Specifically, CoOp [73] demonstrates that a suitable prompt for improving the recognition performance of CLIP can be learned with very few samples. CoCoOp [62] extends CoOp by learning an input-conditional token for each image to obtain generalizable representations. ProDA [38] captures the distribution of diverse prompts to handle the varying visual representations and provides high-quality task-related content for facilitating recognition. ProGrad [75] aligns the gradient to the general direction with other parameters frozen, which prevents prompt tuning from forgetting the general knowledge learned from VLMs.

While the existing approaches primarily focus on learning prompts for text modality, they overlook the optimization of prompts for vision modality. To address this gap, Visual Prompt Tuning (VPT) [24] achieves remarkable performance gains with only a minimal set of trainable vectors acting as prompts, while keeping the model backbone frozen. Drawing from the previously mentioned approaches, MaPLe [26] introduces a method for multi-modal prompt learning to improve the alignment between representations from vision and text modalities. FG-VPL [58] proposes fine-grained visual prompt learning to induce VLMs to focus on the target object and capture discriminative visual information. In contrast, based on a dual-modality prompt tuning network with asymmetric regularization and domain-split constraint, our method learns the domain invariant prompt for both modalities with vision-language models in an end-to-end manner, resulting in better generalization on image classification.

### B. Domain Generalization

Domain generalization refers to learning a robust model generalized to unseen domains. In this paper, the generalization ability of a model is evaluated from the perspectives of both out-of-domain images and classes, corresponding to conventional domain generalization and base-to-new generalization respectively. Conventional domain generalization mainly evaluates the generalization capacity on unseen image domains. Many approaches [2], [18], [33], [41] have attempted to measure the domain gap between images and learn domain invariant features. In order to acquire a set of parameters capable of generalizing to unseen domains, several methods [4], [31] employ meta-learning to simulate domain shift during training. In this paper, we present a theoretical analysis within the context of episodic training, focusing on the guarantee of generalization in the domain generalization scenario.

Recently, another type of generalization task called base-to-new generalization has emerged, aiming to exploit the generalization ability on unseen classes [7], [63], [66], [67]. Conventional methods [17], [22], [25], [64] learn a semantic space based on auxiliary information. Compared with supervised learning, CLIP-based methods achieve high generalization performance due to more vital transferring ability. CoCoOp [62] tackles this generalization problem with conditional prompt learning. Our study explores the viability of learning the domain invariant prompt for the pre-trained V-L model CLIP [47] and introduces the novel concept of conducting episodic training in an alternate way for the first time.

### C. Meta-Learning

Most existing meta-learning approaches focus on few-shot learning, which can be divided into metric learning methods, memory network methods, and optimization-based methods. Metric learning methods [51], [55], [59], [61] learn a similarity space to extract discriminative meta-features for new classes efficiently. Memory network methods [40], [42], [44], [52] store meta-knowledge by memory models when learning seen tasks and then generalize it to unseen tasks. Optimization-based methods [14], [15], [49], [50] train meta-optimizer that enable fast adaption for new tasks. Works like MAML [1], [14], [16], [74] focus on learning meta-initial parameters of a deep model so that it would perform well on new tasks after only a small number of gradient updates. Drawing on recent advancements, we optimize parameters after every in-domain update to learn robust representations instead of learning the initial parameters of the model. In concrete, we utilize gradients on meta-test subtasks to regularize parameters, i.e., prompts. By imposing a modality-specific constraint, our model performs better on various generalization tasks.

## III. Generalization Bound of Episodic Training

Following previous literature [65], our theoretical analysis is based on the assumption that data is composed of attributes (i.e., factors of variation), such as color, shape, texture, etc., and different domains can be defined by different distributions of attributes. For example, as shown in Fig. 2, a sketch domain corresponds to a color distribution with only two values, black and white. In contrast, a cartoon or natural image domain may correspond to a color distribution with more color values. As a result, we assume that there exists a unified meta-domain distribution $\tau$ containing all possible attributes, where data domains $\mathcal{P} = \{\mathcal{P}_i\}_{i=1}^{N}$ are distributions sampled from this meta-domain with different attribute distributions. Under this assumption, we expect a training strategy to learn invariant features from seen domains and be able to generalize to unseen domains. Specifically, given a training algorithm $\mathbf{F}$ trained on a dataset $\mathbf{D} = \left\{D_i = D_i^s\right\}_{i=1}^{N}$, where $D_i^s$ is the

set of data sampled from a support domain, drawn from a domain distribution $\mathcal{P}_i^M$ containing $M$ training samples (i.e., $D_i^s \overset{i.i.d.}{\sim} \mathcal{P}_i^M$), the generalization error $\mathcal{R}$ obtained by $\mathbf{F}(\mathbf{D})$ is as follows:

$$\mathcal{R}(\mathbf{F}(\mathbf{D}), \tau) = \mathbb{E}_{\mathcal{P}\sim\tau, D^s\sim\mathcal{P}^M, z\sim\mathcal{P}} L(\mathbf{F}(\mathbf{D})(D^s), z). \quad (1)$$

Here $z$ represents an instance sampled from the distribution of data domains $\mathcal{P}$.

To improve the generalization ability of meta-learning algorithms, the pioneering work [61] proposes a training strategy – episodic training strategy, which treats each task as a training instance and updates the inner-task algorithm by episode (task by task). In this paper, we transfer episodic training to the domain generalization scenario by treating each data domain as a training instance and updating the inner-domain algorithm by episode (domain by domain). Specifically, we first update the model on a support domain (i.e., in-domain error). Then the performance of the updated model is measured and optimized on another query domain (i.e., out-of-domain error or episodic training error). As a result, the training error of the episodic training strategy $\hat{\mathcal{R}}_{epi}$ is as follows:

$$\hat{\mathcal{R}}_{epi}(\mathbf{F}(\mathbf{D}), \mathbf{D}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_i^q}\sum_{z_i\in D_i^q}\hat{L}(\mathbf{F}(D_i^s), z_i), \quad (2)$$

where $D_i^q$ is the set of data sampled from a query domain, and $N_i^q$ is the sample number of $D_i^q$. From Eq. 2 we can see that episodic training strategy directly minimizes the out-of-domain testing error, and hence intuitively the in-domain sample number $M$ in the generalization bound vanishes, with the generalization bound only depending on the domain number $N$.

Based on this paradigm, we naturally associate episodic training with domain generalization tasks, aiming to learn invariance from various distributions by creating meta-tasks with domain gaps as episodes. By applying this strategy, the distribution shift between the meta-train and meta-test subtask can be approximately equivalent to that between the original training and test task. The error of the parameter over the meta-test task is exactly the test error of generalization tasks and thereby is an unbiased estimate of the generalization error on unseen domains. Theoretically, following [8], we derive the bound of the generalization gap between these two errors only depending on the domain number $N$, which is formulated by:

$$\mathbb{E}_{\mathbf{F}}[\mathcal{R}(\mathbf{F}(\mathbf{D}), \tau)] \leq \mathbb{E}_{\mathbf{F}}\left[\hat{\mathcal{R}}_{epi}(\mathbf{F}(\mathbf{D}), \mathbf{D})\right] + O\left(\frac{1}{\sqrt{N}}\right). \quad (3)$$

The generalization bound implies a strong generalization guarantee for episodic training algorithms in the few-shot regime, which motivates this paper to adopt episodic training to learn the domain invariant prompt with very few samples.

## IV. Alternate Episodic Training Algorithm

In order to enhance generalization performance on out-of-domain data, we propose an alternate episodic training algorithm. To enhance the performance of this algorithm, we introduce a dual-modality prompt tuning network as the foundation of our approach. As a batch-wise algorithm, our approach conducts an in-domain update with an asymmetric contrastive learning paradigm following a domain-split update with a modality-specific optimization strategy on each batch.

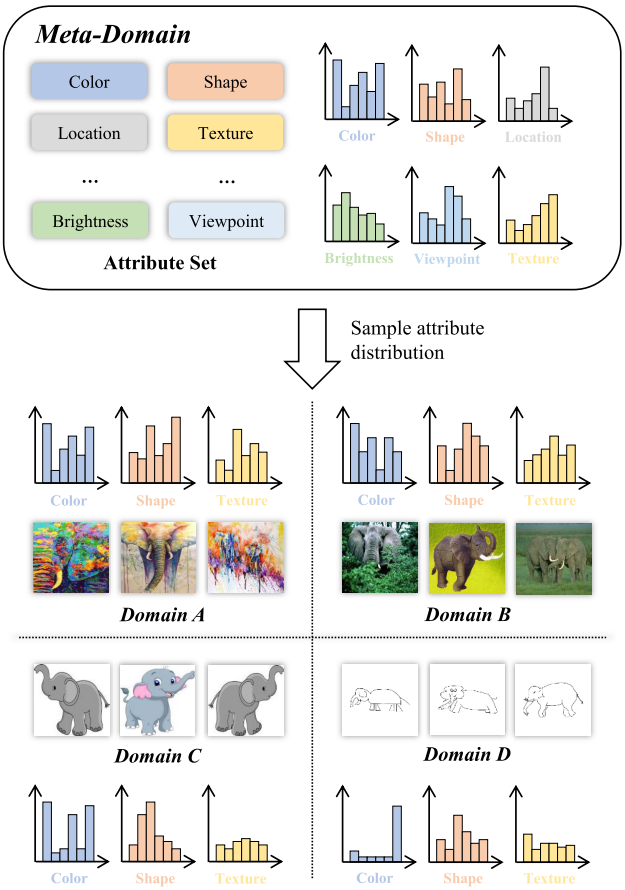



Fig. 2. Input samples are composed of attributes (i.e., factors of variation), such as color, shape, texture, etc., and different domains can be defined by different distributions of attributes.

### A. Dual-Modality Prompt Tuning Network

To enhance the effectiveness of episodic training in prompt tuning and to establish a network that sustains high performance across unseen domains, we demonstrate our framework for prompt tuning on vision-language foundation models, such as CLIP. Among recent works on prompt tuning, prompt vectors can be learned for both text encoder [62], [73] and image encoder [24]. In this section, we first formulate prompt tuning for text and vision modalities as follows:

*1) Textual Prompt Tuning:* We follow CoOp [73] that automatically learns a set of tunable continuous vectors as context tokens that are fed into the text encoder together with the class tokens. Instead of introducing prompts only at the first layer, we expand these vectors at every Transformer layer's input space. Given the textual prompt composed of $P$ vectors for the $i$-th class denoted as $\boldsymbol{t}_i$, the prediction

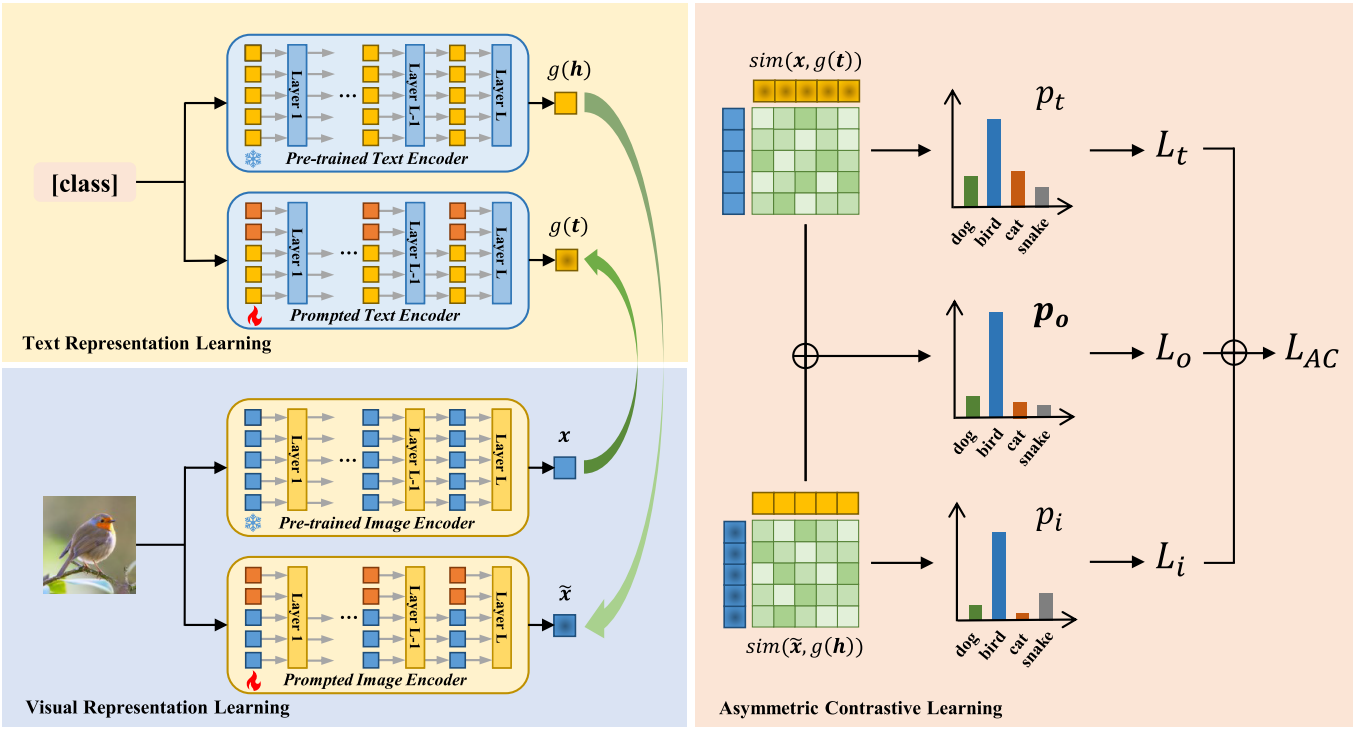



Fig. 3. Our dual-modality prompt tuning network consists of a pre-trained encoder and a prompted encoder for each modality, where we further couple the prompted encoder with the pre-trained encoder from the relative modality. The asymmetric contrastive learning module outputs three probability distributions for the end-to-end training to achieve better recognition performance, where $p_o$ is used for the final prediction.

probability of the $i$-th class can be calculated by:

$$p_t(y=i \mid \boldsymbol{x}) = \frac{\exp\left(\text{sim}\left(\boldsymbol{x}, g\left(\boldsymbol{t}_i\right)\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}\left(\boldsymbol{x}, g\left(\boldsymbol{t}_j\right)\right)/\tau\right)}, \tag{4}$$

where $\boldsymbol{x}$ represents the image representation from the image encoder and $g(\cdot)$ denotes the text encoder.

*2) Visual Prompt Tuning:* We follow VPT-Deep [24] that adopts a similar idea as textual prompt, where extra prompt vectors are automatically learned to be fed into the image encoder. The image patches are firstly embedded into a latent space as the input of the first Transformer layer, and then $P$ learnable vectors are introduced at every Transformer layer's input space as prompts. The output of the Transformer head is considered the final image representation $\widetilde{\boldsymbol{x}}$. The prediction probability of the $i$-th class can be calculated by:

$$p_i(y=i \mid \boldsymbol{x}) = \frac{\exp\left(\text{sim}\left(\widetilde{\boldsymbol{x}}, g\left(\boldsymbol{h}_i\right)\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}\left(\widetilde{\boldsymbol{x}}, g\left(\boldsymbol{h}_j\right)\right)/\tau\right)}, \tag{5}$$

where $g(\cdot)$ denotes the text encoder and $\boldsymbol{h}_i$ denotes the handcrafted prompt for the $i$-th class.

Motivated by previous works on textual and visual prompt tuning, we propose a dual-modality prompt tuning network that jointly learns visual and textual prompts for better recognition performance with in-domain data. As shown in Fig. 3, unlike methods [26], [69] that learn two sets of prompt vectors on a single pair of encoders with cross-entropy loss, we couple each prompted encoder with a pre-trained encoder from the relative modality. By leveraging representations from pre-trained encoders as regularization, the generalization ability of learned prompts can be promised, thereby mitigating the overfitting issue on in-domain data. More details of the implementation will be discussed in the following section.

### B. Asymmetric Contrastive Learning for In-Domain Updates

To achieve good performance on in-domain training samples while preventing the learned prompt vectors from the overfitting issue (especially in a few-shot setting), we propose a novel asymmetric contrastive learning paradigm for in-domain updates. This paradigm employs representations from the pre-trained encoder, renowned for its robust transferability, to serve as guidance for enhancing the generalization ability of prompts in the prompted encoder. Specifically, instead of concurrently training prompted encoders from both modalities in a single pair using cross-entropy loss, we opt for independent training, where prompted representations of one modality are aligned with pre-trained ones of another modality, as shown in Fig. 3.

With this asymmetric contrastive learning paradigm, we have two probabilities $p_t$ and $p_i$, corresponding to textual and visual prompts with Eq. 4 and Eq. 5. We average them to obtain an overall probability $p_o$. In the training phase, we employ the cross-entropy loss to minimize the distance between the ground-truth label $y$ and three probabilities $p_t$, $p_i$ and $p_o$. We denote the losses associated with these probabilities as $\mathcal{L}_t$, $\mathcal{L}_i$ and $\mathcal{L}_o$, respectively. As a result, the final asymmetric contrastive loss function $\mathcal{L}_{AC}$ can be expressed as the sum of three losses:

$$\mathcal{L}_{AC} = \mathcal{L}_o + \mathcal{L}_t + \mathcal{L}_i. \tag{6}$$

During inference, the probability $p_o$ is used for prediction.

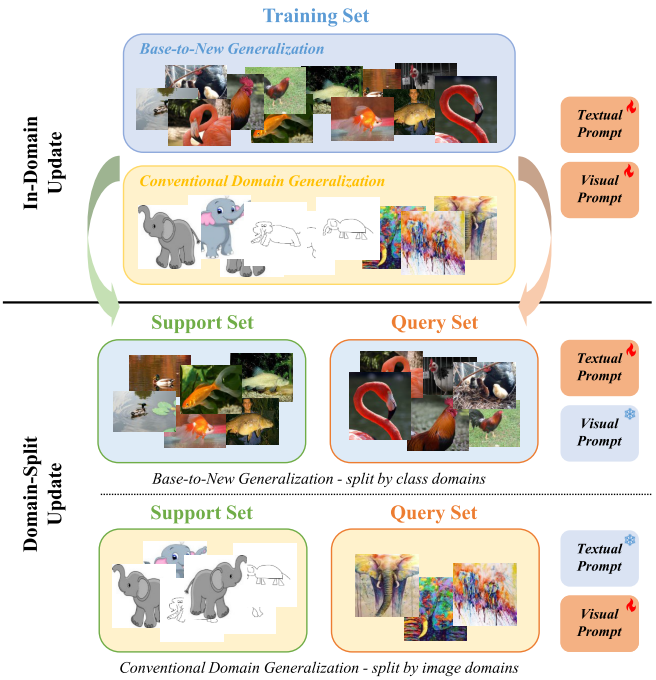



Fig. 4. Given a batch of training data containing samples from different domains, we conduct an in-domain update and a domain-split update. For in-domain updates, we use all samples of the batch for training. For domain-split updates, we split the samples into a support set and a query set by their class domains or image domains based on corresponding tasks.

### C. Domain-Split Optimization for Domain-Split Updates

Motivated by the analysis from Section III, we propose a domain-split optimization strategy for prompt tuning. Based on the generalization bound derived from Eq. 3, the generalization gap only depends on the domain number $N$, which indicates the feasibility of conducting meta-updates by splitting samples according to their domains. Compared with in-domain updates which focus on learning robust representations with asymmetric regularization using full in-domain samples, domain-split updates explicitly enhance the generalization capacity on the out-of-domain distribution based on meta-learning. The performance of this update is only related to the characteristics of the dataset itself instead of the amount of training samples. Given a batch of training data containing samples from various domains generated from the meta-domain, we split it into a support set and a query set based on domains. Our domain-split optimization aims to regularize learnable prompts with a constraint that narrows the gap between training errors on the support and query set.

Specifically, given a batch of $N$ datasets sampled from $N$ domains at the $t$-th time step denoted as $\mathcal{D}_t = \{D_i\}_{i=1}^{N}$, where $D_i \sim \mathcal{P}_i$ and $\mathcal{P}_i$ indicates the distribution of the $i$-th data domains, we split the set by grouping samples from some selected domains as the query set $D_j^q$, and samples from the rest as the support set $D_j^s$, where the index $j$ denotes the $j$-th split. Note that, for domain generalization, since it is clear which domain each sample belongs to, the query and support set can be easily split. However, for base-to-new generalization, there is no explicit definition of which domain each sample belongs to. Hence, we randomly split the query and support set based on the class label of each sample. By imposing various separations, we provide a unified episodic generation paradigm for different generalization tasks, as shown in Fig. 4.

Based on our dual-modality prompt tuning network, we propose a modality-specific optimization strategy, where the prompts of only the task-specific modality are tuned during domain-split updates. For example, when conducting base-to-new generalization on Flowers102, differentiating between the semantics of flower names such as “pink primrose” and “hard-leaved pocket orchid” becomes crucial. This underscores the necessity of tuning invariant textual prompts to accommodate diverse classes within the topic of flowers. On the other hand, in the context of domain generalization tasks, utilizing invariant visual prompts to extract common semantics across diverse domains enhances recognition performance. Based on the aforementioned, we apply constraints on visual prompts concerning cross-domain tasks with $\mathcal{L}_i$ and on textual prompts concerning cross-class tasks with $\mathcal{L}_t$.

During domain-split updates, the learnable prompt $\theta$ from the task-specific modality is updated with the samples on the support set $D_j^s$ to get the updated prompt $\theta_j'$. Then the generalization error of the updated prompt $\theta_j'$ is measured by the cross-entropy loss on the query set $D_j^q$, whose corresponding gradients are back-propagated to update the original prompt $\theta$. Since this update involves second-order gradient computation with high complexity, in our implementation, we design a first-order approximating method. The parameter $\theta$ is updated as follows:

$$\theta \leftarrow \theta - \alpha\eta \sum_j \nabla_{\theta_j'} \mathcal{L}(\theta_j'; D_j^q), \tag{7}$$

where $\alpha$ is the meta-step rate, and $\eta$ is the learning rate of the normal training. $\mathcal{L}_{Meta}$ indicates the meta-test loss on the query set for calculating gradients, which is associated with the aforementioned modality-specific loss function. To simplify the training process, our paradigm treats one batch-wise iteration in Eq. 7 as a series of training episodes and conducts several splits of the query and support set within each batch iteration. The detailed implementation of the alternate episodic training algorithm is shown in Alg. 1.

### D. Computational Complexity Analysis

To analyze the computational complexity of our batch-wise episodic training, we consider the number of operations required for both in-domain and domain-split updates. We denote the batch size of images and the number of class names as $N_i$ and $N_t$. $N_j$ indicates the domain number, where $N_j$ equals to 2 especially for base-to-new generalization. When we conduct independent V-L prompts within one pair of prompted encoders, the complexity of a batch step is $O(N_i + N_t)$.

During in-domain updates, we feed all samples in a batch into our dual-modality prompt tuning network. Considering that the pre-trained encoder is frozen during training, we pre-cache the representations for alignment with those of the prompted encoder. Therefore, the overall complexity of in-domain updates is $O(N_i + N_t)$. During domain-split updates,

**Algorithm 1** Batch-Wise Episodic Training

**Require:** Domain number $N$, split number $N_j$, learning rate $\eta$, meta-step rate $\alpha$, dataset $\mathcal{D}$, loss functions $\mathcal{L}_{AC}$, $\mathcal{L}_t$, $\mathcal{L}_i$

**Ensure:** Prompt parameters $\Theta$

Randomly initialize $\Theta_0 = \{\theta^I, \theta^T\}$
**for** $t$ in iterations **do**
  Randomly sample a batch $\mathcal{D}_t = \{D_i\}_{i=1}^{N}$ from $\mathcal{D}$
  **In-Domain Update:**
  Update $\Theta_t$ w.r.t. $\mathcal{L}_{AC}$:
    $\Theta_t \leftarrow \Theta_t - \eta \nabla_{\Theta_t} \mathcal{L}_{AC}\left(\Theta_t; \mathcal{D}_t\right)$
  **Domain-Split Update:**
  **if** base-to-new generalization **then**
    $\theta \leftarrow \theta_t^T$, $\mathcal{L}_{Meta} \leftarrow \mathcal{L}_t$
    $\{(D_j^s, D_j^q)\}_{j=1}^{N_j} \leftarrow group\_by\_class(\mathcal{D}_t)$
  **else if** conventional domain generalization **then**
    $\theta \leftarrow \theta_t^I$, $\mathcal{L}_{Meta} \leftarrow \mathcal{L}_i$
    $\{(D_j^s, D_j^q)\}_{j=1}^{N_j} \leftarrow group\_by\_domain(\mathcal{D}_t)$
  **end if**
  **for** $j = 1$ to $N_j$ **do**
    $\theta_j' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{Meta}\left(\theta; D_j^s\right)$
    $g_j \leftarrow \nabla_{\theta_j'} \mathcal{L}_{Meta}\left(\theta_j'; D_j^q\right)$
  **end for**
  Update $\theta$ with gradients:
    $\theta \leftarrow \theta - \alpha\eta \sum_{j=1}^{N_j} g_j$
  **if** base-to-new generalization **then**
    $\Theta_{t+1} \leftarrow \{\theta_t^I, \theta\}$
  **else if** conventional domain generalization **then**
    $\Theta_{t+1} \leftarrow \{\theta, \theta_t^T\}$
  **end if**
**end for**

we split samples in the batch into a support set and a query set based on domains. As we feed them into encoders one after the other and compute gradients with first-order approximating, the complexity of one split is also $O(N_i + N_t)$. Since our domain-split optimization contains $N_j$ episodes for calculating the meta-gradients, the overall complexity of domain-split updates is $O(N_j(N_i + N_t))$.

In summary, considering the overall complexity of in-domain and domain-split updates is $O((N_j + 1)(N_i + N_t))$. Despite the increase in computation compared to the baseline, our algorithm does not introduce additional sub-networks to increase the computational burden. Furthermore, since we perform multiple computations for the same batch in an epoch, the overall number of epochs can be fewer compared to other methods, thus compensating for the additional computational consumption brought about by episodic learning.

## V. Experiments

We evaluate our approach mainly in the two generalization settings, i.e. base-to-new generalization and conventional domain generalization. In our experiments, we use the open-source CLIP [47] as the foundation vision-language model. Here we elaborate on the experimental configurations.

*a) Datasets:* For base-to-new generalization, we follow Zhou et al. [73] and evaluate the performance of our method using 11 image recognition datasets, which cover a wide range of recognition tasks. Specifically, the benchmark includes ImageNet [10] and Caltech101 [13] for classification on generic objects; OxfordPets [45], StanfordCars [28], Flowers102 [43], Food101 [5] and FGVCAircraft [39] for fine-grained classification; SUN397 [68] for scene recognition; UCF101 [56] for action recognition; DTD [9] for texture classification; and finally EuroSAT [20] for satellite imagery recognition. For each dataset, we split the classes equally into two groups as base and new classes. We train the model only on base classes in a few-shot setting, while evaluation is conducted independently on base and new classes.

For conventional domain generalization experiments, we select four real-world datasets from the DomainBed benchmark, including VLCS [12], PACS [32], OfficeHome [60], DomainNet [46]. We conduct experiments with the leave-one-out strategy, where one of the domains is selected as the target domain at a time, and other domains are used as the source domains. We train the model on the source domains in few-shot, while evaluation is conducted on the target domain.

*b) Implementation Details:* Our implementation is based on dassl [72], a well-designed PyTorch toolbox for domain generalization. We apply prompt tuning on the pre-trained CLIP model with ViT-B/16 as the visual backbone. Both prompts are randomly initialized from the Gaussian distribution with a mean of 0 and a standard deviation of 0.02. We adopt SGD optimization with an initial learning rate of 0.0015, decayed by the cosine annealing rule, and the meta-step rate $\alpha$ is set to 0.2. The warming-up trick is adopted during the first epoch with a fixed learning rate of $10^{-5}$.

For base-to-new generalization, the maximum epoch is set to 8 for all datasets with a batch size of 16. The prompt length $P$ of visual and textual prompts is set to 2. We set the split number $N_j$ to 2 for domain-split optimization, where we evenly divide the samples of a batch into two groups based on their classes during every split. Following Zhou et al. [73], we use the few-shot evaluation protocol that selects 16 shots for training and leverages the whole test set for evaluation.

For conventional domain generalization, the maximum epoch is set to 6 for all datasets with a batch size of 32. The prompt length $P$ of visual and textual prompts is set to 4. We set the split number $N_j$ the same as the domain number $N$. The leave-one-out strategy is adopted, wherein samples from one domain are grouped as the query set at a time, while samples from other domains are grouped as the support set. We adopt 1-shot and 5-shot settings for each source domain during training and evaluate our model on all samples of the target domain. For the hyper-parameter selection of our implementation, we share the same hyper-parameters instead of searching for each dataset.

### A. Base-to-New Generalization

The performance of our MetaPrompt in base-to-new generalization setting on 11 image recognition datasets is shown in Table I. We compare its performance with zero-shot CLIP using hand-crafted prompts as the input, and recent

TABLE I

COMPARISON OF CLIP, COOP, COCOOP, MAPLE, AND OUR METAPROMPT ON BASE-TO-NEW GENERALIZATION BENCHMARKS. OUR EXPERIMENTS ARE REPEATED THREE TIMES USING DIFFERENT RANDOM SEEDS. METAPROMPT OUTPERFORMS ALL OTHER METHODS ON BOTH BASE AND NEW CLASSES AND DEMONSTRATES STRONG GENERALIZATION PERFORMANCE ON 11 IMAGE RECOGNITION DATASETS. H: HARMONIC MEAN (TO HIGHLIGHT THE GENERALIZATION TRADE-OFF)

**(a) Average over 11 datasets.**

| | Base | New | H |
|---|---|---|---|
| CLIP | 69.34 | 74.22 | 71.70 |
| CoOp | 82.69 | 63.22 | 71.66 |
| CoCoOp | 80.47 | 71.69 | 75.83 |
| MaPLe | 82.28 | 75.14 | 78.55 |
| MetaPrompt | **83.38** | **76.09** | **79.57** |
| vs. MaPLe | +1.10 | +0.95 | +1.02 |

(b) ImageNet.

| | Base | New | H |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | 76.47 | 67.88 | 71.92 |
| CoCoOp | 75.98 | 70.43 | 73.10 |
| MaPLe | 76.66 | 70.54 | 73.47 |
| MetaPrompt | **77.39** | **71.06** | **74.09** |
| vs. MaPLe | +0.73 | +0.52 | +0.62 |

(c) Caltech101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 96.84 | 94.00 | 95.40 |
| CoOp | 98.00 | 89.81 | 93.73 |
| CoCoOp | 97.96 | 93.81 | 95.84 |
| MaPLe | 97.74 | 94.36 | 96.02 |
| MetaPrompt | **98.28** | **94.58** | **96.39** |
| vs. MaPLe | +0.54 | +0.22 | +0.37 |

(d) OxfordPets.

| | Base | New | H |
|---|---|---|---|
| CLIP | 91.17 | 97.26 | 94.12 |
| CoOp | 93.67 | 95.29 | 94.47 |
| CoCoOp | 95.20 | 97.69 | 96.43 |
| MaPLe | 95.43 | **97.76** | **96.58** |
| MetaPrompt | **95.71** | 96.98 | 96.34 |
| vs. MaPLe | +0.28 | -0.78 | -0.24 |

(e) StanfordCars.

| | Base | New | H |
|---|---|---|---|
| CLIP | 63.37 | **74.89** | 68.65 |
| CoOp | **78.12** | 60.40 | 68.13 |
| CoCoOp | 70.49 | 73.59 | 72.01 |
| MaPLe | 72.94 | 74.00 | 73.47 |
| MetaPrompt | 75.43 | 74.43 | **74.93** |
| vs. MaPLe | +2.49 | +0.43 | +1.46 |

(f) Flowers102.

| | Base | New | H |
|---|---|---|---|
| CLIP | 72.08 | **77.80** | 74.83 |
| CoOp | **97.60** | 59.67 | 74.06 |
| CoCoOp | 94.87 | 71.75 | 81.71 |
| MaPLe | 95.92 | 72.46 | 82.56 |
| MetaPrompt | 97.53 | 74.54 | **84.50** |
| vs. MaPLe | +1.61 | +2.08 | +1.94 |

(g) Food101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 90.10 | 91.22 | 90.66 |
| CoOp | 88.33 | 82.26 | 85.19 |
| CoCoOp | 90.70 | 91.29 | 90.99 |
| MaPLe | 90.71 | **92.05** | **91.38** |
| MetaPrompt | **90.76** | 91.77 | 91.26 |
| vs. MaPLe | +0.05 | -0.28 | -0.12 |

(h) FGVCAircraft.

| | Base | New | H |
|---|---|---|---|
| CLIP | 27.19 | 36.29 | 31.09 |
| CoOp | **40.44** | 22.30 | 28.75 |
| CoCoOp | 33.41 | 23.71 | 27.74 |
| MaPLe | 37.44 | 35.61 | 36.50 |
| MetaPrompt | 39.38 | **37.59** | **38.46** |
| vs. MaPLe | +1.94 | +1.98 | +1.96 |

(i) SUN397.

| | Base | New | H |
|---|---|---|---|
| CLIP | 69.36 | 75.35 | 72.23 |
| CoOp | 80.60 | 65.89 | 72.51 |
| CoCoOp | 79.74 | 76.86 | 78.27 |
| MaPLe | 80.82 | 78.70 | 79.75 |
| MetaPrompt | **82.10** | **79.01** | **80.53** |
| vs. MaPLe | +1.28 | +0.31 | +0.78 |

(j) DTD.

| | Base | New | H |
|---|---|---|---|
| CLIP | 53.24 | 59.90 | 56.37 |
| CoOp | 79.44 | 41.18 | 54.24 |
| CoCoOp | 77.01 | 56.00 | 64.85 |
| MaPLe | 80.36 | 59.18 | 68.16 |
| MetaPrompt | **82.52** | **60.10** | **69.55** |
| vs. MaPLe | +2.16 | +0.92 | +1.39 |

(k) EuroSAT.

| | Base | New | H |
|---|---|---|---|
| CLIP | 56.48 | 64.05 | 60.03 |
| CoOp | 92.19 | 54.74 | 68.69 |
| CoCoOp | 87.49 | 60.04 | 71.21 |
| MaPLe | **94.07** | 73.23 | 82.35 |
| MetaPrompt | 93.37 | **78.34** | **85.20** |
| vs. MaPLe | -0.70 | +5.11 | +2.85 |

(l) UCF101.

| | Base | New | H |
|---|---|---|---|
| CLIP | 70.53 | 77.50 | 73.85 |
| CoOp | 84.69 | 56.05 | 67.46 |
| CoCoOp | 82.33 | 73.45 | 77.64 |
| MaPLe | 83.00 | **78.66** | 80.77 |
| MetaPrompt | **84.70** | 78.56 | **81.51** |
| vs. MaPLe | +1.70 | -0.10 | +0.74 |

prompt learning methods, including CoOp, CoCoOp, and MaPLe.

*1) Generalization to Unseen Classes:* In comparison with the state-of-the-art prompt tuning method MaPLe, MetaPrompt obtains an overall improvement to 76.09% in terms of the average accuracy of new classes over 11 datasets with our episodic training strategy that explicitly constrains the prompt to generalize to out-of-domain classes. When considering both base and new classes, MetaPrompt shows an absolute average gain of 1.02% on the harmonic mean over MaPLe. The results strongly prove that our method of learning the domain invariant prompt improves the generalization ability.

*2) Performance Gain in Seen Classes:* While our approach achieves excellent performance on generalizing to unseen classes, it still maintains high accuracy on seen classes compared with other methods optimized to fit in-domain data, even better than MaPLe by 1.10%. While the performance on EuroSAT is inferior to MaPLe on seen classes, the substantial improvement exceeding 5% on unseen classes implies that our approach exhibits remarkable generalization capabilities.

*3) Explanation of our Better Trade-Off:* MetaPrompt achieves a good trade-off between in-domain and out-of-domain data for two reasons. Firstly, our multi-modal prompts improve the recognition accuracy from two modalities concurrently and independently. With in-domain updates where the pre-trained vision-language model assumes supervision, we obtain a stable boost in fitting both in-domain and out-of-domain data. Secondly, from the perspective of training strategies, MaPLe does not explicitly consider the in-domain and out-domain trade-off and achieving good generalization at the expense of lower in-domain accuracy, while our approach proposes an explicit constraint during domain-split updates to optimize prompts for both seen and unseen classes.

TABLE II

COMPARISON OF DOMAIN GENERALIZATION METHODS AND OUR METAPROMPT ON DOMAIN GENERALIZATION BENCHMARKS. CLIP (TEMPLATE) INDICATES USING 'A PHOTO OF A {CLASS NAME}' PROMPT. 'ENSEMBLE' AND 'CLIP' INDICATE ENSEMBLE AND CLIP-BASED METHODS. OUR EXPERIMENTS ARE REPEATED THREE TIMES USING DIFFERENT RANDOM SEEDS. ALTHOUGH OUR METHOD IS BASED ON *few-shot* SETTING, IT ACHIEVES COMPETITIVE RESULTS AGAINST FULL-TRAINING METHODS AND DEMONSTRATES STRONG PERFORMANCE ON DOMAIN GENERALIZATION BENCHMARKS

| Method | Setting | | Category | | Accuracy(%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Zero-shot | Few-shot | Ensemble | CLIP | PACS | VLCS | OfficeHome | DomainNet |
| ERM [19] | | | | | 84.2 ± 0.1 | 77.3 ± 0.1 | 67.6 ± 0.2 | 44.0 ± 0.1 |
| MLDG [31] | | | | | 84.8 ± 0.6 | 77.1 ± 0.4 | 68.2 ± 0.1 | 41.8 ± 0.4 |
| Fish [54] | | | | | 85.5 ± 0.3 | 77.8 ± 0.3 | 68.6 ± 0.4 | 42.7 ± 0.2 |
| CORAL [57] | | | | | 86.2 ± 0.3 | 78.8 ± 0.6 | 68.7 ± 0.3 | 41.5 ± 0.1 |
| SWAD [6] | | | ✓ | | 88.1 ± 0.1 | 79.1 ± 0.1 | 70.6 ± 0.2 | 46.5 ± 0.1 |
| EoA [3] | | | ✓ | | 95.8 ± 0.0 | 81.1 ± 0.0 | 83.9 ± 0.0 | 60.9 ± 0.0 |
| SEDGE [35] | | | ✓ | | 96.1 ± 0.0 | 82.2 ± 0.0 | 80.7 ± 0.2 | 54.7 ± 0.1 |
| CLIP [47] | ✓ | | | ✓ | 95.7 ± 0.0 | 75.9 ± 0.0 | 79.4 ± 0.0 | 56.8 ± 0.0 |
| CLIP (template) | ✓ | | | ✓ | 96.1 ± 0.0 | 82.3 ± 0.0 | 82.1 ± 0.0 | 56.9 ± 0.0 |
| CoCoOp [62] (5-shot) | | ✓ | | ✓ | 96.7 ± 0.4 | 78.3 ± 1.0 | 84.1 ± 0.1 | 61.1 ± 0.2 |
| MetaPrompt (1-shot) | | ✓ | | ✓ | 96.9 ± 0.5 | 81.1 ± 0.3 | 84.1 ± 0.3 | 61.2 ± 0.3 |
| MetaPrompt (5-shot) | | ✓ | | ✓ | **97.0 ± 0.2** | **82.6 ± 0.6** | **85.2 ± 0.3** | **61.8 ± 0.2** |

*4) Failure in Some Datasets:* Nevertheless, it is still noteworthy that in some datasets, there exists a gap compared to previous methods on base or new classes. In OxfordPets, with fewer classes than most datasets, the effectiveness of the domain-split optimization is slightly limited due to the poor diversity of categories. In Food101, the good performance of zero-shot CLIP indicates the small difference between distributions of this dataset and pre-trained data, thus leading to a potential risk of overfitting during training. For other datasets like EuroSAT and UCF101, the performance trade-off on base and new classes should be better balanced.

## B. Conventional Domain Generalization

The performance of our MetaPrompt in conventional domain generalization setting on four benchmarks is shown in Table II. We compare its performance with different categories of domain generalization methods, including the non-ensemble methods like ERM [19], MLDG [31], Fish [54], CORAL [57], the ensemble methods like SWAD [6], EoA [3], SEDGE [35], as well as zero-shot CLIP and CoCoOp in domain generalization setting. Since extracting domain invariant features is the mainstream idea in traditional domain generalization tasks, we follow this idea for CLIP-based learning to train the domain invariant prompt.

In comparison with traditional domain generalization methods, CLIP-based methods demonstrate outstanding generalization capabilities, attributed to the strong transfer learning ability acquired from pre-trained knowledge. Despite using a limited number of training samples, our MetaPrompt yields competitive results in domain generalization benchmarks. It outperforms alternative methods, including the conditional prompt tuning approach CoCoOp, across all datasets when considering average accuracy in the 5-shot setting. Moreover, it achieves comparable performance even in the 1-shot setting. By simulating the generalization error between different domains with domain-split optimization, our domain invariant prompt has a stronger generalization capacity than a conditional-based prompt generator training independently with domains.

## C. Further Analysis

*1) Influence of Model Components:* We analyze the influence of components in our model and conduct an ablation study on various combinations of them, as shown in Table III. The baseline method (the first row) simultaneously trains both textual and visual prompts with a conventional gradient descent optimizer. The results show that our alternate episodic training algorithm with both in-domain updates and domain-split updates positively affects generalization to unseen domains. Among them, in-domain updates achieve an absolute performance gain on new class domains and an overall boost on new image domains, which shows the effectiveness of leveraging representations of pre-trained vision-language foundation models. Our domain-split updates with a novel optimization strategy also play an important role in boosting the ability of generalization, which will be analyzed in the subsequent section. In addition, our modality-specific optimization strategy during domain-split updates further improves performance on both tasks.

*2) Influence of Model Architectures:* We perform an ablation study on diverse model architectures and evaluate the efficacy of our proposed approach on domain generalization. We conduct experiments utilizing three commonly employed architectures for the visual encoder: ViT-B/32, ViT-B/16, and ViT-L/14. We evaluate the performance of both zero-shot CLIP and our method in 1-shot and 5-shot settings. As demonstrated in Table IV, under the 1-shot setting, our approach consistently outperforms zero-shot CLIP in terms of out-of-domain performance. The only exception is VLCS, where there is a slight lag attributable to the limitations imposed by sample size and significant domain shift. In the 5-shot setting, our method demonstrates notably greater progress, yielding a substantial 6.0% improvement on DomainNet when employing the ViT-L/14 as our model

TABLE III

ABLATION ON DIFFERENT COMPONENTS. 'ID-UPDATE' AND 'DS-UPDATE' DENOTE OUR IN-DOMAIN UPDATES AND DOMAIN-SPLIT UPDATES. 'MOS' INDICATES USING OUR MODALITY-SPECIFIC OPTIMIZATION STRATEGY INSTEAD OF REGULARIZING PROMPTS FOR BOTH MODALITIES IN BOTH TASKS DURING DOMAIN-SPLIT UPDATES. FOR DOMAIN GENERALIZATION, WE USE THE 5-SHOT ACCURACY AS THE EVALUATION METRIC

(a) Base-to-New Generalization.

| ID-Update | DS-Update | MOS | Base | New | H |
|---|---|---|---|---|---|
| | | | 82.58 | 72.81 | 77.39 |
| ✓ | | | 82.89 | 74.87 | 78.68 |
| | ✓ | | 82.68 | 75.16 | 78.74 |
| ✓ | ✓ | | 83.20 | 75.82 | 79.34 |
| ✓ | ✓ | ✓ | **83.38** | **76.09** | **79.57** |

(b) Domain Generalization.

| ID-Update | DS-Update | MOS | P | V | O | D |
|---|---|---|---|---|---|---|
| | | | 96.6 | 77.2 | 83.8 | 61.2 |
| ✓ | | | 96.6 | 79.6 | 84.7 | 61.5 |
| | ✓ | | 96.8 | 80.9 | 84.6 | 61.4 |
| ✓ | ✓ | | 96.9 | 82.1 | 84.9 | **61.8** |
| ✓ | ✓ | ✓ | **97.0** | **82.6** | **85.2** | **61.8** |

TABLE IV

ABLATION ON DIFFERENT MODEL ARCHITECTURES

| Backbone | Setting | P | V | O | D | *Avg.* |
|---|---|---|---|---|---|---|
| ViT-B/32 | CLIP | 94.7 | **80.1** | 78.2 | 53.5 | 76.6 |
| | 1-shot | 94.7 | 79.0 | 80.7 | 57.0 | 77.9 |
| | 5-shot | **95.0** | 79.8 | **81.9** | **57.6** | **78.6** |
| ViT-B/16 | CLIP | 96.1 | 82.3 | 80.7 | 56.9 | 79.0 |
| | 1-shot | 96.9 | 81.1 | 84.1 | 61.4 | 80.9 |
| | 5-shot | **97.0** | **82.6** | **85.2** | **61.8** | **81.7** |
| ViT-L/14 | CLIP | 98.4 | 81.9 | 85.7 | 61.2 | 81.8 |
| | 1-shot | 98.1 | 81.4 | 88.8 | 66.7 | 83.8 |
| | 5-shot | **98.7** | **82.2** | **89.5** | **67.2** | **84.4** |

architecture. This illustrates the efficacy of our method across various model architectures.

*3) Visualization of Image Embeddings:* We randomly select three datasets to analyze the t-SNE plots of image embedding, as shown in Fig. 5. Our MetaPrompt demonstrates superior inter-class separability and intra-class cohesiveness across both base and new classes. We attribute the strong performance of our method to the utilization of visual prompts, which are acquired under the guidance of pre-trained textual representations. Because these representations remain constant during the training process, embeddings with visual concepts can be more effectively aligned with their corresponding textual labels, thus tending to form distinct clusters On the other hand, the pre-trained CLIP model possesses a robust capability for semantic representation. Under the guidance of distinct textual semantics, image embeddings from various classes can be better separated.

*4) Influence of Prompt Length:* We conduct an ablation study on prompt length in both generalization settings. Specifically, we examine prompt vectors of 1, 2, 4, 8, 16, and 32 in each layer for both modalities, all initialized randomly, as summarized in Table V. For base-to-new generalization, it is evident that models with longer prompt lengths perform better

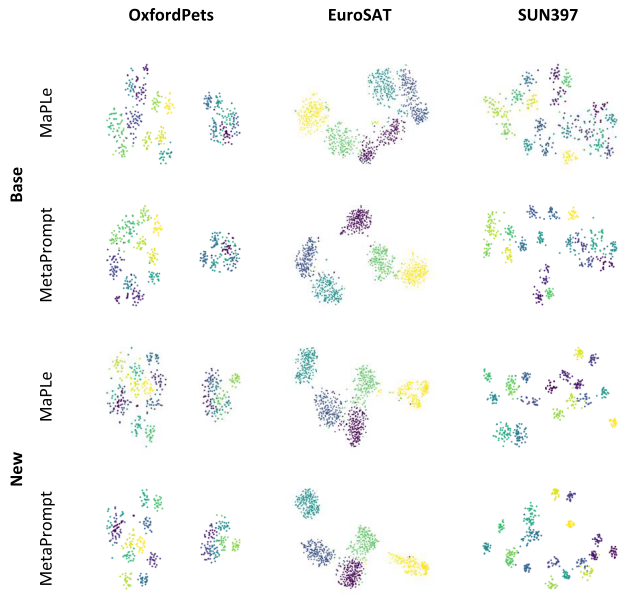



Fig. 5. T-SNE plots of image embeddings in previous methods MaPLe and our method MetaPrompt on diverse image recognition datasets. Points with the same color represent image embeddings of the same class.

TABLE V

ABLATION ON DIFFERENT PROMPT LENGTHS

(a) Base-to-New Generalization.

| Length | Base | New | H |
|---|---|---|---|
| 1 | 82.88 | 75.28 | 78.90 |
| 2 | 83.38 | **76.09** | **79.57** |
| 4 | 83.62 | 75.82 | 79.53 |
| 8 | 83.89 | 75.54 | 79.50 |
| 16 | **83.91** | 75.19 | 79.31 |
| 32 | 83.86 | 74.62 | 78.97 |

(b) Domain Generalization.

| Length | P | V | O | D | *Avg.* |
|---|---|---|---|---|---|
| 1 | 96.6 | 80.5 | 84.5 | 61.5 | 80.8 |
| 2 | 96.8 | 81.3 | 85.0 | 61.6 | 81.2 |
| 4 | **97.0** | **82.6** | **85.2** | **61.8** | **81.7** |
| 8 | **97.0** | 81.5 | 84.6 | 61.6 | 81.2 |
| 16 | 96.9 | 80.4 | 84.5 | 61.5 | 80.8 |
| 32 | 96.8 | 79.7 | 84.5 | 61.6 | 80.7 |

on base classes, while the opposite trend emerges on the new classes When applying our training strategy, the difference in performance on the harmonic mean is relatively small, except for 32 prompt vectors with a dramatic drop. These results suggest that employing 2 prompt vectors is the optimal choice when considering the accuracy of both base and new classes. For conventional domain generalization, a shorter prompt proves insufficient for recognizing visual concepts effectively, whereas a longer prompt appears prone to overfitting on in-domain samples. Our method demonstrates promising results in terms of overall performance with a prompt length of 4.

*5) Influence of Domain-Split Optimization:* We investigate the influence of our proposed domain-split optimization strategy. Fig. 6 illustrates a consistent performance improvement across datasets for both generalization tasks. Specifically, our optimization strategy leads to an approximate 3%

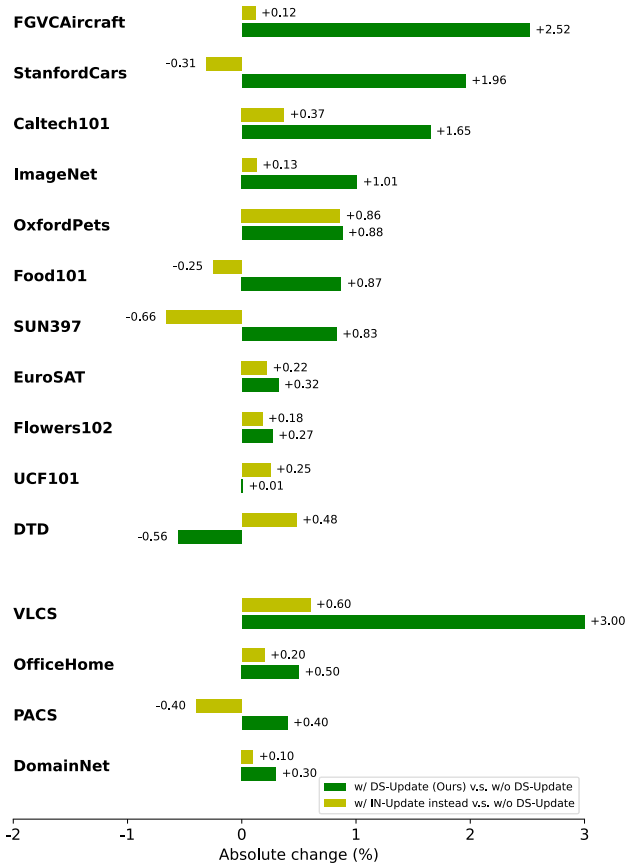



Fig. 6. Performance change with our proposed domain-split optimization strategy over datasets for base-to-new generalization and domain generalization. We compare the performance of our model with domain-split updates and with in-domain updates instead (equivalent to doubling the learning rate of in-domain updates).

increase in accuracy for both FGVCAircraft in base-to-new generalization and VLCS in conventional domain generalization. The consistent improvements provide evidence that our domain-split optimization significantly mitigates failures on out-of-domain data and enhances robustness to new classes, underscoring its excellent generalization capability. Comparing the replacement of domain-split updates with in-domain updates, the observed improvements are not statistically significant, thus demonstrating the effectiveness of alternate updates.

### D. Limitation and Bias

Although achieving significant experimental results compared with previous methods, it is noteworthy that our experimental design may still have some limitations and biases. From the aspect of dataset selection, despite the selected image recognition datasets covering a wide range of tasks, we only randomly sample a small amount of data from the whole dataset. It may be a tricky challenge to apply the method to real *few-shot* datasets, like medical images. From the aspect of metric limitations, beyond the aforementioned generalization tasks, experiments on more complex settings of generalization, like from different tasks with larger domain shifts, may also validate the effectiveness of domain invariant prompts.

## VI. Conclusion

We introduce MetaPrompt, a novel approach for learning the domain invariant prompt with the vision-language model CLIP to address the challenge of generalization. Our theoretical analysis demonstrates that the episodic training strategy provides a robust generalization guarantee for domain generalization tasks. Utilizing this analysis as a foundation, we devise an innovative episodic training algorithm, which alternates between in-domain updates and domain-split updates for prompt tuning. Through the application of asymmetric regularization and modality-specific optimization, our dual-modality prompt tuning network enables prompt learning in few-shot scenarios, showing remarkable generalization to unseen classes and domains. Extensive experiments on base-to-new generalization and domain generalization consistently validate the superior performance of our approach over existing methods.

While traditional prompt learning approaches frequently lead to a degradation in generalization performance, our method offers valuable insights into accessing the inherent relationship between domains and presents a viable solution for acquiring the invariant prompt, thus mitigating poor performance on unseen tasks. In the future, we will attempt to utilize the power of LLMs to acquire linguistic knowledge for learning domain-invariant as well as domain-specific prompts to fully capture semantic information to assist downstream recognition tasks. In addition, we will aim to apply domain invariant prompt learning for dense prediction, including semantic segmentation and depth estimation, etc., to enhance the generalization performance on other tasks.

## References

[1] A. Antoniou, H. Edwards, and A. Storkey, "How to train your MAML," 2018, *arXiv:1810.09502*.
[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.
[3] D. Arpit, H. Wang, Y. Zhou, and C. Xiong, "Ensemble of averages: Improving model selection and boosting performance in domain generalization," 2021, *arXiv:2110.10832*.
[4] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[5] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 446–461.
[6] J. Cha et al., "SWAD: Domain generalization by seeking flat minima," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22405–22418.
[7] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 52–68.
[8] J. Chen, X.-M. Wu, Y. Li, Q. Li, L.-M. Zhan, and F.-L. Chung, "A closer look at the training strategy for modern meta-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 396–406.
[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[12] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1657–1664.

[13] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[15] C. Finn, K. Xu, and S. Levine, "Probabilistic model-agnostic meta-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[16] S. Flennerhag, A. A. Rusu, R. Pascanu, F. Visin, H. Yin, and R. Hadsell, "Meta-learning with warped gradient descent," 2019, *arXiv:1909.00025*.

[17] A. Frome et al., "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[18] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[19] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," 2020, *arXiv:2007.01434*.

[20] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[21] I. Higgins et al., "$\beta$-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[22] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4483–4493.

[23] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 4904–4916.

[24] M. Jia et al., "Visual prompt tuning," 2022, *arXiv:2203.12119*.

[25] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11487–11496.

[26] M. Uzair Khattak, H. Rasheed, M. Maaz, S. Khan, and F. Shahbaz Khan, "MaPLe: Multi-modal prompt learning," 2022, *arXiv:2210.03117*.

[27] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2649–2658.

[28] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.

[29] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, "Improving zero-shot visual question answering via large language models with reasoning question prompts," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4389–4400.

[30] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021, *arXiv:2104.08691*.

[31] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[32] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5542–5550.

[33] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.

[34] X. Lisa Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, *arXiv:2101.00190*.

[35] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li, "Domain generalization using pretrained models without fine-tuning," 2022, *arXiv:2203.04600*.

[36] P. Liu, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.

[37] X. Liu et al., "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," 2021, *arXiv:2110.07602*.

[38] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5196–5205.

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

[40] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," 2017, *arXiv:1707.03141*.

[41] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5716–5726.

[42] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2554–2563.

[43] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[44] B. Oreshkin, P. R. López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[45] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3498–3505.

[46] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1406–1415.

[47] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[49] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, "Meta-learning with implicit gradients," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[50] A. Srinivasan, A. Bharadwaj, M. Sathyan, and S. Natarajan, "Optimization of image embeddings for few shot learning," in *Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, 2021, pp. 1–6.

[51] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 121–138.

[52] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1842–1850.

[53] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14974–14983.

[54] Y. Shi et al., "Gradient matching for domain generalization," 2021, *arXiv:2104.09937*.

[55] J. Wang and Y. Zhai, "Prototypical Siamese networks for few-shot learning," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 178–181.

[56] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[57] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–450.

[58] H. Sun, X. He, J. Zhou, and Y. Peng, "Fine-grained visual prompt learning of vision-language models for image recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5828–5836.

[59] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[60] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5018–5027.

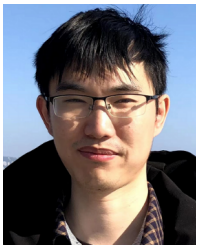
[61] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[62] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16816–16825.

[63] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, Mar. 2019.
[64] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6857–6866.
[65] O. Wiles et al., "A fine-grained analysis on distribution shift," 2021, *arXiv:2110.11328*.
[66] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.
[67] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3077–3086.
[68] J. Xiao, J. Hays, K. Ehinger, A. Olivia, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3485–3492.
[69] Y. Xing et al., "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Trans. Multimedia*, vol. 26, pp. 2056–2068, 2023.
[70] M. Yuan, G. Jia, and B.-K. Bao, "GPT-based knowledge guiding network for commonsense video captioning," *IEEE Trans. Multimedia*, early access, Nov. 6, 2023, doi: 10.1109/TMM.2023.3330070.
[71] Z. Zheng, X. Yue, K. Wang, and Y. You, "Prompt vision transformer for domain generalization," 2022, *arXiv:2208.08914*.
[72] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.
[73] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
[74] P. Zhou, X. Yuan, H. Xu, S. Yan, and J. Feng, "Efficient meta learning via minibatch proximal update," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[75] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15659–15669.

**Cairong Zhao** (Member, IEEE) received the B.S. degree from Jilin University in 2003, the M.S. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology, in 2011. He is currently a Professor with the College of Electronic and Information Engineering, Tongji University. He works on visual and intelligent learning, including computer vision, pattern recognition, and visual surveillance. He has published more than 40 top-rank international conferences and journals in the field, including CVPR, ICCV, ICLR, AAAI, ACM MM, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Pattern Recognition*. He holds prestigious positions, such as the Deputy Secretary-General of the Pattern Recognition and Machine Intelligence Committee, Chinese Association of Automation; the Chairperson of the Computer Vision Special Committee, Shanghai Computer Society; an Outstanding Member of the China Computer Federation; and a Senior Member of the China Graphics Society. He also serves as a reviewer for more than ten AI-related international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, NIPS, ICML, and AAAI.

**Yubin Wang** received the B.E. degree in data science and big data technology from Tongji University, China, in 2022, where he is currently pursuing the master's degree. His main research interests include prompt learning, multi-modal learning, and person re-identification.



**Xinyang Jiang** received the B.E. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2012 and 2017, respectively. He is currently a Researcher with Microsoft Research Asia (MSRA). Before joining MSRA, he was a Researcher with Tencent Youtu Laboratory. His main research interests include computer vision, including person re-identification, vector graphics recognition, and video enhancement and recognition.

**Yifei Shen** (Graduate Student Member, IEEE) received the Ph.D. degree from The Hong Kong University of Science and Technology. He is currently with Microsoft Research Asia.

**Kaitao Song** received the B.S. and Ph.D. degrees in computer science and technology from the Nanjing University of Science and Technology, China, in 2015 and 2021, respectively. He has published more than 20 academic papers in top-tier international journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, ICML, NeurIPS, ACL, KDD, ICCV, AAAI, IJCAI, InterSpeech, and ICASSP. His current research interests include natural language processing, multimodal analysis, deep learning, speech recognition, and machine learning. He has served as a PC Member for ICML, NeurIPS, ICLR, ACL, and EMNLP.

**Dongsheng Li** received the B.E. degree from the University of Science and Technology of China, China, in 2007, and the Ph.D. degree from Fudan University, Shanghai, China, in 2012. Since April 2015, he has been a Research Staff Member of IBM Research, China. Since February 2020, he has also been the Principal Research Manager with Microsoft Research Asia (MSRA). He is currently an Adjunct Professor with the School of Computer Science, Fudan University. His research interests include recommender systems and machine learning applications. His work on the cognitive recommendation engine received the 2018 IBM Corporate Award.

**Duoqian Miao** was born in 1964. He is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. He serves as the Vice President of the International Rough Set Society, the Executive Manager for the Chinese Association for Artificial Intelligence (CAAI), the Chair of the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member of the Chinese Computer Federation (CCF), the Vice President for the Shanghai Computer Federation, and the Vice President for the Shanghai Association for Artificial Intelligence. He serves as an Associate Editor for the *International Journal of Approximate Reasoning* and an Editor of the *Journal of Computer Research and Development* (in Chinese).