Computers and Electrical Engineering 117 (2024) 109270

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Modal-aware prompt tuning with deep adaptive feature enhancement

Haonan Wang, Mingwen Shao *, Xiaodong Tan, Lixu Zhang

Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China

ARTICLE INFO

Keywords:
Prompt learning
Vision-language models
Multi-modal
Few-shot image classification

ABSTRACT

Prompt learning has recently emerged as a promising method for fine-tuning vision-language models. By introducing prompts in the text encoder or image encoder, the pre-trained model can quickly adapt to downstream tasks without updating the pre-trained weights. However, prior multi-modal prompt tuning works do not consider the difference in feature distributions between text and images, and adopt the same prompts for both encoders, thus achieving sub-optimal performance in the downstream few-shot learning. In this paper, we propose Modal-Aware Prompt (MAP) to alleviate this issue. Specifically, considering the stability of text features, we design text-specific prompts, which can acquire text class-related information from a general template (i.e., "a photo of a <category>") by unidirectional attention-based interaction. Additionally, considering the diversity of image features, we design visual-specific prompts to acquire image class-related information and adjust the image features by bidirectional attention-based interaction. To learn hierarchical prompt representations and reinforce the prompt features, we further propose a Deep Adaptive Feature Enhancement (DAFE) module to adaptively utilize the prompt output of the former layer, which can combine instance-level and task-level information simultaneously. Combining the above two designs, our method MAP-DAFE obtains state-of-the-art results on 11 image recognition datasets and has the fastest convergence rate. This proves our MAP-DAFE is effective and efficient.

1. Introduction

Vision-language models (VLMs) have obtained remarkable success in downstream image recognition. These models undergo training on extensive datasets that encompass paired images and text descriptions, such as 400 million text–image pairs for CLIP. Therefore, they exhibit powerful generalization capabilities that can be used to execute open-vocabulary [1] image recognition tasks. During inference, general prompts, connected with downstream task labels, are passed to the text encoder to acquire the text embeddings. The resulting text embedding is then computed for similarity with the visual embedding to forecast the probability distribution.

Despite the remarkable generalization ability of these VLMs, how to better adapt them to downstream tasks remains a challenge. Due to their large scale and the scarcity of training data, fine-tuning all parameters carries a high risk of overfitting in downstream tasks and requires significant storage space. Inspired by Natural Language Processing (NLP) field, prompt learning [2–6] has been introduced in VLMs. It can make pre-trained VLMs efficiently adapt to downstream tasks by introducing only a few trainable parameters while keeping the pre-trained parameters frozen. Existing works can be classified into two categories: single-modal

* Corresponding author.
E-mail addresses: hnwang2024@163.com (H. Wang), smw278@126.com (M. Shao), reyes.tan@foxmail.com (X. Tan), z2216842477@163.com (L. Zhang).

https://doi.org/10.1016/j.compeleceng.2024.109270
Received 20 January 2024; Received in revised form 16 April 2024; Accepted 27 April 2024
Available online 8 May 2024

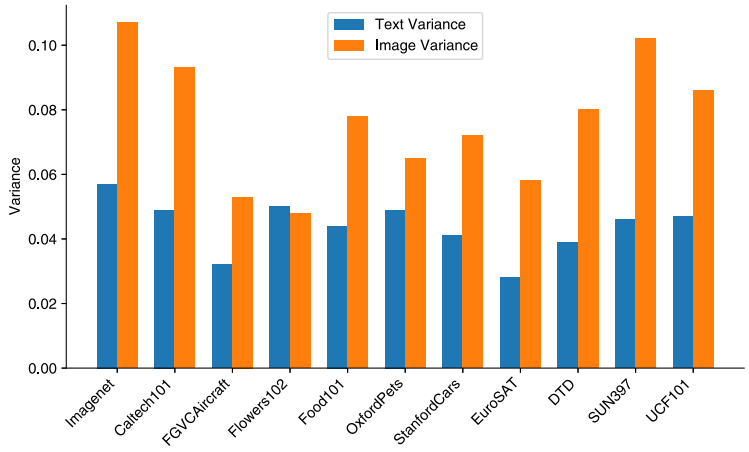


Fig. 1. Variance comparison across datasets. In most datasets, the variance of the images is significantly greater than that of the text.

Table 1
Compared to prior works, our proposed MAP-DAFE is more efficient, obtaining **better performance and faster convergence.**

Methods	Prompts	Accuracy			Train-epoch
		Base	New	H	
CLIP [1]	general template	69.34	74.22	71.70	–
CoOp [7]	single-modal	82.63	67.99	74.60	100 epoch
CoCoOp [10]	multi-modal	80.47	71.69	75.83	10 epoch
ProGrad [11]	single-modal	82.48	70.75	76.16	100 epoch
KgCoOp [12]	single-modal	80.73	73.6	77.0	100 epoch
RPO [8]	multi-modal	81.13	75.00	77.78	15 epoch
Ours	multi-modal	**82.80**	**76.37**	**79.45**	**8 epoch**

prompt learning and multi-modal prompt learning. Single-modal prompt learning such as Context Optimization (CoOp) [7] introduces prompts in text encoder, to automatically find the template that best fits the specific task. Considering the multi-modal nature of CLIP [1], multi-modal prompt learning such as RPO [8] proposes learning read-only prompts in text and image encoder to prevent representation shift, achieving better adaptability and generalization ability.

Prior multi-modal prompt learning works overlook the distribution characteristics of image and text features, and adopt the same prompts for text encoder and image encoder, resulting in sub-optimal performance in the base-to-novel setting, as shown in Table 1. Intuitively, text and image behave differently. The distribution of image features is more diverse and complex than that of text features. We calculate the variances of image features and text features (i.e., "a photo of a <category>") on 11 downstream datasets, as shown in Fig. 1. The variance of image features is significantly larger than that of text features, further supporting our intuition. Therefore, it is necessary to design modal-specific prompts for text and image encoders.

To alleviate the problem mentioned above, we propose Modal-Aware Prompt (MAP). For the general template (i.e., "a photo of a <category>"), the only thing that changes for different templates is the class name "<category>". The distribution of text features in general templates is more stable and simplex. Therefore, we devise the text-specific prompts in the text encoder, which is capable of acquiring abundant text class-related information through unidirectional attention-based [9] interaction with the general templates. In contrast, the distribution of image features varies greatly, even among images within the same category. Through only unidirectional attention-based information interaction from image features to prompts, prompts cannot obtain enough information from image features and may even acquire incorrect information due to the diversity of image features. Meanwhile, for some datasets with large domain shifts from pre-training data, we need to guide the pre-trained image features to adapt to the specific downstream task. Therefore, we introduce the image-specific prompts in the image encoder, which is capable of acquiring image class-related information and adjusting image features through bidirectional attention-based [9] interaction between image features and prompts.

To learn hierarchical prompt representations and reinforce the prompt features, we propose a Deep Adaptive Feature Enhancement (DAFE) module. Unlike traditional deep prompt learning, we combine the prompt output of the former layer which contains abundant instance-level information with the newly inserted prompts in the current layer. We further introduce layer-wise adaptive parameters to control how much the prompt output of the former layer contributes to the newly inserted prompts in the current layer. The layer-wise adaptive parameters encode the task-level information that varies with the specific task, depending on the characteristics of the task. As a result, our architecture takes into account instance-level and task-level information simultaneously, resulting in better adaptation ability to downstream tasks and better generalization ability to unseen classes. The following are the main contributions:

- We investigate the different distribution characteristics of the text features and image features and then propose text-specific prompts and image-specific prompts. As far as we know, we are the first to design modal-specific prompts for different encoders from the perspective of attention-based interaction.
- We propose a Deep Adaptive Feature Enhancement (DAFE) module, which combines instance-level and task-level information simultaneously.
- We conduct evaluations of our method on 11 image recognition datasets, attaining SOTA average results. The results presented in Table 1 indicate that our MAP-DAFE is a more efficient method, achieving higher performance while utilizing less training time.

The remainder of this paper is arranged as follows. In Section 2, we provide a brief overview of related works, including visual-language pre-training and prompt Learning in VLMs. Section 3 details the overall framework and computational flow of our proposal. Exhaustive experimental results and ablation analysis are provided in Section 4. Finally, we summarize the article and present some thoughts on future work in Section 5.

2. Related work

2.1. Few-shot learning

Few-shot learning (FSL) is an area related to our research that aims to learn from several labeled base classes to be able to generalize to novel classes. Conventional FSL methods include transfer learning [13], meta-learning [14–16], and metric learning [17,18]. Transfer learning-based methods start by training a backbone network using base classes to gain generic knowledge, which is then fine-tuned on novel classes. Literature [13] combines high and low-resolution features when training the backbone network, resulting in excellent object detection results. Meta-learning achieves better generalization to new tasks by learning a meta-learner on multiple tasks. Literature [16] uses meta-batch training to optimize the scaling and shifting operations, thus learning a better backbone network. The metric learning-based methods attempt to learn a feature space that brings samples of the same category as close together as possible and pushes samples of different categories as far apart as possible. The literature [17] presents a modified episodic learning algorithm that increases the number of negative samples within the same batch, thereby increasing the efficiency of metric learning. However, the above methods require rich data in the base classes for training, which limits their scalability. The success of VLMs provides another answer to the question. They can achieve excellent performance in downstream tasks in a zero-shot manner without the need for base class datasets. Prompt tuning can further improve their performance.

2.2. Visual-language pre-training

Pre-trained on paired text–image corpora [1,19–21] for modeling multi-modal information have shown great performance in downstream vision and language tasks. Compared with pre-training only with image supervision, the model combined with language supervision could catch more comprehensive multi-modal representations and therefore have great transfer ability in downstream tasks. The pre-training methods mainly contain masked language modeling [22,23] and masked-region modeling [24,25], contrastive learning [1,21,26] and image–text matching [22,24]. Among these methods, we concentrate on the VLMs pre-trained with contrastive learning, such as CLIP [1], ALIGN [21]. CLIP has a typical dual-encoder architecture which contains a text encoder and image encoder. By pre-training on 400M large-scale web datasets consisting of paired images and text descriptions, CLIP could encode text and images into a joint embedding space. Therefore, CLIP exhibits excellent performance across a broad range of tasks, including few-shot and zero-shot tasks. Although pre-trained VLMs have excellent transfer ability, adapting them efficiently to a specific downstream task remains challenging. Many prior works have attempted to adapt VLMs for few-shot image classification [27–29], object detection [30–32], and segmentation [33–35]. However, fine-tuning the whole pre-trained model to a specific task could be inefficient and even lead to a high risk of overfitting. In this study, we develop a new prompt tuning approach called MAP-DAFE to effectively adapt the pre-trained CLIP to few-shot image recognition tasks.

2.3. Prompt tuning in VLMs

Due to the massive scale of pre-trained models and the limit of downstream data, fine-tuning the whole pre-trained model is inefficient and even leads to a high risk of overfitting on specific downstream tasks. Inspired by the NLP field, prompt tuning [2–5,36] has been proven a promising approach to fine-tune the pre-trained VLMs efficiently. During training, only a few learnable parameters are introduced, and the pre-trained weights do not need to be updated. For each task, we only need to store the task-related prompts, not the entire model parameters and therefore prompt tuning has been seen as a parameter-efficient tuning approach. The application of prompt learning to VLMs has been extensively researched such as single-modal prompt learning [7,10–12,37–40] and multi-modal prompt learning [8,41–44]. As the first work to introduce prompt learning into VLMs, Context Optimization (CoOp) [7] replaces the hand-craft template “a photo of a” with trainable prompt tokens to find the best prompt template for a specific task. Considering the generalization problem that CoOp often overfits in base classes and faces a performance degeneration in unseen classes, Conditional Context Optimization (CoCoOp) [10] presents to construct instance-level prompts for each image by learning a meta-net, which could alleviate the overfitting problem. Prompt-aligned Gradient for Prompt Tuning (ProGrad) [11] considers the overfitting problem as

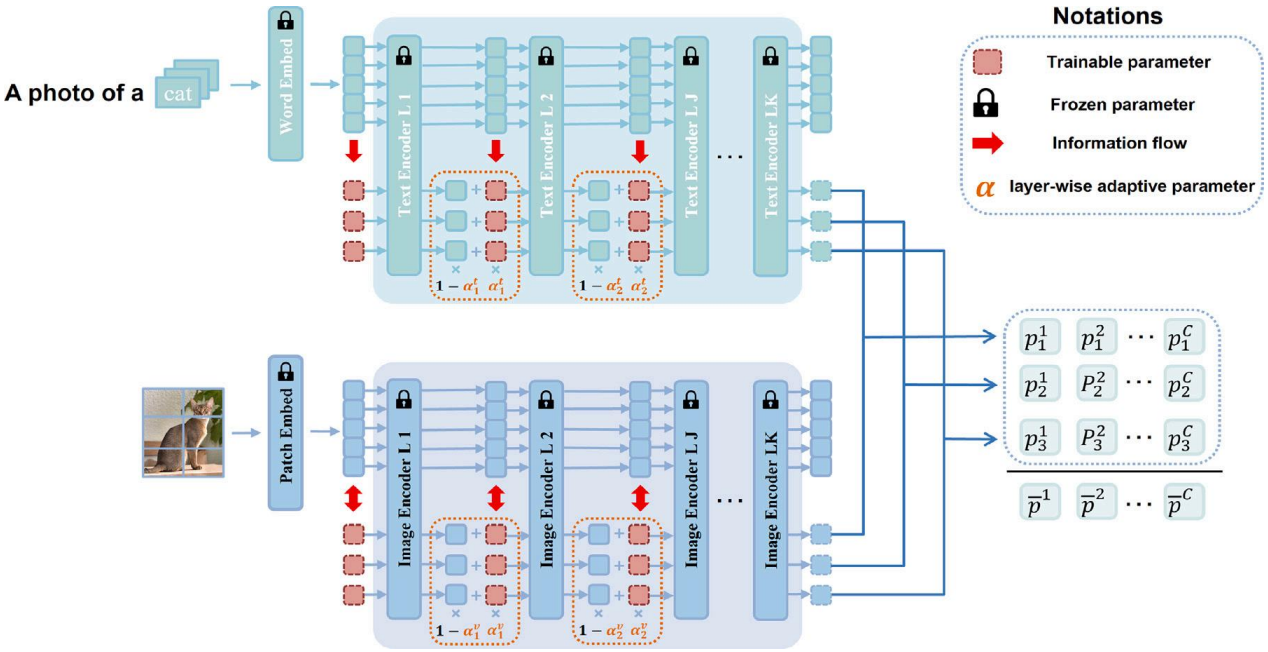


Fig. 2. Overall framework of MAP-DAFE. We adopt the general prompt "A photo of a <category>" across all datasets. Our modal-specific prompts are then introduced in each layer of both encoders until a special layer J. For the text encoder, we adopt a textual attention mask for unidirectional attention-based interaction to make the text-specific prompts only acquire information from the original features. For the image encoder, there is no limit to attention-based interaction between the image features and prompts. Therefore, the image-specific prompts can acquire and adjust information of the image features. The prompt output from the former layer is combined with the newly inserted prompts by layer-wise adaptive parameter α_l. Finally, the classification score is generated by averaging the similarity scores between the outputs of the encoders corresponding to each b prompt. where C denotes the number of classes.

a problem of forgetting pre-trained general knowledge, and therefore ensures the gradient of the prompts is updated in a direction that does not deviate from the direction of general knowledge. To better retain the general knowledge acquired by pre-training, KgCoOp [12] proposes to ensure the learnable prompts differ little from the general prompts (i.e., "a photo of a <category>"). To prevent internal representation shift, RPO [8] introduces read-only prompts in text and image encoder, achieving better adaptability and generalization ability. To better utilize the discarded prompt outputs, Progressive Visual Prompt (ProVP) [37] designs a new structure that combines the prompt outputs from adjacent layers by a constant parameter.

Compared with previous works, RPO and ProVP are most relevant to ours. Compared with RPO, we take into account the diversity difference between the text and image features, therefore designing text-specific prompts and image-specific prompts. ProVP combines the prompt outputs from adjacent layers which introduces instance-level information. However, ProVP only performs these operations on the visual encoder and connects the prompts from adjacent layers by a constant parameter, which ignores the importance of the text encoder and the difference in the amount of information between different layers. Compared with the above two methods, our proposed method takes into account the different distribution characteristics of text features and image features and combines the task-level and instance-level information simultaneously. Furthermore, the experimental results indicate the superiority of our method.

3. Method

In this section, we will introduce our method in detail. Fig. 2 shows our overall architecture. We devise Modal-Aware Prompt (MAP) based on the distribution characteristics of image features and text features for more flexible adaptation. Meanwhile, inspired by the success of deep prompt tuning, we design a new Deep Adaptive Feature Enhancement (DAFE) module to improve performance further. We use a pre-trained VIT-B/16 CLIP [1] as the foundational framework, and its specific architecture is shown in Fig. 3. During training, only the introduced text-specific prompts, vision-specific prompts, and layer-wise adaptive parameters can be updated. All pre-trained parameters in CLIP are frozen. Below, we will first revisit the architecture of pre-trained CLIP and then describe our proposed method in detail.

3.1. Revisit CLIP

CLIP [1] is a multi-modal model with a typical dual-encoder architecture, which contains a text encoder and a vision encoder. During training, CLIP adopts a contrastive loss to narrow the cosine distance between an image and its textual description, while pushing away the cosine distance of the mismatched text descriptions. By pre-training on a large-scale web dataset consisting of 400M paired images and text descriptions, CLIP can project an image and its textual description into a joint feature space. Therefore, CLIP can be utilized for open-vocabulary image recognition by calculating the similarity between features of a general template like

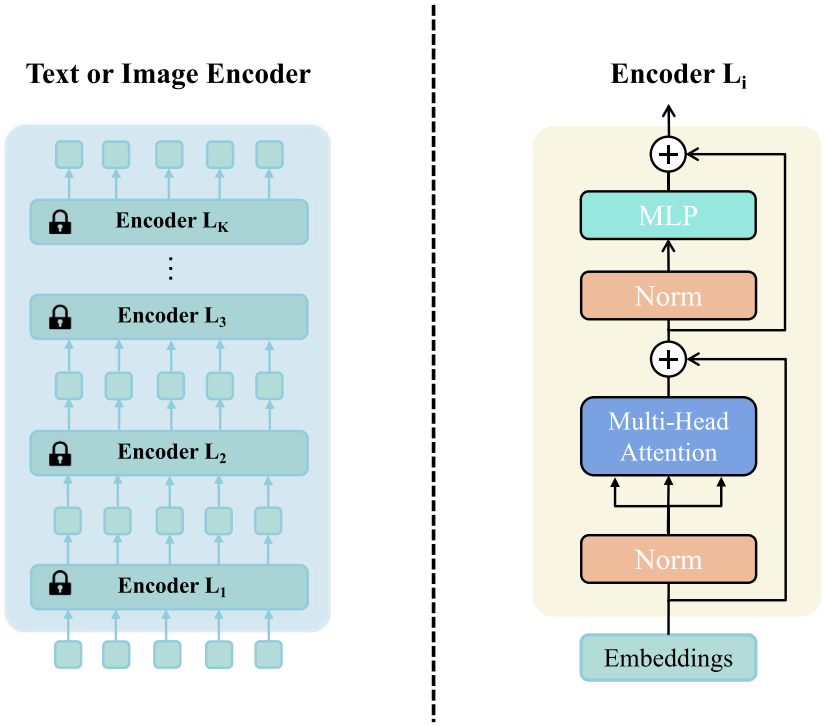


Fig. 3. Detailed architecture of text encoder and image encoder. Right: we adopt VIT-B/16 CLIP as the base framework, so both the text encoder and image encoder have similar transformer architecture, and their inputs are word embeddings and patch embeddings respectively. Left: a detailed view of each transformer layer, the semantic information of the context can be learned by self-attention computation in each layer's embeddings.

"a photo of a <category>" and image feature. Let $\{y_i\}_{i=1}^{C}$ be the label set of the downstream task, y_i denotes the ith class. the general prompt filled with y_i will be fed into text encoder $\mathcal{T}$. Given the image x and image encoder $\mathcal{F}$, CLIP computes the prediction probabilities [7] according to the cosine similarity between the image embedding and C text embeddings:

$$p\left(y_i|\mathbf{x}\right) = \frac{\exp(sim(\mathcal{F}(x), \mathcal{T}(w_i))/\tau)}{\sum_{j=1}^{C} \exp(sim(\mathcal{F}(x), \mathcal{T}(w_j))/\tau)}, \tag{1}$$

where $sim(\cdot)$ denotes cosine similarity, w_i denotes the handcraft prompt filled with y_i, and τ denotes the learned temperature parameter.

3.2. Modal-aware prompt

Prior multi-modal prompt learning works have explored introducing prompts in both text encoder and image encoder for better aligning vision-language modalities. Maple [41] introduces prompts in both encoders and further designs a coupling function to encourage mutual synergy between vision prompts and text prompts. RPO [8] proposes read-only prompts in text encoder and image encoder which can acquire information from the features of pre-trained CLIP to prevent representation shift during adaptation, achieving a robust and generalizable adaptation. However, they both do not take into account the distribution discrepancy between text features and image features and adopt the same prompts for both modalities, thus achieving a sub-optimal performance. We underline the importance of designing modal-specific prompts according to the different distribution characteristics of image features and text features. Therefore, we present MAP for more flexible adaption of pre-trained VLMs.

We design the text-specific prompts for the text encoder to perform unidirectional attention-based [9] interaction with the general templates due to the stability of template features. Considering the image features' diversity, we further design image-specific prompts for the image encoder to perform bidirectional attention-based [9] interaction with image features. We first concatenate the introduced prompts with image patch embedding or word embedding. The algorithm works as follows:

$$\begin{aligned} [O_1^t,\ W_1^t,\ H_1^t] &= \mathcal{T}_1([P_1^t, W_0^t, H_0^t]), \\ [O_1^v,\ W_1^v,\ H_1^v] &= \mathcal{F}_1([P_1^v, W_0^v, H_0^v]), \end{aligned} \tag{2}$$

where $\mathcal{F}$, $\mathcal{T}$ denote the image encoder and text encoder. $W_0^t = \{w_i\}_{i=1}^{N_t}, W_0^v = \{w_i\}_{i=1}^{N_v}$ denote the text and visual input embeddings of the first layer, while N_t, N_v denote the length of feature tokens. W_0^t is initialized with a general template filled with category such as "a photo of a <CAT>", $P_1^t = \{p_i\}_{i=1}^{b}, P_1^v = \{p_i\}_{i=1}^{b}$ denote the b prompt tokens introduced in text encoder and image encoder for the first layer, O_1^t, O_1^v denote the prompt output of the first layer, and H_0^t, H_0^v denote the special token of text encoder and image encoder.

To achieve unidirectional attention-based information interaction from text features to prompts, we adopt textual attention mask $M_t \in \mathbb{R}^{D_t \times D_t}$, where $D_t = 1 + N_t + b$. The definition of the textual mask [8] is as follows:

$$M_t^{i,j} = \begin{cases} -\infty, & \text{if } j > 1 + N_t \text{ or } i > j \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $M_t^{i,j}$ represents the element located in the ith row and in the jth column of the mask. To achieve bidirectional attention-based interaction between image features and prompts, we adopt visual attention mask $M_v \in \mathbb{R}^{D_v \times D_v}$, where $D_v = 1 + N_v + b$. The definition of the visual mask is as follows:

$$M_v^{i,j} = 0. \tag{4}$$

The direction of information flow between prompts and input embeddings is controlled by introducing the mask into the process of attention calculation. The self-attention formula [9] has been redefined as:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}} + M\right)V, \tag{5}$$

where M denotes the textual or visual attention mask.

Through the subsequent layers of processing, the final prompt features can be calculated by the following equations:

$$\begin{aligned} z &= \text{TextProj}(O^t), \\ r &= \text{ImageProj}(O^v), \end{aligned} \tag{6}$$

where O^t, and O^v denote the text and image prompt outputs of the last layer. TextProj, ImageProj denote the text and image projection matrix.

3.3. Deep prompt tuning

Deep prompt tuning [41,45,46] denotes introducing prompts in each transformer layer. Compared with introducing prompts only in the first layer, deep prompt tuning has better flexibility to guide features within different hierarchies. Moreover, adding prompts to deeper layers has a more direct effect on model predictions. Finally, there are more trainable task-specific parameters to allow a better adaption ability to the downstream tasks. Formally, the formulation is defined as:

$$[\_,\ W_i,\ H_i] = \mathcal{L}_i([P_i, W_{i-1}, H_{i-1}]) \quad i = 1, 2, \ldots, K, \tag{7}$$

where $W_{i-1} = \quad [w^1, w^2, \ldots, w^N]$ denotes the input token of the ith layer, H_{i-1} denotes the special token which is used to aggregate the global feature, and $[\ ,\ ]$ denotes the concatenation operation. $\mathcal{L}_i$ denotes the ith transformer layer. $P_i = \{p_j\}_{j=1}^{b}$ denotes the introduced b prompt tokens in the ith layer and the prompt output of the former layer is discarded.

3.4. Deep adaptive feature enhancement

In deep prompt tuning, the prompt outputs are discarded, and the newly introduced prompts for the current layer only guide the features of the current layer. The prompt outputs contain a wealth of pre-trained knowledge. Utilizing this knowledge appropriately could help guide the pre-trained model to better adapt to downstream tasks. However, combining the prompt output of the former layer with the newly inserted prompts by a constant alone cannot achieve optimal performance. This is because this approach overlooks the fact that prompt outputs of different layers contain varying amounts of information. Additionally, different downstream tasks further affect the distribution of information across different layers. To solve the problems mentioned above, we further propose DAFE which can take advantage of the prompt outputs adaptively. The prompt output of the former layer will be combined with the newly introduced prompts for the current layer by a layer-wise adaptive parameter α_i. The processing of the algorithm is as follows:

$$[O_1,\ W_1,\ H_1] = \mathcal{L}_1([P_1, W_0, H_0]). \tag{8}$$

Except for the first layer, the prompts for subsequent layers are a combination of the prompt output of the former layer and the newly introduced prompts. The algorithm works as follows:

$$[O_i,\ W_i,\ H_i] = \mathcal{L}_i([(1 - \alpha_i) * O_{i-1} + \alpha_i * P_i, W_{i-1}, H_{i-1}]), \tag{9}$$

where α_i is a layer-wise adaptive parameter introduced in ith layer. It is calculated by β_i according to a sigmoid function as follows:

$$\alpha_i = \frac{1}{1 + e^{-\beta_i}} \in \mathbb{R}, \tag{10}$$

and β_i could be updated according to the gradient.

To further improve the efficiency of parameters, we introduce prompts in the front J layers of the transformer instead of every layer. After Jth layer, the subsequent processing is defined as:

$$[O_i,\ W_i,\ H_i] = \mathcal{L}_i([O_{i-1}, W_{i-1}, H_{i-1}]) \quad i = J+1, \ldots, K, \tag{11}$$

where O_{i-1} denotes the prompt output of the former layer.

Our experiments are based on the CLIP [1] model which contains an image encoder and a text encoder, therefore we do the same operation for both encoders with our proposed DAFE. Compared with deep prompt tuning, our proposed DAFE considers the instance-level and task-level information for the downstream tasks simultaneously. The prompt outputs contain abundant knowledge of the image embeddings or text embeddings and vary with inputs by attention-based interaction. Thus, combining the previous output with the newly introduced prompts by $(1-\alpha_i) * O_{i-1} + \alpha_i * P_i$ will make our architecture have instance-level information which could avoid overfitting within the seen classes. The introduced layer-wise adaptive parameter α_i can control how much of the previous output O_{i-1} contributes to the newly introduced prompt P_i. By training on each task, the layer-wise adaptive parameter α_i will encode task-level information which contains the distribution of information across different layers. Therefore, prompt output with more amounts of information will have greater weight and then merge with the newly introduced prompts, and vice versa. Finally, our architecture will include task-level information as a supplement to instance-level information. This enhances our architecture's adaptability to specific tasks and improves its generalizability to unseen classes simultaneously.

3.5. Loss function

There are b pairs prompts that acquire abundant information from the text features and image features respectively. Therefore, there are two designs for the loss function to calculate the logits. The first strategy is prompt-based average which calculates the logits by averaging the prompts output from the last layer. The second strategy is logit-based average which calculates the b logits separately, then uses the average as the final logits. Experiments have shown that the second strategy is more effective. Given image x and label y, the processing of the algorithm is as follows:

$$p_i^k = \frac{\exp(sim(r_i, z_i^k)/\tau)}{\sum_{j=1}^{C} \exp(sim(r_i, z_i^j)/\tau)}, \tag{12}$$

$$p(y_k|x) = \frac{1}{b} \sum_{i=1}^{b} p_i^k, \tag{13}$$

$$Loss_{ce}(x) = -\sum_{k=1}^{C} y_k \log p(y_k|x), \tag{14}$$

where $sim(\cdot)$ denotes cosine similarity. r_i, z_i denote the ith visual and text prompt feature respectively, and z_i^k denotes the kth class feature in text prompt feature.

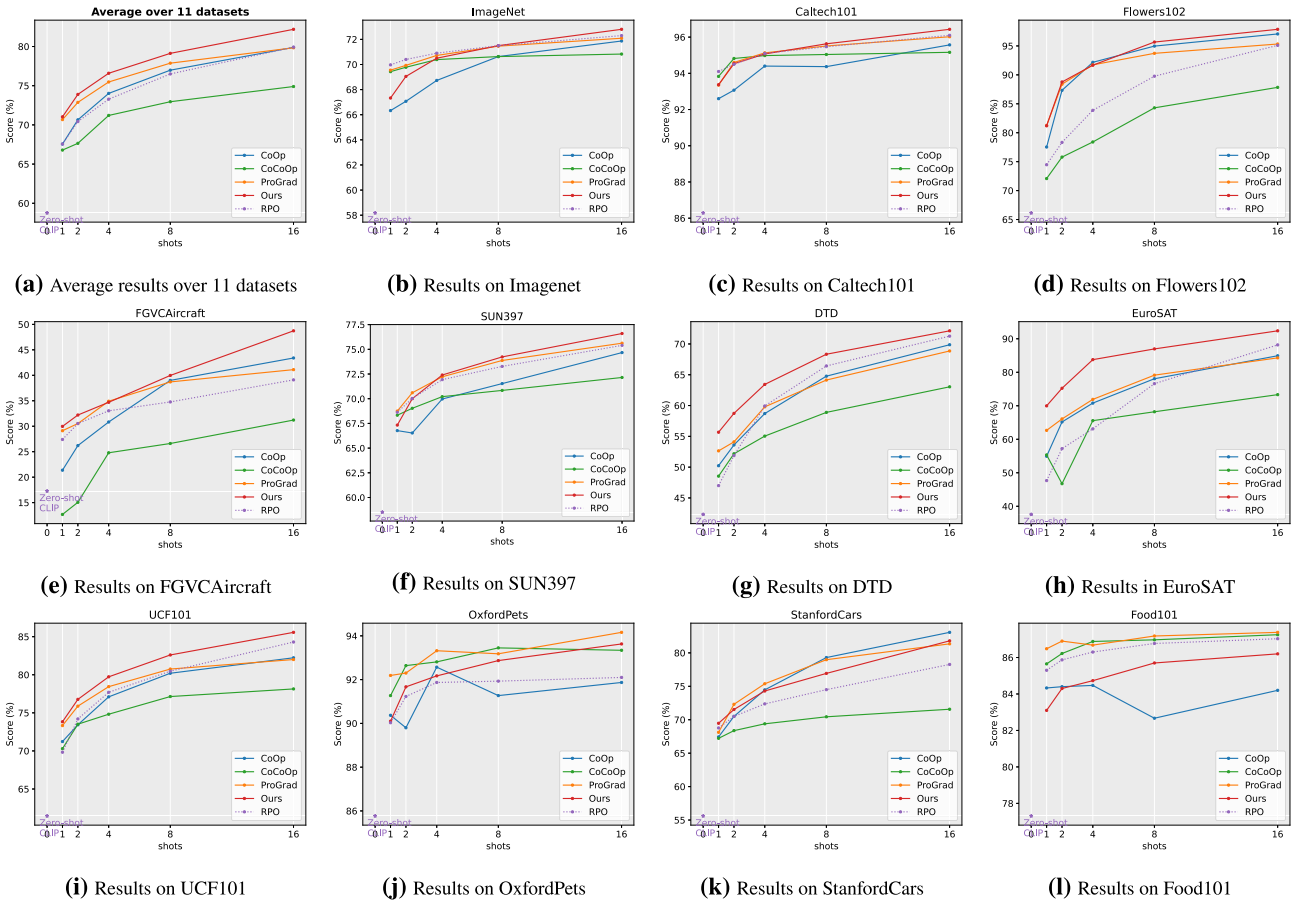
4. Experiment

Building on prior research, we verify the validity of our method in two experimental scenarios: 1) Base-to-novel setting. The dataset will be divided into base and novel classes using random seeds. The model will undergo training in base classes and subsequently be evaluated in novel classes. This will allow for the evaluation of the generalization of our proposed method. 2) Few-shot setting. It is used to evaluate the adaption ability of our MAP-DAFE. We train the model using a limited number of labeled images and evaluate it on the test set which has the same classes as the train set.

Datasets: Following the prior works, we employ 11 image recognition datasets in both scenarios, these datasets cover a broad range of classification scenarios, including animals, plants, and scene classification. The ImageNet dataset [47] is a commonly used collection of 14,197,122 images from 1000 categories for visual tasks. For the base-to-novel setting, the training and test sets of this dataset will be divided into two parts: 500 categories as base classes and another 500 categories as novel classes. Caltech101 [48] is a dataset for generic object recognition that contains 101 categories, with 40 to 800 images per category. In the base-to-novel setting, we will use 50 of these categories as base categories and the other 50 as new categories. For the other datasets we will use the same division in the base-to-novel setting. For example, FGVCAircraft [49], Flowers102 [50], Food101 [51], OxfordPets [52], StanfordCars [53] for fine-grained object recognition, EuroSAT [54] for satellite image recognition, DTD [55] for texture image recognition, SUN397 [56] for scene recognition and UCF101 [57] for action recognition. Several images from base classes in the training set are randomly selected according to a random seed for few-shot training. The novel classes images in the testing set are used for testing. To ensure a fair comparison for base-to-novel setting, we train the model in a 16-shot setting. The results are expressed as the mean value derived from three iterations, corresponding to three different random seeds (1, 2, and 3).

For the few-shot setting, we will not divide the categories of the dataset into two parts. In accordance with previous studies, we conduct model training using 1, 2, 4, 8, and 16 shots. We then evaluate the model on whole test sets.

Training Details: The code utilizes the PyTorch framework, which is based on the Python language. The workstation is equipped with an Intel(R) Xeon(R) Gold 6133 processor, 64 GB RAM, and an NVIDIA A6000 graphics card with 48 GB GPU memory. All our experiments are conducted using a pre-trained VIT-B/16 CLIP [1] model as the foundational framework. This model comprises a text encoder and an image encoder, both of which have a similar architecture consisting of 12 layers of transformers. The text encoder takes a 512-dimensional word embedding as input, while the image encoder takes a 224 × 224 image as input and transforms it into a 768-dimensional patch embedding. All experiments are trained with a batch size of 4 and a learning rate of 0.1 with SGD



(a) Average results over 11 datasets **(b)** Results on Imagenet **(c)** Results on Caltech101 **(d)** Results on Flowers102

(e) Results on FGVCAircraft **(f)** Results on SUN397 **(g)** Results on DTD **(h)** Results in EuroSAT

(i) Results on UCF101 **(j)** Results on OxfordPets **(k)** Results on StanfordCars **(l)** Results on Food101

Fig. 4. A comprehensive comparison of MAP-DAFE with previous methods in few-shot learning reveals notable advancements. MAP-DAFE demonstrates substantial improvements across 8 out of 11 datasets, leading to a significant enhancement in overall performance averages.

optimizer. The layer-wise adaptive parameter α_l is calculated by β_l with a sigmoid function, and β_l is randomly initialized for each layer from a normal distribution. We report the base-to-novel and few-shot class accuracies averaged over 3 runs.

For the base-to-novel setting, we set prompt length (the number of prompt tokens in each layer) to 3 and prompt depth (the number of layers to introduce prompts) J to 9. The maximum epoch is set to 8. For the few-shot setting, we configure the prompt length to 8 and prompt depth J to 9. The maximum epoch is configured to be 50 for the 16-shot scenario and 30 for other shot configurations. Specifically for ImageNet, which involves 1000 classes, the maximum epoch is uniformly set to 30 across all shot scenarios.

4.1. Few-shot learning

Few-shot learning is a scenario in which models are trained using a limited number of labeled images and subsequently tested on a dataset consisting of the consistent label set with the training set. Fig. 4 demonstrates the whole results on 11 datasets. The sub-figure in the top-left corner illustrates the average performance of five methods. Compared with other existing methods, our proposed approach demonstrates a substantial enhancement across all shot settings, outperforming all previous works. With an increase in the number of training samples, the average performance improves substantially, suggesting that our method is more effective when dealing with relatively large numbers of shots. There is a similar finding in few-shot setting with base-to-novel setting. Compared with the previous SOTA (ProGrad) [11], our method obtains a more remarkable improvement in 16-shot setting on datasets with large domain shift with pre-training data, such as EuroSAT (8.07%), DTD (3.27%), FGVCAircraft (7.62%), UCF101 (3.57%), which further proves the validity of Modal-Aware Prompt. The observed average performance gains across all shots confirm the capability of our MAP-DAFE to enhance prompt learning in a data-efficient manner.

Fig. 5 ranks the absolute improvements obtained by our MAP-DAFE over RPO [8] in a 16-shot setting. Our MAP-DAFE demonstrates a distinct advantage over RPO in 10 out of 11 datasets, especially in datasets characterized by substantial domain shifts from the pre-training data. For instance, in FGVCAircraft and EuroSAT, the performance gains reach 9.63% and 4.20%, respectively. On some fine-grained datasets, MAP-DAFE also gains obvious increases, such as StanfordCars (3.53%), and Flowers102 (2.77%). In contrast, the improvement on Food101 is less appealing. This phenomenon may be a result of the noisy training data of Food101. Nonetheless, the overall results indicate that our approach has better adaptability than RPO in a data-efficient manner.

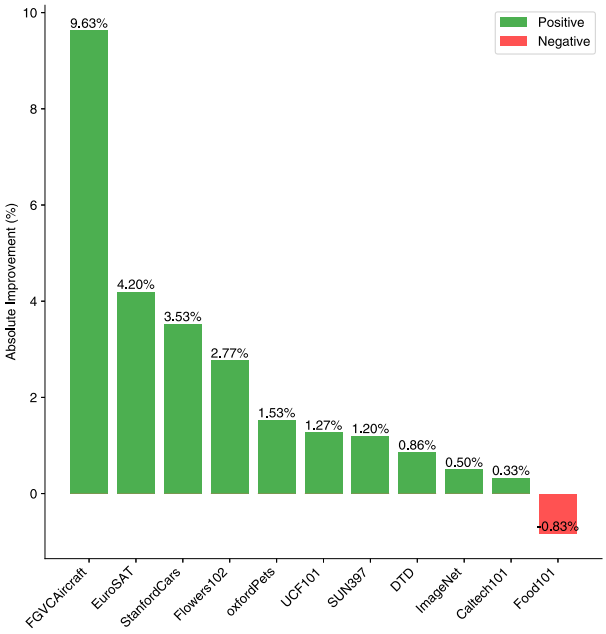


Fig. 5. Comparison between MAP-DAFE and RPO in few-shot setting (16-shot setting).

Table 2
Comparison of varying shot settings. The average results across all 11 datasets are presented. The blue font indicates the gain relative to the previous SOTA method (RPO) on different settings..

Backbones	Methods	1-shot			4-shot			8-shot			16-shot		
		Base	Novel	H	Base	Novel	H	Base	Novel	H	Base	Novel	H
ViT-B/16	CoOp [7]	71.68	67.11	69.32	78.43	68.03	72.44	**80.73**	68.39	73.5	82.63	67.99	74.60
	CoCoOp [10]	71.62	71.42	71.52	76.72	73.34	74.85	78.56	72.0	74.9	80.47	71.69	75.83
	ProGrad [11]	73.54	70.97	72.23	79.18	71.14	74.62	80.62	71.02	75.2	82.48	70.75	76.16
	KgCoOp [12]	**74.87**	72.94	**73.89**	**79.92**	73.11	75.90	78.36	73.89	76.06	80.73	73.6	77.0
	RPO [8]	70.04	73.80	71.87	74.13	74. 31	74.21	77.09	74.39	75.71	81.13	75.00	77.78
	Ours	72.21	**73.94**	73.06	77.26	**75.70**	**76.47**	80.27	**75.85**	**77.99**	**82.80**	**76.37**	**79.45**
		+2.17	+0.14	+1.19	+3.13	+1.39	+2.26	+3.18	+1.46	+2.28	+1.67	+1.37	+1.67

4.2. Base-to-novel generalization

To verify the generalization of our proposed MAP-DAFE, we evaluate our method in base-to-novel setting, in which the dataset will be split into two nonoverlapping groups: base classes and novel classes. Our model uses the base classes as the train set for prompt tuning and is evaluated on the novel classes. To analyze the performance of our proposed method from the whole and details more intuitively, we conduct base-to-novel experiments with various few-shot settings. The Base metric in the table indicates the accuracy of MAP-DAFE on the base classes, reflecting the model's ability to adapt to downstream tasks, while the Novel metric indicates the accuracy on new classes, reflecting the model's ability to generalize to new concepts. The challenge remains to reach a good trade-off between base class performance and novel class performance. As shown in Tables 2 and 3, the novel class accuracy is significantly lower than the base class accuracy. Due to this discrepancy, we adopt the H metric (Harmonic mean) to highlight the generalization trade-off, and $H = 2/(\frac{1}{base} + \frac{1}{novel})$. Table 2 illustrates the average results with varying few-shot configurations. Table 3 presents a detailed performance analysis of all 11 datasets. Both experiments are conducted on the VIT-B/16 CLIP [1] as the foundational framework.

Holistic Analysis: As shown in Table 2, our proposed MAP-DAFE has been significantly improved on all settings in terms of harmonic mean and novel accuracy compared with existing methods, proving its advantages in generalizing from base to novel classes. KgCoOp [12] achieves the best base performance of 74.87% and 79.92% for the 1-shot and 4-shot settings. It is possible that the small sample size prevented the parameters from learning the distribution of information between the different layers, resulting in a decrease in performance on the base class. Nevertheless, our proposed method achieves a significant improvement on novel classes accuracy compared with KgCoOp, e.g., obtains the improvement of 1.00% and 2.59% upon KgCoOp for novel accuracy. The results demonstrate the better generalization of our proposed method. Furthermore, as the sample size increases, our proposed method achieves optimal performance for both the base classes and the novel classes. CoOp [7] obtains a best base performance of 80.73% for 8-shot setting, but the novel accuracy decreases dramatically, e.g., achieves the worst performance of 68.39%. We hold the opinion that the CoOp has a serious overfitting on the base classes, resulting in a dramatic drop in performance on the novel classes. Compared to CoOp, our proposed method shows a slight decrease in performance on the base classes but achieves

Table 3
Comparison with existing methods. The metric H represents the harmonic mean and the blue font indicates the gain relative to previous SOTA method (RPO).

(a) Average over 11 datasets.

Methods	Base	Novel	H
CLIP [1]	69.34	74.22	71.70
CoOp [7]	82.63	67.99	74.60
CoCoOp [10]	80.47	71.69	75.83
ProGrad [11]	82.48	70.75	76.16
KgCoOp [12]	80.73	73.6	77.0
RPO [8]	81.13	75.00	77.78
Ours	**82.80**	**76.37**	**79.45**
	+1.67	+1.37	+1.67

(b) ImageNet.

Methods	Base	Novel	H
CLIP	72.43	68.14	70.22
CoOp	76.46	66.31	71.02
CoCoOp	75.98	70.43	73.10
ProGrad	**77.02**	66.66	71.46
KgCoOp	75.83	69.96	72.78
RPO	76.60	**71.57**	**74.00**
Ours	76.50	71.00	73.64
	-0.10	-0.57	-0.36

(c) Food101.

Methods	Base	Novel	H
CLIP	90.10	91.22	90.66
CoOp	89.44	87.50	88.46
CoCoOp	**90.70**	91.29	90.99
ProGrad	90.37	89.59	89.98
KgCoOp	90.5	**91.7**	**91.09**
RPO	90.33	90.83	90.58
Ours	90.40	91.50	90.94
	+0.07	+0.67	+0.36

(d) FGVCAircraft.

Methods	Base	Novel	H
CLIP	27.19	**36.29**	31.09
CoOp	39.24	30.49	34.30
CoCoOp	33.41	23.71	27.74
ProGrad	**40.54**	27.57	32.82
KgCoOp	36.21	33.55	34.83
RPO	37.33	34.20	35.70
Ours	39.3	34.57	**36.78**
	+1.97	+0.37	+1.08

(e) SUN397.

Methods	Base	Novel	H
CLIP	69.36	75.35	72.23
CoOp	80.85	68.34	74.07
CoCoOp	79.74	76.86	78.27
ProGrad	**81.26**	74.17	77.55
KgCoOp	80.29	76.53	78.36
RPO	80.60	77.80	79.18
Ours	81.0	**78.03**	**79.48**
	+0.40	+0.23	+0.30

(f) DTD.

Methods	Base	Novel	H
CLIP	53.24	59.90	56.37
CoOp	80.17	47.54	59.68
CoCoOp	77.01	56.00	64.85
ProGrad	77.35	52.35	62.45
KgCoOp	77.55	54.99	64.35
RPO	76.70	62.13	68.61
Ours	**80.67**	**63.93**	**71.33**
	+3.97	+1.80	+2.72

(g) EuroSAT.

Methods	Base	Novel	H
CLIP	56.48	64.05	60.03
CoOp	91.54	54.44	68.27
CoCoOp	87.49	60.04	71.21
ProGrad	90.11	60.89	72.67
KgCoOp	85.64	64.34	73.48
RPO	86.63	68.97	76.79
Ours	**94.57**	**82.60**	**88.18**
	+7.94	+13.63	+11.39

(h) UCF101.

Methods	Base	Novel	H
CLIP	70.53	77.50	73.85
CoOp	**85.14**	64.47	73.37
CoCoOp	82.33	73.45	77.64
ProGrad	84.33	74.94	79.35
KgCoOp	82.89	76.67	79.65
RPO	83.67	75.43	79.34
Ours	84.97	**79.03**	**81.89**
	+1.30	+3.60	+2.55

(i) OxfordPets.

Methods	Base	Novel	H
CLIP	91.17	97.26	94.12
CoOp	94.24	96.66	95.43
CoCoOp	95.20	97.69	96.43
ProGrad	95.07	97.63	96.33
KgCoOp	94.65	**97.76**	96.18
RPO	94.63	97.50	96.05
Ours	**95.60**	97.53	**96.55**
	+0.97	+0.03	+0.50

(j) Caltech101.

Methods	Base	Novel	H
CLIP	96.84	94.00	95.40
CoOp	98.11	93.52	95.76
CoCoOp	97.96	93.81	95.84
ProGrad	98.02	93.89	95.91
KgCoOp	97.72	**94.39**	96.03
RPO	97.97	94.37	96.03
Ours	**98.37**	94.00	**96.13**
	+0.40	-0.37	+0.10

(k) StanfordCars.

Methods	Base	Novel	H
CLIP	63.37	74.89	68.65
CoOp	76.20	69.14	72.49
CoCoOp	70.49	73.59	72.01
ProGrad	**77.68**	68.63	72.88
KgCoOp	71.76	75.04	73.36
RPO	73.87	**75.53**	**74.69**
Ours	73.27	73.63	73.44
	-0.60	-1.90	-1.25

(l) Flowers102.

Methods	Base	Novel	H
CLIP	72.08	**77.08**	74.83
CoOp	**97.63**	69.55	81.23
CoCoOp	94.87	71.75	81.71
ProGrad	95.54	71.87	82.03
KgCoOp	95.00	74.73	83.65
RPO	94.13	76.67	**84.50**
Ours	96.20	74.37	83.88
	+2.07	-2.30	-0.62

a significant improvement of 7.49% and 4.49% in novel accuracy and harmonic mean, respectively. For the 16-shot setting, our proposed method achieves the highest performance with base accuracy, novel accuracy, and harmonic mean of 82.8%, 76.33%, and 79.93%, respectively, compared to all existing methods.

RPO [8] achieves excellent performance on novel classes, second only to our proposed method. However, our proposed method obtains a significant improvement on base classes of 2.17%, 3.13%, 3.18%, and 1.67% upon RPO for the four scenarios. The demonstrated outstanding results serve as evidence that MAP-DAFE can effectually guide the pre-trained VLM, concurrently improving accuracy for both base and novel classes.

Detailed Analysis: Table 3 presents the detailed performance of MAP-DAFE in base-to-novel setting on 11 image recognition datasets. In a comparison of average performance across 11 datasets, our proposed method showcases a remarkable improvement over all prior research, achieving enhanced accuracy in both base and novel classes. Specifically, MAP-DAFE gains a significant enhancement compared with the previous SOTA (RPO), e.g., obtaining the increases of 1.67% and 1.37% for the base accuracy and novel accuracy. We have a key finding: MAP-DAFE achieves a substantial absolute improvement on datasets characterized by significant domain shifts from the pre-trained data. Notably, there are improvements on FGVCAircraft (1.08%), UCF101 (2.55%), DTD (2.72%), and particularly EuroSAT (11.39%). The experiment results further prove the validity of our proposed MAP. For the datasets that have large domain shifts from the pre-trained data, it is necessary to design modal-specific for different modalities.

Comparison with RPO: RPO [8] adopts a similar design to MAP-DAFE, which introduces prompts in both encoders. It obtains an excellent performance on novel classes and gains the previous SOTA performance on the harmonic mean. Despite the architectural

Table 4
Effectiveness of Proposed Components. We add components one by one to verify the validity of each component.

Methods	Base	Novel	H
baseline	79.84	74.33	76.98
w/ deep	81.42	74.87	78.00
w/ modal-aware	80.10	75.00	77.46
w/ deep/adaptive	81.89	75.27	78.44
w/ modal-aware/deep	82.50	75.90	79.06
MAP-DAFE	82.80	76.37	**79.45**

Table 5
Comparison of computational complexity between previous works. Our proposed method has excellent computational efficiency and fastest convergence and achieves optimal performance.

Method	Params (K)	Params % CLIP	FLOPS (G)	Epochs	HM
CoOp	2.05	0.002	20.96	100	74.60
CoCoOp	35.36	0.043	20.96	10	75.83
ProGrad	8.19	0.09	20.96	100	76.16
KgCoOp	2.05	0.002	20.96	100	77.00
RPO	30.72	0.037	22.32	15	77.78
Ours	34.57	0.042	21.13	8	**79.45**

similarity, out of 11 recognition datasets, MAP-DAFE outperforms RPO in 9 out of the base classes and 7 out of the novel classes in terms of performance. For average performance, MAP-DAFE achieves an obvious improvement of 1.67% and 1.37% for base accuracy and novel accuracy upon RPO. RPO obtains excellent performance on novel classes performance, however, the performance on base classes is not advantageous. We hold the opinion that RPO overlooks the discrepancy between different modalities and then adopts the same prompts for both encoders. In contrast to RPO, our proposed method achieves SOTA performance in terms of both base classes accuracy and novel classes accuracy, surpassing all prior works. This exhibits excellent adaptability and generalization ability.

4.3. Ablation experiments

Effectiveness of Proposed Components: In our proposed architecture, the modal-aware prompt, deep prompt, and layer-wise adaptive components serve as core elements. To confirm the impact of each component, we perform experiments on all 11 datasets by sequentially adding each component and observing the performance changes on both base and novel classes. We strip out all the components and used them as the baseline model. When the deep prompt component is added, we set the layer-wise adaptive parameters to a constant of 0.6 in the situation of stripping out the adaptive component. The detailed experiment results are in Table 4. Incorporating the modal-aware prompt contributes to enhanced accuracy across base, novel, and harmonic mean, with increases of 0.26%, 0.67%, and 0.48%, respectively, compared to the baseline. The result reveals that modal-aware prompt not only helps improve the adaption ability to downstream tasks, but also helps improve the generalizability to unseen classes. By adding the deep prompt, there is an absolute improvement of 1.58% on the base classes compared to the baseline. This proves that the deep prompt has better adaptability compared to the shallow one, which only introduces prompts in the first layer. Compared to the baseline, the classification accuracies for base classes, novel classes, and harmonic mean are boosted by 2.96%, 2.04%, and 2.47%, respectively, with the combination of modal-aware prompt, deep prompt, and layer-wise adaptive parameters. These obvious gains prove that the combination of all components results in better adaption and generalization performance.

Prompting complexity:

Table 5 shows the comparison of complexity between our algorithm and other algorithms in base-to-novel setting. Compared to CoCoOp, our proposed algorithm only exceeds the FLOPS of CoCoOp by 0.8%. However, it improves the overall performance of the base to novel experiments by 4.8% and has fewer parameters and faster convergence than CoCoOp. Compared to the previous SOTA RPO, our proposed algorithm slightly increases the number of parameters, but at the same time we significantly reduce the computational FLOPS, increase the computational efficiency, and remarkably improve the overall performance and convergence speed. For CoOp, Prograd, and KgCoOp, our proposed algorithm does not dominate in terms of the number of parameters. However, as the number of parameters increases, we observe only a slight improvement in computational FLOPS. Furthermore, in comparison to these three algorithms, our algorithm has faster convergence and requires only one-tenth of the training epoch (8 vs 100 epochs). The experimental results indicate that our method exhibits excellent computational efficiency, the fastest convergence, and the best performance.

Ablation on prompt depth and length: The impact of prompt depth and length on performance is illustrated in Fig. 6. Specially, we conduct experiments on 11 image recognition datasets, by varying the number of prompt tokens or varying the depth at which the prompts are introduced into the transformer layers. As the prompt depth J increases, the performance on the harmonic mean gradually improves. But when we further introduce prompts at deeper layers such as all layers (1~12), performance drops

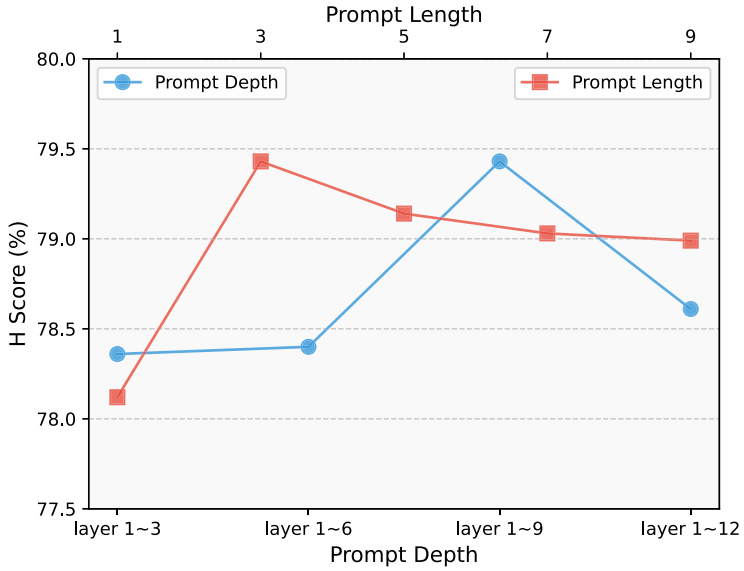


Fig. 6. Investigations on prompt's context length and introduced depth.

Table 6
Different α in base-to-novel generalization. Compared with the settings that set a constant α, introducing layer-wise adaptive parameters achieves the best performance.

α	Base	Novel	H
0.0	80.10	75.00	77.46
0.2	81.99	75.62	78.67
0.4	82.68	75.71	79.04
0.6	82.50	75.90	79.06
0.8	81.89	75.56	78.59
1.0	81.35	75.14	78.12
adaptive	**82.80**	**76.37**	**79.45**

dramatically. We believe that the deeper features are consummate enough so that they do not need to be guided by additional prompts. Therefore, we ended up setting the prompt depth to 9. As the prompt length increases, the performance curve on the harmonic mean is similar to the curve of prompt depth. The best performance on harmonic mean is achieved when the prompt length is set to 3. We hold the opinion that the longer prompt length leads to more severe overfitting on base classes as more parameters are learned, and thus the performance on novel classes drops dramatically.

Effectiveness of layer-wise adaptive parameters: To verify whether the layer-wise adaptive parameters can introduce task-level information into our model and improve performance, we set constant α_l for each layer of both encoders and conduct experiments on all 11 datasets in the base-to-novel setting. We compare these results with those obtained using the layer-wise adaptive parameters. Table 6 demonstrates the average performance on all 11 datasets. Through careful analysis of the experimental results, we have three key findings: 1) Setting the α_l to 0 is equivalent to introducing prompts only at the first layer, which obtains the worst performance compared with the other settings. Therefore, the experiments further prove the effectiveness of deep prompt which introduce prompts in each layer. 2) When we set the α_l to 1, the prompt outputs are completely discarded. There is a significant drop in performance on both base classes and novel classes compared with other settings that utilize the prompt output of the former layer according to the α_l. The experiment results further prove that combining the previous outputs which contain abundant information is beneficial in improving performance. 3) Our proposed method achieves the highest performance of 82.8% and 76.37% on base and novel classes, respectively, using the layer-wise adaptive parameters. This reveals that layer-wise adaptive parameters introduce task-level information to the model, guiding the pre-trained VLMs more effectively. Table 7 shows the detailed performance on various downstream tasks at different settings. Our method obtains the best performance on each dataset compared to other settings, which indicates that layer-wise adaptive parameters are capable of introducing task-specific information for each task by training on specific downstream tasks.

Friedman's rank test: To better evaluate the differences between our algorithm and previous works, we used Friedman's rank test and visualized the results in Fig. 7. Firstly, we calculated the average rank of the 6 algorithms on 11 datasets as [1.27, 2.82, 2.73, 4.0, 4.64, 5.55]. These values are visualized on the horizontal axis of Fig. 7. The p-value was calculated as 6.42e−7, which is much less than 0.05. Therefore, the original hypothesis that all algorithms perform equally was rejected. And the average rank of our algorithm is much more than the other works which proves that our algorithm outperforms other algorithms in most of the datasets and hence demonstrating its superior performance. Furthermore, we used a post-hoc test (Nemenyi test) to distinguish between the algorithms, the results are visualized in Fig. 7. The algorithms connected by the bold horizontal lines in the figure indicate

Table 7
Different α in various tasks. By introducing layer-wise adaptive parameters, our method obtains the best results across various tasks.

α	DTD			EuroSAT			UCF101		
	Base	Novel	H	Base	Novel	H	Base	Novel	H
0	73.37	**65.40**	69.15	92.53	72.50	81.29	81.33	75.97	78.55
0.2	77.57	61.67	68.71	94.77	77.10	85.02	83.60	77.97	80.68
0.4	79.17	61.73	70.09	94.77	69.37	86.21	83.03	78.70	80.80
0.6	79.77	63.13	70.48	94.43	81.37	87.41	82.90	79.27	81.04
0.8	78.37	61.47	68.89	94.30	80.23	86.69	82.40	79.13	80.73
1.0	77.63	61.77	68.79	92.47	81.33	86.54	82.80	**79.53**	81.13
adaptive	**80.67**	63.93	**71.33**	94.57	**82.60**	**88.18**	**84.97**	79.03	**81.89**

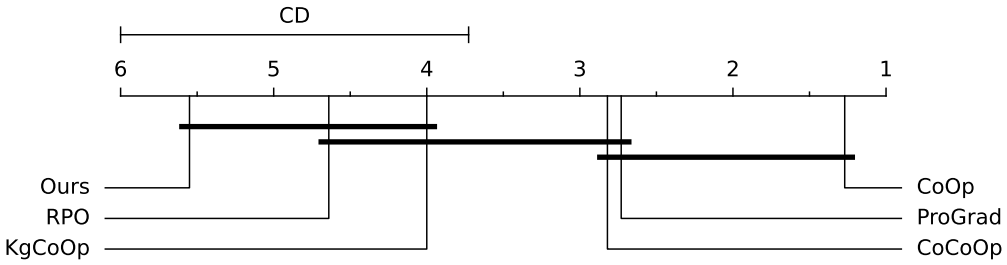


Fig. 7. Evaluating algorithm performance using Friedman's rank test and Post-hoc Nemenyi test result visualization.

that there is no statistically significant difference between them. Our algorithm is connected to RPO, KgCoOp, which shows that the difference in average rank between them is not significant enough and the performance rankings are relatively close, i.e., they are not statistically significantly different from each other. However, objectively speaking, our algorithm outperforms the others in terms of performance metrics.

5. Conclusion

The large-scale VLMs like CLIP [1], and FLIP [58] have exhibited outstanding generalization capabilities on image recognition tasks. However, how to better adapt them to downstream tasks poses a significant challenge. Due to its massive scale and the scarcity of training data, fine-tuning the entire model to a specific task would increase the risk of overfitting and require huge storage space, resulting in inefficiency. Prompt learning has recently emerged as a promising method to mitigate these challenges, but existing methods do not take into account the distribution discrepancy between the features of text and images, and thus obtain a sub-optimal performance. To mitigate this problem, we propose a novel prompt learning approach called MAP-DAFE. We first devise modal-specific prompts for each encoder, considering the diversity differences between different modalities. To learn hierarchical prompt representations and reinforce the prompt features, we further present a Deep Adaptive Feature Enhancement (DAFE) module. This module dynamically leverages the prompt outputs, allowing the simultaneous integration of both instance-level and task-level information. Through comprehensive experiments, we have shown that our MAP-DAFE surpasses existing methods in both base-to-novel generalization and few-shot settings, exhibiting improved performance and faster convergence.

Although the performance of our proposed method has gained excellent performance, there are certain aspects that need to be further studied. Firstly, we devise text-specific prompts to acquire text class-related information from a general template (i.e., "a photo of a <Abyssinian>"), however, richer information about this breed of cat is ignored. Therefore, in future work, we will consider using a large language model to extend the generic templates and generate richer textual descriptions. Secondly, the image-specific prompts extract class-related information from image patches. However, since different patches have varying levels of importance for classification, we plan to explore the addition of an extra loss to constrain the optimization of the image-specific prompts and prioritize task-related content in future work. Optimization of computational efficiency and number of parameters will also be further investigated.

CRediT authorship contribution statement

Haonan Wang: Investigation, Writing – original draft, Read and approved the final manuscript. **Mingwen Shao:** Conceptualization, Supervision, Validation, Read and approved the final manuscript. **Xiaodong Tan:** Methodology, Data curation. **Lixu Zhang:** Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. All authors read and approved the final manuscript.

This work was supported by National Natural Science Foundation of China (Nos. 62376285, 61673396), Natural Science Foundation of Shandong Province, China (No. ZR2022MF260).

References

[1] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021, p. 8748–63.
[2] Li XL, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. 2021, http://dx.doi.org/10.18653/v1/2021.acl-long.353, arXiv preprint arXiv:2101.00190.
[3] Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT understands, too. AI Open 2023. http://dx.doi.org/10.48550/arXiv.2103.10385.
[4] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. 2021, http://dx.doi.org/10.18653/v1/2021.emnlp-main.243, arXiv preprint arXiv:2104.08691.
[5] Gu Y, Han X, Liu Z, Huang M. Ppt: Pre-trained prompt tuning for few-shot learning. 2021, http://dx.doi.org/10.18653/v1/2022.acl-long.576, arXiv preprint arXiv:2109.04332.
[6] Ping Z, Sang G, Liu Z, Zhang Y. Aspect category sentiment analysis based on prompt-based learning with attention mechanism. Neurocomputing 2024;565:126994. http://dx.doi.org/10.1016/j.neucom.2023.126994, URL: https://www.sciencedirect.com/science/article/pii/S0925231223011177.
[7] Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis 2022;130(9):2337–48. http://dx.doi.org/10.1007/s11263-022-01653-1.
[8] Lee D, Song S, Suh J, Choi J, Lee S, Kim HJ. Read-only prompt optimization for vision-language few-shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 1401–11. http://dx.doi.org/10.48550/arXiv.2308.14960.
[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30. http://dx.doi.org/10.48550/arXiv.1706.03762.
[10] Zhou K, Yang J, Loy CC, Liu Z. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 16816–25. http://dx.doi.org/10.1109/CVPR52688.2022.01631.
[11] Zhu B, Niu Y, Han Y, Wu Y, Zhang H. Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 15659–69.
[12] Yao H, Zhang R, Xu C. Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 6757–67. http://dx.doi.org/10.1109/CVPR52729.2023.00653.
[13] Li H, Ge S, Gao C, Gao H. Few-shot object detection via high-and-low resolution representation. Comput Electr Eng 2022;104:108438. http://dx.doi.org/10.1016/j.compeleceng.2022.108438, URL: https://www.sciencedirect.com/science/article/pii/S004579062200653X.
[14] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. 2017, URL: https://api.semanticscholar.org/CorpusID:6719686.
[15] Chen Z, Fu Y, Wang YX, Ma L, Liu W, Hebert M. Image deformation meta-networks for one-shot learning. In: 2019 IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 8672–81, URL: https://api.semanticscholar.org/CorpusID:67926490.
[16] Lei S, Dong B, Shan A, Li Y, Zhang W, Xiao F. Attention meta-transfer learning approach for few-shot iris recognition. Comput Electr Eng 2022;99:107848. http://dx.doi.org/10.1016/j.compeleceng.2022.107848, URL: https://www.sciencedirect.com/science/article/pii/S0045790622001409.
[17] Zeng X, Huang B, Jia K, Jia L, Zhao K. Adaptive few-shot learning with a fair priori distribution. Comput Electr Eng 2022;102:108133. http://dx.doi.org/10.1016/j.compeleceng.2022.108133, URL: https://www.sciencedirect.com/science/article/pii/S0045790622003834.
[18] Wang W, Duan L, En Q, Zhang B, Liang F. TPSN: Transformer-based multi-prototype search network for few-shot semantic segmentation. Comput Electr Eng 2022;103:108326. http://dx.doi.org/10.1016/j.compeleceng.2022.108326, URL: https://www.sciencedirect.com/science/article/pii/S004579062200547X.
[19] Desai K, Kaul G, Aysola Z, Johnson J. RedCaps: Web-curated image-text data created by the people, for the people. 2021, http://dx.doi.org/10.48550/arXiv.2111.11431, arXiv preprint arXiv:2111.11431.
[20] Srinivasan K, Raman K, Chen J, Bendersky M, Najork M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021, p. 2443–9.
[21] Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. PMLR; 2021, p. 4904–16.
[22] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning. PMLR; 2021, p. 5583–94.
[23] Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv Neural Inf Process Syst 2019;32.
[24] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers. 2019, arXiv preprint arXiv:1908.07490.
[25] Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, et al. Vl-bert: Pre-training of generic visual-linguistic representations. 2019, arXiv preprint arXiv:1908.08530.
[26] Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. Adv Neural Inf Process Syst 2021;34:9694–705.
[27] Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. Clip-adapter: Better vision-language models with feature adapters. Int J Comput Vis 2023;1–15. http://dx.doi.org/10.1007/s11263-023-01891-x.
[28] Kim K, Laskin M, Mordatch I, Pathak D. How to adapt your large-scale vision-and-language model. 2021.
[29] Zhang R, Fang R, Zhang W, Gao P, Li K, Dai J, et al. Tip-adapter: Training-free clip-adapter for better vision-language modeling. 2021, arXiv preprint arXiv:2111.03930.
[30] Feng C, Zhong Y, Jie Z, Chu X, Ren H, Wei X, et al. Promptdet: Towards open-vocabulary detection using uncurated images. In: European conference on computer vision. Springer; 2022, p. 701–17. http://dx.doi.org/10.1007/978-3-031-20077-9_41.
[31] Gu X, Lin TY, Kuo W, Cui Y. Open-vocabulary object detection via vision and language knowledge distillation. 2021, http://dx.doi.org/10.48550/arXiv.2104.13921, arXiv preprint arXiv:2104.13921.

[32] Maaz M, Rasheed H, Khan S, Khan FS, Anwer RM, Yang MH. Class-agnostic object detection with multi-modal transformer. In: European conference on computer vision. Springer; 2022, p. 512–31. http://dx.doi.org/10.1007/978-3-031-20080-9_30.
[33] Ding J, Xue N, Xia GS, Dai D. Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11583–92. http://dx.doi.org/10.1109/CVPR52688.2022.01129.
[34] Lüddecke T, Ecker A. Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 7086–96. http://dx.doi.org/10.1109/CVPR52688.2022.00695.
[35] Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, et al. Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 18082–91. http://dx.doi.org/10.1109/CVPR52688.2022.01755.
[36] Zhang Y, Sun B, He J, Yu L, Zhao X. Multi-level neural prompt for zero-shot weakly supervised group activity recognition. Neurocomputing 2024;571:127135. http://dx.doi.org/10.1016/j.neucom.2023.127135, URL: https://www.sciencedirect.com/science/article/pii/S0925231223012584.
[37] Xu C, Shen H, Shi F, Chen B, Liao Y, Chen X, et al. Progressive visual prompt learning with contrastive feature re-formation. 2023, http://dx.doi.org/10.48550/arXiv.2304.08386, arXiv preprint arXiv:2304.08386.
[38] Lu Y, Liu J, Zhang Y, Liu Y, Tian X. Prompt distribution learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 5206–15. http://dx.doi.org/10.1109/CVPR52688.2022.00514.
[39] Zhang Y, Fei H, Li D, Yu T, Li P. Prompting through prototype: A prototype-based prompt learning on pretrained vision-language models. 2022, http://dx.doi.org/10.48550/arXiv.2210.10841, arXiv preprint arXiv:2210.10841.
[40] Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: European conference on computer vision. Springer; 2022, p. 709–27. http://dx.doi.org/10.1007/978-3-031-19827-4_41.
[41] Khattak MU, Rasheed H, Maaz M, Khan S, Khan FS. Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 19113–22. http://dx.doi.org/10.1109/CVPR52729.2023.01832.
[42] Liu X, Tang W, Lu J, Zhao R, Guo Z, Tan F. Deeply coupled cross-modal prompt learning. 2023, http://dx.doi.org/10.18653/v1/2023.findings-acl.504, arXiv preprint arXiv:2305.17903.
[43] Cho E, Kim J, Kim HJ. Distribution-aware prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 22004–13. http://dx.doi.org/10.48550/arXiv.2309.03406.
[44] Miao Y, Li S, Tang J, Wang T. MuDPT: Multi-modal deep-symphysis prompt tuning for large pre-trained vision-language models. 2023, http://dx.doi.org/10.1109/ICME55011.2023.00013, arXiv preprint arXiv:2306.11400.
[45] Liu X, Ji K, Fu Y, Tam WL, Du Z, Yang Z, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. 2021, http://dx.doi.org/10.48550/arXiv.2110.07602, arXiv preprint arXiv:2110.07602.
[46] Sun T, He Z, Zhu Q, Qiu X, Huang XJ. Multitask pre-training of modular prompt for chinese few-shot learning. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). 2023, p. 11156–72. http://dx.doi.org/10.18653/v1/2023.acl-long.625.
[47] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009, p. 248–55. http://dx.doi.org/10.1109/cvpr.2009.5206848.
[48] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. IEEE; 2004, p. 178. http://dx.doi.org/10.1016/j.cviu.2005.09.012.
[49] Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. 2013, arXiv preprint arXiv:1306.5151.
[50] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. In: 2008 sixth Indian conference on computer vision, graphics & image processing. IEEE; 2008, p. 722–9.
[51] Bossard L, Guillaumin M, Van Gool L. Food-101–mining discriminative components with random forests. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13. Springer; 2014, p. 446–61.
[52] Parkhi OM, Vedaldi A, Zisserman A, Jawahar C. Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE; 2012, p. 3498–505.
[53] Krause J, Stark M, Deng J, Fei-Fei L. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. 2013, p. 554–61. http://dx.doi.org/10.1109/ICCVW.2013.77.
[54] Helber P, Bischke B, Dengel A, Borth D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J Sel Top Appl Earth Obs Remote Sens 2019;12(7):2217–26. http://dx.doi.org/10.1109/JSTARS.2019.2918242.
[55] Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, p. 3606–13. http://dx.doi.org/10.1109/CVPR.2014.461.
[56] Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE; 2010, p. 3485–92. http://dx.doi.org/10.1109/cvpr.2010.5539970.
[57] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. 2012, arXiv preprint arXiv:1212.0402.
[58] Yao L, Huang R, Hou L, Lu G, Niu M, Xu H, et al. Filip: Fine-grained interactive language-image pre-training. 2021, http://dx.doi.org/10.48550/arXiv.2111.07783, arXiv preprint arXiv:2111.07783.