Self-Prompt Mechanism for Few-Shot Image Recognition

Mingchen Song*, Huiqiang Wang*, Guoqiang Zhong†

College of Computer Science and Technology, Ocean University of China
songmingchen@stu.ouc.edu.cn, wanghuiqiang@stu.ouc.edu.cn, gqzhong@ouc.edu.cn

Abstract

Few-shot learning poses a formidable challenge as it necessitates effective recognition of novel classes based on a limited set of examples. Recent studies have sought to address the challenge of rare samples by tuning visual features through the utilization of external text prompts. However, the performance of these methods is constrained due to the inherent modality gap between the prompt text and image features. Instead of naively utilizing the external semantic information generated from text to guide the training of the image encoder, we propose a novel self-prompt mechanism (SPM) to adaptively adjust the neural network according to unseen data. Specifically, SPM involves a systematic selection of intrinsic semantic features generated by the image encoder across spatial and channel dimensions, thereby engendering self-prompt information. Subsequently, upon backpropagation of this self-prompt information to the deeper layers of the neural network, it effectively steers the network toward the learning and adaptation of new samples. Meanwhile, we propose a novel parameter-efficient tuning method that exclusively fine-tunes the parameters relevant to self-prompt (prompts are no more than 2% of the total parameters), and the incorporation of additional learnable parameters as self-prompt ensures the retention of prior knowledge through frozen encoder weights. Therefore, our method is highly suited for few-shot recognition tasks that require both information retention and adaptive adjustment of network parameters with limited labeling data constraints. Extensive experiments demonstrate the effectiveness of the proposed SPM in both 5-way 1-shot and 5-way 5-shot settings for standard single-domain and cross-domain few-shot recognition datasets, respectively. Our code is available at https://github.com/codeshop715/SPM.

Introduction

Despite the significant advancements achieved by deep learning in computer vision, it typically relies on a vast number of labeled samples, which deviates from the human learning process. Few-shot learning (Finn, Abbeel, and Levine 2017; Munkhdalai et al. 2018; Antoniou, Edwards, and Storkey 2018) aims to bridge the gap between human intelligence and learning machines by addressing the challenge of learning from a limited amount of labeled training data and generalizing to unseen data. Few-shot image recognition is indeed an extensive research task in the field of few-shot learning algorithms. The objective of few-shot image recognition is to develop models that can effectively adapt to recognize and classify unseen classes with limited sample data. This task is particularly challenging as it requires learning discriminative features from a few labeled data (Zhang et al. 2022; Afrasiyabi et al. 2022).

Due to the scarcity of labeled samples in new classes, a simple method is to utilize information from other modalities as auxiliary guidance. Recently, with the introduction of the CLIP model (Radford et al. 2021), a series of text-

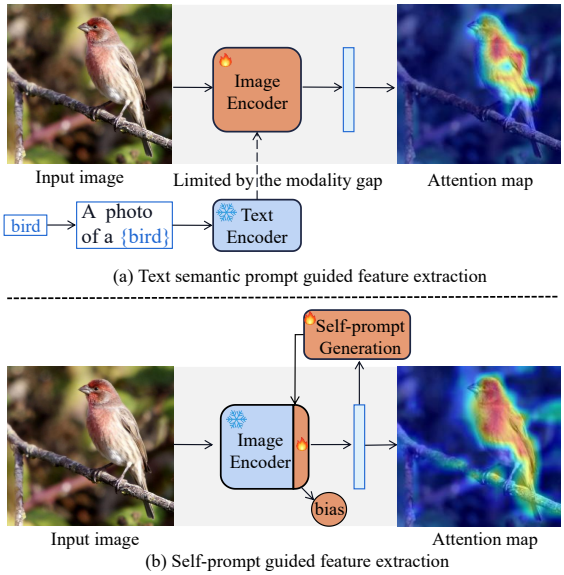

Figure 1: Semantic prompt and self-prompt mechanism for few-shot image recognition. (a): The semantic prompt methods require additional text information and text encoder to generate the external prompts. (b): Our proposed self-prompt mechanism does not necessitate additional information, and only a tiny amount of parameters need to be fine-tuned to generate prompts derived from the intrinsic semantic information of the image encoder.

*These authors contributed equally.
†Corresponding author.

based semantic prompt methods (Chen et al. 2023; Zhu et al. 2023; Jeong et al. 2023) have emerged to guide the training of the visual modules. As shown in Figure 1 (a), these methods typically follow the training paradigm of CLIP, where separate text encoder and image encoder are employed to generate discriminative image features based on text embeddings. Despite the significant success of text-based semantic prompt methods in the field of few-shot learning, most of the mentioned methods suffer from the following issues. Firstly, semantic prompts rely on generated or manually authored textual information. Although large language models such as BERT (Devlin et al. 2019) and GPT (Radford et al. 2018) can extract rich textual information from class names, the diversity in textual descriptions for the same class results in inaccurately generated semantic prompts. Secondly, text-based prompt methods require additional text encoders to extract features from textual information, leading to additional computational overhead. Thirdly, the information gap resulting from distinct modalities of text and images restricts the effectiveness of text features in providing optimal external semantic prompts for visual feature learning, due to misalignment between text and visual features generated by the network.

To address the aforementioned three issues, inspired by the human cognitive process (LEE 2002; Yu and Dayan 2004; Baifeng, Trevor, and Xin 2023) and human metacognitive ability (Salles et al. 2016), we propose a novel self-prompt mechanism to guide the training of visual networks. Intuitively, humans have the metacognitive ability to summarize based on past experiences and provide self-prompt when encountering similar problems or tasks (Fleming and Dolan 2012), allowing them to modify their strategies or directions of action in order to align explicitly with the goal of the tasks. We propose a novel scheme that leverages this human mechanism by applying the self-prompt mechanism into the few-shot learning process, as illustrated in Figure 1 (b). Specifically, the guidance of the learning process is achieved through a top-down approach. We perform spatial and channel selection on the deep layer features of the image encoder to generate intrinsic self-prompt information for unseen classes or domains, and then transmit the generated prompt information back to the deep layers of the network to adaptive adjustment of the feature extraction process. By prompting the calculation process of self-attention, our proposed self-prompt mechanism can guide the training of the image encoder to extract discriminative features from unseen data.

Furthermore, since different unseen classes or domains present distinctive feature requirements (Li, Liu, and Bilen 2022), it is imperative for the network to possess a versatile and efficient adaptation mechanism capable of effectively handling the significantly diverse semantic characteristics of unseen classes or domains. Simultaneously, the network should be parameter-efficient to adjust adaptive parameters when confronting unseen classes or domains that have only a limited number of labeled data. To address these challenges, we propose a novel parameter-efficient tuning method that exclusively fine-tunes the parameters relevant to self-prompt according to unseen data, requiring adjustments of no more than 2% of the total network parameters. Therefore, this method also ensures the retention of prior knowledge in the form of frozen encoder weights, which is particularly suitable in the context of limited data availability. Meanwhile, our proposed method offers a unified adaptive method designed for few-shot image recognition tasks in both single-domain and cross-domain scenarios. Our main contributions can be summarized as follows:

- We propose a novel self-prompt mechanism for few-shot image recognition. This mechanism is inspired by the human cognitive process and aims to adaptively tune the network to learn discriminative features according to the self-prompts.
- We have devised a feature selection strategy across spatial and channel dimensions to proficiently generate intrinsic self-prompt information, which is harnessed to guide self-attention computation.
- We propose a novel parameter-efficient tuning method that exclusively fine-tunes the parameters relevant to self-prompt (prompts are no more than 2% of the total parameters), and the incorporation of additional learnable parameters as self-prompt ensures the retention of prior knowledge through frozen encoder weights.
- We have evaluated our proposed self-prompt mechanism for few-shot image recognition (for short, SPM) on both the single-domain and cross-domain benchmark datasets, including Mini-ImageNet, CIFAIR-FS, and CDFSL. SPM achieves promising results, improving the state-of-the-art 1-shot and 5-shot recognition accuracy by 1.97% and 1.45% on average, respectively. Additionally, ablation experiments demonstrate the effectiveness of the proposed feature selection strategy and parameter-efficient tuning scheme.

Related Works

Few-shot image recognition. Few-shot image recognition is a significant subarea within the field of few-shot learning. Unlike common recognition tasks, few-shot image recognition tasks involve a task distribution shift between the training and test sets. Typically, few-shot recognition tasks can be categorized into two different scenarios. The first type is in a single-domain scenario, where a category shift exists between the training and test sets. There are two main streams of learning methods in this scenario, optimization-based and metric-based. For example, as a representative of optimization-based methods, MAML (Finn, Abbeel, and Levine 2017) and its variants (Sun et al. 2019) aim to learn a proficient model initialization capable of swift adaptation to novel classes within a limited number of optimization steps. Alternatively, metric-based methods aim to represent the samples in an appropriate feature space and then calculate the distance between a query and the centroid of a set of support examples (Vinyals et al. 2016; Hu et al. 2022; Afrasiyabi et al. 2022). The second type involves a cross-domain scenario, which is more challenging compared to the single-domain scenario. In addition to the category shift, there is also a domain shift between the training and test

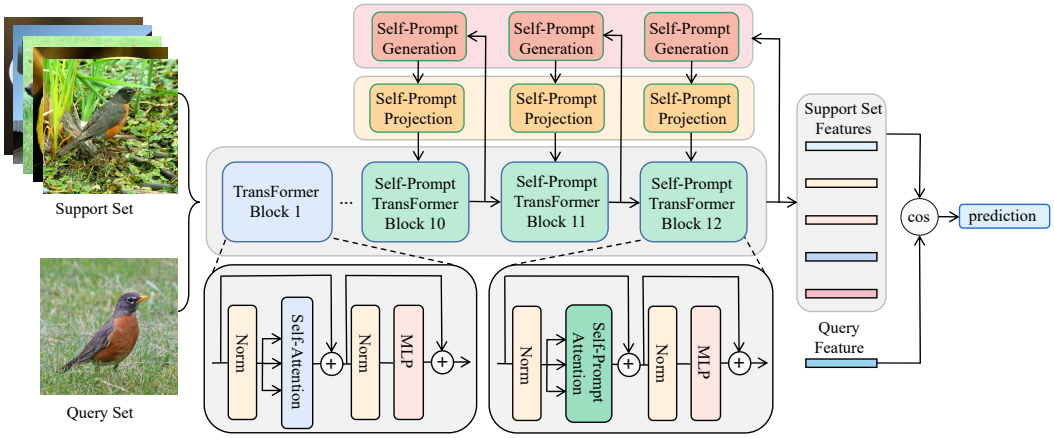

Figure 2: Our proposed self-prompt mechanism generates self-prompt information by selecting deep layer features of the network in spatial and channel dimensions and then transmits this self-prompt information to the deeper layers of the network, thereby modifying the calculation process of self-attention to guide network training.

sets. This kind of task is mainly dealt with through adaptive adjustment of network parameters (Luo, Xu, and Xu 2022; Zhao, Zhang, and Tian 2023; Yi et al. 2023). In contrast to the aforementioned methods, we propose a self-prompt mechanism to adaptively generate features suitable for unseen classes or domains. Our method is applicable not only to few-shot image recognition tasks in single-domain scenarios but also to those in cross-domain scenarios.

Prompt learning. Prompt learning (Liu et al. 2023) has emerged as a highly efficient technique for adapting the Transformer models in the field of computer vision. By incorporating a set of learnable parameters into the input and intermediate representations of a pre-trained model, Transformer can be adapted to specific tasks and domains. Recent works (Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2021) propose to treat the prompts as class-specific continuous vectors and directly optimize them via gradients during fine-tuning. These studies underscore the potential of harnessing the intrinsic capabilities of Transformer to enhance adaptation methods across a broad spectrum of computer vision tasks. Simultaneously, VPT (Jia et al. 2022) introduces learnable tokens at each layer of Transformer, enabling interaction with patch and class tokens. These learnable tokens and the classifier head are jointly optimized to achieve effective adaptation. Additionally, (Chen et al. 2023) employs large language models and leverages new text information to guide the training of visual models for few-shot image recognition. However, these methods often rely on information from other modalities or need to generate additional external prompts to participate in self-attention calculations, leading to increased computational costs due to the quadratic complexity of the self-attention layer. In contrast, our proposed self-prompt mechanism generates intrinsic self-prompt information that is transmitted to the deep layers of Transformer and is parameter-efficient, which only needs to fine-tune a tiny amount of parameters to be suitable for few-shot image recognition tasks in different scenarios.

Method

Overview

The core of the proposed self-prompt mechanism is to adaptively adjust the model parameters according to unseen classes or domains. The complete pipeline of our proposed method is illustrated in Figure 2. First of all, we input support set images and query set images into the model and extract features by Vision Transformer (ViT) (Dosovitskiy et al. 2020). Simultaneously, we perform feature selection on the semantic features from the deep layers of the network, generating self-prompt information. Meanwhile, we propagate the generated self-prompt information to guide the training process of the calculation process of self-attention in the deep layers of the network. It is worth noting that during meta-training, we train the parameters of the image encoder, while during meta-testing, we utilize the proposed parameter-efficient tuning method to fine-tune only a tiny amount of parameters.

The Self-Prompt Mechanism

Self-Prompt Generation and Projection. Humans can be capable of summarizing experiences and lessons learned from previous tasks, enabling them to adjust strategies and correct directions based on past experiences when encountering similar or related tasks (LEE 2002; Yu and Dayan 2004). The proposed self-prompt mechanism simulates this human process by extracting and refining deep layer features of the network and fine-tuning the network in a top-down manner to adapt to different unseen classes. Specifically, we employ a standard ViT model as our backbone and apply the self-prompt mechanism to the last three layers of the Transformer structure. To ensure the accuracy of the self-prompt information extracted, we perform feature selection on both the spatial and channel dimensions of the network. Specifically, we first apply spatial dimension selection to the feature $\boldsymbol{F} \in R^{N \times D}$, where N represents the number of tokens and D represents the dimension of the feature to

which each token is mapped. Subsequently, a learnable spatial prompt vector $\boldsymbol{s} \in R^D$ is trained and normalized, which is then element-wise multiplied with the same normalized deep-layer features $\boldsymbol{F} \in R^{N \times D}$, resulting in vector $\boldsymbol{m}$:

$$\boldsymbol{m} = Norm(\boldsymbol{F}) \times Norm(\boldsymbol{s}). \tag{1}$$

And then, we can obtain the mask vector $\boldsymbol{m} \in R^N$ of the spatial dimension through rounding operation according to the following formula:

$$\boldsymbol{m} = \begin{cases} 0, & \boldsymbol{m}_i \leq 0.5; \\ 1, & \boldsymbol{m}_i > 0.5, \end{cases} \tag{2}$$

where the value of i ranges from 1 to N. Finally, $\boldsymbol{m} \in R^N$ is used to perform mask operation on the deep feature $\boldsymbol{F} \in R^{N \times D}$, thus realizing the feature selection of the spatial dimension. The specific calculation process is as follows:

$$\boldsymbol{F}_s = \boldsymbol{F} \odot \boldsymbol{m}, \tag{3}$$

where $\odot$ is the broadcasted element-wise product and $\boldsymbol{F}_s$ is the feature obtained after spatial selection. Meanwhile, we also define a learnable matrix $\boldsymbol{C} \in R^{N \times D}$, which is multiplied with the matrix after spatial mask processing to select the channel-wise features, thereby generating features obtained after channel selection $\boldsymbol{F}_c$. The specific calculation process is as follows:

$$\boldsymbol{F}_c = \boldsymbol{F}_s \times \boldsymbol{C}. \tag{4}$$

It is noteworthy that both the spatial prompt vector $\boldsymbol{s} \in R^D$ and the channel prompt matrix $\boldsymbol{C} \in R^{D \times D}$ mentioned above are trainable. Moreover, our network is capable of adaptively adjusting the features of different classes or domains, in order to adapt to the few-shot recognition tasks in single-domain or cross-domain scenarios. Furthermore, the varying depths of the network demonstrate specialization in capturing and emphasizing distinct sets of features. The shallow layers primarily emphasize the texture and details of the image, while the deeper layers focus more on semantic information. Meanwhile, for classification tasks, the semantic information contained in the images is of vital importance. As a result, during the training process of the network, we generate self-prompt information for the last three layers of the backbone network.

Self-Prompt Projection. We employ three optional projection methods: identity, linear, and MLP mapping, respectively. Taking MLP projection as an example, the self-prompt projection process is as follows:

$$\boldsymbol{P} = MLP(\boldsymbol{F}_c), \tag{5}$$

where $\boldsymbol{P} \in R^{N \times D}$ is the ultimate self-prompt matrix we generate. By further projection of the self-prompt information, the information guiding the network training process can be adjusted to adapt to unseen classes or domains. It is worth noting that different mapping methods correspond to different amounts of adjustable parameters, we will further introduce the selection of projection methods in the experimental section. In this paper, the identity mapping method is adopted by default.

Self-Prompt Attention. Inspired by the cognitive process of humans, where individuals can adapt their strategies according to different task requirements, we similarly guide the initialization of the query vectors in the self-attention calculation process at deeper network layers. Specifically, we require the network to have knowledge of 'what to query'. Therefore, we reshape the generated self-prompt information $\boldsymbol{P} \in R^{N \times D}$ to match the dimensions of the query vectors, and then add them together. This modification aims to guide the learning process of the network. Concretely, the self-attention calculation (Vaswani et al. 2017) is modified as follows:

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{W}_Q(\boldsymbol{X}+\boldsymbol{P}), \boldsymbol{W}_K\boldsymbol{X}, \boldsymbol{W}_V\boldsymbol{X}, \tag{6}$$

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax(\frac{\boldsymbol{Q}(\boldsymbol{K})^T}{\sqrt{d_k}})\boldsymbol{V}, \tag{7}$$

where $\boldsymbol{X} \in R^{N \times D}$ is the input to the original self-attention calculation process, $\boldsymbol{P} \in R^{N \times D}$ represents the self-prompt matrix generated in the previous section, and $\sqrt{d_k}$ is the scaling factor. By employing self-prompt information as guidance in the training process of self-attention, the network can adaptively adjust its learning process based on the varying feature requirements for new classes or domains. This enables the network to learn purposefully and adapt to few-shot recognition tasks in different scenarios during the learning process.

Training Procedure

Meta-training. We employ the unsupervised pre-training model as the initial weights of the model training. In the meta-training stage, we employ the strategy of episodic training (Snell, Swersky, and Zemel 2017), which simulates the few-shot scenario on the base training dataset. Specifically, we randomly sample K-way-N-shot and Q-queries, for a K-way-N-shot task. Generally, we define

$$\boldsymbol{c}_k = \frac{1}{N_k} \sum_{i:y_i=k} f(\boldsymbol{x}_i), \tag{8}$$

where f is a backbone network, $N_k = \sum_{i:y_i=k}$ is the size of class k in the support set, and $\boldsymbol{c}_k$ is the prototype of class k in the support set. Whereafter, we leverage a softmax function to compute the probability of a query image $\boldsymbol{x}_q$ belonging to class k:

$$p(y=k|\boldsymbol{x}_q) = \frac{exp\big(-d(f(\boldsymbol{x}_q), \boldsymbol{c}_k)\big)}{\sum_i^K exp\big(-d(f(\boldsymbol{x}_q), \boldsymbol{c}_i)\big)}, \tag{9}$$

where K is defined as the number of categories in the support set. Note that, the prototypes can be computed regardless of the value of K. This enables our model to be trained under various-way-various-shot settings. Finally, we update the parameters of the network after computing the cross-entropy loss:

$$\ell_{pre} = -\sum_{\boldsymbol{x}_i \in \boldsymbol{X}_{batch}} log\big(p(y=y_i|\boldsymbol{x}_i)\big), \tag{10}$$

where y_i is the target output corresponding to the instance x_i of the query set.

Method	backbone	1-shot	5-shot
DeepEMD (Zhang et al. 2022)	ResNet-12	65.9	82.4
RS-FSL (Afham et al. 2021)	ResNet-12	65.3	-
PLCM (Huang et al. 2021)	ResNet-12	70.1	83.7
SetFeat12 (Afrasiyabi et al. 2022)	ResNet-12	68.3	82.7
GEL (Wang et al. 2023)	ResNet-12	68.3	83.1
DeepEMD-BERT (Yan et al. 2021)	ResNet-12	67.0	83.7
CNAPS + FETI (Bateni et al. 2022)	ResNet-18	79.9	91.5
EPNet + SSL (Rodríguez et al. 2020)	WRN-28-10	79.2	88.1
SIB (Hu et al. 2020)	WRN-28-10	70.0	79.2
SCR (Wu, Tian, and Zhong 2022)	Swin-T	66.8	83.2
SP-CLIP (Chen et al. 2023)	Visformer-T	72.3	83.4
SUN (Dong et al. 2022)	Visformer-S	67.8	83.3
SP-CLIP* (Chen et al. 2023)	ViT-small	93.4	98.1
PMF (Hu et al. 2022)	ViT-small	93.1	98.0
SPM (Ours)	ViT-small	**93.7**	**98.3**

Table 1: Accuracy (%) of 5-way 1-shot/5-shot setting trained on the Mini-ImageNet dataset. Marked in bold are the best results, * denotes results are reported by us.

Method	backbone	1-shot	5-shot
Relation Networks (Sung et al. 2018)	CNN-4-64	55.0	69.3
R2D2 (Bertinetto et al. 2018)	CNN-4-64	65.3	79.4
SIB (Hu et al. 2020)	CNN-4-64	68.7	77.1
MetaOpt-SVM (Lee et al. 2019)	ResNet-12	72.0	84.3
PLCM (Huang et al. 2021)	ResNet-12	77.6	86.1
GEL (Wang et al. 2023)	ResNet-12	76.7	87.6
SIB (Hu et al. 2020)	WRN-28-10	80.0	85.3
Fine-tuning (Dhillon et al. 2019)	WRN-28-10	76.6	85.8
CC+rot (Gidaris et al. 2019)	WRN-28-10	76.1	87.8
SCR (Wu, Tian, and Zhong 2022)	Swin-T	76.4	88.1
SP-CLIP (Chen et al. 2023)	Visformer-T	82.2	88.3
SUN (Dong et al. 2022)	Visformer-S	78.4	88.8
SP-CLIP* (Chen et al. 2023)	ViT-small	81.9	92.8
PMF (Hu et al. 2022)	ViT-small	81.1	92.5
SPM (Ours)	ViT-small	**82.4**	**93.1**

Table 2: Accuracy (%) of 5-way 1-shot/5-shot setting trained on the CFAIR-FS dataset. Marked in bold are the best results, * denotes results are reported by us.

Parameter-efficient tuning. After meta-training on a training set $\boldsymbol{D}_{train}$, our SPM model is evaluated on unseen data $\boldsymbol{T}_{test}$, with a provided support set $\boldsymbol{S}_{\boldsymbol{T}_{test}}$. Here, we propose a parameter-efficient tuning method by fine-tuning the self-prompt parameters and the corresponding biases of the deeper layers leveraging the support set $\boldsymbol{S}_{\boldsymbol{T}_{test}}$. For this, we simulate episodic meta-learning by randomly partitioning the support set into a sub-support set $\boldsymbol{S}^*$ and a sub-query set $\boldsymbol{Q}^*$, such that $\boldsymbol{S}_{\boldsymbol{T}_{test}} = \boldsymbol{S}^* \cup \boldsymbol{Q}^*$. The process of parameter-efficient tuning in $\boldsymbol{S}_{\boldsymbol{T}_{test}}$ can be expressed as the following formula:

$$\min_{\theta_T} E_{\boldsymbol{S}^*, \boldsymbol{Q}^* \sim \boldsymbol{S}_{\boldsymbol{T}_{test}}} \Big[\frac{1}{\boldsymbol{Q}^*} \sum_{\boldsymbol{X}_q, \boldsymbol{Y}_q \in \boldsymbol{Q}^*} L(\boldsymbol{Y}_q, f(\boldsymbol{X}_q; \boldsymbol{S}^*)) \Big], \quad (11)$$

where θ_T denotes self-prompt parameters and the corresponding biases of the deeper layers, and the calculation process of loss L is consistent with the process described in the meta-training phase as shown in Equations 8, 9, and 10. Meanwhile, the portion of parameters to be fine-tuned is negligible (no more than 2% of the total parameters), so that SPM can quickly and adaptively adjust on the small support set $\boldsymbol{S}_{\boldsymbol{T}_{test}}$. After fine-tuning, SPM is evaluated by predicting the label of unseen query images using the support set $\boldsymbol{S}_{\boldsymbol{T}_{test}}$. Furthermore, the incorporation of additional learnable parameters as self-prompt ensures the retention of prior knowledge through frozen encoder weights and learn discriminative features according to unseen data. Therefore, our method is highly suited for few-shot recognition tasks that require both information retention and adaptive adjustment of network parameters within limited labeling data constraints.

Experiments

Experimental Settings

Single-domain datasets. We employ two standard benchmarks to evaluate our proposed SPM method, including Mini-ImageNet (Vinyals et al. 2016) and CIFAR-FS (Bertinetto et al. 2018). Mini-ImageNet contains 100 classes, which is divided into 64 classes for training, 16 for validation, and 20 for testing. CIFAR-FS is a few-shot image recognition dataset built on CIFAR100. We follow the split division proposed by (Hu et al. 2022), where the dataset is divided into 64 classes for training, 16 for validation, and 20 for testing. Each class comprises 100 images.

Cross-domain datasets. We conduct extensive experiments under cross-domain settings, using four few-shot image recognition datasets: CropDiseases, EuroSAT, ISIC, and ChestX, which are introduced by (Guo et al. 2020). Each dataset consists of train/val/test splits. We employ the Mini-ImageNet domain as the single source domain, and select the model parameters with the best accuracy on the validation set of the Mini-ImageNet for model evaluation, where 4 out-of-domain datasets are considered. The results are reported under 5-way 1/5/20-shot settings.

Training details. We employ ViT as our backbone network, and the backbone is trained for 20 epochs using ViT-small and 80 epochs using ViT-base, each epoch consisting of 2000 episodes. Our learning rate schedule incorporates warm-up and cosine annealing, with the learning rate commencing at 10^{-6}, surging to 5×10^{-5} in 5 epochs, and gradually tapering off to 10^{-6} via cosine annealing. In order to attain the finest test outcomes, we utilize the early stop strategy to train our model. We use a single Nvidia GeForce 4090 for all the experiments.

Comparison With the State-of-the-Art

Table 1 and Table 2 present a comparison between our proposed SPM and other state-of-the-art methods on single-domain datasets. Our method surpasses the current state-of-the-art methods both on 1-shot and 5-shot settings. Specifically, we not only outperform the traditional CNN-based methods but also outperform recent Transformer-based methods. In particular, we modify the backbone of the method that utilized text prompts for guidance (Chen et al.

	ChextX			ISIC			EuroSAT			CropDisease		
	1-shot	5-shot	20-shot	1-shot	5-shot	20-shot	1-shot	5-shot	20-shot	1-shot	5-shot	20-shot
ProtoNet+FWT (Tseng et al. 2020) (RN10)	-	23.77	26.87	-	38.87	43.78	-	67.34	75.74	-	72.72	85.82
RelationNet (Sung et al. 2018) (RN10)	21.95	22.96	26.63	30.53	39.41	41.77	49.08	61.31	74.43	53.58	68.99	80.45
MetaOptNet (Lee et al. 2019) (RN10)	-	22.53	25.53	-	36.28	49.42	-	64.44	79.19	-	68.41	82.89
STARTUP (Phoo and Hariharan 2021) (RN10)	-	26.94	33.19	-	47.22	58.63	-	82.29	89.26	-	93.02	97.51
FT-All (Guo et al. 2020) (RN10)	-	25.97	31.32	-	48.11	59.31	-	79.08	87.64	-	89.25	95.51
TPN+ATA (Wang and Deng 2021) (RN10)	22.45	24.74	-	**35.55**	49.83	-	70.84	85.47	-	82.47	93.56	-
DeepCluster2 (Caron et al. 2020) (RN50)	-	26.51	31.51		40.73	49.91	-	88.39	92.02	-	93.63	96.63
PMF (Hu et al. 2022) (RN50)	-	27.13	31.57	-	43.78	54.06	-	89.18	93.08	-	95.06	97.25
PMF (Hu et al. 2022) (ViT-small)	21.73	27.27	35.33	30.63	50.12	63.78	70.74	85.98	91.32	80.79	92.96	98.12
StyleAdv-FT (Fu et al. 2023) (ViT-small)	22.92	26.97	-	33.99	**51.23**	-	74.93	**90.12**	-	84.11	95.99	-
SPM (ViT-small)	**22.96**	**27.35**	**35.71**	31.45	50.95	**63.91**	**74.97**	89.72	**94.30**	**84.43**	**96.11**	**98.32**

Table 3: Comparison on cross-domain few-shot image recognition. We train the models on Mini-ImageNet, and evaluate them on CDFSL. Marked in bold are the best results in each block.

SS	CS	PT	SPA	single-domain		cross-domain	
				Mini	CIFAR-FS	EuroSAT	CropDisease
✗	✗	✗	✗	93.09	81.10	70.74	80.79
✓	✗	✓	✓	93.26	81.95	71.31	81.39
✗	✓	✓	✓	93.35	82.02	73.87	83.66
✓	✓	✗	✓	93.22	81.54	71.17	81.22
✓	✓	✓	✓	**93.72**	**82.42**	**74.97**	**84.43**

Table 4: Ablation study on four datasets under the 5-way 1-shot setting. SS means spatial selection, CS means channel interaction, PT means parameter-efficient tuning, and SPA means self-prompt attention.

Projection method	Mini-Imagenet		CFAIR-FS	
	meta-train	meta-test	meta-train	meta-test
MLP	92.45	93.25	76.17	81.22
Linear	92.10	93.35	75.47	81.31
Identity	91.98	**93.72**	75.29	**82.42**

Table 5: Different projection methods correspond to train and test results. We report 5-way 1-shot accuracy (%) on the validation set and test set of Mini-ImageNet and CIFAR-FS during the meta-training process and meta-testing process.

2023) to conduct a fair comparison. Experimental results indicate that utilizing text information as prompts brings certain performance improvements. However, due to the modality gap between textual and visual information, its effectiveness is suboptimal compared to our proposed SPM model(as shown in Tables 1, 2). SPM can selectively harness the intrinsic information within the pre-trained visual model, enabling performance enhancement of the network without introducing extra modality information.

In addition, we conduct experiments on four cross-domain datasets. The results are shown in Table 3, where we achieve an average performance improvement of 3.65% and 2.33% on the EuroSAT and CropDisease datasets, demonstrating the ability of our proposed method that can adapt to the task in different scenarios. Furthermore, we only obtain an average performance improvement of 0.56% and 0.59%

Model	Params	FLOPs	Fine-tuning	EuroSAT	CropDisease
PMF (ViT-small)	21.66M	1×	21.6M (100%)	70.74	80.79
PMF (ViT-base)	85.80M	4×	21.6M (100%)	72.83	82.15
SPM (ViT-small)	21.96M	1.2×	0.30M(1.34%)	**74.97**	**84.43**

Table 6: Compared with the number of parameters and computation of PMF, the proposed SPM method improves the performance of the model.

on the Chext and ISIC datasets. This is because that both the Chext and ISIC datasets belong to the medical field, which significantly differs from the source domain. Therefore, due to the constraints of inherent information within the pre-trained visual model, our proposed SPM, while harnessing the inherent information from both the source domain and the pre-trained visual model, still results in marginal performance improvements. In general, these findings provide substantial evidence supporting the effectiveness of our proposed SPM method.

Model Analysis

Ablation study. The results of the ablation study are presented in Table 4. We conduct ablation experiments on SS (spatial selection), CS (channel selection), PT (parameter-efficient tuning), and SPA (self-prompt attention) in both single-domain and cross-domain scenarios. As shown in Table 4, the SS and CS components of the self-prompt generation module demonstrate certain effects, leading to an increase of 0.55% and 1.80% in accuracy under the 1-shot setting, respectively. It is worth noting that CS improves by an average of 3% in cross-domain scenarios, which is consistent with the results obtained in (Luo, Xu, and Xu 2022). Meanwhile, PT also provides a performance boost compared to fine tuning all parameters. Furthermore, the combination of SPA with SS, SC, and PT leads to a further improvement in the model performance, thereby indicating the efficacy of our proposed SPM in the task of few-shot recognition.

Layer selection. Theoretically, self-prompt information can guide the model to learn more discriminative features.

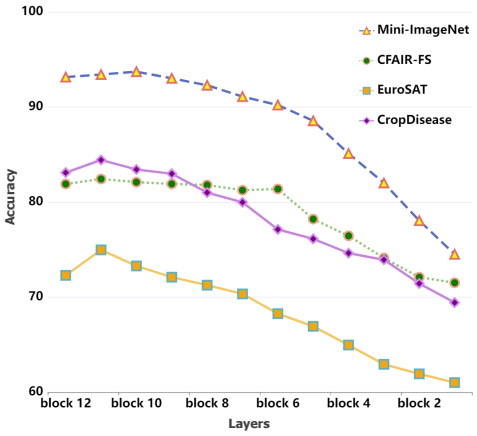

Figure 3: Accuracy vs. different layers to apply SPM. We report the results of two single-domain datasets and two cross-domain datasets respectively under 5-way 1-shot settings.

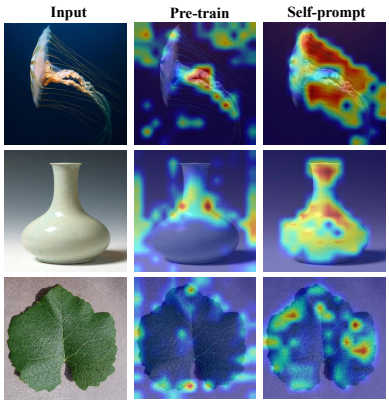

Figure 4: Visualization of attention maps for different domains/tasks: the first two rows depict images from the single-domain dataset, while the third row showcases visualizations from the cross-domain dataset.

Our proposed self-prompt mechanism can be flexibly applied into various layers of Transformer. However, we observe that the depth at which the self-prompt mechanism is applied significantly impacts the model performance. As shown in Figure 3, we apply the self-prompt mechanism to different layers of Transformer and found that applying it to deeper layers achieves better results, with slight variation in the optimal layers for different datasets. We attribute this result to the crucial role of high-level semantic information in few-shot recognition tasks, whereas shallow layers lack such semantic information. Consequently, incorporating the self-prompt mechanism into deeper layers of the network improves the model performance. To simplify architecture design, we default to inserting the self-prompt mechanism into the last three layers of the model.

Self-prompt projection selection. We attempt three different ways to map the generated self-prompt information: identity, linear, and MLP. The experimental results are shown in Table 5. When employing MLP as the mapping method, it achieves higher accuracy during the meta-training process, but results in suboptimal performance during the fine-tuning (meta-testing) process. However, despite exhibiting lower accuracy on the validation set during the meta-training process compared to MLP, the identity mapping method yields superior accuracy during the fine-tuning process. We argue that in the testing process of the few-shot learning tasks, there is limited labeled data available for fine-tuning, and MLP mapping would increase the number of parameters that require adjustment during the fine-tuning process, resulting in the model inability to adapt effectively to unseen data. Therefore, controlling the number of parameters during the fine-tuning phase is of paramount importance, which also indirectly underscores the significance of the parameter efficiency of our proposed SPM.

Computation cost analysis. We analyze the complexity of our model. As shown in Table 6, due to the fact that our proposed SPM method solely entails learning self-prompt information, it results in a tiny parameter overhead. Additionally, there is a slight computational cost associated with the secondary self-prompt forward process. However, the PMF (Hu et al. 2022) method requires fine-tuning of all model parameters, whereas our proposed self-prompt mechanism and parameter-efficient tuning method only requires adjustment of self-prompt related parameters and corresponding bias parameters in deep layers, resulting in a significantly smaller number of parameters that need to be fine-tuned. Consequently, our proposed SPM method reduces the computational overhead of the fine-tuning process, and it still outperforms fine-tuned PMF (ViT-base) which has four times as many Params and FLOPs as SPM. Hence, our proposed self-prompt mechanism not only maintains parameter efficiency but also enhances model performance, making it more suitable for few-shot recognition tasks.

Visualization To demonstrate the adaptive adjustment of our proposed method for different few-shot recognition tasks, in this section, we visualize the attention maps by computing the dot product between the output feature and the feature vector at each location. As shown in Figure 4, the pre-trained models are cluttered with background information. However, our method can focus on intrinsic semantic features according to the self-prompt mechanism.

Conclusion

In this paper, we propose a novel method called the self-prompt mechanism (SPM) for few-shot learning, which generates and utilizes the intrinsic semantic features to steer the network toward the adaptation of unseen data. Meanwhile, we propose a novel parameter-efficient tuning method enhancing the model ability to extract discriminative features through fine-tuning a tiny amount of parameters. The effectiveness of the proposed scheme is evaluated on both single-domain and cross-domain datasets. Moreover, we hope that our proposed self-prompt mechanism could inspire and facilitate follow-up works with potential. We will further investigate efficient self-prompt methods in our future work.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, HY Project under Grant No. LZY2022033004, the Natural Science Foundation of Shandong Province under Grants No. ZR2020MF131 and No. ZR2021ZD19, Project of the Marine Science and Technology cooperative Innovation Center under Grant No. 22-05-CXZX-04-03-17, the Science and Technology Program of Qingdao under Grant No. 21-1-4-ny-19-nsh, and Project of Associative Training of Ocean University of China under Grant No. 202265007.

References

Afham, M.; Khan, S.; Khan, M. H.; Naseer, M.; and Khan, F. S. 2021. Rich semantics improve few-shot learning. *arXiv preprint arXiv:2104.12709*.

Afrasiyabi, A.; Larochelle, H.; Lalonde, J.-F.; and Gagné, C. 2022. Matching Feature Sets for Few-Shot Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9014–9024.

Antoniou, A.; Edwards, H.; and Storkey, A. 2018. How to train your MAML. *arXiv preprint arXiv:1810.09502*.

Baifeng, S.; Trevor, D.; and Xin, W. 2023. Top-Down Visual Attention from Analysis by Synthesis.

Bateni, P.; Barber, J.; van de Meent, J.-W.; and Wood, F. 2022. Enhancing few-shot image classification with unlabelled examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2796–2805.

Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2018. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Le Centre pour la Communication Scientifique Directe - HAL - Université Paris Descartes,Le Centre pour la Communication Scientifique Directe - HAL - Université Paris Descartes*.

Chen, W.; Si, C.; Zhang, Z.; Wang, L.; Wang, Z.; and Tan, T. 2023. Semantic Prompt for Few-Shot Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23581–23591.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*.

Dhillon, G. S.; Chaudhari, P.; Ravichandran, A.; and Soatto, S. 2019. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*.

Dong, B.; Zhou, P.; Yan, S.; and Zuo, W. 2022. Self-Promoted Supervision for Few-Shot Transformer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Fleming, S. M.; and Dolan, R. J. 2012. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594): 1338–1349.

Fu, Y.; Xie, Y.; Fu, Y.; and Jiang, Y.-G. 2023. Meta Style Adversarial Training for Cross-Domain Few-Shot Learning.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2019. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8059–8068.

Guo, Y.; Codella, N. C.; Karlinsky, L.; Codella, J. V.; Smith, J. R.; Saenko, K.; Rosing, T.; and Feris, R. 2020. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, 124–141. Springer.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9068–9077.

Hu, S. X.; Moreno, P. G.; Xiao, Y.; Shen, X.; Obozinski, G.; Lawrence, N. D.; and Damianou, A. 2020. Empirical bayes transductive meta-learning with synthetic gradients. *arXiv preprint arXiv:2004.12696*.

Huang, K.; Geng, J.; Jiang, W.; Deng, X.; and Xu, Z. 2021. Pseudo-loss confidence metric for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8671–8680.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10657–10665.

LEE, T. 2002. Top-down influence in early visual processing: a Bayesian perspective. *Physiology amp; Behavior*, 645–650.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Li, W.-H.; Liu, X.; and Bilen, H. 2022. Cross-domain Few-shot Learning with Task-specific Adapters.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 1–35.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. *Cornell University - arXiv,Cornell University - arXiv*.

Luo, X.; Xu, J.; and Xu, Z. 2022. Channel Importance Matters in Few-Shot Image Classification.

Munkhdalai, T.; Yuan, X.; Mehri, S.; and Trischler, A. 2018. Rapid adaptation with conditionally shifted neurons. In *International Conference on Machine Learning*, 3664–3673. PMLR.

Phoo, C.; and Hariharan, B. 2021. Self-training For Few-shot Transfer Across Extreme Task Differences. *International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training.

Rodríguez, P.; Laradji, I.; Drouin, A.; and Lacoste, A. 2020. Embedding propagation: Smoother manifold for few-shot classification. In *European Conference on Computer Vision*, 121–138. Springer.

Salles, A.; Ais, J.; Semelman, M.; Sigman, M.; and Calero, C. I. 2016. The metacognitive abilities of children and adults. *Cognitive Development*, 40: 101–110.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Sun, Q.; Liu, Y.; Chua, T.-S.; and Schiele, B. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 403–412.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; and Yang, M.-H. 2020. Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation. *Cornell University - arXiv,Learning*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wang, H.; and Deng, Z.-H. 2021. Cross-domain few-shot classification via adversarial task augmentation. *arXiv preprint arXiv:2104.14385*.

Wang, H.; Pang, G.; Wang, P.; Zhang, L.; Wei, W.; and Zhang, Y. 2023. Glocal Energy-based Learning for Few-Shot Open-Set Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7507–7516.

Wu, J.; Tian, X.; and Zhong, G. 2022. Supervised Contrastive Representation Embedding Based on Transformer for Few-Shot Classification. In *Journal of Physics: Conference Series*, volume 2278, 012022. IOP Publishing.

Yan, K.; Bouraoui, Z.; Wang, P.; Jameel, S.; and Schockaert, S. 2021. Aligning Visual Prototypes with BERT Embeddings for Few-Shot Learning. *Cornell University - arXiv,Cornell University - arXiv*.

Yi, Q.; Zhang, R.; Peng, S.; Guo, J.; Gao, Y.; Yuan, K.; Chen, R.; Lan, S.; Hu, X.; Du, Z.; Zhang, X.; Guo, Q.; and Chen, Y. 2023. Online Prototype Alignment for Few-shot Policy Transfer.

Yu, A.; and Dayan, P. 2004. Inference, Attention, and Decision in a Bayesian Neural Architecture. *Neural Information Processing Systems,Neural Information Processing Systems*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2022. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, Y.; Zhang, T.; and Tian, Y. 2023. Dual Adaptive Representation Alignment for Cross-Domain Few-Shot Learning.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement.