MMTP: Meta-learning-based Multi-Textual Prompt Tuning for Visual-Language Models

Fangtong Sun†, Junjie Zhu†, Zunlin Fan, Yiying Li*, Zhiyuan Wang*, Ke Yang*

Intelligent Game and Decision Lab (IGDL), Beijing, 100091, China
Academy of Military Science, Beijing, 100091, China

***Abstract*—Pre-trained Visual-Language Models (VLMs) have demonstrated powerful performance on various downstream tasks. Recently, many prompt tuning methods represented by Context Optimization (CoOp) have effectively adapted VLMs to few-shot tasks. However, the CoOp-based methods suffer from overfitting to base classes, which impairs the model's generalization to new classes. Considering that meta-learning excels at generalizing to new classes, we combine meta-learning with CoOp-like vision-language model fine-tuning methods to improve performance on few-shot generation tasks. In this paper, we present a novel Meta-learning-based Multi-Textual Prompt tuning (MMTP) method, which learns multiple textual prompts leveraging meta-learning to enhance the visual-language model's representation and generalization capabilities. Specifically, we introduce multi-textual prompts to enhance the representation of the model for improving the recognition of base classes. Simultaneously, we employ meta-learning to optimize prompt training, bolstering the model's generalization to new classes. Extensive experiments demonstrate the superiority of our method under base-to-new generalization and cross-domain generalization settings. Furthermore, we also conduct ablation studies to validate the effectiveness of each component.**

***Index Terms*—prompt tuning, meta-learning, few-shot tasks, visual-language models**

I. Introduction

Recently, large-scale pre-trained Visual-Language Models (VLMs), such as CLIP [1] and ALIGN [2], have exhibited impressive performance across a variety of downstream tasks. However, some particular tasks like few-shot tasks [3] are still universally acknowledged as highly challenging due to the limited number of samples available. Additionally, limitations in computational resources make it difficult for ordinary users to adapt VLMs to few-shot tasks. Fortunately, researchers have proposed many parameter-efficient fine-tuning techniques [4]–[6], among which prompt tuning stands out as a straightforward and effective approach [7]–[9].

Prompt tuning is initially introduced in the field of natural language processing [10]–[12], and gradually applied to fine-tune visual-language models like CLIP. In practice, the original prompt of CLIP proves to be inaccurate because the hand-crafted template lacks the necessary context information of the current task. To this end, prompt tuning methods like Context Optimization (CoOp) [13], have emerged as a new paradigm, employing a modest set of learnable vectors (soft prompts) instead of manual prompts (Fig.1(a)). The soft prompts are learned from a few training samplers with the pre-trained parameters fixed, which achieves significant improvement.

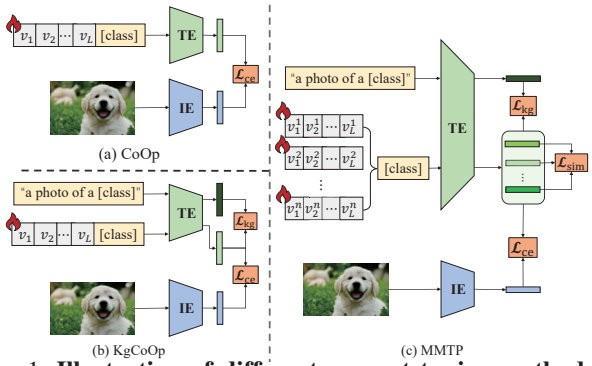

Fig. 1: **Illustration of different prompt tuning methods.** (a)CoOp [13], (b)KgCoOp [14] and (c)our MMTP. Our method is based on KgCoOp and introduces a new constraint of multiple prompts. (TE: Text Encoder, IE: Image Encoder)

However, the learned prompts are prone to overfitting to the base classes, which hampers their generalization ability when applied to new classes, particularly on few-shot tasks. Recently, several methods have focused on tackling this issue. Building upon CoOp, CoCoOp [15] incorporates the image-conditional feature as a prompt bias to improve the generalizability to new classes. ProGrad [16] only updates the aligned prompt, thereby alleviating the general knowledge forgetting. Furthermore, KgCoOp [14] preserves the original CLIP generalization ability by introducing an auxiliary loss to reduce the difference between the learnable prompts and the hand-crafted prompts. Unfortunately, these methods still exhibit limitations in generalizing to new classes, with several performing less effectively than the zero-shot CLIP model. How to balance the performance between the base and new classes on few-shot tasks remains a significant challenge.

To address this challenge, in this paper, we propose a novel **M**eta-learning-based **M**ulti-**T**extual **P**rompt tuning (MMTP) method for visual-language models as shown in Fig.1(c). MMTP learns multiple textual prompts leveraging meta-learning to enhance the model's performance on few-shot tasks. Specifically, considering the single prompt struggles to comprehensively depict the image content and descriptions of same category can be elaborated from different perspectives,

† Equal Contributions. * Corresponding Authors. This work is supported by the National Natural Science Foundation of China under Grants 62006241, 62106280, and 62206307.

we introduce multiple learnable prompts to enhance the representation of the model. Meanwhile, we employ the meta-learning to optimize the prompt parameters. Meta-learning acquires better parameter initialization in a learn-to-learn manner, thereby improving the model's generalization capability to adapt to new tasks.

We have conducted experiments in different settings. In the base-to-new class generalization setting, our method improves the harmonic mean accuracy of CoOp and KgCoOp by 8.44% and 3.1%, respectively. In the cross-domain generalization setting, our method attains the best average performance compared with others. Additionally, we conduct ablation studies to substantiate the effectiveness of our method.

Overall, our paper makes the following key contributions:

- We propose a novel prompt tuning method for the visual-language model by incorporating multi-textual prompt and meta-learning optimization, which greatly improves the performance both on base and new classes.
- We evaluate our method on various few-shot generalization tasks. Experiment results show MMTP achieves competitive performance among all compared approaches.

II. Method

A. Preliminaries

Existing CoOp-based methods are proposed based on Contrastive Language-Image Pre-training (CLIP) [1] which incorporates two types of encoders: image encoder $f(\cdot)$ and text encoder $g(\cdot)$. CLIP produces the textual embedding $\mathbf{w}_i^{clip} = g(p_i^{clip})$, where p_i^{clip} is the hand-crafted prompt as "a photo of a $[class_i]$", and the class token is the specific class name.

Different from the manual prompt of CLIP, soft prompt represented by CoOp [13] is $p_i^{coop} = \{v_1, v_2, ..., v_L, class_i\}$, where $class_i$ is the *i*-th class name and each vector $v_j (j \in \{1, 2, ..., L\})$ has the same dimension. The predicted probability of image $\mathbf{x}$ for class *i* is formulated as

$$\mathbf{P}(y=i|\mathbf{x}) = \frac{\exp(\mathbf{sim}[f(\mathbf{x}), g(p_i^{coop})]/\tau)}{\sum_{j=1}^{N_c} \exp(\mathbf{sim}[f(\mathbf{x}), g(p_j^{coop})]/\tau)}, \quad (1)$$

where $\mathbf{sim}[\cdot,\cdot]$ denotes the cosine similarity, N_c is the number of classes, and τ is the temperature parameter learned by CLIP.

B. Multi-Textual Prompt Tuning

In order to arouse better representation of the model, we propose a multi-textual prompt tuning method. The multi-textual prompt is formulated as

$$\mathbf{p}_i^{mmtp} = \{v_1, v_2, ..., v_L, class_i\}^{N_p}, \quad (2)$$

where L is the length of a prompt and N_p is the number of multiple prompts. These learnable context vectors $\{v_1, v_2, ..., v_L\}^{N_p}$ are shared with all classes. Here, for the n_p-th learnable prompt, we can specific it as

$$p_i^{mmtp\_n_p} = \{\{v_1, v_2, ..., v_L\}_{n_p}, class_i\}, \quad (3)$$

The textual embeddings based on each prompt are represented as $\mathbf{w}_i^{mmtp\_n_p} = g(p_i^{mmtp\_n_p})$, so the formula (1) can be rewritten as

$$\mathbf{P}^{n_p}(y=i|\mathbf{x}) = \frac{\exp(\mathbf{sim}[f(\mathbf{x}), \mathbf{w}_i^{mmtp\_n_p})]/\tau)}{\sum_{j=1}^{N_c} \exp(\mathbf{sim}[f(\mathbf{x}), \mathbf{w}_j^{mmtp\_n_p})]/\tau)}, \quad (4)$$

where $n_p \in \{1, 2, ..., N_p\}$, $\mathbf{P}^{n_p}(y=i|\mathbf{x})$ represents the predicted probability of image $\mathbf{x}$ for class i under n_p-th set of prompts.

We have improved the loss function of CoOp-based method to adapt on our multi-textual prompt tuning method. Formally, the Cross-Entropy loss which applied to calculate the discrepancy of multiple category prediction $\mathbf{P}$ and label y, is calculated using all N_P prompts as

$$\mathcal{L}_{cls} = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathcal{L}_{CE}(\mathbf{P}^i, y). \quad (5)$$

To further promote the diversity description of multiple learnable prompts, we design similarity regularization constraints between multiple sets of category predictions:

$$\mathcal{L}_{sim} = \frac{1}{C_{N_p}^2} \sum_{i=1}^{N_p} \sum_{j=i+1}^{N_p} \mathbf{sim}[\mathbf{P}^i, \mathbf{P}^j], \quad (6)$$

where C is the abbreviation of $Combination$, thus $C_{N_p}^2 = N_p(N_p-1)/2$.

Inspired by KgCoOp [14], we also use $\mathcal{L}_{kg}$ to reserve original knowledge from manual CLIP prompt as a regularization:

$$\mathcal{L}_{kg} = \frac{1}{N_c \cdot N_p} \sum_{i=1}^{N_c} \sum_{n_p=1}^{N_p} ||\mathbf{w}_i^{mmtp\_n_p} - \mathbf{w}_i^{clip}||_2^2, \quad (7)$$

Overall, the loss for optimization of prompt tuning is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{sim} + \lambda_2 \mathcal{L}_{kg}, \quad (8)$$

where λ_1 and λ_2 are loss balancing hyper-parameters.

In the inference phase, we get multiple category predictions of the image, and the final category prediction is calculated as the mean value of multiple predictions:

$$\mathbf{P} = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{P}^i. \quad (9)$$

C. Meta-learning Optimization

We first initialize the multi-textual prompt ($\{v_1, v_2, ..., v_L\}^{N_p}$) via contrastive learning between image and text while keeping the backbone frozen. Specifically, we design a novel batch sampling method with random number of classes in each batch. Such sampling method exposes the model confronting with new tasks in each iteration, aiming to learn better initialization.

Subsequently, we construct meta-tasks $\mathcal{T}$ which denote the set of M subtask τ_i with training dataset$\{(D_{sup}, D_{que})^{(i)}\}_{i=1}^{M}$, where each task has both support and query data. The classes in the support and query data are non-overlapping. D_{sup} and D_{que} are randomly split from the base class dataset D_{base} in each iteration. D_{sup} is regarded as the data in simulated base classes, and D_{que} is data in simulated new classes. We adopt a bi-level optimization [17] method to enhance the generalization to new class tasks, this meta-learning process in each meta-task can be formalized as

$$\begin{aligned} \theta^* &= \arg\min_{\theta'} \mathcal{L}(\theta'; D_{que}) \\ s.t. \quad \theta' &= \arg\min_{\theta} \mathcal{L}(\theta; D_{sup}), \end{aligned} \quad (10)$$

where we refer θ as the simplification of the learnable $\{v_1, v_2, ..., v_L\}^{N_p}$ and the calculation of $\mathcal{L}$ is as Eq.(8). Each optimization iteration has the inner loop and outer loop. In the inner loop, the parameters are updated on D_{sup} and the computation graph of network is remained. Then in the outer loop, we hope the parameters outputted from the inner loop

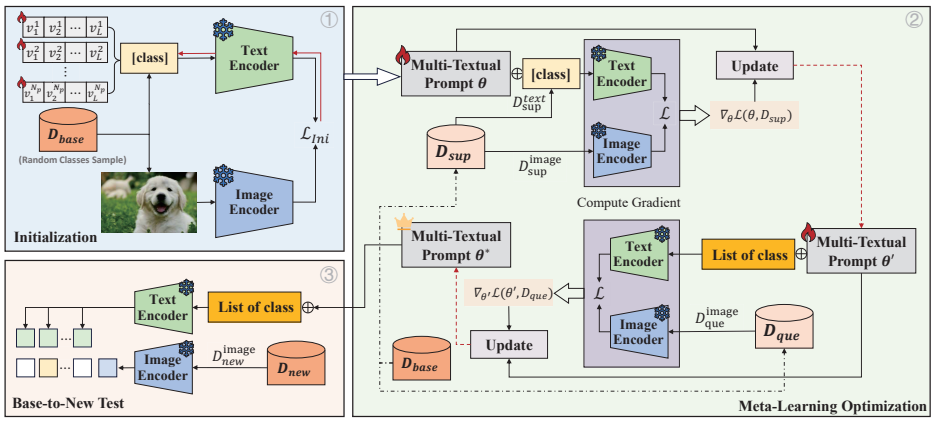

Fig. 2: **An overview of our proposed MMTP.** First, we use contrastive learning to initialization by random classes sample on base class. Subsequently, the model learns better initialization and undergoes fine-tuning through meta-learning optimization. Finally, we test the model's generalization on new classes.

(θ') can have satisfied performance on D_{que}, and finally get the optimized θ^*.

The overall meta-learning process is to ask itself:"Did the parameter learned on simulated base classes can help to improve the generalization on simulated new classes?" We use it as meta-objective to update θ as:

$$\min_\theta \sum_{\tau_i \in \mathcal{T}} \mathcal{L}(\theta - \alpha \nabla_\theta \mathcal{L}_{\tau_i}(\theta, D_{sup}), D_{que}). \quad (11)$$

After the meta-learning optimization, the model can be extended to tasks with new classes. The details are in Alg.1.

Algorithm 1 Meta-learning Optimization

Input: Meta-tasks $\mathcal{T}$; Learning rate α and β
Output: Prompt parameters θ^*

Initialize θ with contrastive learning
while no converge or reach max steps **do**
  Sample batch of tasks τ_i from $\mathcal{T}$
  for all τ_i **do**
    Sample mini-batch d_{sup}^i in D_{sup}^i from τ_i
    Compute $\mathcal{L}_{\tau_i}$ based on d_{sup}^i according to Eq.(8)
    Update $\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\tau_i}(\theta, d_{sup}^i)$
    Sample mini-batch d_{que}^i in D_{que}^i from τ_i
  end for
  Update $\theta^* \leftarrow \theta - \beta \nabla_\theta \sum_{\tau_i \in \mathcal{T}} \mathcal{L}_{\tau_i}(\theta_i', d_{que}^i)$
end while

III. Experiments

A. Experimental Setup

We validate the effectiveness of MMTP on base-to-new generalization and cross-domain generalization setting.

Datasets. For base-to-new generalization, we evaluate on 11 benchmarks datasets: ImageNet [18], Caltech [19], OxfordPets [20], StanfordCars [21], Flowers [22], Food101 [23], FGVCAircraft [24], EuroSAT [25], UCF101 [26], DTD [27], and SUN397 [28]. For cross-domain generalization [29], we use the ImageNet as the source domain and its variants (ImageNet-V2 [30], ImageNet-Sketch [31], ImageNet-A [32] and ImageNet-R [33]) as the target domains.

Baselines. We use existing CoOp-based methods for comparison, *e.g.*, CoOp [13], CoCoOp [15], ProGrad [16], KgCoOp [14], PLOT [34], LASP [35], MaPLe [36] and TCP [37]. Recently, there have been some excellent works, but they have introduced other large models like ChatGPT [38] to get additional knowledge, so we do not report them in the results(*e.g.*, PromptKD [39], CoPrompt [40], HPT [41]).

Training details. Our implementation is based on KgCoOp's code[1]. We use ViT-B/16 CLIP as the backbone model and report the accuracy averaged over 3 runs. The prompt's length L is set as 8 and the number of prompts N_p is 3. λ_1 and λ_2 are set to 1.0 and 8.0. In meta-learning optimization, the number of meta task is 25. We randomly divide the support set and the query set on the base classes data in a 1:1 ratio during each iteration. α and β are set to 1.0e-3 and 3.5e-3. Other settings are maintained consistent with KgCoOp. We rent HPS with RTX 4090 as the experiment platform.

B. Base-to-New Class Generalization

The performance for base-to-new class generalization on 11 image recognition datasets with 16-shot samples is shown in Table I. MMTP demonstrates superior performance over all methods in the harmonic mean of base and new classes, achieving the highest accuracy of 80.10%. Moreover, MMTP also performs best on new classes on 7 out of 11 datasets, and improves the average accuracy from 76.11% to 76.92% surpassing LASP for new classes. The results prove MMTP improves base-to-new generalization under few-shot setting.

C. Cross-Domain Generalization

To further verify the generalization of MMTP, we conduct the cross-domain generalization experiment. Only minor modifications were made to the base-to-new experiment setting, we train the model with all classes of ImageNet in D_{sup} and D_{que} in 16-shot during meta-learning and test on four variants of ImageNet dataset. Compared to the basic method, our method is better able to adapt to various domain through the "learn to learn" approach. We consider the diverse styles of ImageNet which inherently possesses some shifting data help for the domain generalization. In Table II, MMTP generalizes to cross-domain data with the best average accuracy 60.41%.

[1]https://github.com/htyao89/KgCoOp

TABLE I: **Comparison with existing methods on base-to-new generalization.** The best results are in bold and the second-best results are underlined. (HM:Harmonic Mean)

Dataset	Sets	CoOp (IJCV22)	CoCoOp (CVPR22)	ProGrad (ICCV23)	KgCoOp (CVPR23)	PLOT (ICLR23)	LASP (CVPR23)	MaPLe (CVPR23)	TCP (CVPR24)	MMTP (Ours)
Average	Base	82.69	80.47	82.48	80.73	83.98	83.18	82.28	**84.13**	83.56
	New	63.22	71.67	70.75	73.60	71.72	76.11	75.14	75.36	**76.92**
	HM	71.66	75.83	76.16	77.00	77.37	79.48	78.55	79.51	**80.10**
ImageNet	Base	76.47	75.98	77.02	75.83	77.30	76.25	76.66	77.27	**77.56**
	New	67.88	70.43	66.66	69.66	69.37	**71.17**	70.54	69.87	69.80
	HM	71.92	73.10	71.46	72.78	73.40	**73.62**	73.47	73.38	73.47
Caltech101	Base	98.00	97.96	98.02	97.72	**98.53**	98.17	97.74	98.23	98.45
	New	89.81	93.81	93.89	94.39	92.80	94.33	94.36	94.67	**95.30**
	HM	93.73	95.84	95.91	96.03	95.58	96.21	96.02	96.42	**96.85**
OxfordPets	Base	93.67	95.20	95.07	94.65	94.50	95.73	95.43	94.67	**95.84**
	New	95.29	97.69	97.63	97.76	96.83	97.87	97.76	97.20	**97.96**
	HM	94.47	96.43	96.33	96.18	95.65	96.79	96.58	95.92	**96.89**
StanfordCars	Base	78.12	70.49	77.68	71.76	79.07	75.23	72.94	**80.80**	77.96
	New	60.40	73.59	68.63	75.04	74.80	71.77	74.00	74.13	**75.11**
	HM	68.13	72.01	72.88	73.36	76.88	73.46	73.47	**77.32**	76.51
Flower102	Base	97.60	94.87	95.54	95.00	97.93	97.17	95.92	97.73	**98.12**
	New	59.67	71.75	71.87	74.73	73.53	73.53	72.46	75.57	**77.60**
	HM	74.06	81.71	82.03	83.65	83.99	83.71	82.56	85.23	**86.67**
Food101	Base	88.33	90.70	90.37	90.05	89.80	**91.20**	90.71	90.57	90.85
	New	82.26	91.29	89.59	91.70	91.37	91.90	**92.05**	91.37	91.98
	HM	85.19	90.99	89.98	91.09	90.58	**91.54**	91.38	90.97	91.41
FGVCAircraft	Base	40.44	33.41	40.54	36.21	42.13	38.05	37.44	41.97	**42.31**
	New	22.30	23.71	27.57	33.55	33.73	33.20	35.61	34.43	**38.15**
	HM	28.75	27.74	32.82	34.83	37.46	35.46	36.50	37.83	**40.12**
SUN397	Base	80.60	79.74	81.26	80.29	82.20	80.70	80.82	**82.63**	82.17
	New	65.89	76.86	74.17	76.53	73.63	**79.30**	78.70	78.20	78.57
	HM	72.51	78.27	77.55	78.36	77.68	80.00	79.75	**80.35**	80.33
DTD	Base	79.44	77.01	77.35	77.55	81.67	81.10	80.36	**82.77**	78.87
	New	41.18	56.00	52.35	54.99	43.80	62.57	59.18	58.07	**62.76**
	HM	54.24	64.85	62.45	64.35	57.09	**70.64**	68.16	68.25	69.90
EuroSAT	Base	92.19	87.49	90.11	85.64	93.70	**95.00**	94.07	91.63	91.23
	New	54.74	60.04	60.89	64.34	62.67	**83.37**	73.23	74.73	76.97
	HM	68.69	71.21	72.67	73.48	75.11	**88.86**	82.35	82.32	83.50
UCF101	Base	84.69	82.33	84.33	82.89	86.60	85.53	83.00	**87.13**	85.84
	New	56.05	73.45	74.94	76.67	75.90	78.20	78.66	80.77	**81.92**
	HM	67.46	77.67	79.35	79.65	80.90	81.70	80.77	83.83	**83.84**

TABLE II: **Comparison with existing methods on Cross-domain generalization.** ('*' denotes the performance obtained by our implementation)

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
PLOT	63.01	55.11	33.00	21.86	55.61	41.39
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.90
ProGrad	72.24	**64.73**	47.61	49.39	74.58	59.07
TCP*	70.68	64.43	48.66	50.40	74.93	59.70
KgCoOp	71.20	64.10	48.67	50.69	76.70	60.11
LASP	71.10	63.96	49.01	50.70	**77.07**	60.19
MaPLe	70.72	64.07	**49.15**	50.90	76.98	60.26
MMTP	70.84	64.51	48.86	**51.24**	77.03	**60.41**

TABLE III: **Ablation on different components.**

	Multi-Textual Prompt	Meta-Learning Optimization	Base	New	HM
(a)			80.73	73.60	77.00
(b)		✓	81.26	76.65	78.89
(c)	✓		82.41	74.81	78.37
(d)	✓	✓	**83.56**	**76.92**	**80.10**

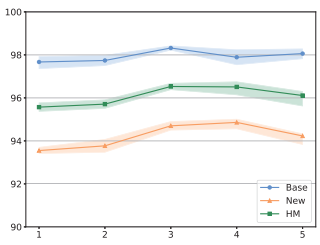

Fig. 3: Ablation on the number of prompts

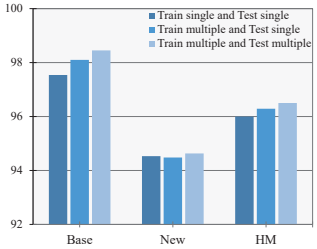

Fig. 4: Ablation on different train-test paradigm

D. Ablation study

Effect of different components. We remove different components of the MMTP as illustrated in Table III. The baseline is KgCoOp [14]. The Meta-learning Optimization demonstrates remarkable improvements in new classes. Moreover, Multi-textual Prompt also plays a vital role in enhancing the performance. Examining the last row reveals that all combined components contribute to the performance of MMTP.

Effect of the number of prompts. We calculate the variance for drawing Fig.3. The results demonstrate that increasing the number of prompts beyond three leads to diminishing benefits, because redundancy among the prompts reduces their ability to provide diverse perspectives. Furthermore, considering the training cost, we think that three is the optimal number.

Effect of different train-test paradigm. As shown in Fig. 4, we compared three prompt train-test paradigms: 1) single prompt training and testing, 2) multiple prompt training with single prompt testing, and 3) multiple prompt training and testing. The last showed the best results, highlighting the performance and validating the efficacy of MMTP.

IV. Conclusion

In this paper, we propose a novel Meta-learning-based Multi-Textual Prompt tuning (MMTP) method for VLMs to improve the discrimination and generalization of the learnable prompt. Specifically, we adopt meta-learning to learn multiple textual prompts for enhance the model's performance on few-shot tasks. Extensive experiments across 11 datasets for base-to-new class generalization and 4 datasets for cross-domain generalization demonstrate the effectiveness of our method.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[3] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, pp. 1–40, 2023.

[4] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," *ICLR*, 2022.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, pp. 2337–2348, 2022.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ICLR*, 2022.

[7] J. Zhu, Y. Li, K. Yang, N. Guan, Z. Fan, C. Qiu, and X. Yi, "Mvp: Meta visual prompt tuning for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[8] J. Zhu, K. Yang, N. Guan, X. Yi, and C. Qiu, "Hcpnet: Learning discriminative prototypes for few-shot remote sensing image scene classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 123, p. 103447, 2023.

[9] T. Feng, Q. Li, X. Wang, M. Wang, G. Li, and W. Zhu, "Multi-weather cross-view geo-localization using denoising diffusion models," in *Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, 2024, p. 35–39.

[10] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP*, 2021.

[11] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[12] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," in *ACL*, 2022.

[13] K. Zhou, J. Yang, and C. C. L. . Z. Liu, "Learning to prompt for vision-language models," *IJCV*, pp. 2337–2348, 2022.

[14] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *CVPR*, 2023, pp. 6757–6767.

[15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022, pp. 16 816–16 825.

[16] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *ICCV*, 2023, pp. 15 659–15 669.

[17] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in *ICML*, 2018.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPR-W*, 2004, pp. 178–178.

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *CVPR*, 2012, pp. 3498–3505.

[21] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV-W*, 2013, pp. 554–561.

[22] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008, pp. 722–729.

[23] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014, pp. 446–461.

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

[25] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 2217–2226, 2019.

[26] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014, pp. 3606–3613.

[28] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010, pp. 3485–3492.

[29] Y. Li, Y. Yang, W. Zhou, and T. Hospedales, "Feature-critic networks for heterogeneous domain generalization," in *ICML*. PMLR, 2019, pp. 3915–3924.

[30] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *ICML*, 2019, pp. 5389–5400.

[31] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *NeurIPS*, 2019, p. 10506–10518.

[32] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *CVPR*, 2021, pp. 15 262–15 271.

[33] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021, pp. 8340–8349.

[34] G. Chen, W. Yao, X. Song, X. Li, Y. Rao, and K. Zhang, "Plot: prompt learning with optimal transport for vision-language models," in *ICLR*, 2023.

[35] A. Bulat and G. Tzimiropoulos, "Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models," in *CVPR*, 2023, pp. 23 232–23 241.

[36] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023, pp. 19 113–19 122.

[37] H. Yao, R. Zhang, and C. Xu, "Tcp: Textual-based class-aware prompt tuning for visual-language model," in *CVPR*, 2024, pp. 23 438–23 448.

[38] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[39] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, "Promptkd: Unsupervised prompt distillation for vision-language models," in *CVPR*, 2024, pp. 26 617–26 626.

[40] S. Roy and A. Etemad, "Consistency-guided prompt learning for vision-language models," in *ICLR*, 2024.

[41] Y. Wang, X. Jiang, D. Cheng, D. Li, and C. Zhao, "Learning hierarchical prompt with structured linguistic knowledge for vision-language models," in *AAAI*, 2024, pp. 5749–5757.