PromptTA: Prompt-driven Text Adapter for Source-free Domain Generalization

Haoran Zhang[1⋆], Shuanghao Bai[1⋆], Wanqi Zhou[1], Jingwen Fu[1], Badong Chen[1†]
[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
{zhr2001, baishuanghao, zwq785915792, fu1371252069}@stu.xjtu.edu.cn, chenbd@mail.xjtu.edu.cn

***Abstract*—Source-free domain generalization (SFDG) tackles the challenge of adapting models to unseen target domains without access to source domain data. To deal with this challenging task, recent advances in SFDG have primarily focused on leveraging the text modality of vision-language models such as CLIP. These methods involve developing a transferable linear classifier based on diverse style features extracted from the text and learned prompts or deriving domain-unified text representations from domain banks. However, both style features and domain banks have limitations in capturing comprehensive domain knowledge. In this work, we propose Prompt-Driven Text Adapter (PromptTA) method, which is designed to better capture the distribution of style features and employ resampling to ensure thorough coverage of domain knowledge. To further leverage this rich domain information, we introduce a text adapter that learns from these style features for efficient domain information storage. Extensive experiments conducted on four benchmark datasets demonstrate that PromptTA achieves state-of-the-art performance. The code is available at https://github.com/zhanghr2001/PromptTA.**

***Index Terms*—Source-free domain generalization, text adapter, vision-language models.**

I. Introduction

Modern deep neural networks often rely on the oversimplified assumption that training and testing data are independent and identically distributed (i.i.d.), which makes them susceptible to out-of-distribution (OOD) data. To address the challenge of distribution shift, domain generalization (DG) has been introduced [1]–[4]. DG aims to develop models which can generalize well to unseen target domains from one or several related but distinct source domains [5]. The emergence of vision-language models (VLMs), like CLIP [6], has catalyzed the development of promising DG methods, notably those employing prompt tuning [7]–[9] and adapter tuning [10]–[12], as illustrated in Fig. 1 (a) and (b), respectively. However, DG assumes access to source domain data, which may not be feasible in scenarios involving confidentiality, privacy concerns, or data transmission limitations.

DG介绍
和缺点

催化 有前途的 特别是

Recently, source-free domain generalization (SFDG) has been proposed to enable the model to make predictions on unseen target domains without requiring source domain data for training. Existing SFDG methods primarily focus on leveraging prompt engineering and prompt tuning within VLMs [13], [14]. These methods enhance generalization capabilities by utilizing augmented target task definitions (e.g., class names) rather than relying on image data. For instance, Niu et al. [14] introduced the concept of domain bank to incorporate textual domain knowledge into soft prompts. As shown in Fig. 1 (c), Cho et al. [13] proposed PromptStyler, which learns a transferable linear classifier in joint vision-language space by integrating textual style features that encompass diverse domain knowledge. However, both the textual domain bank and style features have limitations in fully capturing comprehensive domain knowledge, due to constraints in the number of domains represented in the domain bank and the quantity of style features.

SFDG介绍
两个之前的
方法

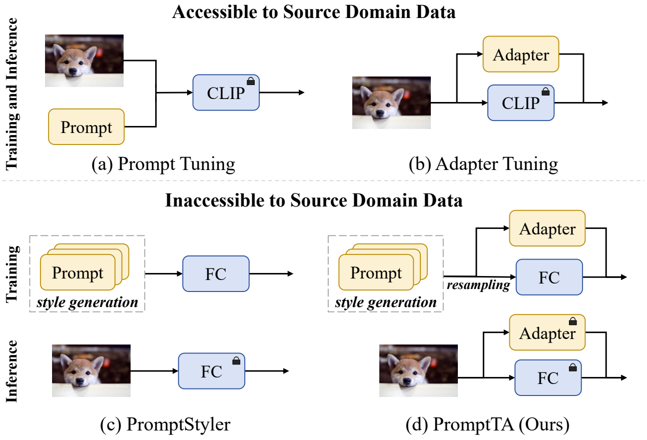

Fig. 1. Comparison of CLIP-based domain generalization methods. (a) and (b) require source domain data for fine-tuning, while (c) [13] and our method operate without such data. Our method uniquely leverages diverse domain information through style feature resampling and a text adapter.

包含

In this paper, we propose a method that incorporates a prompt-driven text adapter for source-free domain generalization, namely PromptTA. Building upon the style features design of PromptStyler [13], we establish distributions for these style features and implement resampling to ensure the representation of highly diverse domain knowledge. To further leverage this rich domain information, we introduce a text-based adapter that learns from these style features for efficient domain information storage. The text adapter is initialized with the template "a [DOM] of a [CLS]", where [DOM] represents the domain name and [CLS] represents the class name. Both the resampled and original style features are utilized in training the text adapter and the linear classifier. This approach addresses the previous limitations of fully capturing comprehensive domain knowledge and effectively utilizes available domain information.

介绍本
文工作

Our contributions are summarized as follows:

- We propose PromptTA, a novel adapter-based framework for SFDG that incorporates a text adapter to effectively leverage rich domain information.
- We introduce style feature resampling that ensures comprehensive coverage of textual domain knowledge.
- Extensive experiments demonstrate that our PromptTA achieves the state of the art on DG benchmarks.

II. Related Work

Domain Generalization. DG aims to train a model on source domain data that effectively generalizes to unseen target domains. Most DG methods address distribution shifts through three main

⋆Equal contribution. †Corresponding author. This work was supported by the National Natural Science Foundation of China with grant numbers (U21A20485, 62436005, and 62088102), and the Fundamental Research Funds for Xi'an Jiaotong University under Grants xzy022024012.

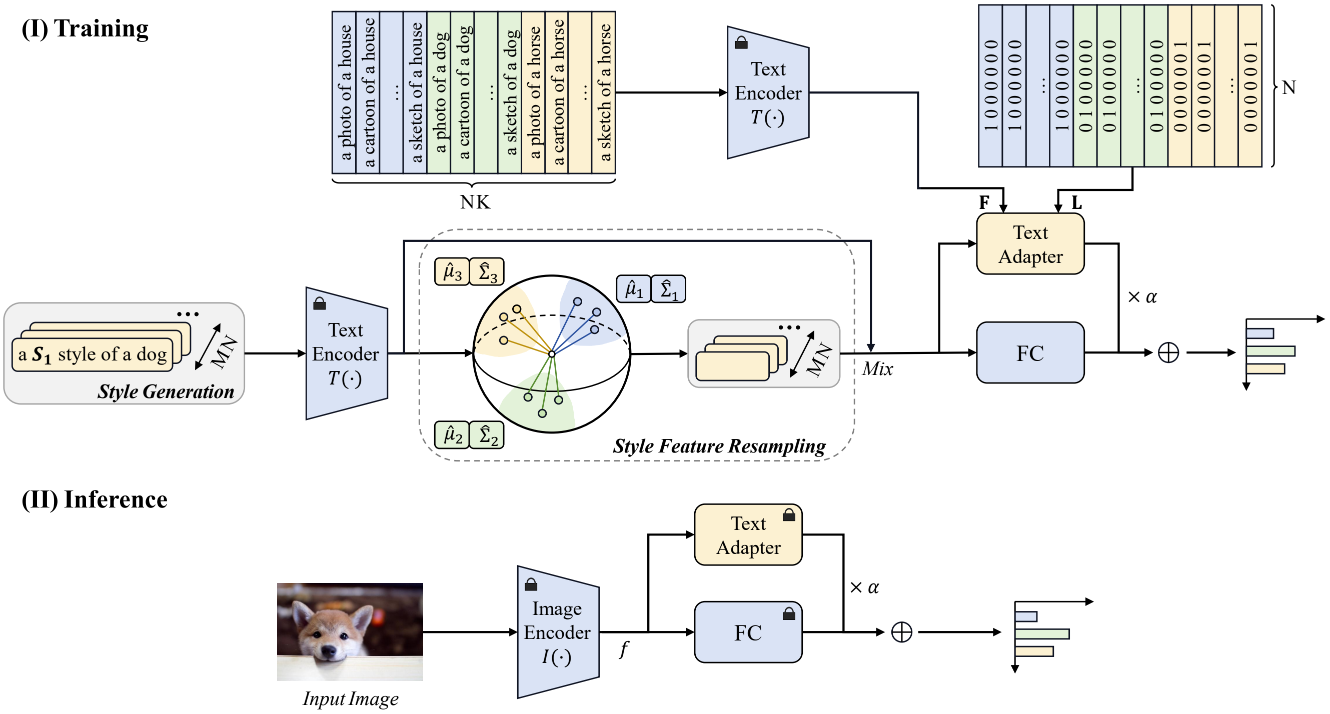

Fig. 2. Overall framework of PromptTA. Initially, the style generation process yields a fixed set of style features. These are then enhanced through Style Feature Resampling to capture comprehensive domain knowledge. Both the original style features and resampled style features are utilized to train a linear classifier and a text adapter. Note that the encoders are derived from CLIP model [6].

DG实现的三种方法，本文与第三种类似

perspectives [5]. Data augmentation enriches training datasets to help models learn more generalizable features [15], [16]. Representation learning designs feature extractors and regularizers to learn domain-invariant features [17], [18]. Classifier debiasing (去偏) retrains classifiers on balanced datasets to enhance generalization [19], [20]. Our method leverages the vision-language alignment properties of CLIP [6], conceptually aligning more closely with the third perspective by learning a robust classifier in the joint vision-language space.

SFDG：prompt tuning和adapter tuning

Adaptation of Vision language Models. Large-scale VLMs [6] have shown impressive zero-shot capabilities, but their size renders (使得) full fine-tuning impractical. Parameter-efficient methods like prompt tuning and adapter tuning have emerged to address this issue. Prompt tuning methods, such as CoOp [7] and CoCoOp [8], replace fixed prompt contexts with learnable vectors, significantly outperforming hand-crafted prompts. Adapter tuning methods introduce a small number of additional parameters via residual connections. For instance, CLIP-adapter [21] employs residual feature blending with original CLIP-encoded features, while Tip-adapter [10] utilizes a learnable key-value cache model to refine predictions. Our method differs from previous works that used adapters composed of visual features. Instead, we address the SFDG task by employing an adapter constructed from text features. We further fine-tune the text adapter with a diverse set of learned prompts to enhance the adaptability (适应性).

Source-free Domain Generalization. Based on the setting of DG, SFDG tackles a more challenging scenario where source domain data is unavailable. Current SFDG methods predominantly leverage the alignment between textual and visual features in VLMs, which enhance visual recognition by utilizing rich domain information to learn generalizable text representations. Niu et al. [14] employ a pre-defined domain bank to obtain domain-unified text representations, while PromptStyler [13] learns style features covering diverse domains to train a robust linear classifier, which can be transferred for image classification. However, these methods are constrained by the limited number of domains in the domain bank or the quantity of style features. To address these limitations, we propose style feature resampling to capture comprehensive domain knowledge, coupled with a text adapter for efficient domain information storage.

III. Method

The overall framework of the proposed PromptTA method is illustrated in Fig. 2. Our method utilizes the aligned image-text representations from CLIP [6] to tackle the DG problem without using any images. We introduce our method as follows.

Style Generation. Following Algorithm 1 of PromptStyler [13], we employ pseudo-word S_i as a placeholder within the prompt template P_i, represented as "a S_i style of a [CLS]," where [CLS] denotes the class name. We define P_i^{dom} as "a S_i style of a" and P_j^{cls} as "CLS", where $j \in 1, 2, ..., N$, and N denotes the number of class names. The pseudo-word S_i is replaced with M learnable style word vectors $\{s_i\}_{i=1}^{M}$ that represent diverse domain information. Thus we can derive MN style features corresponding to $\{\{P_{i,j}\}_{i=1}^{M}\}_{j=1}^{N}$. To optimize these learnable vectors $\{s_i\}_{i=1}^{M}$, a style diversity loss $\mathcal{L}_{\text{style}}$ (Equation 1 in [13]) is adopted to maximize the variance of M text features of P_i^{dom}. And a content consistency loss $\mathcal{L}_{\text{content}}$ (Equation 3 in [13]) is used to ensure the class information remains consistent by minimizing the distance of text features between $P_{i,j}$ and its corresponding P_j^{cls}. Finally, we obtain a set of fixed learned style word vectors $\{s_i\}_{i=1}^{M}$. The i-th style feature of j-th class is denoted as $T(P_{i,j})$, where $T(\cdot)$ denotes the text encoder.

Style Feature Resampling. While PromptStyler [13] employs a fixed set of M style features to train a linear classifier, this approach has limitations in fully capturing comprehensive domain knowledge. To this end, we propose Style Feature Resampling (SFR), a module

TABLE I

Comparison with the state-of-the-art (SOTA) methods on four domain generalization benchmark datasets for multi-source DG. † represents our reproduced results for PromptStyler, which trains the linear classifier with cross-entropy loss.

Method	Venue	Source-free	PACS	VLCS	OfficeHome	DomainNet	Avg.
ResNet-50 [22] with pre-trained weights on ImageNet [23].							
RSC [24]	ECCV 2020	✗	85.2±0.9	77.1±0.5	65.5±0.9	38.9±0.5	66.7
SagNet [25]	CVPR 2021	✗	86.3±0.2	77.8±0.5	68.1±0.1	40.3±0.1	68.1
MIRO [2]	ECCV 2022	✗	85.4±0.4	79.0±0.0	70.5±0.4	44.3±0.2	69.8
SWAD [26]	NeurIPS 2021	✗	**88.1**±0.1	**79.1**±0.1	**70.6**±0.2	**46.5**±0.1	**71.1**
ResNet-50 [22] with pre-trained weights from CLIP [6].							
ZS-CLIP (C) [6]	ICML 2021	✓	91.0±0.0	81.2±0.0	67.1±0.0	45.9±0.0	71.3
CAD [27]	ICLR 2022	✗	90.0±0.6	81.2±0.6	70.5±0.3	45.5±2.1	71.8
ZS-CLIP (PC) [6]	ICML 2021	✓	90.8±0.0	81.4±0.0	70.8±0.0	46.7±0.0	72.4
PromptStyler† [13]	ICCV 2023	✓	92.5±0.1	82.2±0.1	72.3±0.1	48.6±0.0	73.9
PromptTA (Ours)	–	✓	**93.8**±0.0	**83.2**±0.1	**73.2**±0.1	**49.2**±0.0	**74.9**
ViT-B/16 [28] with pre-trained weights from CLIP [6].							
ZS-CLIP (C) [6]	ICML 2021	✓	95.8±0.0	76.5±0.0	79.3±0.0	57.3±0.0	77.2
MIRO [2]	ECCV 2022	✗	95.6	82.2	82.5	54.0	78.6
ZS-CLIP (PC) [6]	ICML 2021	✓	96.1±0.0	83.4±0.0	81.8±0.0	57.2±0.0	79.6
PromptStyler† [13]	ICCV 2023	✓	97.2±0.1	83.4±0.3	82.5±0.2	58.3±0.0	80.4
PromptTA (Ours)	–	✓	**97.3**±0.1	**83.6**±0.3	**82.9**±0.0	**59.4**±0.0	**80.8**
ViT-L/14 [28] with pre-trained weights from CLIP [6].							
ZS-CLIP (C) [6]	ICML 2021	✓	97.7±0.0	79.1±0.0	85.6±0.0	62.8±0.0	81.3
ZS-CLIP (PC) [6]	ICML 2021	✓	**98.6**±0.0	82.6±0.0	86.7±0.0	63.4±0.0	82.8
PromptStyler† [13]	ICCV 2023	✓	**98.6**±0.0	82.9±0.5	88.4±0.1	64.5±0.0	83.6
PromptTA (Ours)	–	✓	**98.6**±0.0	**83.3**±0.3	**88.5**±0.0	**65.2**±0.0	**83.9**

that incorporates more diverse domain information by dynamically regenerating sampled features during each training epoch, thereby enhancing generalization capabilities. Building upon the previous subsection, we leverage M learned style word vectors and N classes to derive a total of MN style features. Specifically, the style feature corresponding to the i-th style and j-th class is denoted as $T(P_{i,j})$. We posit that for a given class j, the prompt features generated by combining each class with various style word vectors follow a Gaussian distribution, as expressed as follows:

$$\{T(\{P_{i,j}\}_{i=1}^{M})\} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{1}$$

The ground-truth feature distribution of the j-th class can be approximated by computing mean and standard deviations from known prompt features as follows:

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{M}\sum_{i=1}^{M} T(\mathcal{P}_{i,j}), \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{M-1}\sum_{i=1}^{M}(T(\mathcal{P}_{i,j})) - \hat{\boldsymbol{\mu}}_j)^2. \tag{3}$$

By drawing new features from this estimated distribution, we incorporate new styles that encompass diverse domains. We resample style features at the onset of each training iteration, which are then combined with the original features to train the model.

Text Adapter. To fully leverage the acquired domain knowledge, we introduce a learnable text adapter that learns from style features. Different from previous works [10], our adapter is constructed using text features rather than visual features, a design tailored to address the challenges of the SFDG task. The adapter stores domain knowledge from learning from style features and generates predictions based on the similarities between input features and adapter features. To initialize the N-way K-shot text adapter, we employ a pre-defined domain bank comprising K distinct domains (e.g., photo, painting, sketch). Following the template "a [DOM] of a [CLS]", we construct KN prompts and extract adapter features using CLIP's text encoder, resulting in $\mathbf{F} \in \mathbb{R}^{NK\times D}$, where D denotes the adapter feature dimension. The classes are encoded as one-hot vectors, forming $\mathbf{L} \in \mathbb{R}^{NK\times N}$. For an input feature $f \in \mathbb{R}^{D}$, we compute its similarities with adapter as follows:

$$\text{Logits}_{\text{Adapter}} = \varphi(f\mathbf{F}^T)\mathbf{L}, \tag{4}$$

$$\varphi(x) = \exp(-\beta(1-x)), \tag{5}$$

where $\varphi(\cdot)$ denotes an exponential function that transforms similarity scores into non-negative values. The parameter β serves to modulate the sharpness of this transformation, effectively controlling the sensitivity of the adapter to input similarities. We retain the linear classifier to operate directly on style features, whose weight can be denoted as W. Thus the final classification logits are computed as a weighted combination of the outputs from both the text adapter and the linear classifier, which can be formulated as:

$$\begin{aligned}\text{Logits} &= \text{Logits}_{\text{FC}} + \alpha\text{Logits}_{\text{Adapter}},\\ &= fW^T + \alpha\varphi(f\mathbf{F}^T)\mathbf{L},\end{aligned} \tag{6}$$

where α denotes the residual ratio. Then we adopt cross-entropy loss as our classification loss to train the linear classifier and text adapter. This approach integrates the classification power of the linear classifier with the potential feature refinement capabilities of the text adapter, enhancing the model's performance across diverse domains.

Inference. During inference, an input image is processed by the CLIP image encoder to produce an image feature, which is then passed through both the trained linear classifier and trained text adapter. The final prediction is also formulated as (6).

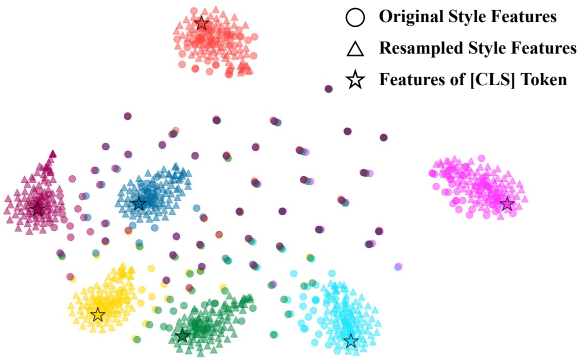

Fig. 3. The t-SNE [29] visualization results for original style features, style features from a single resampling instance, and class token features of PACS dataset. Different colors denote different classes.

TABLE II
ABLATION STUDY ON THE MODEL COMPONENT OF STYLE FEATURE RESAMPLING (SFR) AND TEXT ADAPTER (TA).

SFR	TA	PACS	VLCS	OfficeHome	DomainNet
		92.5	82.2	72.3	48.6
✓		93.2	82.9	72.7	48.7
	✓	93.6	83.1	73.1	49.0
✓	✓	**93.8**	**83.2**	**73.2**	**49.2**

IV. Experiments

A. Experimental Settings

Evaluation datasets. To validate the effectiveness of our method, we evaluate it on four DG datasets: PACS [30], VLCS [31], OfficeHome [32] and DomainNet [33]. On each dataset, we repeatedly test three times with different random seeds and 采用 adopt the average top-1 accuracy with 标准方差 standard deviations.

Baselines. We compare our method with a series of baselines, including source-free methods like zero-shot CLIP [6], PromptStyler [13], and conventional DG algorithms that require source data: RSC [24], SagNet [25], MIRO [2], SWAD [26], and CAD [27]. Note that ZS-CLIP (C) and ZS-CLIP (PC) denote zero-shot CLIP [6] with a prompt template of "[CLS]" and "a photo of a [CLS]", respectively.

Implementation details. During style generation, we adopt the configuration from PromptStyler [13] to learn $M = 80$ style word vectors initialized from a zero-mean Gaussian distribution with a standard deviation of 0.02. The text adapter incorporates $K = 11$ domains, including photo, cartoon, and painting, etc. For training, we employ SGD with 0.9 momentum and a cosine learning rate scheduler. The learning rates are set to 0.05 for the linear classifier and 0.01 for the adapter on PACS, VLCS, and OfficeHome. For DomainNet, we reduce the adapter's learning rate to 0.001 due to its large sample size. The hyper-parameter α of the text adapter is dataset-specific, ranging from 1 to 5, while β is fixed at 2. For PromptStyler, we train the linear classifier using cross-entropy loss.

B. Evaluations

Main results. As shown in Table I, our PromptTA method achieves SOTA performance across all evaluation datasets with three visual backbones: ResNet-50 [22], ViT-B/16 [28], and ViT-L/14 [28]. Specifically, PromptTA achieves average accuracy improvements of 1.0%, 0.4%, and 0.3%, respectively, over SOTA PromptStyler method. Notably, it surpasses conventional DG baselines despite not utilizing any images during training, outperforming SOTA SWAD and CAD methods by 3.8% and 3.1% in average accuracy, respectively, with ResNet-50 as the backbone.

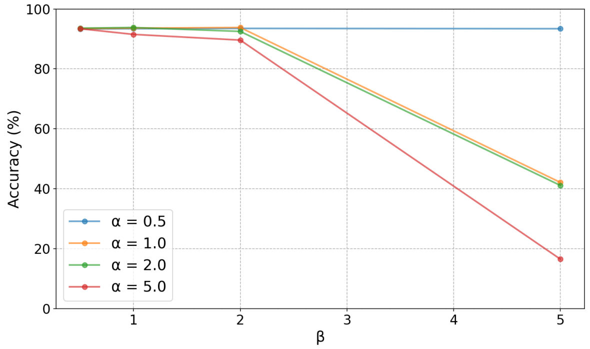

Fig. 4. Sensitivity analysis of hyperparameter α and β on PACS dataset.

TABLE III
COMPARISON OF INITIALIZATION METHODS FOR TEXT ADAPTER.

Initialization	PACS	VLCS	OfficeHome	DomainNet
Random	93.3	82.5	72.7	47.0
Template	**93.8**	**83.2**	**73.2**	**49.2**

T-SNE visualization results. As shown in Fig. 3, for each class, the resampled style features (triangles) align closely with the original style features (circles), indicating that the resampling effectively captures the distribution of the original style features while possibly enhancing generalization. Both the original (circles) and resampled (triangles) style features cluster distinctly by class (stars). This suggests that the resampling process maintains the inherent structure of the style features and does not distort the original class boundaries.

C. More analyses

Ablation study on each module. We perform an ablation study on style feature resampling (SFR) and text adapter (TA) to assess their individual and combined contributions. As shown in Table II, the removal of either SFR or TA leads to a decrease in accuracy. Notably, the highest accuracy is achieved when both components are present, underscoring their synergistic effect on the model's performance.

Sensitivity analysis of hyperparameter α and β. As shown in Fig. 4. We vary α and β from 0.5 to 5.0. The model exhibits robust performance when $\beta \in [0.5, 2]$ across a wide range of α values ($0.5 \leq \alpha \leq 5$). For $\beta > 2$, robustness is maintained only when $\alpha \approx 0.5$, while performance degrades significantly for $\alpha \in [1, 5]$.

Ablation study on initialization method of text adapter. As shown in Table III, initializing the text adapter with a template (a [DOM] of a [CLS]) as prior knowledge outperforms random initialization. This suggests that incorporating domain-specific and class-specific information during initialization provides a better starting point.

V. Conclusion

In this paper, we propose PromptTA, a novel method that incorporates text adapter to address the challenging SFDG task. We introduce a style feature resampling module to better capture the distribution of style features and employ resampling to ensure comprehensive domain coverage. Then a text adapter is utilized to act as a dynamic repository for domain knowledge, which can be effectively leveraged during inference. Experiments on four benchmarks demonstrate that PromptTA achieves state-of-the-art performance.

References

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

[2] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," in *European conference on computer vision*. Springer, 2022, pp. 440–457.

[3] Z. Huang, A. Zhou, Z. Ling, M. Cai, H. Wang, and Y. J. Lee, "A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 685–11 695.

[4] S. Bai, Y. Zhang, W. Zhou, Z. Luan, and B. Chen, "Soft prompt generation for domain generalization," in *European Conference on Computer Vision*, 2024.

[5] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[6] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[8] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[9] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.

[10] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.

[11] L. Song, R. Xue, H. Wang, H. Sun, Y. Ge, Y. Shan *et al.*, "Meta-adapter: An online few-shot learner for vision-language model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 361–55 374, 2023.

[12] Q. Wang, G. Liu, and B. Wang, "Caps-adapter: Caption-based multimodal adapter in zero-shot classification," *arXiv preprint arXiv:2405.16591*, 2024.

[13] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 702–15 712.

[14] H. Niu, H. Li, F. Zhao, and B. Li, "Domain-unified prompt representations for source-free domain generalization," *arXiv preprint arXiv:2209.14926*, 2022.

[15] R. Volpi and V. Murino, "Addressing model vulnerability to distributional shifts over image transformation sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7980–7989.

[16] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6817–6826.

[17] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 624–639.

[18] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[19] E. Rosenfeld, P. Ravikumar, and A. Risteski, "Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization," *arXiv preprint arXiv:2202.06856*, 2022.

[20] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2427–2440, 2021.

[21] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*. Springer, 2020, pp. 124–140.

[25] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8690–8699.

[26] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 405–22 418, 2021.

[27] Y. Ruan, Y. Dubois, and C. J. Maddison, "Optimal representations for covariate shift," in *International Conference on Learning Representations*, 2022.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1657–1664.

[32] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027.

[33] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.