ProMEA: Prompt-driven Expansion and Alignment for Single Domain Generalization

Yunyun Wang[1*] , **Yi Guo**[1] , **Xiaodong Liu**[1] and **Songcan Chen**[2,3]
[1]School of Computer Science, University of Posts and Telecommunications
[2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[3] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics
{wangyunyun, 1022040813, 1022040815}@njupt.edu.cn, s.chen@nuaa.edu.cn

Abstract

In single Domain Generalization (single-DG), data scarcity in the single source domain hampers the learning for invariant features, leading to overfitting over source domain and poor generalization to unseen target domains. Existing single-DG methods primarily augment the source domain by adversarial generation. However, there are still two key challenges. i) With simple feature perturbation to confuse the classifier, it may generate unnatural samples with semantic ambiguity or distortion. ii) It is still difficult to cover the sufficient shift in a real domain by generating indistinguishable samples from source data, thus the learning model is inescapable from overfitting to the single source domain. To this end, we turn to augment the domain prompt, considering that text prompt perturbation is easier to generate and generalize. Then the source domain is expanded with the guidance of augmented text prompts, which are learnable with both semantic consistency and style diversity. Specifically, we propose a ProMpt-driven Expansion and Alignment (ProMEA) method for single-DG, in which a Domain Prompt Expansion module is first developed to expand the single source domain with frequency features of augmented text prompts, in which the amplitude spectrum predominantly harbors the domain style information. With source prompts, a Domain Prompt Alignment module is further designed in inference for adapting target samples to the expanded source domains, in order to reduce the domain discrepancy. Finally, empirically results over single-DG benchmarks demonstrate the superiority of our proposal.

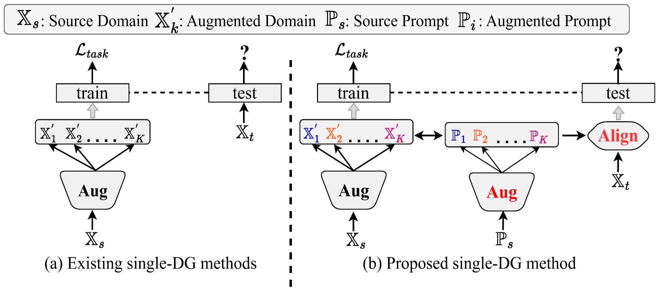
$\mathbb{X}_s$: Source Domain $\mathbb{X}'_k$: Augmented Domain $\mathbb{P}_s$: Source Prompt $\mathbb{P}_i$: Augmented Prompt
$\mathcal{L}_{task}$
?
train
test
$\mathbb{X}'_1$ $\mathbb{X}'_2$ .... $\mathbb{X}'_K$
$\mathbb{X}_t$
Aug
$\mathbb{X}_s$
(a) Existing single-DG methods
$\mathcal{L}_{task}$
?
train
test
$\mathbb{X}'_1$ $\mathbb{X}'_2$ .... $\mathbb{X}'_K$
$\mathbb{P}_1$ $\mathbb{P}_2$ .... $\mathbb{P}_K$
Align
Aug
Aug
$\mathbb{X}_s$
$\mathbb{P}_s$
$\mathbb{X}_t$
(b) Proposed single-DG method

Figure 1: Difference between previous single-DG methods and our proposed ProMEA. (a) Existing methods primarily focus on expanding source domain by feature perturbation or adversarial generation. (b) ProMEA leverages augmented text prompts to guide the feature expansion in training, and align the target samples to source domains in inference.

1 Introduction

In the past years, deep learning has indeed made remarkable achievements over various tasks. However, these accomplishments and efforts are all built on a basic assumption that the test data share the same data distribution with the training data. In practice, this assumption can be easily violated since the distribution discrepancy between training and target data is often tremendous [Chen *et al.*, 2020; Chen *et al.*, 2021b; Long *et al.*, 2018; Luo *et al.*, 2020]. Such a phenomenon is termed as distribution shift and usually leads to severe performance degradation. To tackle this problem, a series of research has been proposed, primarily focusing on domain adaptation (DA) and domain generalization (DG). DA endeavors to learn transferable knowledge from the labeled source data to unlabeled target data, in which target data are accessible in model learning. DG mainly attempts to solve a more challenging problem, that is, generalize the learning model to an unseen target domain by learning from multiple source domains. Both DA and DG exhibit impressive generalization performance over out-of-distribution data.

In real tasks, however, collecting and annotating data from different domains or environments is commonly time-consuming and resource-intensive, which hinders the application of DA and DG methods. Recently, the paradigm of single domain generalization (single-DG) has been introduced, which learns robust representation for generalizing from only one single source domain. State-of-the-art single-DG methods commonly generate diverse training samples with feature perturbation or adversarial data augmentation, in order to expand the representation scope of the source domain, and consequently enhance the generalization ability of source model, as shown in Fig. 1 (a). However, it is usually hard to guarantee that the augmented samples effectively assist the model in performing well on target domain, since the target data is inaccessible in training and the generation is actually unoriented. First, through simply modifying features by confusing the classifier, it may generate unnatural samples with semantic ambiguity or distortion. Second, it is difficult to cover a real domain with sufficient shift by generating indistinguishable samples from the source data, thus the model will still be stuck in overfitting to the single source domain, and thus contribute slightly to generalization over unknown domains.

Unlike typical single-DG methods that aim to enlarge the scope of single-source domain with adversarial sample generation, we propose a text-to-image generation to augment the text prompts of domain first, that is, we learn multiple domain prompts with both semantic consistency and style diversity, and then use the generated text prompts to guide the data expansion in the frequency domain. In real applications, the visual perturbations are usually complex and difficult to control, sometimes a slight change in visual perturbation may lead to unpredictable results [Schiappa *et al.*, 2022], while the text prompts perturbations are easier to generate and generalize. To the end, we propose a novel ProMpt-driven Expansion and Alignment (ProMEA for short) method for single-DG, which utilizes augmented text prompts to compensate for the distribution discrepancy between source and target domains in both training and testing stage.

In ProMEA, we design a Domain Prompt Expansion (DPE) module in the training process, in which a prompt generator is first trained to augment the text prompt, in order for more reliable feature expansion. The prompt generator creates class prompts that are filled with specific category words and expanded domain or stylistic words, thus helps preserve semantic information accompanied with diverse domain styles. Then, the distribution of source domain is expanded under the guidance of frequency features from class prompts, as shown in Fig. 1 (b). For instance, for a given class of dog, utilizing an augmented text prompt such as “a sketch of dog” or “a painting a dog”, an expanded feature domain comprising “sketch dogs” or “painting dogs” will be generated. Here, the class prompts are learnable for both semantic consistency and style diversity. Further, to reduce the discrepancy between source and unseen target domains in inference, an distribution alignment module named Domain Prompt Alignment (DPA) is developed in inference. Specifically, the prompt generator also extracts domain prompts containing stylistic information for individual domains. When target samples come, the similarities between the target images and source domain prompts will be calculated. Ultimately, the target domain prompt will be represented as a linear combination of source prompts during the testing phase, in order to adapt the target data to the convex hull of source domains.

Different from previous single-DG works that employ adversarial data augmentation to enlarge the distribution of source data, ProMEA utilizes augmented text-prompts to guide the data expansion, as well as style transfer for target samples, in order to narrow the distribution gap between the source and target domains. Besides, some multi-source DG methods [Liu and Wang, 2023; Cho *et al.*, 2023; Cheng *et al.*, 2024] treat text features directly as training samples, and guide the feature extraction of image samples with a unified representation space. However, off-the-shelf solutions to multi-domain DG are not applicable to single-DG, since the former relies on domain identifiers as supervision signals for learning domain invariant representation. Moreover, different from directly utilizing the text features as training data, ProMEA aims to guide the expansion of source domain using augmented text prompts, so as to boost the diversity of source distribution with semantic consistency. To summarize, our main contributions are list as follows,

- We propose a novel ProMEA method for single-DG, in which a Domain Prompt Expansion module is developed for text-to-image expansion. It generates diverse text prompts embodying both semantic consistency and domain diversity, then source expansion is guided with the augmented text prompts in training for robust domain generalization.
- A Domain Prompt Alignment module is designed to adaptively align the target data to source domain using source prompts in the inference stage, so as to further bridge the distribution gap between domains.
- Extensive experiments have been conducted over several DG benchmarks (PACS, Office-Home, and VLCS) to demonstrate the competitiveness of ProMEA. Furthermore, we conduct qualitative analysis and ablation studies to verify the effectiveness of proposal.

2 Related Work

2.1 Domain Generalization

Domain generalization (DG) aims to train a learning model from source domains that can generalize to unknown target domains, so as to keep robust performance on out-of-distribution data. DG tasks can be divided into multi-source and single-source settings in terms of the number of source domains for training. Various methods have been proposed for multi-source DG, such as learning domain-invariant representation [Chen *et al.*, 2021a; Li *et al.*, 2017a; Seo *et al.*, 2020; Dayal *et al.*, 2024], feature disentanglement [Xu *et al.*, 2014; Li *et al.*, 2017b; Yang *et al.*, 2023] or meta-learning [Jia and Zhang, 2024].

In single-DG, there is only one source domain available, thus current state-of-the-art single-DG methods typically use perturbation-based approaches to expand source domain distribution, in order to enhance the generalization ability of model. For example, Cugu et al. [Cugu *et al.*, 2022] utilize visual corruptions for data augmentation to expand the domain distribution while maintain the class semantic consistency. Wang et al. [Wang *et al.*, 2021] adopt a bound of mutual information (MI) between domains to extract semantic features and generate diverse images. Gokhale et al. [Gokhale *et al.*, 2023] use adversarial neural network and a diversity module to generate new samples with both diversity and hardness. Chen et al. [Chen *et al.*, 2023] employ a novel angular center loss to push the augmented samples away from the class centers. Liu et al. [Liu *et al.*, 2024b] use stylization and destylization module for style transfer, in order to learn domain-invariant representation among the gener-

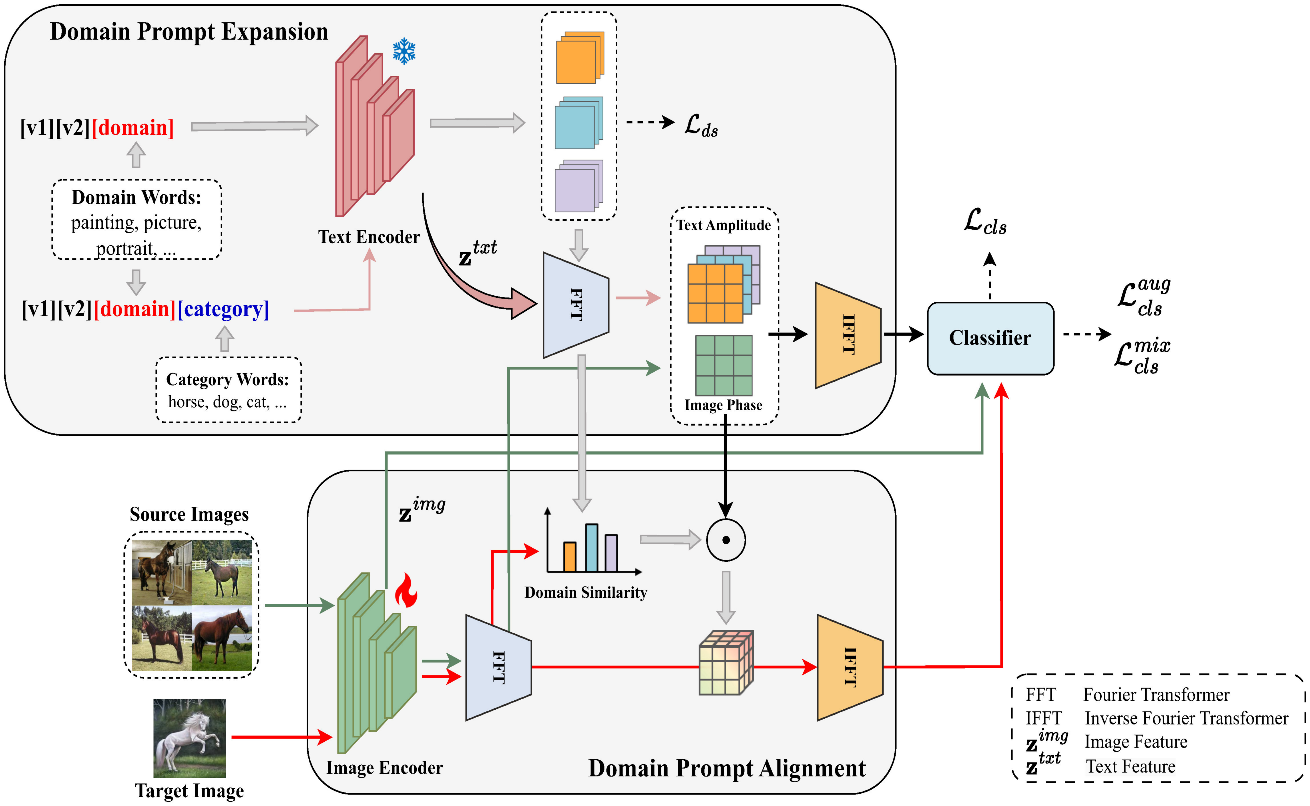

Figure 2: The model overview of ProMEA, in which Domain Prompt Expansion adaptively learns text prompts for guiding the source expansion, so as to preserve both semantic consistency and domain diversity. Further, Domain Prompt Alignment is introduced in inference, in order to further align the target data to source domain for better generalization.

ated domains with different styles. These methods inevitably face the issues of semantic distortion or insufficient diversity caused by feature perturbations. Moreover, due to the invisibility of the target domain, they can not guarantee that the augmented source domain is close to the unknown target domain.

2.2 Vision and Language Pre-training

Pre-trained Vision Language Models (VLMs), trained on vast amounts of data, have gained widespread application in natural language processing and computer vision tasks due to their proficiency in handling multi-modal data. VLMs significantly reduce data requirements for specific tasks, making them invaluable in resource-constrained scenarios. Recently, VLMs models, such as CLIP, have been introduced into DG learning tasks. Liu et al. [Liu and Wang, 2023] exploit CLIP to project texts and images into a common space, and then text features are actually treated as training data, in order to expand the sample distribution of multi-source DG. Cho et al. [Cho *et al.*, 2023] synthesize diverse style features via learnable word vectors, in order to simulate various distributions via adopting prompts for source-free DG. Niu et al. Niu et al. [Niu *et al.*, 2022] yield domain-unified representations with prompts generated by VLMs, in order to cope with samples from open-world domains. Cheng et al. [Cheng *et al.*, 2024] disentangle text prompt into domain-specific and domain-invariant descriptions first by a large language model, and then enable the learning of domain-invariant features more effectively. However, previous methods commonly treat text features directly as training samples in multi-source DG, so as to guide the feature extraction for image samples from multiple domains. While ProMEA adopts text prompts to guide the source expansion and distribution alignment for single-DG, it aims to achieve both semantic coherence and distribution diversity, which will be described in the next section.

3 Methodology

3.1 Preliminary and Model Structure

Assuming that we have a single source domain $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes the i-th sample from source domain $\mathcal{D}_s$, and y_i is the corresponding label from the label space $y \in \{1, \ldots, N_c\}$, N and N_c denote the number of source samples and classes, respectively. For each input image $x_i \in \mathbb{R}^{C \times H \times W}$, where C, H and W represent the channel number, height and width of the image. The goal of ProMEA is to train a learning model over $\mathcal{D}_s$, so that it performs well on an unknown target domain $\mathcal{D}_t$. The source and target domains have the same label space.

To address the generalization issue in single-DG, we propose ProMEA, and the method overview is shown in Fig. 2. In ProMEA, Domain Prompt Expansion is first proposed in training for source domain expansion, which utilizes learnable text prompts to guide the source expansion. Furthermore, unlike previous work treating text features directly as training samples to learn feature extractor, Fourier transformation (FFT) [Nussbaumer and Nussbaumer, 1982] is adopted to decouple text prompts into phase and amplitude, capitalized on its well-established property that the phase component of the Fourier spectrum preserves high-level semantics of the original signal, while the amplitude component contains low-level statistics [Oppenheim *et al.*, 1979; Yang and Soatto, 2020]. The text amplitude is actually domain style information, which are combined with image phase of semantic information through inverse Fourier transformation, in order for prompt-guided expansion of source domain. In inference, there is domain prompt alignment between source and target domains. Specifically, the similarities of target image and domain prompt features are calculated. With the similarity vector, the alignment prompt will be introduced into the target image, so as to align the target data to the convex hull of source domains.

3.2 Domain Prompt Expansion

For single-DG, learning content-invariant representation is a critical component. However, due to the scarcity of source domain samples, data augmentation becomes essential. To maintain semantic consistency with respect to categories while achieving distribution diversity with respect to domains in source expansion, we propose the DPE module to expand the source domain with augmented text prompts.

Learning for Augmented Text Prompts. We adopt two types of text prompt here, which is shown in Fig. 2. The class prompt contains domain-related words and class-specific words, as well as domain-agnostic and domain-specific context variables [Ge *et al.*, 2023], which are learnable in training, since a slight change of the prompt often causes a huge variance in performance [Zhou *et al.*, 2022a]. The domain-agnostic context is shared across all domains, while domain-specific context is specific for each domain. The domain prompts are similar to the class prompts, except for the lack of class-specific words. Specifically, the class prompts and domain prompts follow an unified style as,

$$z_{kt}^{txt} = [a]_1 \dots [a]_{M_1} [b]_1 \dots [b]_{M_2} [Domain]_k [Class]_t \, , \tag{1}$$

and

$$z_k^{style} = [a]_1 \dots [a]_{M_1} [b]_1 \dots [b]_{M_2} [Domain]_k \, , \tag{2}$$

where $[a]_{m_1}, m_1 \in \{1, 2, \dots, M_1\}$ is the domain-agnostic context with the same dimension as word embedding, M_1 is the number of context tokens applied in our class prompts. $[b]_{m_2}, m_2 \in \{1, \dots, M_2\}$ is the domain-specific context, M_2 is the number of domain-specific tokens. $[Domain]_k, k \in \{1, \dots, K\}$ and $[Class]_t, t \in \{1, \dots, N_c\}$ denote the vocabularies for domains and classes, respectively.

To obtain a sufficient amount of domain-related words, which have inherent connections with the source domain while significantly different styles, we introduce a pretrained lexical substitution model [Arefyev *et al.*, 2022] in DPE. It replaces any word in a given sentence with various semantically similar alternatives. Further, these domain prompts generated still exhibit strong underlying correlations with each other. As a result, the text encoder might not be able to identify these prompts when extracting features from them [Zhou *et al.*, 2022b]. However, the style of the target domain may differ entirely from that of the source domain. To alleviate this issue, we leverage the learnable part of prompts to maximize the diversity between the generated domain prompts and the source prompt, so that the style of the augmented domains has a sufficiently large distributional difference from source domain,

$$\mathcal{L}_{ds} = -\sum_{k=1}^{K} KL\left(E_{txt}(z_s^{style}/T) | E_{txt}(z_k^{style}/T)\right), \tag{3}$$

where E_{txt} is the text encoder from CLIP, z_s^{style} represents the source domain prompt and T is the temperature of a softened softmax.

Domain Expansion with Text Prompts. To maintain semantic consistency during data augmentation, we attempt to decouple the semantic and domain information from both image samples and text prompts, which is exactly what the Fourier transformation accomplishes. For a given x, the Fourier transformation $\mathcal{F}(x)$ decouples each channel into two components, i.e., phase spectrum $\mathcal{P}(x)$ and amplitude spectrum $\mathcal{A}(x)$, which can be treated as semantic and domain information, respectively. Specifically,

$$\mathcal{F}(x)(u,v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) \, e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}, \tag{4}$$

where u and v are the frequency domain indices representing the horizontal and vertical frequency components, respectively. Then

$$\mathcal{A}(x) = \left[R^2(x)(u,v) + I^2(x)(u,v)\right]^{1/2}, \tag{5}$$

and

$$\mathcal{P}(x) = \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \tag{6}$$

where $R(x)$ and $I(x)$ represent the real and imaginary parts of $\mathcal{F}(x)$.

We decouple the image features and the prompt features extracted by text encoder into their corresponding amplitude and phase parts. With the separated phase and amplitude spectrum features, we reconstruct the augmented samples for the kth domain with the inverse FFT algorithm,

$$x_{ik}^{aug} = \mathcal{F}^{-1}\left(\mathcal{A}\left(z_{kt}^{txt}\right) * e^{-j*\mathcal{P}(x_i)}\right). \tag{7}$$

where z_{kt}^{txt} shares the same class semantics with x_i, or x_i belongs to the tth class.

Finally, we incorporate these augmented samples into the training phase to enhance the generalization ability of model. The loss function can be written as,

$$\mathcal{L}_{cls}^{aug} = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_i \log\left(f\left(x_{ik}^{aug}; \theta\right)\right). \tag{8}$$

3.3 Domain Prompt Alignment

The DPE module improves generalization ability by generating diverse source representations for training. At the same time, since we actually have no clue of the real target distribution, it is difficult to guarantee that the augmented samples perfectly match the target data. Inspired by testing-time style shifting [Park *et al.*, 2023], we further align the distribution of target data with that of the training domains during the testing time without model updating.

In order to effectively align each target sample with the training domains, we try to transfer the target style to that of the training domains to better align with the classifier. To achieve this goal, we compute the similarity vector ω_{ik} using cosine similarity between the features extracted from each x_i^{target} and those from the K source domains during training, where $k \in 1, \dots, K$. In terms of the similarities, we can obtain the weighted amplitude features over all source domains,

$$\mathcal{A}_i^{mix} = \frac{1}{K}\sum_{k=1}^{K}\omega_{ik}\mathcal{A}\left(z_k^{style}\right), \tag{9}$$

which helps transfer the style of the target samples towards that of the training domains. Then for each target sample, we decouple it into $\mathcal{A}\left(x_i^{target}\right)$ and $\mathcal{P}\left(x_i^{target}\right)$, and then reconstruct it with,

$$\hat{x}_i^{mix} = \mathcal{F}^{-1}\left(\mathcal{A}_i^{mix} * e^{-j*\mathcal{P}\left(x_i^{target}\right)}\right). \tag{10}$$

Further, we adopt an adaptive alignment in terms of the similarity, since in practical situations, not all target samples have a large style discrepancy from source domain. When the target style is similar to that of source domain, for example, an "art" source and a "photo" target, we directly use the original target data in inference, allowing the model to adapt flexibly to varying levels of style discrepancy between domains.

$$p_i = \begin{cases} f(\hat{x}_i^{mix};\theta), & \text{if } ||\mathcal{A}\left(x_i^{target}\right) - \mathcal{A}_i^{mix}||^2 \geq \tau \\ f(x_i^{target};\theta), & \text{otherwise.} \end{cases} \tag{11}$$

where τ is the average cosine similarity between amplitudes of the target sample and training domain prototypes.

3.4 Optimization Objective

To make full use of training data, we also reconstruct x_i^{mix} with $\mathcal{P}\left(x_i\right)$ and $\mathcal{A}_i^{mix}$ during training as a supplement to the scarce source data, and formulate the classification loss as,

$$\mathcal{L}_{cls}^{mix} = -\sum_{i=1}^{N} y_i \log\left(f\left(x_i^{mix};\theta\right)\right). \tag{12}$$

Finally, the total training loss for our ProMEA model to backward can be formulated as,

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{cls}^{aug} + \beta\mathcal{L}_{cls}^{mix} + \lambda\mathcal{L}_{ds}, \tag{13}$$

where $\mathcal{L}_{cls}$ is the classification loss of original images with standard cross entropy, α, β and λ are trade-off parameters. The learning process is summarized in Algorithm 1.

Algorithm 1 Training of ProMEA

```
Input: Source domain $\mathcal{D}_s$, domain words $W_d = \{d_1, \dots, d_K\}$, prompt templates $Prompts$, and CLIP text encoder $E_{\text{txt}}(\cdot)$
Output: Learned model weights $\theta$
for $e \in MaxEpoch$ do
    $z_{kt}^{\text{txt}}, z_k^{\text{style}} \sim Prompts, W_d$   ▷ initialize text prompts
    Calculate $\mathcal{L}_{ds}$ according to Eq. 3
    $x_{ik}^{aug} \leftarrow \mathcal{F}^{-1}(\mathcal{A}(z_{kt}^{txt}), \mathcal{P}(x_i))$ ▷ synthesize samples
    for domain word $d_k \in W_d$ do
        $\omega_{ik} \leftarrow similarity(x_i, z_k^{\text{style}})$
    end for
    $\mathcal{A}_i^{mix} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\omega_{ik}\mathcal{A}\left(z_k^{style}\right)$
    $x_i^{mix} \leftarrow \mathcal{F}^{-1}((\mathcal{A}_i^{mix}), \mathcal{P}(x_i))$
    Calculate $\mathcal{L}_{cls}^{aug}, \mathcal{L}_{cls}^{mix}$ according to Eq. 8,12
    if warm up then
        update $\theta$ to minimize $\mathcal{L}_{cls} + \alpha\mathcal{L}_{cls}^{aug} + \lambda\mathcal{L}_{ds}$
    else
        update $\theta$ to minimize $\mathcal{L}$ according to Eq. 13
    end if
end for
```

4 Experiments

In this section, we evaluate our model ProMEA on three challenging benchmark datasets for single-DG.

4.1 Datasets

PACS [Li *et al.*, 2017a] is composed of four domains (Photo, Art, Cartoon, Sketch) and 7 classes (dog, elephant, giraffe, guitar, horse and person), and there is large distribution discrepancy between different domains. **VLCS** [Fang *et al.*, 2013] consists of 10,729 images across 5 classes (bird, car, chair, dog, and person), sourced from 4 real-world datasets (VOC2007, LabelMe, Caltech, SUN09). The scenes captured vary from urban to rural. **OfficeHome** [Venkateswara *et al.*, 2017] is a challenging dataset. It contains 30,475 images belonging to 65 classes, originating from four different domains (Art, Clipart, Product, Real), where the domain shift mainly stems from differences in image styles and viewpoints.

4.2 Comparison Methods

We compare ProMEA with state-of-the-art methods: ERM [Koltchinskii, 2011], ADA [Volpi *et al.*, 2018], Augmix [Hendrycks *et al.*, 2019], pAdaIn [Nuriel *et al.*, 2021], SagNet [Nam *et al.*, 2021], MixStyle [Zhou *et al.*, 2021], L2D [Wang *et al.*, 2021], FACT [Xu *et al.*, 2021], ACVC [Cugu *et al.*, 2022], Pro-RandConv [Choi *et al.*, 2023], ACVC/MAD [Qu *et al.*, 2023], CADA [Chen *et al.*, 2023], UniFreqSDG [Liu *et al.*, 2024a] and StyDeSty [Liu *et al.*, 2024b]. ERM is the baseline method, MixStyle, SagNet and FACT are multi-source DG methods, and the others are single-DG methods. All compared methods are conducted over the same datasets and backbone networks.

4.3 Implementation details

Following previous works, we adopt the pre-trained CLIP model to extract text prompts. Images are uniformly resized

Methods	Art	Cartoon	Sketch	Photo	Avg.
ERM	70.90	76.50	53.10	44.20	40.70
ADA	73.20	71.97	44.63	45.73	58.70
Augmix	72.73	76.83	46.22	46.32	60.22
SagNet	73.2	75.67	48.53	50.07	61.9
L2D	76.91	77.88	52.29	53.66	65.18
RandConv+AugMix	76.70	79.30	61.6	54.60	68.10
ACVC	73.68	77.39	55.30	48.05	63.61
FACT	75.73	78.43	62.83	50.7	66.92
Pro-RandConv	76.98	78.54	62.89	57.11	68.88
ACVC/MAD	52.95	75.51	**77.25**	57.75	65.87
CADA	76.33	79.08	61.59	56.65	68.41
UniFreqSDG	78.94	79.51	68.92	56.97	70.59
StyDeSty	80.06	79.86	63.24	**62.22**	71.35
ProMEA	**82.58**	**81.43**	62.52	61.09	**71.91**

Table 1: single-DG accuracy over PACS dataset. Each model is trained over ResNet18.

to 224 × 224. For PACS and VLCS, we select ResNet-18 as our backbone. We use SGD optimizer for training, set the batch size to 16, the weight decay to 5e-4, and train for 75 epochs. The initial learning rate is 1e-3, and the warm-up period consists of 25 epochs. For OfficeHome, we choose ResNet-50 as the backbone and set the weight decay to 5e-5, training for 50 epochs. We train with SGD optimizer and set the batch size to 16. The initial learning rate is 1e-3 and decays by 0.1 at 80% of the total epochs. The warm-up period lasts for 10 epochs. When the number of current epoch is less than the warm-up epoch, we merely train the model without performing testing-time alignment.

The model is trained on a single source domain and evaluated on the remaining domains. To ensure a fair comparison, we adopt the same backbone network as previous approaches. For PACS and OfficeHome, we follow the data splits from Shu et al. [Shu *et al.*, 2021]. For VLCS, we randomly split each domain into 90% training and 10% validation subsets.

4.4 Comparison Results

The comparison results are shown in Tables 1-3, in which the best performance among the compared methods is indicated by bold values, and the second best performance is indicated by underlines.

Results on PACS From Table 1, among the four generalization tasks, ProMEA achieves the best performance over two tasks and the second-optimal performance over one task. Overall, ProMEA attains the highest average accuracy, with an improvement of 0.56%. Moreover, compared with the suboptimal method StyDeSty, the performance of ProMEA is improved by 2.52% and 1.57% over the Art and Cartoon domains respectively. As a result, by conducting domain prompt expansion for source domain and domain prompt alignment for target data, ProMEA can effectively enhance the performance of single-DG.

Results on VLCS From Table 2, ProMEA achieves the best performance in three tasks and the second-optimal performance in the other task. Finally, ProMEA achieves the best average accuracy. Compared with the suboptimal StyDeSty method, ProMEA achieves significant improvements in three domains (VOC2007, LabelMe, Caltech), with respective increases of 2.2%, 9.1% and 5.29%, and the overall performance improvement achieves 2.78%.

Methods	V	L	C	S	Avg.
ERM	76.72	58.86	44.95	57.71	59.56
Augmix	75.52	59.52	45.90	57.43	59.53
pAdaIn	76.03	65.21	43.17	57.94	60.59
MixStyle	75.73	61.29	44.66	56.57	59.56
ACVC	76.15	61.23	47.43	60.18	61.25
ACVC/MAD	76.15	69.36	48.04	61.74	63.82
Pro-RandConv	-	-	-	-	53.35
StyDeSty	76.87	62.87	53.73	**65.41**	64.72
ProMEA	**79.07**	**71.97**	**59.02**	59.94	**67.50**

Table 2: single-DG accuracy over VLCS dataset. Each model is trained with ResNet18.

Methods	Art	Clipart	Product	Real	Avg.
ERM	57.08	54.59	51.48	60.86	56.00
SagNet	55.18	52.48	51.16	60.83	54.91
Augmix	59.69	56.82	54.71	62.54	58.44
RandConv+AugMix	54.22	52.22	50.16	59.51	54.03
ALT	59.87	55.89	54.72	64.66	58.79
Pro-RandConv	-	-	-	-	59.20
StyDeSty	59.09	**57.78**	54.73	65.02	59.16
ProMEA	**62.52**	56.98	**56.02**	**65.64**	**60.29**

Table 3: single-DG accuracy over OfficeHome dataset. Each model is trained with ResNet50.

Results on OfficeHome From Table 3, ProMEA also achieves the best performance in three tasks and the second-optimal performance in the other task. Finally, ProMEA achieves the best average accuracy with an improvement of 1.09%. These results further validate the superiority of our proposal.

4.5 Ablation Study

To evaluate each component in ProMEA, we conduct ablation experiments over the PACS dataset.

Methods	$\mathcal{L}_{cls}^{aug}$	$\mathcal{L}_{cls}^{mix}$	$\mathcal{L}_{ds}$	Art	Cartoon	Sketch	Photo	Avg.
baseline				75.90	79.41	58.31	51.15	66.44
	✓			79.16	80.34	58.94	58.01	69.11
		✓		78.89	78.92	60.97	56.34	68.78
		✓	✓	80.97	80.23	57.00	57.77	68.99
	✓		✓	80.43	81.38	62.21	59.01	70.76
Ours	✓	✓	✓	**82.58**	**81.43**	**62.52**	**61.09**	**71.91**

Table 4: Ablation Study over different training losses. We train ProMEA on PACS and adopt RestNet18 as the backbone.

Effect of Training Loss. We conduct an ablation study to

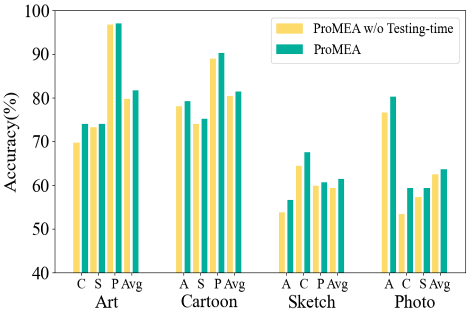

Figure 3: Accuracy analysis for Testing-time Alignment over PACS dataset.

investigate the effect of each training loss component of the ProMEA model, as shown in Table 4. First, with $\mathcal{L}_{cls}^{aug}$, the model achieves a significant improvement by using the generated domain data. However, when we only use the original images and mixed data from prompt domains for training, the model performance improves sightly because of insufficient domain augmentation. Furthermore, by incorporating $\mathcal{L}_{ds}$ into the optimization loss, the classification accuracy of ProMEA has improved significantly, indicating that it plays a crucial role in improving the generalization capabilities, which aligns well with our expectations.

Analysis of Testing-time Alignment. In ProMEA, the PDA module plays a crucial role in addressing domain shift across different domains. By aligning target data with source domains during the testing phase, PDA enables the model to classify target samples more effectively. To analyze the impact of testing-time alignment, we select one domain as the source and the remaining three as targets on the PACS dataset. As illustrated in Fig. 3, the performance with testing-time alignment surpasses that without alignment in most scenarios. This improvement is particularly pronounced in the Sketch domain, which contains less stylistic information, highlighting the effectiveness of PDA in domains with limited style diversity. These results confirm that the distribution alignment at testing time can indeed enhance the performance in target domains.

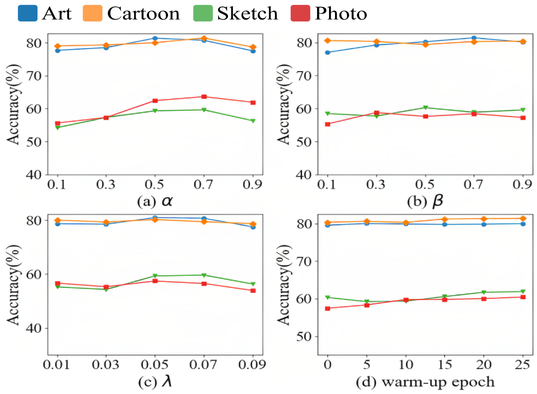

Figure 4: Hyperparameter analysis of ProMEA with respect to parameters α, β, λ and warm-up epoch over PACS dataset.

4.6 Hyperparameter Analysis

In this subsection, we perform hyperparameter analysis for α, β and λ, and the warm-up epoch on PACS dataset, as illustrated in Fig. 4. In the experiments, we initially set $\alpha = 0.3$, $\beta = 0.5$, $\lambda = 0.01$ and the warm-up epoch at 10. When analyzing the sensitivity of a specific parameter, we fix the values of the other parameters. From Fig. 4(a), the performance of ProMEA improves progressively as more generated samples are integrated. From Fig. 4 (b) and (c), we observe that β and λ have a relatively limited impact on the Cartoon and Art domains, while their changes cause larger fluctuations on Sketch and Photo domains. To ensure reliable style transfer for target samples during the testing phase, ProMEA implements a warm-up epoch to optimize the prompt tokens. According to Fig. 4(d), the warm-up epoch has a limited effect on performance fluctuations over the Sketch domain. This is because the Sketch domain contains relatively less style information compared to other domains. Additionally, insufficient training iterations hinder the effective optimization of both domain-agnostic and domain-specific tokens within the prompts. Consequently, it results in an inadequate representation of prompt styles, reducing the model ability to generate high-quality augmented samples.

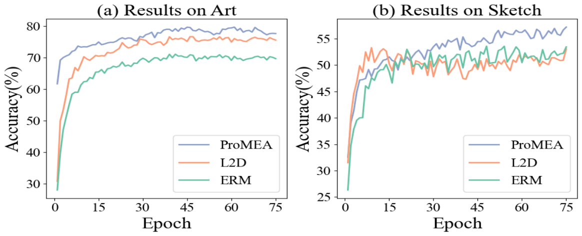

Figure 5: Stability Analysis for ProMEA over PACS dataset. ProMEA is trained on Cartoon and tested on the other domains.

4.7 Stability Analysis

We conduct stability analysis for the performance of ProMEA, comparing it with ERM and L2D in Fig. 5, from which we can observe that ProMEA maintains a relatively stable performance compared to both ERM and L2D. During the training process, while the accuracy curve shows some fluctuations, the overall classification accuracy steadily increases. This trend further validates the effectiveness of our proposed method.

5 Conclusion

In this paper, a prompt-driven expansion and alignment model is proposed for single-DG. Leveraging the ease of generation and generalization in text prompt perturbations, we expand the single source domain using frequency prompts of augmented texts. Further, we design a testing-time alignment module that adapts target samples to the convex hull of source prompt features during inference. It effectively minimizes the discrepancy between the source and target domains. Extensive experiments demonstrate that ProMEA outperforms existing SOTA single-DG methods, highlighting its superiority in handling domain shift challenges with a single source.

References

[Arefyev *et al.*, 2022] Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. *arXiv preprint arXiv:2206.11815*, 2022.

[Chen *et al.*, 2020] Zhi Chen, Jingjing Li, Yadan Luo, Zi Huang, and Yang Yang. Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 874–883, 2020.

[Chen *et al.*, 2021a] Yang Chen, Yu Wang, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. A style and semantic memory mechanism for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9164–9173, 2021.

[Chen *et al.*, 2021b] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8712–8720, 2021.

[Chen *et al.*, 2023] Tianle Chen, Mahsa Baktashmotlagh, Zijian Wang, and Mathieu Salzmann. Center-aware adversarial augmentation for single domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4157–4165, 2023.

[Cheng *et al.*, 2024] De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23595–23604, 2024.

[Cho *et al.*, 2023] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023.

[Choi *et al.*, 2023] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023.

[Cugu *et al.*, 2022] Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4174, 2022.

[Dayal *et al.*, 2024] Aveen Dayal, Vimal KB, Linga Reddy Cenkeramaddi, C Mohan, Abhinav Kumar, and Vineeth N Balasubramanian. Madg: margin-based adversarial learning for domain generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[Fang *et al.*, 2013] Chen Fang, Ye Xu, and Daniel 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*, pages 1657–1664, 2013.

[Ge *et al.*, 2023] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Gokhale *et al.*, 2023] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 434–443, 2023.

[Hendrycks *et al.*, 2019] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[Jia and Zhang, 2024] Chen Jia and Yue Zhang. Meta-learning the invariant representation for domain generalization. *Machine Learning*, 113(4):1661–1681, 2024.

[Koltchinskii, 2011] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[Li *et al.*, 2017a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[Li *et al.*, 2017b] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. Domain generalization and adaptation using low rank exemplar svms. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1114–1127, 2017.

[Liu and Wang, 2023] Geng Liu and Yuxi Wang. Tdg: Text-guided domain generalization. *arXiv preprint arXiv:2308.09931*, 2023.

[Liu *et al.*, 2024a] Chuang Liu, Yichao Cao, Xiu Su, and Haogang Zhu. Universal frequency domain perturbation for single-source domain generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6250–6259, 2024.

[Liu *et al.*, 2024b] Songhua Liu, Xin Jin, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Stydesty: Min-max stylization and destylization for single domain generalization. In *Forty-first International Conference on Machine Learning*, 2024.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[Luo *et al.*, 2020] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *International Conference on Machine Learning*, pages 6468–6478. PMLR, 2020.

[Nam *et al.*, 2021] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[Niu *et al.*, 2022] Hongjing Niu, Hanting Li, Feng Zhao, and Bin Li. Domain-unified prompt representations for source-free domain generalization. *arXiv preprint arXiv:2209.14926*, 2022.

[Nuriel *et al.*, 2021] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9482–9491, 2021.

[Nussbaumer and Nussbaumer, 1982] Henri J Nussbaumer and Henri J Nussbaumer. *The fast Fourier transform*. Springer, 1982.

[Oppenheim *et al.*, 1979] A. Oppenheim, Jae Lim, G. Kopec, and S. Pohlig. Phase in speech and pictures. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637, 1979.

[Park *et al.*, 2023] Jungwuk Park, Dong-Jun Han, Soyeong Kim, and Jaekyun Moon. Test-time style shifting: Handling arbitrary styles in domain generalization. In *International Conference on Machine Learning*, pages 27114–27131. PMLR, 2023.

[Qu *et al.*, 2023] Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. Modality-agnostic debiasing for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24142–24151, 2023.

[Schiappa *et al.*, 2022] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, 35:34405–34420, 2022.

[Seo *et al.*, 2020] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020.

[Shu *et al.*, 2021] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[Volpi *et al.*, 2018] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[Wang *et al.*, 2021] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021.

[Xu *et al.*, 2014] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 628–643. Springer, 2014.

[Xu *et al.*, 2021] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021.

[Yang and Soatto, 2020] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020.

[Yang *et al.*, 2023] Qingyue Yang, Hongjing Niu, Pengfei Xia, Wei Zhang, and Bin Li. Frequency decomposition to tap the potential of single domain for generalization. *arXiv preprint arXiv:2304.07261*, 2023.

[Zhou *et al.*, 2021] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.